pISSN:2349-5707 Volume:13,Issue 2, February -2026 eISSN:2349-5715

TRANSFORMER-BASED SPAM DETECTION FOR UZBEK LANGUAGE: A COMPARATIVE STUDY WITH TRADITIONAL MACHINE LEARNING MODELS

Atajonov Muzaffar Ne'matjon ugli
Teacher, Jaloliddin Manguberdi Military-Academic Lyceum
Tashkent, Uzbekistan
Email: muzaffar19910627@gmail.com

Abstract: Spam detection remains a critical challenge in natural language processing, particularly for low-resource languages such as Uzbek. While traditional machine learning approaches have been widely applied to text classification tasks, their reliance on handcrafted features limits contextual understanding. Recent advances in Transformer-based architectures, especially BERT (Bidirectional Encoder Representations from Transformers), have demonstrated superior performance in capturing semantic relationships within text.

This study proposes a BERT-based spam detection model for Uzbek SMS messages and compares its effectiveness with conventional machine learning models including Naïve Bayes, Support Vector Machine (SVM), and Logistic Regression. A labeled Uzbek SMS dataset was utilized, divided into training and testing subsets. Traditional models were trained using TF-IDF feature extraction, while a multilingual BERT model was fine-tuned for binary classification.

Experimental results indicate that the Transformer-based model significantly outperforms classical approaches in terms of accuracy, precision, recall, and F1-score. The findings confirm that contextual embeddings are highly effective for spam detection in morphologically rich and low-resource languages.

The research contributes to Uzbek NLP by providing one of the first systematic comparative analyses between deep contextual models and traditional machine learning techniques for spam classification. Practical implications include the potential integration of the proposed model into mobile communication platforms. Limitations include dataset size and computational requirements, which future studies may address using lightweight Transformer architectures.

Keywords: Uzbek NLP; Spam Detection; BERT; Machine Learning; Text Classification; Transformer; Low-Resource Language

INTRODUCTION

Spam messages remain a major challenge in digital communication systems, posing risks to privacy and information security. With the expansion of mobile messaging platforms, spam has evolved into sophisticated phishing and fraud schemes. Spam detection is therefore a critical application of machine learning and NLP in cybersecurity [1].

Traditional approaches rely on classical classifiers such as Naïve Bayes and Support Vector Machines (SVM) [1], [2]. These models typically employ TF-IDF or bag-of-words representations, which convert text into frequency-based numerical vectors. Although effective for high-resource languages, such approaches struggle in morphologically rich and low-resource languages [3].

The Uzbek language is agglutinative, with complex suffixation and word formation patterns. Frequency-based representations often fail to capture semantic dependencies in such languages. Transformer-based architectures, introduced by Vaswani et al. [4], address this limitation through self-attention mechanisms. BERT, proposed by Devlin et al. [5], enables bidirectional contextual encoding and has achieved state-of-the-art results in text classification tasks [6].

Multilingual models such as mBERT and cross-lingual pretraining methods [7] have shown promising results for low-resource languages. However, systematic comparative research for Uzbek spam detection remains limited.

This study aims to compare traditional machine learning models with a fine-tuned BERT model for Uzbek spam detection and evaluate their performance using statistical metrics.

METHODOLOGY

Research Framework and Experimental Design – This research adopts a quantitative experimental framework to evaluate the effectiveness of Transformer-based deep learning models compared to traditional machine learning approaches for Uzbek-language spam detection. The study is structured as a supervised binary text classification problem, where each SMS message is categorized into one of two classes: *spam* (1) or *non-spam* (0).

The experimental workflow consists of the following sequential stages:

1. Dataset preparation and annotation
2. Text preprocessing and normalization
3. Feature extraction (TF-IDF for classical models)
4. Model training and hyperparameter optimization
5. Fine-tuning of a pre-trained BERT model
6. Performance evaluation and statistical validation

This structured pipeline ensures methodological consistency and reproducibility of results.

Dataset Collection and Annotation – The dataset used in this study consists of Uzbek-language SMS messages collected from available digital communication sources and manually curated samples. Each message was carefully reviewed and labeled by domain knowledge criteria to ensure classification reliability.

Messages were assigned to one of two categories:

- Spam: Promotional, fraudulent, phishing, or unsolicited advertising messages.
- Non-spam: Personal, informational, or legitimate communication messages.

To minimize annotation bias, labeling was conducted following predefined criteria focusing on semantic intent and contextual meaning rather than keyword presence alone.

The dataset was analyzed for class distribution to prevent imbalance-related bias. Stratified sampling was applied to divide the dataset into training (80%) and testing (20%) subsets, ensuring proportional representation of both classes.

Text Preprocessing Strategy – Given the agglutinative structure of the Uzbek language, preprocessing was carefully adapted to the requirements of each modeling approach.

For classical algorithms, preprocessing included:

- Conversion to lowercase
- Removal of punctuation and special characters

- Tokenization
- Stop-word filtering
- Optional morphological normalization
- TF-IDF vectorization

The TF-IDF representation was computed as:

$$TF\!-\!IDF(t, d) = TF(t, d) \times log\left(\frac{N}{DF(t)}\right) \text{ (1)}$$

where: $TF(t, d)$ = term frequency of term t in document d, $DF(t)$ = number of documents containing term t *and* N = total number of documents. This approach converts textual data into high-dimensional sparse vectors suitable for statistical classifiers.

For the Transformer-based approach, minimal preprocessing was applied in order to preserve contextual integrity. Since BERT relies on subword tokenization and contextual embedding mechanisms, aggressive text cleaning (such as stop-word removal) was avoided.

The following steps were performed:

- Unicode normalization
- WordPiece tokenization
- Maximum sequence length truncation (128 tokens)
- Padding and attention mask generation

This preprocessing strategy enables the model to capture semantic relationships and morphological variations without manual feature engineering.

Traditional Machine Learning Models – Three widely used supervised classification algorithms were implemented for comparative analysis:

Naïve Bayes is a probabilistic classifier based on Bayes' theorem with conditional independence assumptions. The posterior probability is calculated as:

$P(C|X) \propto P(C) \prod P(x_i \mid C)$ (2)

Despite its simplicity, Naïve Bayes performs efficiently in high-dimensional sparse text data.

Support Vector Machine (SVM) – A linear SVM classifier was employed due to its effectiveness in text classification tasks. The optimization objective is expressed as:

$min \, ½ \, ||w||^2 + C\Sigma \, \xi_i$ (3)

where w is the weight vector, C is the regularization parameter, and ξ_i are slack variables. SVM aims to maximize the margin between classes while minimizing classification error.

Logistic Regression estimates class probabilities using the sigmoid function:

$P(y=1|x) = 1 / (1 + e\^(-w^Tx))$ (4)

Hyperparameters for all traditional models were optimized using Grid Search combined with 5-fold cross-validation to ensure robust performance evaluation.

Transformer-Based Model Architecture and Fine-Tuning – A pre-trained multilingual BERT (mBERT) model was fine-tuned for binary spam classification. BERT utilizes a bidirectional self-attention mechanism to encode contextual information from both left and right text segments simultaneously.

To capture contextual and semantic relationships within Uzbek-language SMS messages, a pre-trained multilingual BERT (mBERT) model was fine-tuned for binary spam classification. Transformer-based architectures rely on bidirectional self-attention mechanisms, enabling deep contextual representation of textual input by considering both preceding and succeeding tokens simultaneously.

The adopted BERT model consists of 12 Transformer encoder layers, 12 attention heads, and a hidden representation size of 768 dimensions. Each encoder layer includes multi-head self-attention, layer normalization, residual connections, and position-wise feed-forward networks. This architecture allows the model to learn complex syntactic and semantic dependencies inherent in morphologically rich languages such as Uzbek.

For classification purposes, the final hidden representation corresponding to the special classification token ([CLS]) was passed through a fully connected dense layer followed by a Softmax activation function to produce class probabilities.

Fine-tuning was conducted using supervised learning on the labeled Uzbek SMS dataset. The training configuration included:

- Optimizer: AdamW
- Learning rate: 2×10^{-5}
- Batch size: 16
- Number of training epochs: 3–5
- Weight decay coefficient: 0.01
- Dropout rate: 0.1

The loss function used for optimization was Binary Cross-Entropy:

$$L = -1/N \Sigma [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (5)$$

where y_i represents the true label and p_i denotes the predicted probability.

To enhance generalization and prevent overfitting, early stopping was applied based on validation loss monitoring. The model's convergence behavior was tracked throughout the training process, and the best-performing checkpoint was selected for final evaluation on the test dataset.

This fine-tuning approach allows the pre-trained Transformer model to adapt effectively to domain-specific spam detection tasks while preserving the benefits of large-scale contextual pre-training.

RESULTS

Classification Performance Comparison – The performance of the implemented models was evaluated on the held-out test dataset using Accuracy, Precision, Recall, and F1-score metrics. Traditional machine learning models (Naïve Bayes, SVM, and Logistic Regression) were compared against the fine-tuned BERT model.

Table 1 presents the averaged results obtained from 5-fold cross-validation for traditional models and the final evaluation results for the BERT-based model.

Model	Accuracy (%)	Precision	Recall	F1-score
Naïve Bayes	87.4 ± 1.2	0.85	0.88	0.86
Logistic Regression	89.1 ± 0.9	0.88	0.90	0.89
SVM (Linear)	91.3 ± 0.8	0.90	0.92	0.91
BERT (Fine-tuned)	96.5	0.96	0.97	0.96

The BERT model achieved the highest accuracy (96.5%), outperforming SVM (91.3%), Logistic Regression (89.1%), and Naïve Bayes (87.4%).

ROC-AUC analysis showed values of 0.89 (Naïve Bayes), 0.92 (Logistic Regression), 0.94 (SVM), and 0.98 (BERT). The improvement was statistically significant ($p < 0.01$).

The BERT model significantly reduced both false positives and false negatives compared to classical methods.

DISCUSSION

The experimental results demonstrate that contextual Transformer-based architectures significantly outperform traditional machine learning approaches in Uzbek spam detection. While classical models such as Naïve Bayes and SVM rely on TF-IDF frequency representations [1], [2], they are limited in capturing semantic dependencies and morphological variations typical of agglutinative languages like Uzbek.

The superior performance of BERT can be attributed to its bidirectional contextual encoding mechanism [4], [5], which allows the model to learn deep semantic relationships between tokens. Unlike bag-of-words models, BERT considers the entire sentence context simultaneously, improving its ability to distinguish between legitimate and spam messages even when lexical overlap exists.

The ROC-AUC improvement (0.98) confirms the robustness of contextual embeddings in binary text classification tasks. These findings are consistent with previous studies demonstrating the effectiveness of Transformer architectures in multilingual and low-resource language environments [6], [7].

However, the computational cost of fine-tuning large Transformer models remains a limitation. Compared to classical models, BERT requires greater memory and processing power. For deployment in resource-constrained environments, lightweight architectures such as DistilBERT or quantized Transformer models may be considered [9].

Overall, the results indicate that contextual language modeling provides a statistically significant advantage for spam detection in morphologically rich languages.

CONCLUSION

This study presented a comparative evaluation of traditional machine learning models and a fine-tuned BERT model for Uzbek spam message classification.

The results show that BERT achieves superior accuracy, precision, recall, F1-score, and ROC-AUC compared to Naïve Bayes, Logistic Regression, and Support Vector Machine classifiers. The improvement is statistically significant and highlights the importance of contextual semantic modeling in low-resource and agglutinative languages.

The research contributes to the development of Uzbek NLP resources and AI-based spam protection systems. The findings demonstrate that Transformer-based models can effectively enhance cybersecurity mechanisms in mobile and messaging platforms.

Future work will focus on expanding the Uzbek spam dataset, exploring lightweight Transformer variants, and developing real-time deployment strategies for mobile applications.

REFERENCES

1. Drucker, H.; Wu, D.; Vapnik, V.N. Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1048–1054, 1999. DOI: 10.1109/72.788645.

2. Androutsopoulos, I.; Koutsias, J.; Chandrinos, K.; Spyropoulos, C. An evaluation of Naïve Bayesian anti-spam filtering. *Proceedings of the Workshop on Machine Learning in the New Information Age*, pp. 9–17, 2000.

3. Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T. Bag of Tricks for Efficient Text Classification. *EACL*, pp. 427–431, 2017. DOI: 10.18653/v1/E17-2068.

4. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017. DOI: 10.48550/arXiv.1706.03762.

5. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*, pp. 4171–4186, 2019. DOI: 10.18653/v1/N19-1423.

6. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to Fine-Tune BERT for Text Classification? *China National Conference on Chinese Computational Linguistics*, 2019. DOI: 10.48550/arXiv.1905.05583.

7. Conneau, A.; Lample, G.; et al. Cross-lingual Language Model Pretraining. *NeurIPS*, 2019. DOI: 10.48550/arXiv.1901.07291.

8. Liu, Y.; Ott, M.; Goyal, N.; et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 2019. DOI: 10.48550/arXiv.1907.11692.

9. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv*, 2019. DOI: 10.48550/arXiv.1910.01108.

10. Kim, Y. Convolutional Neural Networks for Sentence Classification. *EMNLP*, pp. 1746–1751, 2014. DOI: 10.3115/v1/D14-1181.