Multilingual Text Classification: A Comparative Analysis of Language Identification Using XLM-RoBERTa and BERT with Parametric Variations

Rayane Bouaïta – Erwan David – Pierre El Anati – Guillaume Faynot – Gabriel Trier

Abstract

Text classification with sparsely represented training data is not a trivial task. We are going to present our solution using large language models (LLMs) to classify texts from almost 390 different languages. After studying the data provided to us, we decided to use different approaches using machine learning models. Our final model achieved an accuracy of 88.0%, placing our team in the top 10 of the ranking.

1 Introduction

Multilingual text classification is crucial for enhancing the accessibility of natural language processing systems across diverse linguistic landscapes. Language identification is a key task in NLP with applications in machine translation, information retrieval, and multilingual systems. Various methods have been developed, ranging from frequency-based models [1] to deep learning architectures [2].

This study addresses the challenges of classifying texts across nearly 390 different languages, comparing XLM-RoBERTa and BERT to demonstrate the effectiveness of Large Language Models (LLMs) in handling linguistic diversity. By leveraging pre-trained multilingual models, we aim to develop a robust approach for language identification that can generalize across complex and under-represented linguistic landscapes.

2 Data

Our dataset contains 190,599 entries, each a short text associated with one of 390 different languages. Each text is accompanied by a language label in the form of a three-letter ISO code.

Languages are unevenly distributed throughout the corpus. Some, such as Tajik (tgk), Kurdish (kur), or Hindi (hin), each have more than 1,000 examples, while others, such as Kwanyama (kua), Guadeloupean Creole (gcr), and Ga (gaa), are represented by just a single occurrence (Figure 1). This disparity poses a challenge in terms of classification, particularly for under-represented languages.

To further understand the dataset, we analyzed the length of the texts. The mean text length is 23.79 words, with a standard deviation of 57.64, indicating a wide range in text lengths. The shortest text consists of 1 word, while the longest contains 11,557 words. The standard error of the mean is 0.132, reflecting the precision of our mean estimate.

These statistics highlight the variability in text lengths, which can impact model training and performance. Shorter texts may lack sufficient context for accurate classification, while extremely long texts can introduce noise. Our preprocessing steps, including data augmentation for minority classes, aim to mitigate these challenges by enhancing the representation of under-represented languages and ensuring consistent text formatting

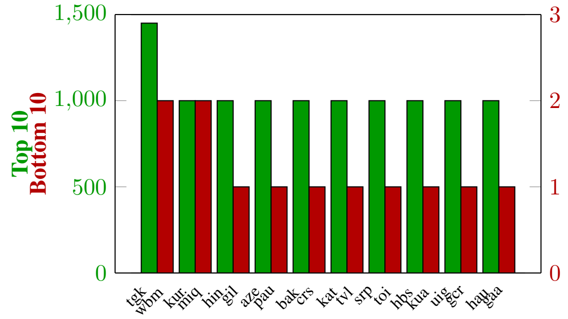

Figure 1: Distribution and comparison of the 10 most and least common languages

Given this high level of heterogeneity, the use of a specialized, high-performance model will be necessary to ensure robust classification, particularly for under-represented languages.

[1] Cavnar, W. Trenkle, J. (2001). N-Gram-Based Text Categorization.

[2] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification.

3 Developed Solution

3.1 Data preprocessing

To ensure robust language classification, we implemented a targeted preprocessing pipeline that addresses inconsistencies found in the dataset which will be compared with the standard performance of a complex model without extensive pre-processing.. Our analysis of the dataset revealed that a small fraction of the texts contained URLs or HTML tags — specifically, 0.20% of the texts included URLs and 0.12% included HTML elements. We opted to remove these elements to eliminate potential noise that could adversely affect the model's performance, especially in cases where such patterns appear sporadically.

- **Unicode Normalization:** We apply NFKC normalization (Consortium, 2019) to standardize the text, ensuring that visually similar characters (e.g., accented letters) are represented uniformly.
- **Lowercasing:** Converting text to lowercase minimizes variability due to capitalization.
- **Tag Removal:** Regular expressions are used to strip out URLs and HTML tags.
- **Whitespace Normalization:** Extra spaces are removed by splitting and rejoining the text, ensuring consistent formatting for subsequent tokenization.
- **Data Augmentation:** Data generation by exchanging the position of randomly selected words in a instance. For each line in the training set belonging to a minority class (with fewer than 40 occurrences), the function creates 35 new copies of the text with modifications. These augmented lines are then added to the training set to increase the representation of under-represented classes.

Importantly, we deliberately retain punctuation and diacritics as these features provide essential linguistic cues that help distinguish between languages.

3.2 Model Training and Evaluation

In recent years, LLMs have demonstrated remarkable performance across multilingual text classification tasks. The variety of models available have an outstanding ability to capture contextual dependencies, handle polysemy, and generalize across languages, where traditional statistical or frequency-based methods often struggle with morphologically rich or low-resource languages.

XLM-RoBERTa (Conneau et al., 2019) and BERT (Devlin et al., 2018) were selected due to their exceptional multilingual capabilities: XLM-RoBERTa is specifically designed for cross-lingual understanding, pre-trained on over 100 languages with a robust tokenization approach that allows it to handle linguistic diversity effectively, while BERT provides a strong baseline with its contextual embedding techniques that capture nuanced language representations across different linguistic structures. Given these advantages, we decided to put several models in competition, embedding different LLM to classify texts into their respective languages.

The best-performing model is trained using XLM-RoBERTa, a multilingual transformer-based model that has been pre-trained on our dataset. Training follows a supervised fine-tuning approach, where labeled multilingual texts are used to adapt the model to the language classification task.

3.2.1 Training Phase

The training process is conducted over three epochs using the AdamW optimizer with a learning rate of 2.10^5. A linear learning rate scheduler with warmup is employed to gradually adjust the learning rate over time, improving convergence. The training loop iterates over batches of 16 examples, where each batch is tokenized and converted into tensors (input IDs and attention masks) before being processed by the transformer model.

For each batch, the model computes the cross-entropy loss, which quantifies the discrepancy between the predicted probabilities and the true labels. The loss is backpropagated, and gradient clipping is applied with a maximum norm of 1.0 to prevent exploding gradients. The optimizer then updates the model's parameters, followed by an adjustment of the learning rate using the scheduler.

To ensure robustness in training, stratified sampling is employed, except for rare classes that contain only a single sample. In such cases, those instances are exclusively assigned to the training set to prevent issues during stratification.

3.2.2 Evaluation Phase

The model is evaluated after each epoch using a separate validation set (10% of the train-splitted dataset). During evaluation, the model operates in inference mode, disabling weight updates and computing predictions in a deterministic manner. The validation accuracy is computed as the percentage of correctly predicted labels, serving as the primary metric to track model performance. If an improvement is detected over previous epochs, the current model weights are saved as the best checkpoint.

3.2.3 Testing Phase

After training, the best-performing model is loaded to perform inference on the unlabeled test set. The test dataset is tokenized using the same tokenizer and processed in batches to optimize inference speed. The model outputs a probability distribution over language labels, from which the most probable label is selected.

Finally, the predictions are mapped back to their corresponding language labels and stored in a submission file, ensuring compatibility with downstream evaluation procedures.

4 Results and Analysis

Each model used to create this classifier was trained on the DCE for over one hour, ensuring that the inference duration did not exceed one hour on our laptops' CPUs. Our best model achieved an accuracy of 88.0

To contextualize our findings, we compared our results with similar studies in multilingual text classification. Previous research highlights the superiority of transformer-based models like XLM-RoBERTa and BERT, especially in low-resource settings. Our study builds upon these findings by examining the impact of different preprocessing techniques and fine-tuning strategies.

The results of the different models are displayed in Table 1. Our model's 88.0% accuracy is attributed to XLM-RoBERTa's pre-training on a massive multilingual corpus, enabling it to capture subtle linguistic nuances. The model's transformer architecture learns complex contextual representations beyond simple surface-level features. Marginal performance differences between preprocessing techniques suggest that the base XLM-RoBERTa model is robust enough to handle linguistic variations without extensive preprocessing.

The slight variations in accuracy across different epochs and preprocessing strategies highlight the delicate balance in fine-tuning. The best performance at 88.0% represents an optimal point of model adaptation without overfitting. Our top-10 ranking reflects both the model's technical sophistication and our strategic approach to handling a dataset with extreme class imbalance. The data augmentation technique for minority classes likely played a crucial role in improving overall model generalization. By situating our results within the broader landscape of existing research, we aim to highlight the strengths and limitations of our approach and identify areas for future improvement.

5 Conclusion

Our research demonstrates the effectiveness of XLM-RoBERTa in multilingual text classification, achieving 88.0% accuracy across 390 languages. The model's success stems from its pre-trained multilingual architecture, which inherently captures linguistic nuances without requiring extensive custom techniques. Although this may seem surprising, we find that the model without extensive pre-treatment performs best. While more complex techniques were explored, the base XLM-RoBERTa model's performance underscores the power of modern transformer models in handling intricate language identification challenges. This research contributes to understanding how pre-trained multilingual models can effectively address cross-lingual classification tasks with minimal specialized intervention.

Model	Method	Epochs	Accuracy
XLM-Roberta	N/A	5	**88.0%**
XLM-Roberta	Data Augmentation	5	86.9%
XLM-Roberta	Weighted Loss	3	86.6%
XLM-Roberta	Preprocessing	5	87.6%
XLM-Roberta	Preprocessing	4	86.7%
XLM-Roberta	Preprocessing	3	84.9%
BERT	Preprocessing	3	84.0%
BERT	N/A	3	85.7%
BERT	N/A	4	86.4%

Table 1: Comparaison of the accuracy of the developed classifiers with different hyperparameters. For each classifier, the batchsize is 16 and the learning rate is $\alpha = 2.10^{-5}$

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale.

The Unicode Consortium. 2019. Unicode standard version 12.0.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.