Cultural Specificity Classification: A Comparative Analysis of Language Model and Feature-Based Approaches

Archit Rastogi **Ejaz Ahmed**
Sapienza University of Rome
rastogi.1982785@studenti.uniroma1.it ahmed.1988401@studenti.uniroma1.it

1 Introduction

We classify Wikidata items into three cultural specificity categories: **cultural agnostic** (CA; universally known), **cultural representative** (CR; culture-specific but globally recognized), and **cultural exclusive** (CE; known primarily within one culture). Comparing fine-tuned mDeBERTa against feature-engineered XGBoost, we find the LM achieves substantially better performance (77.33% vs 63.67%) due to semantic understanding.

2 Methodology

2.1 Dataset Characteristics

The dataset exhibits distribution shift between training (n=6,251) and test (n=300; referred to as "validation" in the official HuggingFace split): CE dominates training (43.0%) but becomes minority in test (25.3%), while CA shows the opposite (29.9% $\rightarrow$ 39.0%). CR and CE classes share semantic overlap, differing primarily in global recognition rather than intrinsic properties.(Figure 1)

2.2 Language Model Approach

Model Selection. We use **mDeBERTa-v3-base** (He et al., 2021): disentangled attention, multilingual pretraining (100+ languages), and balanced size (86M parameters). Larger models showed <2% F1 gain for 3$\times$ training time.

Input Engineering. We structure metadata as: `Item: [name]. Description: [desc]. Type: [type]. Category: [cat].` This outperformed raw concatenation (65.3% $\rightarrow$ 68.7% test F1), helping the model learn distinct semantic roles for different fields.

Addressing Class Imbalance. Class-weighted loss, label smoothing (ϵ=0.1), stratified 5-fold CV. Converged at F1: 0.799 ± 0.004.

Wikipedia Augmentation Failure. Enriching training descriptions with Wikipedia content severely degraded performance (77.33% $\rightarrow$ 52.00% accuracy, 0.758 $\rightarrow$ 0.407 F1), even when both training and validation sets were consistently augmented. We tested this with DeBERTa-v3-large to ensure the failure was not model-dependent. The degradation suggests Wikipedia summaries introduce noise that overwhelms the original semantic signals, regardless of distributional consistency.

2.3 Feature-Based Approach

We train **XGBoost** on Wikipedia/Wikidata metadata, hypothesizing that cultural reach correlates with encyclopedic coverage.

Feature Engineering. We extract 14 raw features via APIs (page length, language count, statements, categories, identifiers, coordinates, cultural flags) and engineer 50+ derived features: log transforms, ratios, interactions, and composite scores. Identical extraction on train and test ensures distributional consistency.

Hyperparameter Optimization. Optuna (Akiba et al., 2019) (50 trials) tuned XGBoost (max_depth=4, lr=0.0146, subsample=0.626), improving F1 from 0.558 to 0.603. Shallow depth suggests simple decision boundaries.

3 Results and Analysis

We evaluate on held-out test set (n=300) using accuracy and macro-F1. Table 1 shows LM outperforms feature-based by 13.7 points in accuracy and 15.5 in macro-F1, demonstrating pretrained representations' value over hand-crafted features.

Approach	Acc.	F1(M)	F1(W)	P(M)	R(M)
mDeBERTa-v3	**77.33**	**0.758**	**0.773**	**0.757**	**0.760**
XGB+Wiki	63.67	0.603	0.636	0.673	0.623

Table 1: Test set performance.

Approach	Class	P	R	F1
LM	Agnostic	0.65	0.75	0.69
	Exclusive	0.5	0.013	0.02
	Representative	0.44	0.62	0.49
Non-LM	Agnostic	0.60	0.96	0.74
	Exclusive	0.64	0.57	0.60
	Representative	0.78	0.34	0.47

Table 2: Failed approaches per-class precision (P), recall (R), and F1.

Per-Class Performance. Both achieve highest (Figure 3) on CA (LM: 0.90, XGBoost: 0.74) universal concepts are most discriminable. LM performs better on CE (0.68 vs 0.60), suggesting semantic context helps disambiguate culture-bound entities.

Error Analysis. Confusion matrix (Figure 2) shows LM: 30.3% CE→CR, 20.6% CR→CE misclassifications. Non-LM has high CR precision (0.78) but poor recall (0.34), over-predicting CA. Errors involve ambiguous scope like regional festivals with UNESCO recognition.

4 Conclusion

Our LM achieves 77.33% accuracy, outperforming the feature baseline (63.67%) through careful handling of class imbalance and distribution shift. Pretrained representations capture cultural signals beyond statistical features. Both confirm CR vs CE distinction is challengingclasses differ in reach, not content. Future work: explicit popularity signals or graph-based approaches leveraging Wikidata's structure.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

A Additional Results

Failed Approach	Acc.	F1(M)	F1(W)	P(M)	R(M)
LM+Wiki Aug.	52.00	0.407	0.456	0.521	0.464
Ensemble (4)	58.67	0.556	0.587	0.598	0.596

Table 3: Failed approaches with complete metrics.

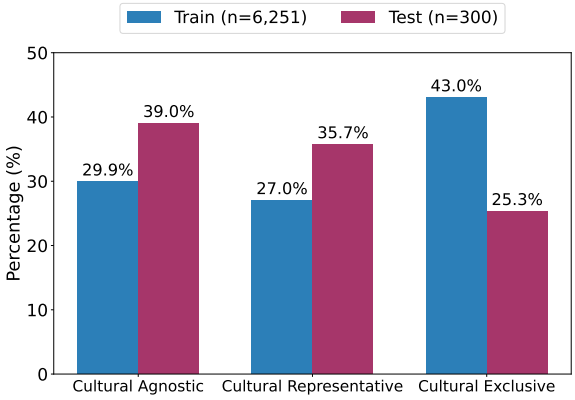

Figure 1: Class distribution shift between train and test.

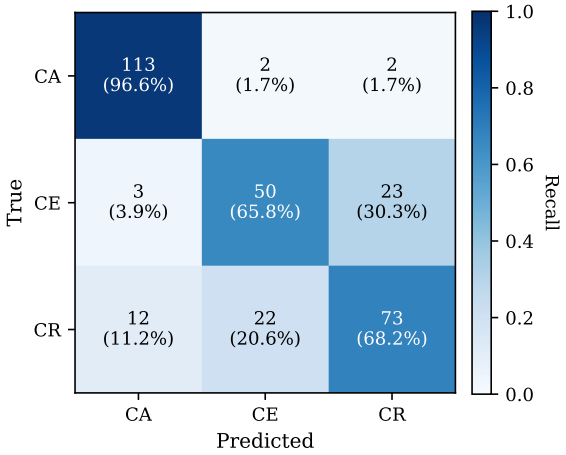

Figure 2: Confusion matrix for best LM approach (mDeBERTa-v3-base). CA=Agnostic, CE=Exclusive, CR=Representative.

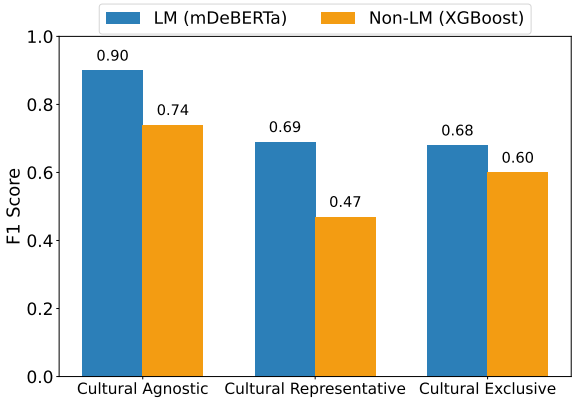

Figure 3: Per-class F1 comparison between LM and Non-LM approaches.