Multilingual Natural Language Processing - Homework 1

Group Name: teXt-Men

Joshua Edwin
vijayanrajendraedwin.1995743@studenti.uniroma1.it
Clemens Kubach
clemens.kubach@gmail.com

1 Introduction

In this project, we have devised two different approaches to the classification of cultural entities: A LM-based and a Rule-Based classifier. The question is whether a certain entity is culturally agnostic, representative, or exclusive.

2 Methodology

2.1 LM-based Approach

For this approach, we consider different BERT variants under the hypothesis that multilingual pretraining is beneficial for the performance in a downstream task with cultural reference. We hypothesized that a LM pretrained on text from a variety of natural languages would have more cultural knowledge because language and culture are closely intertwined. Inspired by this hypothesis, we fine-tuned two[1] pretrained LMs on our downstream task: "BERT" (Conneau et al., 2019) and "Multilingual BERT" (Pires et al., 2019).

Investigating the splits of the given dataset, we have discovered that the two subsets must originate from different data distributions. The frequencies of the three classes are very different between the two subsets, as can be seen in the two histograms in fig. 1 and fig. 2. Furthermore, the classes are not balanced within the respective subsets, but with a standard deviation of 7% and 6% in the train and val subsets, there is no strong imbalance. However, in order to counteract a potential overweighting of the majority class in training on demand, we have written a function for oversampling the other classes. We decided for oversampling for class-aware sampling, as this allows all the available information to be used during training, whereas undersampling with our already not very large datasets could result in too many important samples being omitted. In addition to balancing the train dataset, we also implemented and investigated a method to resample it to the sample class distribution of the validation set. The validation set is explicitly never changed or resampled to retain representativity of the true and test class distributions.

We used the Huggingface transformers and dataset libraries for accessing the data and pretrained models from Huggingface Hub, and fine-tuning the model on the given data for our downstream task. We used the two named BART models as the sequence classifier, consisting of an additional classification head. For reproducibility, we set a fixed random seed. We are using model checkpointing after each epoch to be able to return the best model checkpoint evaluated on the validation subset. We are not using early stopping mainly because of the potential of the double descent effect (Belkin et al., 2019), where the validation loss may initially increase after early improvement but then improves significantly with continued training. However, in our small experimental training, we did not see this phenomenon clearly. Despite that, model checkpointing enabled leveraging improvements made in later training stages. To counteract overfitting, we applied dropout as it is default set in the BART architecture but also applied a dropout rate of 0.4 to the classifier head.

We performed preprocessing by applying the tokenizer associated with the pre-trained model to an aggregated textual representation of each sample. This representation was created by concatenating attribute names with their corresponding content. We evaluated two input configurations: one using all available columns from the training set - "name", "type", "category", "subcategory", and "description" - and another using only the "name" and "description" columns.

[1] Note see appendix A.1

2.2 Non-LM-based Approach

Our non-LM-based system is a hybrid of a deterministic rule-based classifier and a lightweight ML fallback model. It operates entirely on structured Wikidata metadata (Lang et al., 2023) and shallow textual features, designed to be transparent, explainable, and efficient.

The core classification relies on three metadata fields: `heritage_status`, `part_of_culture`, and `instance_of`. Each item is assigned a label based on a priority-driven logic: matches in `heritage_status` lead to a *Cultural Exclusive* label; otherwise, matches in `part_of_culture` result in *Cultural Representative*; failing that, `instance_of` matches are checked for either of the above. If no relevant QIDs are found, the item defaults to *Cultural Agnostic*. This priority-based resolution ensures that high-significance cultural markers override broader class information.

To implement this logic, we created three manually curated "golden" QID lists after multiple phases of enrichment and diagnostics. Initially, we selected culturally meaningful QIDs from correctly classified samples. We then ran error analysis on validation misclassifications to extract frequent but mispredicted QIDs and evaluated them manually using Wikidata context. Overly generic or ambiguous entities were filtered out. This refinement substantially increased the rule-based classifier's F1-score on the validation set to 0.80.

For example, Q1935974 (*UNESCO Intangible Cultural Heritage*) appearing in `heritage_status` triggers a *Cultural Exclusive* label, while Q5043159 (*Indian culture*) in `part_of_culture` leads to *Cultural Representative*. Similarly, Q1107679 (*Kerala-style Hindu temple architecture*) and Q4830453 (*traditional occupation*) are used to classify samples as *Exclusive* and *Representative* respectively when found in `instance_of`.

Despite its high precision, the rule-based model exhibited conservative behavior, often defaulting to *Cultural Agnostic* for edge cases. To address this, we integrated a fallback classifier—a logistic regression model trained on TF-IDF vectors of item names and descriptions. This fallback re-evaluates only those samples initially labeled as agnostic, provided that text data is available. These updated predictions are flagged as `ML` source, while confident rule-based predictions retain a `QID` source.

Finally, the hybrid classifier combines both components: deterministic predictions where confident, and ML-driven correction where metadata is insufficient. This structure strikes a balance between interpretability and broader coverage, yielding a more robust final system on unseen test data.

3 Experiments

3.1 LM-based Approach

We did manual hyperparameter tuning for the learning rate, dropout rate, and the selected columns. As found in our early experiments, using balanced or validation set class-aware sampling is not beneficial in our case. To reduce the number of final experiments, we not further applied them. The same applies to some experiments in which selected parts of our LM-based classifier are frozen at different training epochs to counteract overfitting. For example, we tried to train only the classifier head for the first epochs while the BART part was frozen.

Due to the limited space in the report, we will just look at two experiments, using BERT-base-cased and Multilingual-BERT-base-cased. Both using an AdamW optimizer with a linear learning rate schedule starting at $5e-6$, weight decay 0.01, training using all available columns, and a batch size of 32. The training history is logged via "Weights and Biases" and the best model checkpoint is pushed to the HF Hub.

3.2 Non-LM-based Approach

We conducted iterative experimentation to optimize our rule-based classifier and its integration with a TF-IDF-based logistic regression fallback model. The initial phase focused on manually curating three "golden" QID lists through error analysis and diagnostics on misclassified samples. This yielded a rule-based system that achieved an F1-score of 0.80 on the validation set.

To improve coverage, particularly for items initially labeled as *Cultural Agnostic*, we trained a logistic regression model using TF-IDF features extracted from concatenated item names and descriptions. The model was trained on the enriched and labeled training dataset. Only those items with empty or ambiguous structured metadata but non-empty textual fields were passed to the fallback model.

This hybrid approach—prioritizing rule-based predictions and using the ML fallback selectively—enabled us to resolve previously unclassified samples and correct rule-based mismatches. All predictions were tagged with their origin (`QID`

or ML).

4 Results

4.1 LM-based Approach

Multilingual BERT performs better on our cultural downstream classification task, suggesting the benefits of a multilingual pre-training for transferring cultural knowledge to our classification task. However, this study should not be understood as claiming representativeness, rather as motivation. For a more detailed and complete investigation of our hypothesis, the use of different models would be useful but is beyond the scope of this project. We evaluated the results using various common classification metrics for both models, BERT and mBERT can be seen in table 1. The final model selection from the set of checkpoints is done using the F1-Score on the validation set. We used our fine-tuned multilingual BERT model as requested transformer-based model for the homework and to predict the labels for the given test dataset.

4.2 Non-LM-based Approach

Our rule-based classifier achieved strong performance using structured metadata signals from Wikidata. When tested on the validation set, it produced an F1-score of 0.80 as shown in table 2, with especially high precision on the *Cultural Agnostic* class. The system was later extended into a hybrid classifier by integrating a logistic regression fallback model trained on name and description text fields for cases where the rule-based logic predicted *Cultural Agnostic*. This hybrid model resulted in a modest drop in overall accuracy but improved recall in the minority classes. The class-wise F1-Scores for the rule-based and hybrid non-LM approaches can be found in table 3.

While the hybrid approach improves recall for *Cultural Exclusive* and *Cultural Representative* classes, it introduces overcorrection that reduces the precision of these classes, particularly for *Cultural Exclusive*. Notably, the hybrid model predicts 101 instances as *Cultural Representative* and 102 as *Cultural Exclusive*, a shift from the rule-based model's more conservative predictions.

Comparison between the Systems. The LM-based models, particularly Multilingual BERT, achieve strong F1-scores (up to 0.792) thanks to deep contextual embeddings and multilingual pre-training, making them robust to unseen or ambiguous inputs. Their strength lies in generalization and flexibility in capturing implicit cultural cues.

In contrast, the rule-based classifier, grounded in structured Wikidata properties, provides full interpretability and strong performance (F1-score 0.80) without high computational demands. Its precision-oriented logic, however, may miss culturally relevant items not covered by curated QIDs.

To mitigate this, we introduced a hybrid system that applies a logistic regression fallback when rule-based logic labels an item as *Cultural Agnostic*. This improved coverage while preserving rule-based precision, achieving an F1-score of 0.75.

Overall, LM-based and non-LM-based approaches serve complementary roles offering either broader generalization or interpretable control depending on task priorities. Our work suggests benefits of future work on an approach that combines text-based features by using language models with constraining rules derived from the Wikidata knowledge graph.

References

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

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. *arXiv preprint arXiv:1911.02116*.

Felix Lang, Nikos Mavridis, and Ujwal Gadiraju. 2023. Link topics from q&a platforms using wikidata: A tool for cross-lingual and cross-platform topic classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

A Appendix

A.1 Further Models

In addition, we performed some further experimental fine-tuning with other LMs pretrained on English and multilingual corpora, such as BERT large, XLM RoBERTa (Conneau et al., 2019) base and large. To get those model trainings running, we have looked into the area of training under memory

constraints and used i.e. mixed precision training and gradient checkpointing. However, since we were able to quickly achieve good results with our BERT models with low computational capacities, and since the same architecture but different pretraining corpora were practical, we concentrated on these two BERT models.

A.2 Fallback Model and Extensions

During development, we explored several configurations of the fallback system, including alternative machine learning models (e.g., Random Forest, SVM), additional engineered features (such as instance and culture counts), and confidence-based thresholding. Ultimately, we selected logistic regression due to its strong performance on sparse textual data, low computational overhead, and high reproducibility across platforms like Google Colab. While other models showed marginal gains, logistic regression provided the best trade-off between effectiveness, simplicity, and deployment reliability for our task.

This hybrid integration was key to improving overall coverage while preserving the transparency of rule-based logic.

A.3 Figures and Tables

	BERT	mBERT
F1-Score	0.778	0.792
Accuracy	0.780	0.793
Precision	0.778	0.791
Recall	0.780	0.793

Table 1: Comparison of model performance across different methods. Rules means here the rule-based classifier.

	RBC	Hybrid
F1-Score	0.800	0.757
Accuracy	0.800	0.760
Precision	0.810	0.800
Recall	0.800	0.760

Table 2: Performance of the rule-based and hybrid classifiers on the validation set.

Class	F1 (RBC)	F1 (Hybrid)
C. Agnostic	0.83	0.74
C. Exclusive	0.79	0.74
C. Representative	0.78	0.78

Table 3: Class-wise F1-score comparison between rule-based and hybrid models.

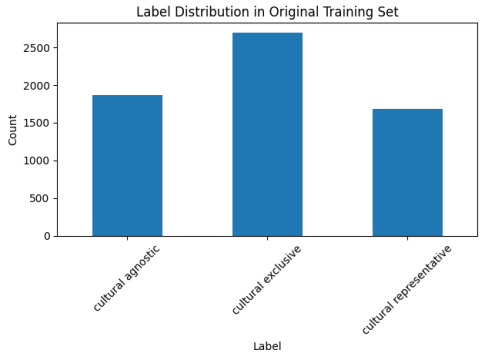

Figure 1: Class distribution of the original training dataset.

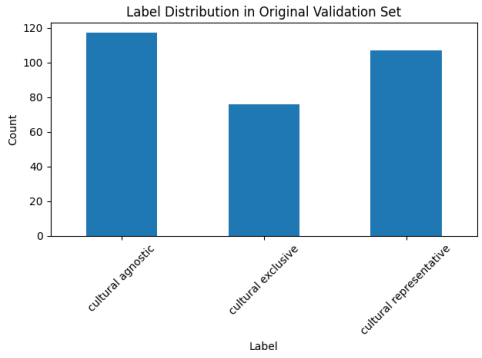

Figure 2: Class distribution of the validation dataset.