Manifestoberta Performance Report

Context Version 2023a (56Topics)

2023-11-01

Summary

- Performance was measured on 198 manifestos, which represent 186276 annotated quasi-sentences
- Overall the model manages to assign the correct category to quasi-sentences in the test data set with an accuracy of 64.00%. In 80.97% of the cases, the true category is among the two most confident predictions of the model, and in 88.15% among the top three. (Table 1)
- Lower macro averaged F1, Precision, and Recall reflect problems with some individual categories, especially rare/exotic categories like 102, 409, or 702 (Table 1 and Table 2)
- The overall distribution and frequency of individual category predictions is closely aligned with the true distribution of categories in our test data set. The model isn't systematically over- or under predicting specific codes (Table 2)
- The model performs considerably well in all countries/languages present in the test data set (Table 3).
- Probability estimates of the model are well calibrated and properly reflect the likelihood of a right prediction. If the model reports a confidence of 95% or higher (which happened for 15.19% of all quasi-sentences in our test set) it was, in fact, right in 94,63% of those cases. (Table 4)
- True rile values and rile values calculated based on model predictions are strongly correlated (Plot 1). However, there are a few cases where large differences occurred (over 30) between true rile and model rile (Plot 2)

Usage Recommendations

Given the capabilities of the model, a number of use cases are conceivable:

- The model's probability estimates make it possible to automatically predict a subset of a document and focus manual labeling efforts only on the more uncertain cases. For instance, if a cutoff of 80% confidence is chosen, almost 40% of the sentences to be coded could be automatically classified with an accuracy of at least 73%.
- To support manual coding, the model predictions can be used as code suggestions. This narrows down the choices during coding and speeds up the process significantly without sacrificing quality, considering an accuracy of 88% for the top 3 suggestions of the model and 94% for the top 5 suggestions.
- It is also conceivable to code documents completely automatically using the model. However, it might then be advisable to at least manually validate parts of the codings for subsequent analysis and/or to robustly integrate the automatically generated codes into the analysis (e.g. multiple variations of random substitutions between top picks, bootstrapping).

Results

accuracy	0.64
top2_acc	0.81
top3_acc	0.88
precision	0.54
recall	0.52
f1macro	0.53
mcc	0.62
cross-entropy	1.15

Table 1: Classification Results - Overall

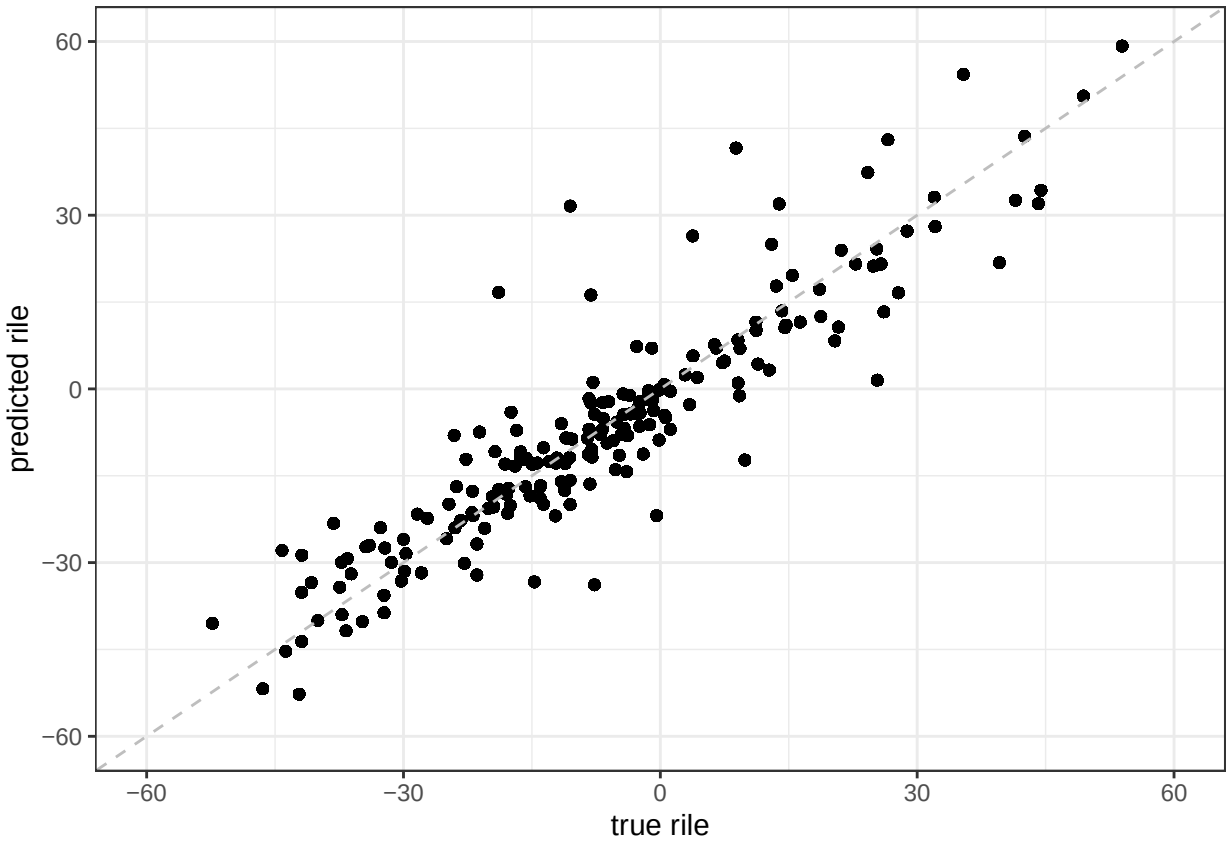

Figure 1: True rile values vs. predicted rile values

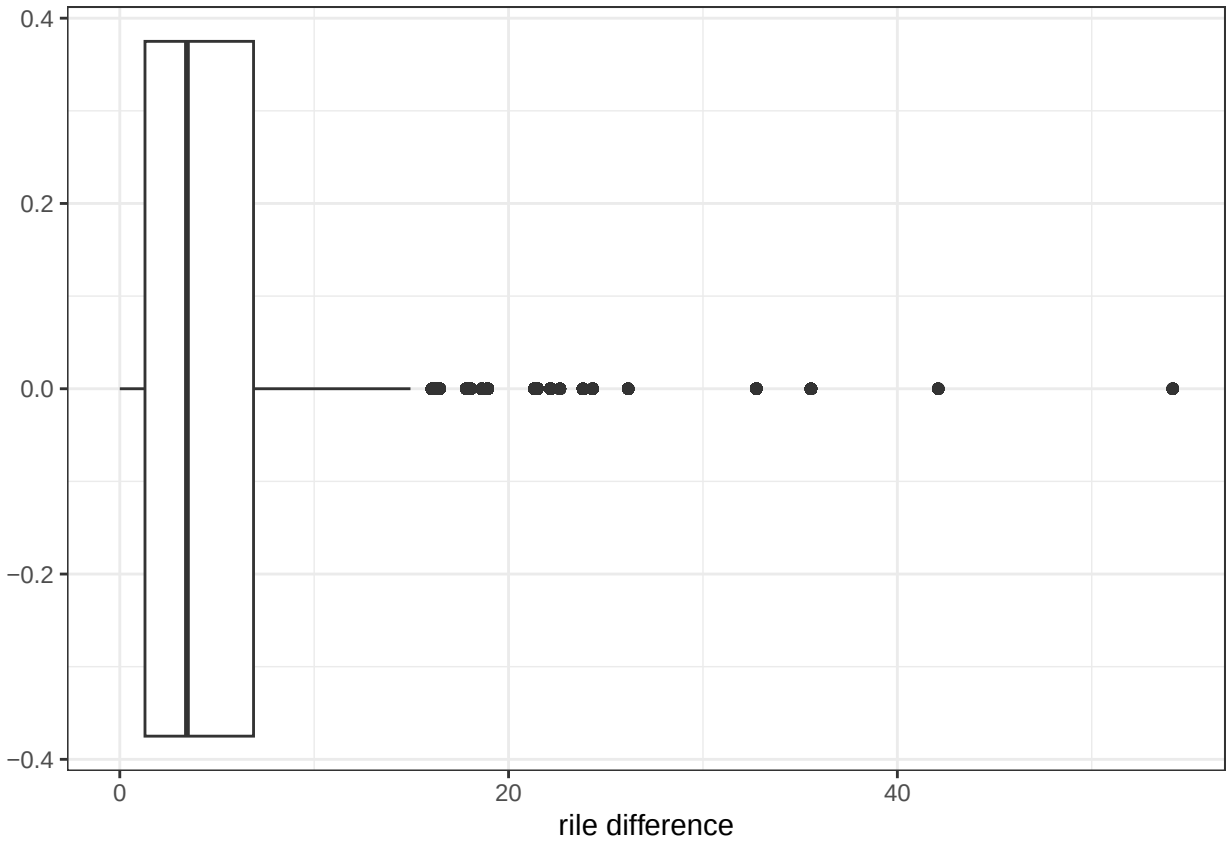

Figure 2: Absolute rile differences between true and predicted cmp codes

Category	Precision	Recall	F1	n(%)	n_predicted(%)
101	0.50	0.48	0.49	0.30%	0.29%
102	0.56	0.61	0.58	0.09%	0.10%
103	0.51	0.36	0.42	0.28%	0.20%
104	0.78	0.81	0.79	1.57%	1.64%
105	0.69	0.70	0.69	0.34%	0.34%
106	0.59	0.57	0.58	0.33%	0.32%
107	0.68	0.66	0.67	2.24%	2.17%
108	0.66	0.68	0.67	1.20%	1.24%
109	0.52	0.39	0.45	0.17%	0.13%
110	0.63	0.68	0.65	0.36%	0.38%
201	0.58	0.59	0.59	2.16%	2.20%
202	0.62	0.63	0.62	3.25%	3.28%
203	0.46	0.47	0.47	0.19%	0.19%
204	0.61	0.37	0.46	0.25%	0.15%
301	0.66	0.71	0.68	2.13%	2.29%
302	0.38	0.25	0.30	0.17%	0.11%
303	0.58	0.60	0.59	5.12%	5.31%
304	0.67	0.65	0.66	1.38%	1.34%
305	0.59	0.57	0.58	2.32%	2.22%
401	0.45	0.36	0.40	1.50%	1.21%
402	0.61	0.58	0.59	2.73%	2.60%
403	0.56	0.51	0.53	3.59%	3.25%
404	0.30	0.15	0.20	0.58%	0.28%
405	0.43	0.51	0.47	0.18%	0.21%
406	0.38	0.46	0.42	0.26%	0.31%
407	0.56	0.52	0.54	0.40%	0.38%
408	0.28	0.17	0.21	1.34%	0.79%
409	0.37	0.21	0.27	0.24%	0.14%
410	0.53	0.50	0.52	2.22%	2.08%
411	0.73	0.75	0.74	8.32%	8.53%
412	0.26	0.20	0.22	0.58%	0.45%
413	0.49	0.63	0.55	0.29%	0.37%
414	0.58	0.55	0.56	1.38%	1.32%
415	0.14	0.23	0.18	0.05%	0.07%
416	0.52	0.49	0.50	2.45%	2.35%
501	0.69	0.78	0.73	4.77%	5.35%
502	0.78	0.84	0.81	3.08%	3.32%
503	0.61	0.63	0.62	5.96%	6.11%
504	0.71	0.76	0.74	10.05%	10.76%
505	0.46	0.37	0.41	0.69%	0.55%
506	0.78	0.82	0.80	5.42%	5.72%
507	0.45	0.26	0.33	0.14%	0.08%
601	0.52	0.46	0.49	1.79%	1.57%
602	0.35	0.34	0.34	0.24%	0.24%
603	0.65	0.68	0.67	1.36%	1.42%
604	0.62	0.48	0.54	0.57%	0.44%
605	0.72	0.74	0.73	4.22%	4.33%
606	0.56	0.48	0.51	1.45%	1.23%
607	0.57	0.67	0.62	1.08%	1.25%
608	0.48	0.48	0.48	0.41%	0.41%
701	0.62	0.66	0.64	3.35%	3.59%
702	0.42	0.30	0.35	0.08%	0.06%
703	0.75	0.87	0.80	2.65%	3.07%
704	0.43	0.32	0.37	0.57%	0.43%
705	0.38	0.33	0.35	0.80%	0.69%
706	0.43	0.37	0.39	1.35%	1.16%

Table 2: Classification Results - Categories

country	accuracy	precision	recall	f1	n
Argentina	60.40%	0.48	0.48	0.46	3,586
Armenia	65.70%	0.65	0.55	0.65	207
Australia	73.01%	0.54	0.50	0.55	10,091
Austria	64.37%	0.48	0.50	0.54	2,366
Belgium	53.96%	0.43	0.42	0.42	26,091
Bosnia-Herzegovina	65.74%	0.45	0.46	0.48	3,193
Brazil	67.36%	0.56	0.54	0.55	2,255
Bulgaria	59.29%	0.42	0.47	0.46	926
Canada	60.74%	0.54	0.42	0.46	2,481
Chile	57.80%	0.39	0.48	0.45	4,810
Colombia	73.05%	0.53	0.56	0.56	590
Costa Rica	72.80%	0.57	0.58	0.57	10,570
Croatia	76.19%	0.62	0.60	0.65	2,579
Cyprus	58.37%	0.53	0.41	0.51	1,057
Czech Republic	62.46%	0.56	0.47	0.47	3,162
Denmark	61.80%	0.48	0.44	0.48	1,631
Ecuador	61.44%	0.58	0.55	0.60	638
Estonia	71.54%	0.66	0.61	0.68	889
Finland	66.06%	0.52	0.53	0.52	2,236
Georgia	62.41%	0.52	0.59	0.61	141
Germany	64.62%	0.55	0.51	0.52	10,650
Greece	53.55%	0.38	0.44	0.44	2,310
Hungary	69.87%	0.59	0.59	0.59	3,332
Iceland	67.58%	0.59	0.50	0.58	910
Israel	70.27%	0.57	0.57	0.57	1,238
Italy	72.38%	0.62	0.56	0.69	210
Latvia	76.67%	0.70	0.68	0.75	60
Lithuania	66.15%	0.50	0.44	0.46	7,595
Luxembourg	63.58%	0.54	0.52	0.52	2,938
Mexico	56.28%	0.41	0.43	0.40	4,636
Montenegro	64.86%	0.58	0.58	0.61	663
Netherlands	62.39%	0.48	0.44	0.46	6,932
New Zealand	62.11%	0.49	0.52	0.49	5,906
North Macedonia	75.61%	0.68	0.60	0.63	906
Norway	67.71%	0.50	0.44	0.47	8,383
Panama	71.29%	0.54	0.48	0.55	1,306
Peru	68.64%	0.55	0.60	0.56	5,090
Poland	66.61%	0.61	0.58	0.60	1,102
Portugal	65.56%	0.49	0.51	0.48	6,440
Romania	58.24%	0.51	0.48	0.51	1,147
Russia	60.49%	0.60	0.57	0.60	567
Serbia	65.32%	0.45	0.55	0.54	571
Slovakia	64.64%	0.56	0.52	0.55	1,120
Slovenia	52.88%	0.39	0.36	0.39	8,113
South Africa	71.21%	0.53	0.62	0.56	2,098
Spain	72.43%	0.60	0.60	0.61	11,986
Sweden	66.89%	0.53	0.48	0.51	4,111
Switzerland	60.82%	0.46	0.53	0.54	1,516
Turkey	63.91%	0.46	0.54	0.52	2,045
Ukraine	59.75%	0.48	0.48	0.50	395
United Kingdom	65.25%	0.52	0.51	0.51	2,501

Table 3: Classification Results - Countries

parfam	accuracy	precision	recall	f1	n
10	58.91%	0.45	0.43	0.43	15,759
20	63.10%	0.52	0.51	0.50	19,603
30	65.09%	0.50	0.50	0.50	41,520
40	61.02%	0.49	0.48	0.48	31,121
50	62.63%	0.51	0.49	0.49	23,334
60	66.98%	0.52	0.54	0.52	29,326
70	65.45%	0.52	0.54	0.52	12,995
80	67.00%	0.59	0.57	0.62	303
90	71.15%	0.65	0.54	0.59	7,568
95	65.16%	0.51	0.46	0.48	3,304
98	73.40%	0.57	0.49	0.58	1,342
999	63.37%	0.63	0.72	0.71	101

Table 4: Classification Results - Parfam

prob_estimates	accuracy	n(%)	cum_n(%)
> 95%	95.48%	15.30%	15.30%
90%-95%	86.40%	10.19%	25.49%
85%-90%	79.84%	8.06%	33.55%
80%-85%	73.76%	6.84%	40.39%
75%-80%	67.83%	6.36%	46.75%
70%-75%	63.79%	6.12%	52.87%
65%-70%	59.94%	6.02%	58.88%
60%-65%	55.33%	6.00%	64.89%
55%-60%	50.58%	6.25%	71.14%
50%-55%	47.59%	6.49%	77.63%
45%-50%	42.26%	6.27%	83.90%
40%-45%	38.23%	5.26%	89.16%
35%-40%	32.70%	4.30%	93.46%
30%-35%	28.79%	3.20%	96.67%
25%-30%	23.01%	2.07%	98.73%
20%-25%	21.22%	0.98%	99.72%
15%-20%	14.23%	0.26%	99.98%
10%-15%	16.28%	0.02%	100.00%

Table 5: Model Calibration - Probability Groups