NLP Kaggle Competition: Text Classifier Development

Andris Oueslati, Marius Boucaut, Yuxian Zuo
CentraleSupélec
{andris.oueslati, marius.boucaut, yuxian.zuo}@student-cs.fr

Abstract

In this project, we fine-tuned mBERT for text classification. The model achieved an accuracy of 85.48%. Despite strong overall performance, the model exhibits challenges with certain underrepresented classes. Future work will focus on hyperparameter tuning and data augmentation to enhance robustness.

1 Introduction

This project aims to develop a high-accuracy text classifier capable of handling multilingual data. To achieve this, we utilize the pretrained multilingual BERT (mBERT) model, which has been trained on over 100 languages and leverages deep contextualized embeddings to generalize effectively across different languages (Devlin et al., 2019). Benchmarks have demonstrated that mBERT performs competitively against state-of-the-art techniques for multilingual text classification (Ahmad et al., 2021; Hu et al., 2020). However, challenges arise due to the imbalance in class distribution within the dataset. The goal of this work is to fine-tune mBERT to enhance classification performance across multiple languages while addressing data imbalance issues.

2 Solution

2.1 Data Exploration and Cleaning

The dataset consists of 190,099 samples and contains 389 unique labels after removing entries with missing labels. A total of 76 duplicate texts were identified, meaning the same text appeared with different labels. This suggests that certain texts may belong to multiple classes, so they were retained.

The dataset is highly imbalanced, with the most frequent class containing 1,500 samples, while some classes have only a single instance. To ensure that each label has reasonable representation in both the training and validation sets, labels with fewer than five instances were filtered out.

Since the dataset is multilingual, an automatic language identification model was applied to detect the language of each text (Joulin et al., 2016). A total of 174 languages were identified, with English and Spanish being the most common, accounting for 13% and 7% of the dataset, respectively.

2.2 Dataset Preparation

A stratified split was applied to divide the dataset into 80% training and 20% validation sets.

The tokenizer of the pretrained mBERT model supports a maximum sequence length of 512. After tokenizing all texts, the length distribution, shown in Figure 1, reveals that most texts are within 256 tokens. Texts exceeding 512 tokens were truncated. To optimize efficiency while preserving information, we set the maximum tokenization length to 256 for both the training and validation datasets.

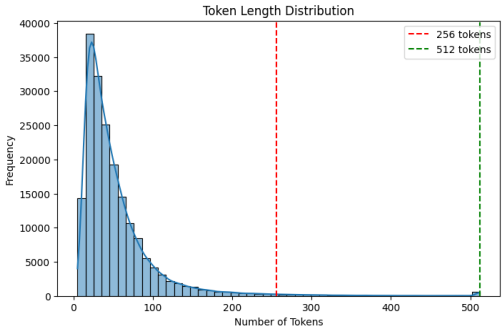

Figure 1: Distribution of tokenized text lengths.

2.3 Training Process

In this project, we selected the BERT-base multilingual cased model, which includes a classification head at the final layer, making it well-suited for our text classification task. We fine-tuned the entire model, which consists of approximately 178 million parameters.

To address class imbalance in the dataset, we computed class weights based on the inverse frequency of each class and incorporated them into

the cross-entropy loss function. This ensured that higher weights were assigned to less frequent labels, preventing the model from being biased toward dominant classes.

For optimization, we used AdamW, a variant of the Adam optimizer designed to improve weight decay regularization. The learning rate was set to 5×10^{-5}, and a batch size of 16 was chosen to balance memory efficiency and training stability.

After each epoch, we evaluated the model's accuracy on the validation set. We implemented checkpoints to save progress. Figure 2 illustrates the validation loss and accuracy curves over five epochs. During the fourth epoch, we observed a slight increase in validation loss and a drop in accuracy, which may indicate overfitting. Based on this observation, we selected the model trained for three epochs as the best-performing version, achieving an accuracy of 85.67% on the validation set.

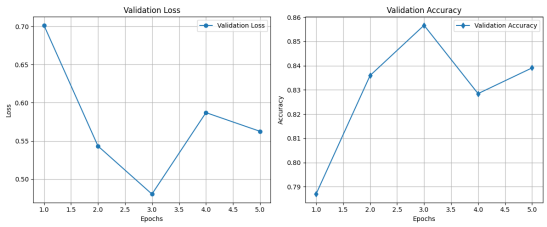

Figure 2: Validation loss and accuracy over five epochs.

3 Results and Analysis

3.1 Performance Metrics

The model achieved a precision of 85.32%, a recall of 86.04%, and an F1-score of 85.33%. These results demonstrate strong classification performance across multiple classes, achieving a well-balanced trade-off between precision and recall. Additionally, the model attained an accuracy of 85.48% in the Kaggle competition.

3.2 Classification Report

Table 1 presents the classification report for the top 5 classes. The results indicate that the model achieves strong performance across most classes. However, certain classes, such as tgk, exhibit lower recall.

3.3 Confusion Matrix

We visualized the confusion matrix for the top 10 classes, as shown in Figure 3. For certain classes, such as tgk, a significant number of samples were misclassified into other categories, indicating potential challenges in distinguishing these classes.

Class	Precision	Recall	F1-score
tgk	0.98	0.33	0.49
bak	0.79	0.76	0.78
tat	0.86	0.67	0.75
crh	0.99	0.90	0.94
srp	1.00	0.62	0.77
Weighted Avg	0.93	0.63	0.72

Table 1: Classification report for the top 5 classes, displaying precision, recall, and F1-score.

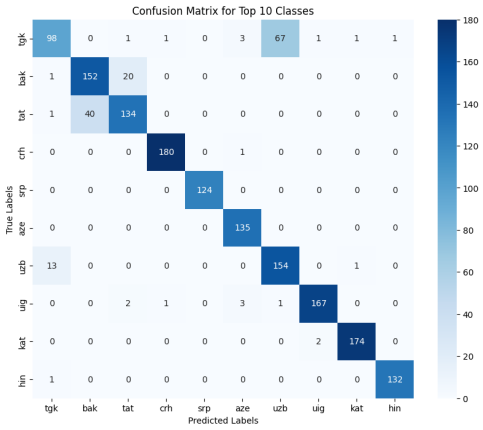

Figure 3: Confusion matrix for the top 10 classes, showing the distribution of correctly and incorrectly classified instances.

4 Conclusion and Future Work

In this work, we fine-tuned a multilingual BERT-based model for text classification, achieving strong performance with an accuracy of 85.48% in the Kaggle competition while maintaining a balance between precision and recall.

4.1 Limitations

Despite its effectiveness, the model has limitations. Its large size makes deployment challenging in resource-constrained environments. Additionally, like other transformer models, mBERT may inherit biases from the training data, potentially impacting classification fairness.

4.2 Future Work

Future improvements include hyperparameter tuning, though fine-tuning remains computationally expensive. Exploring alternative models like XLM-R or DistilBERT may improve efficiency, while data augmentation techniques could mitigate class imbalance and enhance robustness.

References

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual BERT for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.