KazBERT: A Custom BERT Model for the Kazakh Language

with Multilingual Pretraining and Custom Tokenizer Pipelines

Gainulla Eraly

March 30, 2025

Abstract

In this paper, we introduce **KazBERT**, a BERT-based language model specifically tailored for the Kazakh language. Given the scarcity of resources for low-resource languages, we combine Kazakh, English, and Russian datasets, develop a custom WordPiece tokenizer, and fine-tune a BERT-base model using a Masked Language Modeling (MLM) objective. The model and its preprocessing pipelines are fully integrated into the Hugging Face ecosystem (`Eraly-ml/KazBERT`).

1 Introduction

While mBERT remains a strong baseline for Kazakh, it has not been specifically fine-tuned for the language. Kazakh-specific tokenization issues and the relatively small proportion of Kazakh text in mBERT's[5] pretraining data may limit its performance. Similarly, XLM-R[4] has not been specifically evaluated for Kazakh in this study. To explore language-specific improvements, we introduce KazBERT, a BERT-based model trained on a dedicated Kazakh corpus alongside English and Russian texts. Our contributions:

- Combining diverse datasets from different sources.Combining diverse datasets from different sources.
- Training a custom WordPiece tokenizer optimized for Kazakh,
- Fine-tuning a BERT model using the MLM objective.

2 Data Collection and Preprocessing

Our corpus includes:

1. **Kazakh Data**: Extracted from Wikipedia [3],
2. **English Data**: From a curated Parquet file [1],
3. **Russian Data**: Cleaned JSON-lines file of Wikipedia [2].

Data is loaded with `pandas`, merged, and split (80% train, 20% validation).

3 Custom Tokenizer Training

We train a WordPiece tokenizer (vocabulary: 30,000 tokens) on `valid.txt`. The tokenizer is used in all training and inference steps [7].

4 Experiments and Results

4.1 Model Fine-Tuning

Fine-tuning a pre-trained `bert-base-uncased` model with MLM objective:

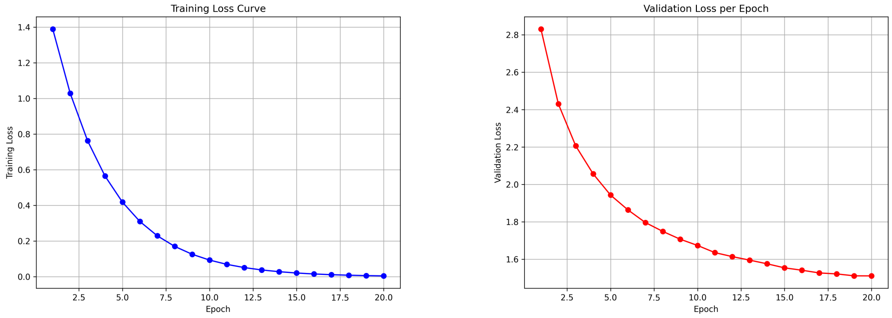
```
from transformers import TrainingArguments

training_args = TrainingArguments(
    output_dir="./results",
    evaluation_strategy="epoch",
    save_strategy="no",
    logging_strategy="epoch",
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=20,
    weight_decay=0.01,
    fp16=True,
    logging_dir="./logs",
    report_to=[]  # Disabled logging to external services
)
```

Listing 1: Model Fine-Tuning Code Snippet

For full training script look here (`Eraly-ml/KazBERT/script.py`).

4.2 Training and Validation Loss Curves



(a) Training Loss Curve over epochs.

(b) Validation Loss Curve over epochs.

Figure 1: Training and validation loss curves over epochs.

5 Evaluation

Perplexity (PPL) is computed as:

$$PPL = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(\hat{w}_i \mid \text{context})\right)$$

Table 1 compares KazBERT to other models.

Table 1: Perplexity Comparison

Model	**500**	**1000**	**1500**	**2000**
KazBERT	3.06	3.39	3.14	3.27
kaz_legal_bert	9.32	9.72	8.92	9.18
mBERT	2.16	2.29	2.13	2.29

6 Future Work

Future directions:

- Evaluating KazBERT on Named Entity Recognition (NER),
- Benchmarking long-context reasoning,

- Extending safety and bias metrics.

7 Conclusion

We present **KazBERT**, a BERT-based model tailored for the Kazakh language, leveraging a custom tokenizer and fine-tuned on a multilingual corpus.

Acknowledgements

We acknowledge contributions from Hugging Face Transformers [6] and Wikipedia contributors [1, 3, 2].

References

[1] English text corpus. Available from internal resources. Details available in the repository documentation.

[2] Russian cleared wikipedia. https://huggingface.co/datasets/Den4ikAI/russian_cleared_wikipedia. Accessed: 2025-03-28.

[3] Amandyk. Kazakh wiki articles dataset. https://huggingface.co/datasets/amandyk/kazakh_wiki_articles. Accessed: 2025-03-28.

[4] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2020.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[6] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clément Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2020.

[7] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maksim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system:

Bridging the gap between human and machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.