# MACHINE LEARNING-BASED FAKE NEWS DETECTION FOR SOCIAL MEDIA

**Khalilova Sarvinoz**

Student, Japan Digital University, Tashkent, Uzbekistan

**Academic Supervisor: Lochinbek Boboev, PhD**

Head of the IT Department, Japan Digital University, Tashkent, Uzbekistan

**ABSTRACT**

With the rapid growth of social media platforms such as Telegram, Facebook, and Instagram, the spread of false, unverified, and manipulative information (fake news) in the Uzbek language has increased significantly, causing social anxiety, misinformation, and public misunderstanding. Existing international fake news detection algorithms are primarily designed for the English language, and due to the rich morphological structure, agglutinative nature of Uzbek, and the scarcity of labeled datasets, direct application of these algorithms yields inefficient results. This paper proposes a hybrid system for automatically determining the veracity of Uzbek-language texts and headlines using TF-IDF and a multilingual BERT (mBERT) model adapted for Uzbek. The system combines classical machine learning algorithms — Logistic Regression, Random Forest, and XGBoost — with transformer-based contextual embeddings. A dataset of more than 2,500 manually labeled Uzbek news articles and social media posts was collected, followed by preprocessing steps including tokenization, lemmatization, and stop-word removal. Expected results indicate that the mBERT-based model will demonstrate higher Accuracy, Precision, Recall, and F1-Score metrics compared to classical algorithms. The proposed system contributes to ensuring information security, combating disinformation, and enriching NLP resources for the Uzbek language.

**Keywords:** *Fake News Detection, Machine Learning, Natural Language Processing, Uzbek Social Media, Information Security, Text Classification.*

## I. INTRODUCTION

In recent years, social media has become the fastest and most widespread means of information exchange. Platforms such as Telegram, Facebook, and Instagram enable millions of users to instantly access news, opinions, and announcements. However, alongside this convenience, a serious problem — the spread of fake news and disinformation — has also increased sharply. Unverified, manipulative, or deliberately distorted information can heighten social tension, undermine public trust, and, in some cases, even pose a threat to national security.

The problem of fake news detection has been actively researched worldwide over the past decade. Machine-learning-based approaches — including Naive Bayes, Support Vector Machine, Random Forest, and XGBoost — have demonstrated high effectiveness in classifying English-language texts [1], [2]. In particular, an XGBoost algorithm combined with the TF-IDF feature extraction method has been reported to achieve an accuracy above 91% [3], while studies employing SVM and the Passive Aggressive Classifier have achieved accuracy levels exceeding 95% [4]. In recent years, transformer-based architectures such as BERT and its multilingual variant mBERT have attracted considerable attention in fake news detection, as they enable bidirectional understanding of textual context [5]. In multilingual settings, models such as mBERT and XLM-RoBERTa have been shown to produce stable results across different languages; however, researchers consistently identify the lack of dedicated datasets and adapted models for low-resource languages as a major challenge [6], [7].

The situation within the Uzbek-language segment is more complex. First, Uzbek is an agglutinative language — meaning that word formation and inflection occur through the extensive use of suffixes — which complicates conventional tokenization and lemmatization methods [8]. Second, high-quality, manually annotated fake news datasets for the Uzbek language are virtually nonexistent; existing NLP resources are primarily designed for topic-based text classification [9] or sentiment analysis [10]. Third, dedicated language models developed for Uzbek —

such as UzBERT [10] and BERTbek [11] — have demonstrated superior performance compared to mBERT in sentiment analysis and text classification tasks, yet they have not yet been extensively tested for the fake news detection task.

Taking the above into account, the aim of this study is to develop a hybrid machine learning system, based on TF-IDF and mBERT contextual embeddings, capable of automatically assessing the veracity of Uzbek-language social media posts and news articles. The study encompasses the following tasks: (1) constructing a dataset of genuine and fake Uzbek-language news items; (2) preprocessing texts in accordance with the morphological characteristics of the Uzbek language; (3) conducting a comparative analysis of classical ML algorithms (Logistic Regression, Random Forest, XGBoost) and a deep learning model (mBERT); and (4) evaluating the results using standard metrics (Accuracy, Precision, Recall, F1-Score). The findings of this study carry practical significance for information security and may be applied in the future by government agencies, media-monitoring organizations, and social media platforms.

## II. MAIN PART

### *2.1. Dataset Collection*

The dataset for this study was compiled from two primary sources: (1) verified, genuine news items obtained from official and reputable Uzbek news portals (Kun.uz, Daryo.uz, Gazeta.uz, and others); and (2) news items that circulated via fact-checking organizations' platforms and social media (Telegram channels, Facebook groups) and were subsequently refuted or remained unverified. Each text was manually annotated into one of two categories (“genuine” and “fake”), with two independent experts participating in the annotation process; in cases of disagreement, a third expert's judgment was sought. The final dataset comprises approximately 2,500 text items, each containing a headline, main text, and label (0 — genuine, 1 — fake).

### *2.2. Text Preprocessing*

Given the morphological complexity of the Uzbek language, the preprocessing stage is of particular importance [8]. This stage consists of the following sequential operations:

1) Cleaning: HTML tags, URLs, emojis, and extraneous whitespace were removed from the text.

2) Tokenization: Text was segmented into word-level units (tokens). Texts containing a mixture of Latin and Cyrillic Uzbek scripts were first transliterated into a unified (Latin) script.

3) Stop-word removal: Auxiliary words carrying minimal semantic weight, such as “va” (and), “bu” (this), “uchun” (for), and “bilan” (with), were filtered out based on a predefined stop-word list.

4) Lemmatization/stemming: Agglutinative suffixes (-lar, -ning, -ga, -dan, etc.) were separated from the word stem to determine the lemma (base form) of each word. Morphological stemmers developed specifically for the Uzbek language were used at this stage [12].

5) Vectorization: For the classical ML models, the TF-IDF (Term Frequency–Inverse Document Frequency) method was applied, incorporating unigram and bigram features. For the mBERT model, texts were segmented into subword units using the WordPiece tokenizer and converted into contextual embeddings [5].

### *2.3. Model Architecture*

A two-stage hybrid approach was adopted in this study:

A) Classical machine learning models. Three algorithms were trained on feature vectors derived from TF-IDF: Logistic Regression, serving as a baseline linear classifier; Random Forest, employed to enhance generalization capability through an ensemble of multiple decision trees; and XGBoost, used to achieve high accuracy through gradient boosting, given that prior research has reported strong performance of XGBoost in fake news classification tasks [3].

B) Transformer-based model (mBERT). The multilingual BERT model (bert-base-multilingual-cased) was fine-tuned on the Uzbek-language dataset. At the input

layer, the model derives an aggregated vector representation of the entire text via the [CLS] token; this vector is then passed to a fully connected layer with a sigmoid activation function, which outputs a probability value indicating whether the text is "genuine" or "fake." This approach has been shown in international research to be effective for multilingual and low-resource-language settings [6], [7].

### *2.4. Evaluation Metrics and Expected Results*

The following standard metrics are used to evaluate model performance:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

Here, TP denotes correctly identified fake news items, TN denotes correctly identified genuine news items, FP denotes genuine news items incorrectly labeled as fake, and FN denotes fake news items incorrectly labeled as genuine.

*Table 1. Expected comparative performance of the models (approximate values, based on the dataset test split)*

| **Model** | **Accuracy** | **Precision** | **Recall** |
|---|---|---|---|
| Logistic Regression (TF-IDF) | ~0.78 | ~0.76 | ~0.75 |
| Random Forest (TF-IDF) | ~0.82 | ~0.81 | ~0.79 |
| XGBoost (TF-IDF) | ~0.85 | ~0.84 | ~0.83 |
| **mBERT (fine-tuned)** | **~0.90** | **~0.89** | **~0.88** |

As shown in the table above, the contextual-embedding-based mBERT model is expected to outperform the classical TF-IDF-based models, owing to its superior ability to capture the contextual meaning of words, intra-sentence relationships, and long-range dependencies. Among the classical models, XGBoost is expected to achieve the highest performance, which can be attributed to its gradient boosting

mechanism's capacity for consistent error reduction [3]. For visual analysis of the results, it is planned to compare the models' F1-Score values using a bar chart and to separately analyze error types (misclassifying fake news as genuine, or vice versa) through a confusion matrix.

### *2.5. Discussion*

The obtained (expected) results indicate that while models based on purely statistical features (TF-IDF) are fast and computationally efficient, they are unable to fully capture the deep semantic meaning of a text. In particular, irony, sarcasm, and polysemous expressions in the Uzbek language pose considerable challenges for TF-IDF-based approaches. The mBERT model, through its deeper contextual understanding, can identify such subtle cases more effectively, but it requires substantially larger training datasets and greater computational power. Consequently, a hybrid system combining both approaches in practice — for example, using TF-IDF combined with XGBoost for rapid initial filtering, followed by mBERT for deeper analysis of more complex cases — may represent the most optimal solution.

## III. CONCLUSION

This study proposed a system for automatically detecting fake news circulating on Uzbek-language social media using machine learning and natural language processing methods. A comparative analysis of TF-IDF-based classical algorithms (Logistic Regression, Random Forest, XGBoost) and the multilingual BERT (mBERT) model indicates that transformer-based deep learning models have the potential to achieve higher accuracy by effectively accounting for the complex morphological characteristics of the Uzbek language. The proposed system carries significant practical value for ensuring information security: it can assist governmental and non-governmental organizations in the rapid detection of disinformation on social media, the protection of public opinion, and the enhancement of media literacy.

Future research is planned to proceed in the following directions: (1) expanding and balancing the dataset across various topics (politics, healthcare, economics); (2)

conducting extensive testing of the Uzbek-specific BERTbek and UzBERT models for the fake news detection task [10], [11]; (3) developing a system capable of analyzing multimodal data, including images and videos, alongside text; and (4) implementing the system as a real-time Telegram bot or web application. Continued research in this direction will contribute to enriching Uzbek-language NLP resources and strengthening regional information security.

## REFERENCES


1. Sharma, A. et al. (2023). A Comprehensive Survey on Fake News Detection with Machine Learning Algorithms. IEEE Xplore. https://ieeexplore.ieee.org/document/10270611/
2. A Survey on Role of Machine Learning and NLP in Fake News Detection on Social Media. IEEE Conference Publication, 2021. https://ieeexplore.ieee.org/document/9573875/
3. Fake News Detection Using Machine Learning Algorithm (XGBoost-based approach). IEEE Conference Publication, 2023. https://ieeexplore.ieee.org/document/10088320/
4. Fake News Detection using Machine Learning (TF-IDF and SVM approach). IEEE Conference Publication, 2021. https://ieeexplore.ieee.org/document/9358890/
5. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805
6. Multilingual Fake News Detection: A Study on Various Models and Training Scenarios (mBERT, XLM-RoBERTa, LASER). Springer Nature Link, 2024. https://link.springer.com/chapter/10.1007/978-3-031-66428-1_5
7. A Transformer-Based Approach to Multilingual Fake News Detection in Low-Resource Languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 2021. https://dl.acm.org/doi/10.1145/3472619

8. Kuriyozov, E. et al. (2023). Text Classification Dataset and Analysis for Uzbek Language. arXiv preprint arXiv:2302.14494. https://arxiv.org/abs/2302.14494
9. Rabbimov, I., Kobilov, S. Multi-class Text Classification of Uzbek News Articles using Machine Learning, as cited in Kuriyozov, E. et al., Text Classification Dataset and Analysis for Uzbek Language, arXiv:2302.14494, 2023. https://arxiv.org/pdf/2302.14494
10. Mansurov, B., Mansurov, A. (2021). UzBERT: Pretraining a BERT Model for Uzbek. arXiv preprint arXiv:2108.09814. https://arxiv.org/abs/2108.09814
11. BERTbek: A Pretrained Language Model for Uzbek. Proceedings of SIGUL 2024 (ACL Anthology). https://aclanthology.org/2024.sigul-1.5.pdf
12. A Feedforward Neural Model for Morphological Stemming of Uzbek Words. ResearchGate, 2025. https://www.researchgate.net/publication/396883399_A_Feedforward_Neural_Model_for_Morphological_Stemming_of_Uzbek_Words