Bakhtiyor Bekmurodov Farkhod o’gli

Farhod Mahmudxo’djayev

Machine Learning and Natural Language Processing

15 December 2025

Encoder-Only Transformer for Sentiment Classification of Uzbek Reviews

Abstract

Sentiment analysis of user-generated reviews plays an important role in understanding customer satisfaction and improving digital services. Traditional approaches such as bag-of-words representations and recurrent neural networks often struggle to capture long-range dependencies and contextual information effectively. In this work, we present an encoder-only Transformer model for sentiment classification of **Uzbek-language** user reviews. The proposed model utilizes multi-head self-attention to model contextual relationships between tokens without relying on recurrence. A **Byte Pair Encoding (BPE)** tokenizer is trained from scratch to handle the morphological characteristics of the language. Experimental results show stable convergence, achieving a final training **loss of 0.1891**. In addition to competitive performance, the attention mechanism enables interpretability by allowing inspection of token-level importance. This work demonstrates that lightweight **Transformer architectures** can be effectively applied to **low-resource language** sentiment analysis tasks.

1. Introduction

With the rapid growth of e-commerce platforms and digital services, large volumes of user-generated reviews are produced daily. Analyzing these reviews is crucial for understanding customer satisfaction, detecting service issues, and supporting data-driven decision-making. Sentiment analysis, a subtask of **natural language processing (NLP),** aims to automatically determine the emotional polarity expressed in textual data.

Early sentiment analysis systems relied on simple frequency-based representations such as **bag-of-words (BoW) and TF-IDF**. While computationally efficient, these approaches ignore word order and contextual relationships. Later, neural network models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) were introduced to capture local and sequential dependencies. However, RNN-based architectures suffer from limited parallelization and difficulty in modeling long-range dependencies.

The Transformer architecture, introduced by **Vaswani et al.**, addressed these limitations by replacing recurrence with self-attention mechanisms. Self-attention enables direct interaction between all tokens in a sequence, allowing the model to capture global contextual information efficiently. Motivated by these advantages, this paper explores the application of **an encoder-only Transformer** model for sentiment classification of **Uzbek-language reviews.**

The main contributions of this work are:

1. Implementation of an encoder-only Transformer trained from scratch for sentiment classification.

2. Construction of a BPE-based tokenizer tailored to the dataset.
3. Empirical evaluation and analysis of model performance and parameter efficiency.
4. Demonstration of attention-based interpretability for text classification.

2. Related Work

Traditional sentiment analysis approaches relied on manually engineered features and frequency-based representations such as bag-of-words and n-grams. While effective for simple tasks, these methods fail to capture semantic meaning and contextual dependencies.

Neural approaches introduced CNNs to capture local patterns and RNNs such as LSTMs and GRUs to model sequential information. Although RNNs improved performance, they are inherently sequential, limiting parallel computation and increasing training time.

The Transformer architecture eliminated recurrence by using self-attention, enabling efficient parallel processing and improved modeling of long-range dependencies. Encoder-only Transformer models, such as BERT, demonstrated strong performance across various NLP tasks, including sentiment analysis. Inspired by these architectures, this work focuses on a simplified encoder-only Transformer suitable for training from scratch on a domain-specific dataset.

3. Dataset and Preprocessing

3.1 Dataset Description

The dataset consists of Uzbek-language user reviews collected from an **Uzum marketplace** platform with approximately **350K customer reviews**. Each review is associated with a rating, which is mapped into three sentiment classes: negative, neutral, and positive. Reviews of **less than or equal to 80 characters** were filtered out to maintain a consistent input length and reduce computational overhead.

3.2 Text Normalization

Text preprocessing plays a critical role in improving model performance. The following normalization steps were applied:

- Removal of unsupported symbols and noise characters
- Retention of **Latin and Cyrillic characters, Russian, digits, and whitespace**
- Mapping of accented and special characters to standardized **Uzbek equivalents**
- Conversion of text into a cleaned, normalized form

This process ensures consistent tokenization and reduces vocabulary fragmentation.

3.3 Tokenization and Padding

A Byte Pair Encoding (BPE) tokenizer was trained from scratch on the cleaned dataset with a **vocabulary size of 30,000.** BPE allows effective handling of rare words and subword units, which is particularly useful for morphologically rich languages.

Each review was tokenized and padded or truncated to a fixed sequence length of **80 tokens.** Padding tokens were added to shorter sequences to ensure uniform input dimensions.

4. Model Architecture

4.1 Encoder Architecture

The proposed model follows an encoder-only Transformer architecture. Each input token is mapped to a dense vector using a trainable embedding layer. Positional embeddings are added to token embeddings to incorporate word order information.

The encoder consists of multiple identical blocks, each containing a multi-head self-attention layer followed by a position-wise feed-forward network. Residual connections and layer normalization are applied to stabilize training.

4.2 Self-Attention Mechanism

The self-attention mechanism computes contextual representations using the following formulation:

$$\mathbf{Attention(Q, K, V) = softmax((Q \cdot K^T) / \sqrt{d\square}\ ) \cdot V}$$

where **Q, K**, and **V** represent the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors. Multi-head attention allows the model to attend to information from different representation subspaces simultaneously.

4.3 Classification Head

After passing through the encoder layers, the token-level representations are aggregated using mean pooling across the sequence dimension. This produces a fixed-size sentence representation without introducing additional parameters. The pooled representation is then passed to a linear classification layer that outputs logits for the three sentiment classes.

4.4 Parameterization

Table 1 summarizes the number of trainable parameters in each component of the model.

Table 1. Model Parameter Breakdown

Components	**Parameters**
Embedding layer	1, 920, 000
Positional embedding	5, 120
Attention projections	65, 792
Feed-Forward block	132, 352
Classifier head	195
Total trainable	**2, 123, 459**

5. Experiments and Results

5.1 Training Setup

The model was trained for **10 epochs** using the AdamW optimizer with a learning rate of **3 x 10^-3.** Training was performed on a **GPU environment,** resulting in an average training time of **approximately 25 minutes.** The **batch size** was set to **64**.

5.2 Results

The training process showed stable convergence, with the final training **loss reaching 0.1891.** Although the evaluation focused primarily on loss minimization, qualitative inspection of predictions indicates reasonable sentiment classification behavior.

6. Discussion

One of the main advantages of the proposed model is the use of self-attention, which enables interpretability by examining attention weights. This allows analysis of which tokens contribute most to the model's predictions, an important feature for real-world applications.

However, several limitations remain, such as mean pooling, while efficient, may dilute important token-level information in some cases.

Future improvements could include incorporating a special classification token, using pretrained embeddings, or evaluating the model on additional datasets.

7. Conclusion

This paper presented an encoder-only Transformer model for sentiment classification of Uzbek-language user reviews. By leveraging self-attention and subword tokenization, the model effectively captures contextual information and achieves stable convergence. The results demonstrate that lightweight Transformer architectures trained from scratch can be successfully applied to sentiment analysis tasks in low-resource language settings. Future work will focus on improving evaluation metrics and enhancing model interpretability through attention visualization.

References

1. Vaswani, A., et al. *Attention Is All You Need*. NeurIPS, 2017.

2. Devlin, J., et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL, 2019.