Evaluating Transformer-Based Models for Sentiment Analysis in the Low-Resource Kazakh Language

Alyara Abilbashar and Almas Kenes

SDU University, Kaskelen, Kazakhstan

Abstract:
Sentiment analysis is now the key to understanding opinions on digital platforms. Languages with few resources, such as Kazakh, have not been studied as much in this area. Although transformer models do well in sentiment analysis, it is not clear how well they work for Kazakh sentiment analysis. We compare three transformer models to see how well they can classify sentiment in Kazakh texts. The models were trained and tested using KazSAnDRA, Kazakh sentiment dataset. We then looked at their performance using common metrics. The results show that performance differs between models and we point out which model works better in low-resource situations. The custom transformer-based model achieved the highest accuracy on the KazSAnDRA dataset, reaching 88.1%. This adds new data to Kazakh NLP research.

Keywords — **Sentiment analysis, Transformer-based models, Kazakh language, Low-resource language**

I. INTRODUCTION

Sentiment analysis (SA) is a rapidly growing field of Natural Language Processing (NLP), largely due to the dominance of social networking platforms. The automatic determination of sentiment in online reviews is now commonly used by marketers, companies, and political analysts [1]. Sentiment is typically evaluated as positive, negative, or neutral. Opinions play a crucial role in decision-making when choosing products or participating in elections. In the past, gathering opinions was a time-consuming and costly process involving polls, surveys, and focus groups, and individuals often depended on recommendations from friends. With the rise of social media, these traditional methods are no longer necessary, and SA has become essential as the volume of available data continues to grow [2].

Kazakhstan, as a developing country, contributes to computer science research, and NLP-particularly sentiment analysis—is an important field for business growth. However, because Kazakh is spoken by only about 11 million people, it is considered a low-resource language. This creates a need for further studies and development in this area.

As noted, Kazakh is a low-resource language in NLP. This research focuses on comparing and evaluating transformer models used for SA and trained on Kazakh dataset. Although there is a large body of research on transformer models for sentiment analysis, the uniqueness of this work lies in its focus on a low-resource language—Kazakh. The core problem we aim to address is determining which transformer model works best for Kazakh sentiment analysis. This study compares three transformer-based models on one Kazakh sentiment analysis dataset to identify which model generalizes better in low-resource settings.

The objective of the present study is to evaluate how well each model performs on the dataset, identify the best-generalizing model on the dataset.

The structure of the paper is organized as follows: Section 2 reviews related work, Section 3 describes the dataset and methodology, Section 4 presents the experimental results, and Section 5 discusses the findings and concludes the study.

II. LITERATURE REVIEW

This study briefly reviews the main research directions related to sentiment analysis, the importance of sentiment classification, and the challenges faced by low-resource languages. We particularly focus on the role of transformer-based models in these settings and on prior studies involving Kazakh sentiment datasets relevant to our work.

A systematic literature review by Yaning Mao and colleagues provides a comprehensive overview of sentiment analysis methods, applications, and emerging technologies. Their work demonstrates that sentiment analysis has wide applicability in diverse domains such as business intelligence, government analytics, healthcare, scientometric studies, and large language models. At the same time, the authors emphasize that despite rapid progress, sentiment analysis still faces several challenges and requires further methodological advances. NLP-based approaches to sentiment analysis are further discussed in the study "Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review." The authors define sentiment analysis as the task of identifying the emotional tone of text—typically positive, negative, or neutral—and highlight its importance in understanding user opinions, customer feedback, and overall experience across digital platforms[3].

Regarding low-resource languages, the paper "Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources" [4] outlines common strategies such as machine translation, word embeddings, and transfer learning. The authors distinguish between non-pretrained and pretrained (transformer-based) deep learning techniques and note that transfer learning and word embeddings are still widely used, especially when datasets are scarce. Social media remains the primary source of sentiment data. The paper also shows that transformer-based models are becoming increasingly prominent in low-resource sentiment analysis due to their strong contextual understanding.

Several studies have investigated sentiment analysis for the Kazakh language. One work on analyzing tourist reviews tested four models—TextBlob, VADER, Stanza, and LCF-BERT—and found that an ensemble method combining predictions achieved the best performance with an accuracy of 0.891 [5]. Another study focused on multi-level sentiment classification using five sentiment categories from –2 to +2. The authors incorporated Kazakh-specific morphological rules to determine sentence polarity and derive overall sentiment[6].

Because Kazakh is considered low resource, we also reviewed related work in other low-resource languages. For example, a study on Arabic sentiment analysis combined transformer-based embeddings with an LSTM architecture. The authors used AraBERT to capture contextual representations, then fed them into an LSTM followed by dense layers to model long-term dependencies, demonstrating strong performance in Arabic text classification[7].

Another relevant study is "Explainable Aspect-Based Sentiment Analysis Using Transformer Models," in which the authors trained several transformer models—BERT, RoBERTa, DistilBERT, and XLNet—on available datasets. They also applied five explainability methods (IME, SHAP, attention visualization, integrated gradients, and Grad-CAM) to demonstrate how transformer models make predictions, which is useful for understanding model behavior[8].

Transformer models are also common in sentiment analysis of other low-resource languages. For example, a study on Bengali (Bernouli) sentiment analysis evaluated five pretrained transformer models, including mBERT, BanglaBERT, Bangla-BERT-Base, DistilmBERT, and XLM-RoBERTa. The authors further introduced a transformer-ensemble model that achieved high performance, reaching 95.97% accuracy and 95.96% F1-score, outperforming previously published methods[9].

Finally, KazSandra [10] is the dataset used in our work. Created in 2022–2023, it contains reviews from Mapping and Market platforms that were manually checked and labeled by experts. The name KazSAnDRA stands for the Kazakh Sentiment Analysis Dataset of Reviews.

III. IMPLEMENTATION AND RESULTS

A. Dataset choice and preparation

In choosing the most suitable dataset for our work we decided to use review based dataset. The dataset is the KazSAnDRA (Kazakh Sentiment Analysis Dataset of Reviews and Attitudes) dataset that contains all we need for sentiment analysis.

KazSAnDRA dataset contains 180,064 entries from 4 different domain sources: Appstore (132,573 entries or 73.6%), Market (29,097 entries or 16%), Mapping (8,492 entries or 4.7%), and Bookstore (4,996 entries or 2.7%). All these entries are split inside the dataset into 3 groups: Train (140,126 _ entries), Test (17,516 entries), and Validation (17,516 entries). The groups are partitioned as 80/10/10 ration[10]. Each entry includes a custom review identifier (custom id), the original review text (text), the pre-processed review text (text cleaned), the corresponding review score (label), and the domain information (domain). For our project we used review text, pre-processed review text and the review score values. Here the main part lies in the review score value which shows the attitude of the review from 1 (lowest, worst attitude) to 5 (highest, best attitude). This specific field can be defined in different ways to differentiate which review is considered negative, positive or neutral. Review scores are as follows. 5 highest attitude (126,628 entries), 4 high attitude (11,394 entries), 3 neutral attitude (7,197 entries), 2 low attitude (4,900 entries), 1 lowest attitude (25,039 entries). The distribution is not good between groups, clearly suggesting some tuning.

B. Model implementation

For this study sentiment analysis was implemented using 3 different models. The first model is Rem- BERT model, which was pre-trained transformer model. The second model is also a pre-trained model XLM-RoBERTa, but it was trained on social media data. The third model was a custom transformer- based neural architecture build within Keras. These three approaches are quite different in a sense that RemBERT needed only some fine-tuning, whereas custom model needed more training and even though XLM-RoBERTa also needed only fine-tuning, it has a different background. Choosing these two specific types of models allows us to compare large-scale pretrained language models (RemBERT and XLM-RoBERTa) with a lightweight model trained from scratch on the same dataset that we used.

1) RemBERT: Pre-Trained Transformer Model

To understand current state of language models, this work uses the RemBERT model. Specific model that we used is open source variant that was released for Kazakh polarity classification on HuggingFace platform[10]. RemBERT as a transformer encoder was trained on large multilingual datasets by using masked language models. This model is specifically good for Kazakh language, because Kazakh language is a

morphologically rich language and it benefits from model's contextualized token representation which can capture complex semantic and syntactic relationships between words and phrases.

1.1) Tokenization and Preprocessing

RemBERT model has native WordPiece tokenizer that was used for this work to process text samples. All inputs were normalized, padded, truncated, lower- cased, and made into same fixed maximum sequence length. The tokenizer ouputs are as follows:

- input ids
- attention mask
- token type_ids

These output representations of sample text were fed into the RemBERT encoder for further work.

1.2) Fine-Tuning Strategy

The pre-trained RemBERT encoder was fine-tuned for the sentiment classification task. The task-specific classification head was placed with the encoder and optimized using the supervised learning. These are key fine-tuning choices: loss function, optimizer, training regime, output classes. They include sparse categorical cross-entropy, early stopping with full- encoder fine-tuning, Positive/Negative/Neutral labeling. RemBERT's general lingustic knowledge is fully adapted by fine-tuning on the domain-specific distribution of Kazakh user-generated textual reviews.

2) XLM-RoBERTa Sentiment Model: Pre-trained

XLM-RoBERTa model was used to introduce some variety to the work. This specific model was pre-trained on social media text that's why it has more to it with normal people speech than just reviews[12]. It was also optimized for multilingual analysis, tweet-based sentiment classification. Although it does not give the same comparison in review data with RemBERT and custom models, it shows more about analyzing general user-generated Kazakh text.

2.1) Pretrained Pipeline Setup

For this model HuggingFace pipeline API was used to ensure authenticity. The AutoTokenizer technology helps converting text into subword units, which is then tokenized by RoBERTa's SentencePiece-based tokenization action. Then classification was used to have sentiment labels output for each entry (positive, negative, neutral). After that the inputs were truncated and padded to have a maximum length of 128 tokens. For efficiency batch inference was performed. At the end the predicted sentiment label was added to the dataset.

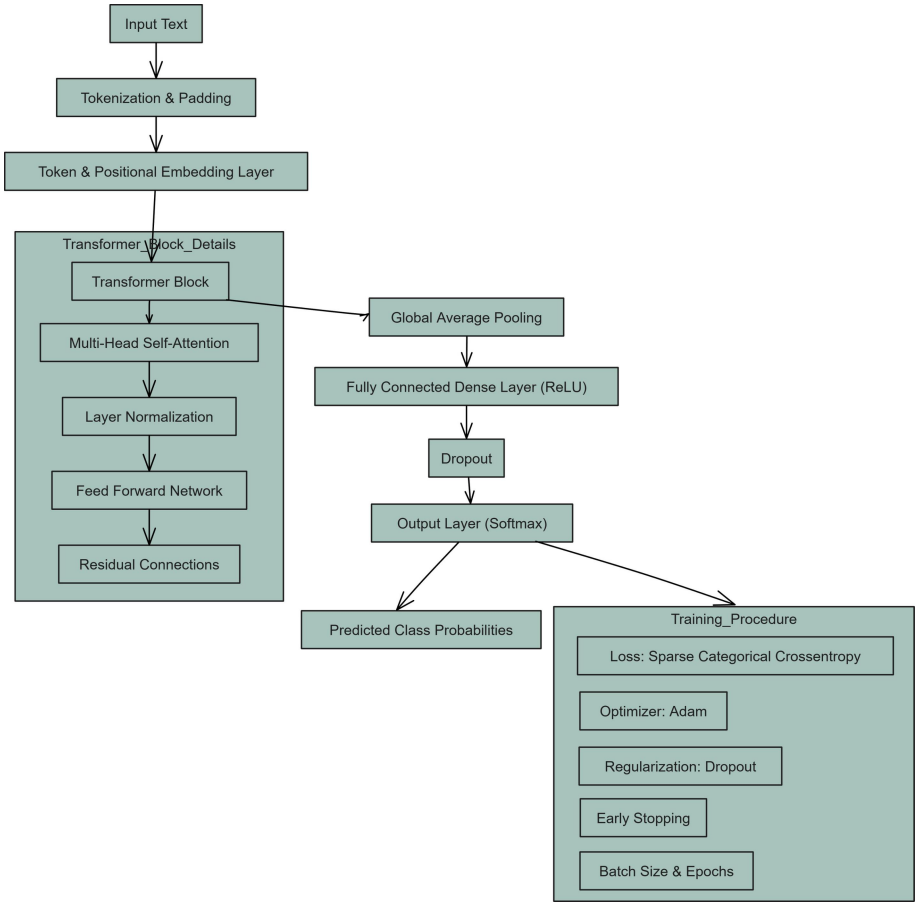

Figure 1: End-to-end flowchart of a Transformer-based text classification model

2.2) Sentiment Extraction

To produce some labels each text sample is passed through the sentiment pipeline. The labels are as follows: "positive", "neutral", "negative". The model output is suitable for short informal text and especially content resembling social-media language.

3) Custom Transformer-Based Model

The work implements a lightweight transformer model [1] by using Keras/Tenserflow. This model provides comparative baseline and helps evaluate the impact of large-scale pre-training before evaluating models on datasets. This specific model does not rely on external pre-trained weights like the RemBERT model, but rather learns task-specific representations from scratch. Above you can find Figure 1 - end-to-end flowchart of a Transformer-based text classification model in Keras, showing data flow from input text through embeddings, transformer block, and classification layers.
(Code available at url:
https://keras.io/examples/nlp/text_classification_with_transformer/)

3.1) Embedding Layer

The first layer of the model is the token embedding layer that maps word indices to dense vector representations of them. This combined with positional encoding ensures that sequence order is preserved.

3.2) Transformer Block

For this study a single transformer block is used, which consists of:

- Multi-Head Self-Attention
- Layer Normalization
- Feed-Forward Network
- Residual connections

With this design the computational power remains efficient compared to deep transformer stacking. Furthermore, it captures dependencies within input text sequences.

3.3) Classification Layers

After the transformer encoder was used global average pooling layer pulls the ouput representations from previous layer and aggregates all into a fixed-sized vector. This part contains two fully connected layers with drouput regularization. Then a softmax output layer follows that produces the final sentiment class probabilities.

3.4) Training Procedure

The model was trained using the following tools: loss, optimizers, batch size and epochs, regularization. They include sparse categorical cross-entropy, adam, dropouts, early stoppings. The given approach on the training gives insights on how much performance accuracy can be achieved without relying on large-scale pre-trainings.

C. Comparison of Approaches

The work uses both large-scale pre-trained models and lightweight model trained specifically for this task. This approach on the problem provides 2 unique perspectives. The work also integrates models specifically pre-trained for the given task of analyzing review texts, freshly trained model from scratch, and a model that was pre-trained for a different task.

RemBERT model evaluates how well a cutting-edge sentiment analysis model can perform for low resource language such as Kazakh language. It's pre-training gives a expectation of great performance and the specific model tuned for Kazakh sentiment classification provides more to the table.

Custom transformer-based model evaluates how a newly trained transformer model performs in the same conditions. This model trained on the same dataset it was being evaluated which does not gives a high expectation, but shows how similar models perform in such environment. XLM-RoBERTa model offers a different perspective on the low-resource language sentiment analysis. It is trained for a slightly different task of analyzing social-media content, but it was evaluated on reviews content. Both of them are user-generated, but have some slight differences. That is why it gives different expectations, but since social-media contents are more resourceful in general the model may show better performance in low-resource environments.

Considering both perspectives of pre-trained and freshly trained models, these implementations demonstrate not only transform models sentiment analysis results on Kazakh language, but tradeoffs between computational efficiency, model complexity and performance for low-resource language sentiment analysis.

IV. RESULTS

Below we can see the results of each implementation:
Table 1 summarizes the performance of the evaluated models on the KazSAnDRA dataset. All models achieve high accuracy in identifying positive sentiment. The Custom Transformer demonstrates competitive performance on both negative and positive classes; however, neutral class metrics are not reported due to the absence of neutral predictions. RemBERT exhibits stable performance for negative sentiment detection, while XLM-RoBERTa provides a more balanced, though less accurate, classification across all three sentiment classes. These results highlight the persistent difficulty of modeling neutral sentiment in low-resource Kazakh-language datasets.

Table 1: Performance comparison of sentiment classification models on the KazSAnDRA dataset

Model	Negative			Neutral			Positive		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
RemBERT (Pre-trained)	0.57	0.77	0.65	0.00	0.00	0.00	0.92	0.94	0.93
XLM-RoBERTa (Pre-trained)	0.34	0.47	0.39	0.10	0.49	0.17	0.95	0.57	0.71
Custom Transformer-based	0.69	0.60	0.64	–	–	–	0.92	0.94	0.93

Note: Neutral class metrics are not reported for the Custom Transformer model due to the absence of neutral samples in the predicted outputs.

Table 2: Overall Accuracy Agreement on Kazakh Sentiment Dataset

Dataset	RemBERT	XLM RoBERTa	Custom Transformer-based
KazSAnDRA	85.36%	54.85%	88.10

In Table 2, the overall accuracy agreement of the implemented models on the dataset is presented.

Considering all metrics, we conclude that the Custom Transformer-Based Keras Model performs the best among the evaluated models for Kazakh-language sentiment analysis.

Figure 2 illustrates the confusion matrix obtained on the validation dataset for the binary sentiment classification task (negative vs. positive). The model correctly identifies a large proportion of positive samples, with 13,000 true positives, indicating a strong ability to recognize positive sentiment. However, a noticeable number of false positives (1,198) and false negatives (801) is observed, suggesting that some negative reviews are misclassified as positive and vice versa. Despite these errors, the overall distribution indicates that the model maintains a reasonable balance between sensitivity and specificity, with a slight bias toward predicting the positive class, which can be attributed to class imbalance in the dataset.

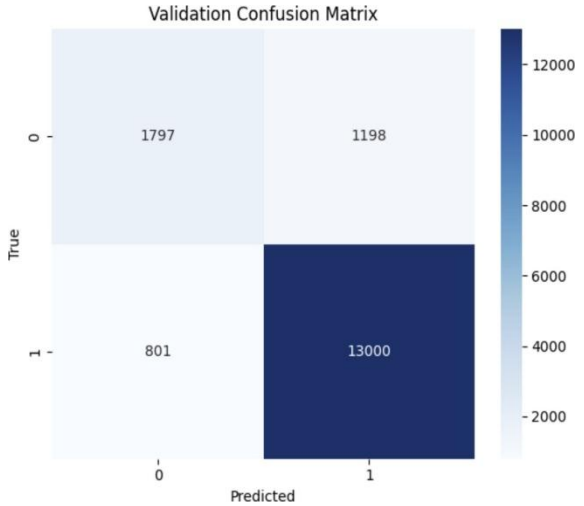

Figure 2: Confusion matrix Custom Transformer Keras model on KazSAnDRA dataset

Additionaly, the dataset itself shows a strong real-world sentiment skew. This particular dataset had 88% of samples labeled as positive, leaving only 12% being negative category. This characteristic highlights a broader challenge in a low-resource language sentiment analysis: the scarcity of balanced, high-quality inputs, and semantically rich datasets. Constructing such datasets s particularly difficult because of limited annotated corpora, high annotation costs, and the challenge of collectic authentic balanced amounts of positive, neutral and negative expressions from real people.

These findings suggest that, for low-resource languages, improving sentiment classification performance may require not only implementing data balancing techniques, but a longer textual inputs with richer contextual representations. Without sufficient input data and diverse sentiment coverage, models are prone to majority-class collapse regardless of dataset balancing efforts.

V. CONCLUSION

This study evaluated three different transformer based models: RemBERT, XML RoBERTa and a custom transformer-based model. These models were used for sentiment analysis in the low-resource Kazakh language using one dataset which is: KazSAnDRA, reviews based dataset. The results show that even though all models perform well consistently on positive sentiment because of class imbalance, predictions on negative sentiment remain challenging. This reflects broader problem observed in low-resource language sentiment analysis, limited annotated examples reducing model discriminative power.

In the dataset the custom Transformer-based model outperformed other models in overall performance, achieving high precision, recall, and F1-score, showing strong generalization despite being trained from scratch. RemBERT reliably performed best on negative sentiment, benefitting mostly from its multilingual pretraining. XML-RoBERTa, on the other hand, provided more balanced predictions overall, despite them being less accurate. This is probably due to its pretraining being on social media data, rather than reviews data like in RemBERT

pretraining.

The findings provided two key implications. First, a custom lightweight models trained on domain specific data can compete and outperform large pretrained models applied on low-resource languages. Second, dataset quality and class distribution inside it greatly influence model performance, particu- larly when it comes to predicting neutral sentiment.

Future research may improve class-balancing strategies, domain adaptive pretraining to enhance performance for sentiment categories in lower percentile. Supporting the development of robust NLP tools for low-resource languages by expanding available Kazakh language datasets.

REFERENCES

[1] Maite Taboada. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2(1):325–347, 2016.Bing Liu. *Sentiment analysis and opinion mining*. Springer Nature, 2022.

[2] Yanying Mao, Qun Liu, and Yu Zhang. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University-Computer and Information Sciences*, 36(4):102048, 2024.

[3] Yusuf Aliyu, Aliza Sarlan, Kamaluddeen Usman Danyaro, Abdullahi Sani BA Rahman, and Mujaheed Abdullahi. Sentiment analysis in low-resource settings: a comprehensive review of approaches, languages, and data sources. *IEEE Access*, 12:66883–66909, 2024.

[4] Banu Yergesh, Gulmira Bekmanova, and Altynbek Sharipbay. Sentiment analysis of kazakh text and their polarity. In *Web Intelligence*, volume 17, pages 9–15. SAGE Publications Sage UK: London, England, 2019.

[5] Aslanbek Murzakhmetov, Maxatbek Satymbekov, Arseniy Bapanov, and Nurbol Beisov. Sentiment analysis of tourist reviews about kazakhstan using a hybrid stacking ensemble approach. *Computation*, 13(10):240, 2025.

[6] Wael Alosaimi, Hager Saleh, Ali A Hamzah, Nora El-Rashidy, Abdullah Alharb, Ahmed Elaraby, and Sherif Mostafa. Arabbert-lstm: improving arabic sentiment analysis based on transformer model and long short-term memory. *Frontiers in Artificial Intelligence*, 7:1408845, 2024.

[7] Isidoros Perikos and Athanasios Diamantopoulos. Explainable aspect-based sentiment analysis using transformer models. *Big Data and Cognitive Computing*, 8(11):141, 2024.

[8] Md Nesarul Hoque, Umme Salma, Md Jamal Uddin, Md Martuza Ahamad, and Sakifa Aktar. Exploring transformer models in the sentiment analysis task for the under-resource bengali language. *Natural Language Processing Journal*, 8:100091, 2024.

[9] Rustem Yeshpanov and Huseyin Atakan Varol. Kazsandra: Kazakh sentiment analysis dataset of reviews and attitudes. *arXiv preprint arXiv:2403.19335*, 2024.

[10] Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 258–266, 2022.