NLP 2025
Homework 1: Multiclass Classification of Cultural Items

Daniele Sabatini, Luca Conti, Riccardo D'Aguanno

1 Introduction

Culture is a complicated topic. It can be represented by semantic similarities, historical events, foods, literature works, art craft, ideologies etc. Culture is important in NLP for mitigating bias and improving the ability of a language model (LM) to accurately perform tasks across different cultural contexts (multicultural generalization). The main goal of this homework was to develop an automatic classifier capable to determine whether an item is specific to a particular culture or not. In particular, the classifier should classify items using the following provided taxonomy:

- a cultural agnostic (C.A.) item is commonly known/used worldwide and no culture claims the item (i.e. bridge, potato, car).
- a cultural representative (C.R.) item is originated in a culture and/or claimed by a culture as their own, but other cultures know/use it or have similar items (i.e. spaghetti, hamburger, baozi).
- a cultural exclusive (C.E.) item is known/used only in a specific culture and it is claimed by that culture (e.g., "pasta alla gricia" or "xiaolongbao").

In order to solve this task, two different approaches were explored: i) a LM-based method, which consisted in fine tuning an encoder Transformer; ii) a not LM-based method. Furthermore, to train and assess the models, two different datasets were provided: a silver annotated (by ChatGPT-03) train set, and a golden annotated (by 5 annotators) validation set. Another set, namely the test set, was provided without labels.

2 Methodology

The first step we took was to explore the training dataset. One of the most important aspects was the imbalance withing the classes distribution (1872 C.A. items, 2691 C.E. items and 1688 C.R. items): this fact led us to use a higher class weight for the C.R. class when training the non LM-based models. Before training and fine-tuning the models, the original datasets were enhanced and some additional features were added; indeed, we believed that the original information alone was not enough to achieve interesting performances. For the LM-based method, we used only the first 1000 characters of each Wikipedia article as input to the encoder. Instead, for the non LM-based method, some features for each entity were computed directly from the original datasets (e.g., "name_length", "uppercase_count_ratio", "special_char_ratio") while others (e.g., "num_images", "total_views") were derived by querying Wikidata. The fist 1000 characters of the Wikipedia descriptions were also exploited: in fact, some words like "may" or "often" are more used in the Wikipedia article texts of the C.A. items, while words like "first", "history" or "most" are more common in the article texts of C.E. and C.R. items. A complete list of the additional features and their corresponding meanings is shown in Table 1.

3 Experiments

For the LM-based approach, we tried several transformer architectures, including BERT, RoBERTa (both base, large and XML) and DeBERTa (both v3 base) models. The configuration that achieved the best performance was RoBERTa large with a batch size of 8, a learning rate of 5×10^{-6}, the first three layers freezed and fine-tuned for 2 epochs to avoid overfitting, which was observed to occur with longer training. No class weighting was applied in this approach. we used only the first 1000 characters of each Wikipedia article as input to the encoder.

For the not LM-based approach, we used a cascading system composed of three models: a Random Forest (RF), a Support Vector Machine (SVM) and a XGBoost (XGB) classifier. The chosen threshold for the cascading system was 0.7. To address class imbalance, we applied a higher class weight to the C.R. class (but not for the XGB). The classifier were trained using the additional numeric features. Each model's hyperparameters were tuned via a grid search: for the SVM classifier, we used a "rbf" kernel and we tested the parameters C and "gamma" ; for the RF and XBG models we tested the "n_estimators" and the "max_depth". The grid search yielded the following best configuration: $C = 10$ and gamma= 0.02 for the SVM, max_depth=None and 150 estimators for the RF and max_depth=5 and 30 estimators for the XGB.

4 Results

The LM-based method achieved an overall accuracy of 79.67% , 80.16% weighted precision, 79.67% weighted recall and 79.81% of weighted F1-score on the validation set. The performance report for the LM-based method is shown Table 2, while the left panel of Figure 1 illustrates the corresponding confusion matrix. The model performs best on the C.A. class, achieving high precision (0.92) and recall (0.89), suggesting it's particularly effective at identifying non-cultural items. For C.E. and C.R. classes, performance is lower, especially in precision (0.67 and 0.75 respectively), which may indicate confusion between cultural categories.

The performance report for the non LM-based method is shown Table 3, while the right panel of Figure 1 illustrates the confusion matrix. Again, the ensemble method performs best on the C.A., with a precision of 0.79 and a high recall of 0.91, resulting in a solid F1-score of 0.84. Anyways, performance is slightly worse for the C.E. and C.R. classes, which show 0.75 and 0.63 respectly of F1-score. In particular, the recall for the C.R. class drops to 0.54: the model tends to miss a notable number of these items. The overall accuracy of 0.753 is balanced across classes, but the gap in recall among classes suggests that class imbalance or overlapping features may be affecting the model's ability to generalize on and C.E. and C.R. entities.

In summary, in both approaches, it seems that the classifiers can perform well on C.A. but has some trouble in the distinction between C.E and C.R items, even using a higher class weight for the the C.R. class (that seems to be the one causing more difficulties to the classifiers). This could be because:

- C.E and C.R items are very similar. The boundary between "representative" and "exclusive" items is ambiguous. This means that even humans annotators may have difficulty in distinguishing between the two classes. So, it's very likely that the the training dataset, that is annotated by a LM, contains a lot of noise. Moreover, even among human annotators, subjective judgment or cultural background may vary, leading to inconsistent labeling of borderline cases. As a result, there may noisy labels even in the validation set, despite it being annotated by five people.

- The features extracted for the non LM-based method may not provide sufficient discriminative information to separate C.E. and C.R. items clearly. Even though class weights were used, the models may still underperform on minority classes if the features are not sufficiently informative. In the Colab notebook, we plotted histograms of each feature by class label, and for many of them (e.g., "special_char"), the distributions appear very similar across classes. This suggests that these features may not be too helpful for the model to differentiate between the classes. Indeed, Figure 2 shows the features importance for the RF classifier: some of the features (e.g., "special_char") has a very little importance.

- We relied only the English wikipedia articles. However, the demographics of Wikipedia editors introduce systemic bias in the content of the site. The greatest number of editors (20%) reside in the United States, followed by Germany (12%) and Russia (7%). This may cause C.R. items to be described in a way that resembles C.E. (and vice versa) or even C.A. ones, leading to misleading inputs for the LM-based classifier.

- The models only had access to the first 1000 characters of the Wikipedia description, which may not include enough cultural context to make a reliable classification.

Feature	Description
name_length	Length of the name of the item in the description.
num_words	Number of words in the name of the item in the description.
wiki_pages	Total number of Wikipedia pages associated with an entity.
wiki_top10_langs_count	Number of Wikipedia pages in the top 10 spoken languages.
special_char	Number of special characters in the item's name.
special_char_ratio	Ratio between special_char and num_words.
uppercase_count	Number of uppercase characters in the item's name.
uppercase_count_ratio	Ratio between uppercase_count and num_words.
description_length	Number of characters in the item's description.
description_word_count	Number of words in the item's description.
keywords_count_cr	Number of words representative of class C.R. in the first 1000 characters of the Wikipedia article.
keywords_count_ca	Number of words representative of class C.A. in the first 1000 characters of the Wikipedia article.
alias_count	Total number of aliases of an entity (in all languages).
external_links_count	Number of external links in the English Wikipedia page of the entity.
label_languages_count	Number of languages with a label (name) for the entity in Wikidata.
description_languages_count	Number of languages with a description of the entity in Wikidata.
total_views	Total number of pageviews of the English Wikipedia page over the last 90 days.
degree_out	Number of outgoing properties from the entity.
degree_in	Number of incoming properties to the entity.
label_length	Length (in characters) of the English label of the entity.
num_languages_with_label	Number of languages in which the entity has a label in Wikidata.
superclass_count	Number of superclasses of the entity, according to property wdt:P279*.
neighbor_count	Number of entities directly connected to the current entity (incoming and outgoing).

Table 1: Overview of the added features used for the not LM-based method.

Class	Precision	Recall	F1-score	Support
Cultural Agnostic	0.93	0.90	0.91	117
Cultural Exclusive	0.67	0.76	0.72	76
Cultural Representative	0.75	0.71	0.73	107
Accuracy		0.80		300
Macro Average	0.79	0.79	0.79	300
Weighted Avg	0.80	0.80	0.80	300

Table 2: Classification report of the LM-based method (RoBERTa).

Class	Precision	Recall	F1-score	Support
Cultural Agnostic	0.79	0.91	0.84	117
Cultural Exclusive	0.70	0.82	0.75	76
Cultural Representative	0.76	0.54	0.63	107
Accuracy		0.75		300
Macro Average	0.75	0.75	0.74	300
Weighted Avg	0.75	0.75	0.74	300

Table 3: Classification report of the non LM-based method).

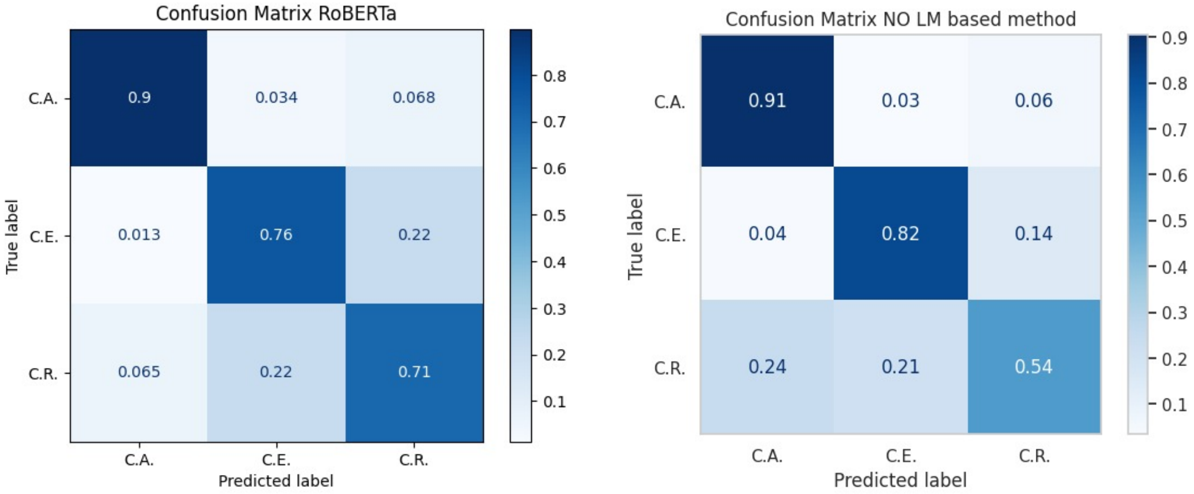

Figure 1: Normalized confusion matrices for the LM-based method (*left*) and for the non LM-based method (*right*).

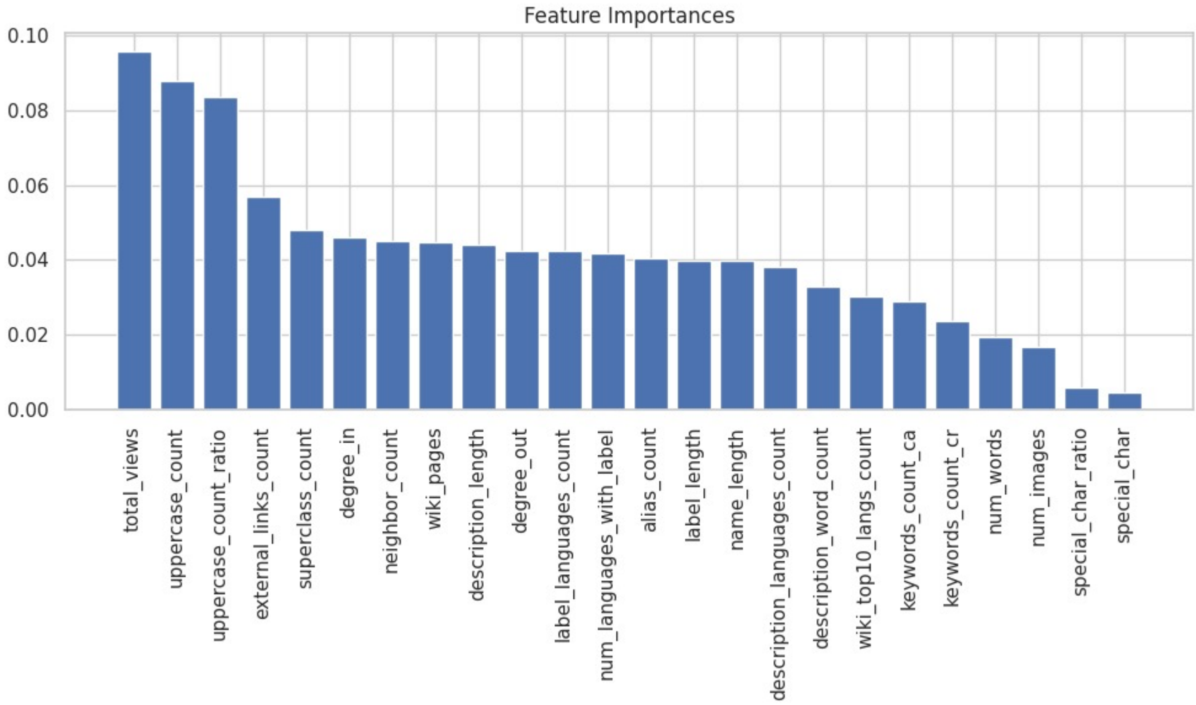

Figure 2: Features importance for the Random Forest (RF) classifier.