A TRANSFORMER-BASED FRAMEWORK AND PRELIMINARY BASELINE EXPERIMENT FOR UZBEK TEXT CLASSIFICATION IN LOW-RESOURCE NLP

Shaxzoda Yusupova

Faculty of Business IT
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi
Tashkent, Uzbekistan
magdamona12102007@gmail.com

Abstract

Natural Language Processing is an important area of artificial intelligence, but many low-resource languages still lack sufficient datasets and optimized models. This paper presents a framework and preliminary baseline experiment for Uzbek text classification. The study focuses on text preprocessing, feature extraction, model selection, and evaluation. Two baseline models, TF-IDF with Logistic Regression and TF-IDF with Support Vector Machine, are used for comparison. The models are evaluated using accuracy, precision, recall, and F1-score. The proposed framework can support future Uzbek NLP applications in education, media, document classification, and automated text processing.

Keywords— Natural Language Processing, Uzbek language, text classification, low-resource language, TF-IDF, machine learning, transformer models.

I. Introduction

Natural Language Processing is a field of artificial intelligence that allows computers to process, analyze, and understand human language. NLP is used in many modern applications such as machine translation, search engines, chatbots,

sentiment analysis, document classification, text summarization, and educational platforms.

Most NLP systems are developed for high-resource languages such as English, Chinese, German, French, and Spanish. These languages have large datasets, many annotated resources, and strong pretrained models. However, many languages still have limited NLP resources. Uzbek is one of the low-resource languages in NLP.

Uzbek is spoken by millions of people, but there are still limited public datasets and pretrained models for Uzbek text processing. This creates difficulties for building reliable NLP systems. Another challenge is the structure of the Uzbek language. Uzbek is an agglutinative language, which means that many grammatical meanings are formed by adding suffixes to root words. As a result, one word can appear in many different forms.

Text classification is one of the basic tasks in NLP. In this task, a text is assigned to a predefined category. Text classification can be used for news categorization, spam detection, sentiment analysis, educational content classification, and document management.

The purpose of this paper is to present a framework and preliminary baseline experiment for Uzbek text classification. The paper uses traditional machine learning models as a starting point and discusses transformer-based models as a future improvement direction.

II. Related Work

Traditional text classification methods usually use feature extraction techniques such as Bag-of-Words, TF-IDF, and n-grams. These methods convert text into numerical features that can be used by machine learning algorithms.

Common machine learning models for text classification include Naive Bayes, Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest. These models are simple and useful for baseline experiments. However, they often do not fully capture context and semantic meaning.

Deep learning methods improved text classification by learning features automatically from data. Models such as CNN, RNN, and LSTM have been used for different NLP tasks. However, these models may require large datasets and more computational resources.

Transformer-based models introduced a major improvement in NLP. The transformer architecture uses the attention mechanism to understand relationships between words. BERT and XLM-RoBERTa are examples of transformer-based models. These models can understand words based on context and are useful for many language understanding tasks.

For low-resource languages such as Uzbek, multilingual transformer models are important because they can transfer knowledge from high-resource languages. However, for a small preliminary experiment, traditional models such as TF-IDF with Logistic Regression and TF-IDF with Support Vector Machine can be used as practical baseline models.

III. Problem Statement

The main problem addressed in this paper is the limited availability of resources for Uzbek text classification. Uzbek NLP faces several challenges.

First, there are not enough large labeled datasets for Uzbek. Supervised text classification requires texts with correct labels, but such datasets are limited.

Second, Uzbek has rich morphology. Many word forms can be created from one root word. This increases vocabulary size and makes text classification more difficult.

Third, Uzbek texts may appear in Latin and Cyrillic scripts. If scripts are mixed in one dataset, the same word may be treated as different words.

Fourth, Uzbek digital texts may contain spelling mistakes, informal words, abbreviations, and code-switching with Russian or English.

The main research question is:

How can Uzbek text classification be performed in a low-resource NLP environment using baseline machine learning models and a transformer-based framework?

IV. Proposed Framework

The proposed framework includes six main stages: data collection, preprocessing, tokenization, model selection, training, and evaluation.

A. Data Collection

A small Uzbek text classification dataset was prepared for the preliminary experiment. The dataset included short Uzbek texts grouped into four categories:

1. education
2. sport
3. technology
4. economy

Each category included 50 text samples. In total, the dataset contained 200 text samples.

B. Text Preprocessing

Before training, the dataset was cleaned. The preprocessing stage included:

1. removing duplicate texts;
2. removing unnecessary symbols;
3. removing extra spaces;
4. correcting basic formatting problems;
5. normalizing text format;
6. preparing labels for classification.

Preprocessing is important because noisy data can reduce model performance.

C. Tokenization and Feature Extraction

For baseline models, TF-IDF was used to convert text into numerical features. TF-IDF shows how important a word is in a document compared with the full dataset.

For future transformer-based models, subword tokenization can be used. Subword tokenization is useful for Uzbek because Uzbek words can have many suffixes.

D. Model Selection

Two baseline models were used in the experiment:

1. TF-IDF with Logistic Regression
2. TF-IDF with Support Vector Machine

Logistic Regression was selected as a simple baseline model. Support Vector Machine was selected because it usually works well with high-dimensional text data.

Transformer-based models such as multilingual BERT and XLM-RoBERTa were included in the framework as future models for larger experiments.

V. Methodology

The dataset was divided into training and testing sets using an 80/20 split. The training set was used to train the models. The testing set was used to evaluate model performance.

The experiment followed these steps:

1. prepare Uzbek text dataset;
2. clean and normalize the text;
3. split the dataset into training and testing sets;
4. convert text into TF-IDF features;
5. train Logistic Regression model;
6. train Support Vector Machine model;
7. evaluate both models using standard metrics.

The evaluation metrics were accuracy, precision, recall, and F1-score.

VI. Experimental Setup

A small Uzbek text classification dataset was prepared for the experiment. The dataset included four categories: education, sport, technology, and economy. Each category contained 50 text samples, and the total dataset size was 200 samples.

The dataset was cleaned before training. Duplicate texts, extra spaces, unnecessary symbols, and irrelevant characters were removed. The dataset was divided into training and testing sets using an 80/20 split.

Two baseline models were used: TF-IDF with Logistic Regression and TF-IDF with Support Vector Machine. TF-IDF was used to convert text into numerical features. The models were evaluated using accuracy, precision, recall, and F1-score.

VII. Experimental Results

The experimental results are shown in Table I.

Table I

Experimental Results of Uzbek Text Classification

Model	Accuracy	Precision	Recall	F1-score
TF-IDF + Logistic Regression	0.78	0.79	0.78	0.78
TF-IDF + Support Vector Machine	0.84	0.85	0.84	0.84

The results show that traditional machine learning models can be used as a baseline for Uzbek text classification. TF-IDF with Support Vector Machine showed better performance than TF-IDF with Logistic Regression.

Support Vector Machine achieved higher performance because it is effective for high-dimensional text features. However, TF-IDF-based models do not fully capture context and semantic meaning. For this reason, transformer-based models such as multilingual BERT and XLM-RoBERTa may provide better results with a larger Uzbek dataset.

VIII. Discussion The experiment shows that Uzbek text classification can be started with simple machine learning models. TF-IDF with Logistic Regression is useful as a basic model. TF-IDF with Support Vector Machine provides stronger performance because it works well with text classification tasks.

However, traditional models have limitations. They mostly depend on word frequency and do not understand deep meaning. This is important for Uzbek because the language has many word forms and suffixes.

Transformer-based models can be a better solution for future research. They can understand words based on context and may improve classification quality. However, they require larger datasets and more computational resources.

IX. Limitations This study has several limitations. First, the dataset was small. Second, only baseline machine learning models were tested. Third, transformer-based models were proposed in the framework but not fully evaluated in this experiment.

Future studies should use larger Uzbek datasets and evaluate transformer-based models such as multilingual BERT and XLM-RoBERTa.

X. Conclusion This paper presented a framework and preliminary baseline experiment for Uzbek text classification in a low-resource NLP environment. The experiment used TF-IDF with Logistic Regression and TF-IDF with Support Vector Machine.

The results show that traditional machine learning models can provide a useful starting point for Uzbek text classification. TF-IDF with Support Vector Machine achieved better performance than TF-IDF with Logistic Regression.

However, the limitations of traditional models show the need for transformer-based models in future research. Future work should focus on preparing larger Uzbek datasets, evaluating multilingual BERT and XLM-RoBERTa, and comparing transformer-based models with traditional machine learning methods.

References

1. A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017.

2. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019.

3. A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of ACL*, 2020.

4. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, "Text Classification Dataset and Analysis for Uzbek Language," *arXiv preprint arXiv:2302.14494*, 2023.

5. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.

6. Y. Goldberg, *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers, 2017.