Manifestoberta Performance Report

Context Version 2024a (56Topics)

2024-07-31

Summary

- Performance was measured on 203 manifestos, which represent 200920 annotated quasi-sentences.
- Overall the model manages to assign the correct category to quasi-sentences in the test data set with an accuracy of 64.23%. In 81.27% of the cases, the true category is among the two most confident predictions of the model, and in 88.37% 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, with the lowest accuracy value of 51.08% in Italy (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 or 15.56% of all quasi-sentences in our test set) it was, in fact, right in 94.67% 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, more than 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	0.55
recall	0.52
f1macro	0.53
mcc	0.63
cross-entropy	1.15

Table 1: Classification Results - Overall

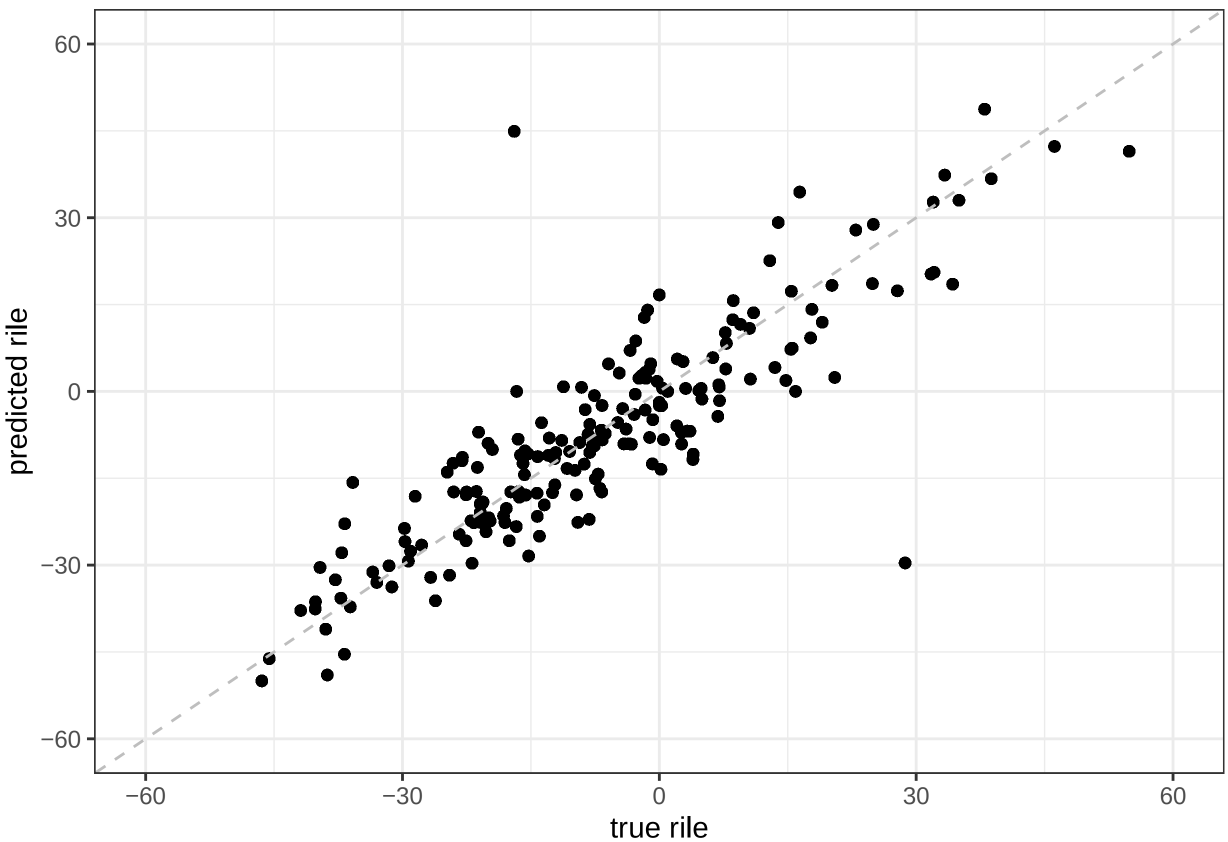

Figure 1: True rile values vs. predicted rile values

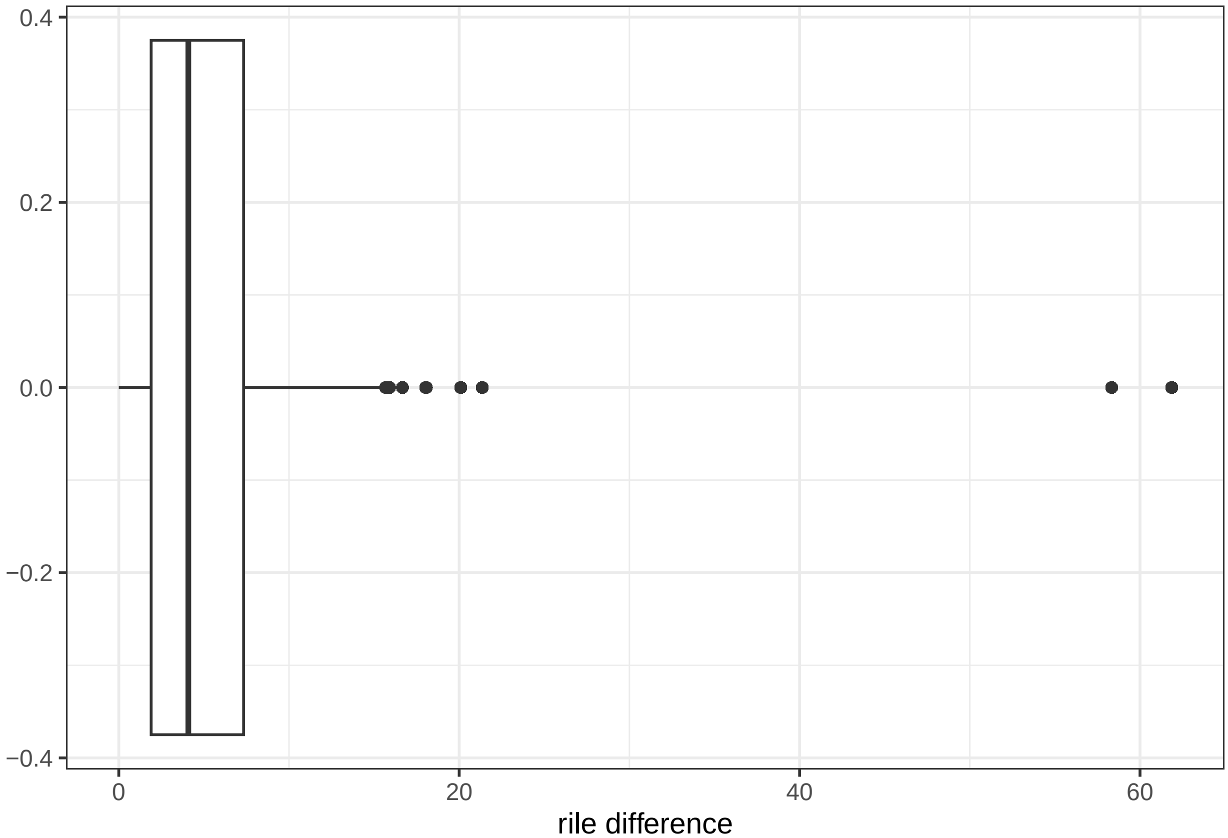

Figure 2: Absolute rile differences between true and predicted cmp codes

Category	Precision	Recall	F1	n(%)	n_predicted(%)
101	0.46	0.53	0.50	0.29%	0.33%
102	0.57	0.49	0.53	0.07%	0.06%
103	0.48	0.43	0.46	0.26%	0.23%
104	0.74	0.79	0.76	1.55%	1.66%
105	0.59	0.74	0.66	0.34%	0.42%
106	0.54	0.67	0.60	0.34%	0.43%
107	0.63	0.66	0.65	2.35%	2.47%
108	0.64	0.67	0.65	1.24%	1.31%
109	0.47	0.31	0.37	0.16%	0.10%
110	0.61	0.62	0.62	0.43%	0.44%
201	0.58	0.59	0.59	2.20%	2.23%
202	0.65	0.59	0.62	3.57%	3.24%
203	0.45	0.38	0.41	0.18%	0.15%
204	0.54	0.58	0.56	0.21%	0.22%
301	0.63	0.65	0.64	2.01%	2.07%
302	0.42	0.23	0.29	0.17%	0.09%
303	0.59	0.56	0.57	4.35%	4.12%
304	0.73	0.64	0.68	1.55%	1.37%
305	0.57	0.54	0.56	2.06%	1.98%
401	0.46	0.35	0.39	1.11%	0.84%
402	0.57	0.57	0.57	2.50%	2.51%
403	0.54	0.50	0.52	3.19%	2.97%
404	0.45	0.28	0.34	0.59%	0.36%
405	0.41	0.28	0.33	0.25%	0.18%
406	0.42	0.36	0.39	0.40%	0.34%
407	0.53	0.51	0.52	0.47%	0.46%
408	0.38	0.25	0.30	1.52%	1.00%
409	0.22	0.12	0.15	0.28%	0.14%
410	0.57	0.53	0.55	2.00%	1.86%
411	0.69	0.76	0.72	8.50%	9.27%
412	0.42	0.18	0.25	0.63%	0.26%
413	0.52	0.69	0.59	0.37%	0.48%
414	0.59	0.58	0.59	1.22%	1.20%
415	0.47	0.37	0.41	0.13%	0.10%
416	0.61	0.44	0.51	3.07%	2.20%
501	0.68	0.82	0.74	5.72%	6.87%
502	0.76	0.86	0.81	3.24%	3.63%
503	0.62	0.61	0.61	6.00%	5.94%
504	0.70	0.77	0.73	9.47%	10.40%
505	0.56	0.39	0.46	0.68%	0.48%
506	0.75	0.82	0.78	5.54%	6.08%
507	0.47	0.19	0.27	0.12%	0.05%
601	0.57	0.53	0.55	1.55%	1.43%
602	0.35	0.31	0.33	0.30%	0.26%
603	0.64	0.69	0.67	1.10%	1.19%
604	0.56	0.62	0.59	0.49%	0.55%
605	0.68	0.75	0.71	3.94%	4.32%
606	0.56	0.48	0.51	1.48%	1.27%
607	0.64	0.66	0.65	1.44%	1.50%
608	0.46	0.47	0.47	0.26%	0.26%
701	0.66	0.67	0.66	3.53%	3.53%
702	0.46	0.37	0.41	0.09%	0.07%
703	0.79	0.86	0.82	3.06%	3.32%
704	0.45	0.26	0.33	0.37%	0.22%
705	0.42	0.25	0.32	0.86%	0.51%
706	0.41	0.35	0.38	1.19%	1.03%

Table 2: Classification Results - Categories

country	accuracy	precision	recall	f1	n
Argentina	64.57%	0.54	0.57	0.55	1,160
Armenia	63.89%	0.57	0.60	0.65	108
Australia	71.89%	0.55	0.56	0.59	6,215
Austria	61.81%	0.46	0.51	0.49	7,798
Belgium	56.50%	0.46	0.43	0.43	16,221
Bolivia	53.79%	0.39	0.38	0.40	924
Bosnia-Herzegovina	53.09%	0.44	0.41	0.41	1,746
Bulgaria	58.61%	0.50	0.48	0.50	1,005
Canada	58.74%	0.48	0.43	0.45	3,551
Chile	54.47%	0.41	0.47	0.43	6,405
Colombia	70.89%	0.41	0.54	0.48	7,534
Costa Rica	72.75%	0.58	0.56	0.57	6,796
Croatia	77.55%	0.62	0.70	0.70	3,194
Cyprus	65.30%	0.48	0.48	0.50	2,977
Czech Republic	63.17%	0.54	0.50	0.51	5,009
Denmark	64.81%	0.54	0.45	0.47	2,376
Dominican Republic	71.66%	0.58	0.59	0.56	4,107
Ecuador	52.78%	0.48	0.42	0.49	360
Estonia	71.13%	0.58	0.52	0.57	2,865
Finland	67.29%	0.55	0.53	0.54	2,975
France	65.03%	0.48	0.50	0.51	2,236
Germany	61.05%	0.51	0.50	0.49	11,226
Greece	64.82%	0.52	0.52	0.52	1,424
Hungary	68.00%	0.58	0.52	0.55	3,650
Iceland	64.27%	0.61	0.55	0.59	375
Israel	71.50%	0.58	0.60	0.61	1,695
Italy	51.08%	0.40	0.34	0.35	7,606
Latvia	80.53%	0.73	0.69	0.85	113
Lithuania	71.34%	0.56	0.53	0.58	2,055
Mexico	53.41%	0.40	0.40	0.41	777
Montenegro	60.68%	0.57	0.52	0.56	562
Netherlands	61.92%	0.50	0.45	0.47	10,835
New Zealand	67.48%	0.55	0.50	0.52	8,819
North Macedonia	70.51%	0.61	0.50	0.56	9,116
Norway	65.43%	0.42	0.48	0.44	5,505
Panama	71.22%	0.55	0.57	0.57	952
Peru	72.28%	0.56	0.59	0.58	4,790
Poland	65.69%	0.54	0.56	0.58	822
Portugal	59.59%	0.49	0.43	0.43	7,290
Romania	63.32%	0.49	0.54	0.56	398
Russia	62.03%	0.61	0.57	0.57	1,164
Serbia	66.49%	0.65	0.58	0.60	382
Slovakia	64.69%	0.55	0.51	0.51	3,302
Slovenia	57.91%	0.43	0.44	0.47	1,668
South Africa	71.82%	0.55	0.56	0.62	802
South Korea	66.61%	0.52	0.50	0.62	626
Spain	70.09%	0.57	0.54	0.55	9,514
Sweden	66.26%	0.55	0.48	0.53	2,398
Switzerland	59.07%	0.55	0.56	0.59	215
Turkey	69.72%	0.56	0.56	0.57	6,601
Ukraine	58.31%	0.49	0.47	0.48	439
United Kingdom	64.21%	0.57	0.54	0.55	2,998
United States	57.67%	0.47	0.48	0.46	3,657
Uruguay	54.22%	0.38	0.39	0.37	3,582

Table 3: Classification Results - Countries

parfam	accuracy	precision	recall	f1	n
10	60.80%	0.48	0.45	0.45	17,701
20	62.19%	0.47	0.46	0.45	25,211
30	65.55%	0.53	0.51	0.51	39,639
40	63.87%	0.53	0.51	0.52	28,942
50	65.32%	0.56	0.51	0.53	22,771
60	66.51%	0.56	0.52	0.52	31,692
70	64.30%	0.51	0.52	0.51	6,810
80	72.69%	0.52	0.57	0.53	5,390
90	62.40%	0.51	0.46	0.49	8,255
95	56.97%	0.50	0.45	0.45	11,726
98	74.80%	0.55	0.53	0.56	2,377
999	71.43%	0.65	0.63	0.66	406

Table 4: Classification Results - Parfam

prob_estimates	accuracy	n(%)	cum_n(%)
> 95%	94.67%	15.56%	15.56%
90%-95%	85.62%	10.36%	25.93%
85%-90%	78.46%	8.13%	34.05%
80%-85%	73.70%	7.07%	41.12%
75%-80%	68.42%	6.34%	47.46%
70%-75%	64.03%	6.13%	53.58%
65%-70%	59.07%	5.98%	59.56%
60%-65%	55.89%	6.03%	65.59%
55%-60%	50.60%	6.22%	71.82%
50%-55%	46.68%	6.42%	78.23%
45%-50%	43.27%	6.25%	84.49%
40%-45%	38.82%	5.15%	89.63%
35%-40%	34.41%	4.21%	93.84%
30%-35%	29.92%	3.08%	96.92%
25%-30%	23.60%	1.91%	98.83%
20%-25%	19.38%	0.92%	99.75%
15%-20%	14.50%	0.23%	99.98%
10%-15%	15.15%	0.02%	100.00%
5%-10%	100.00%	0.00%	100.00%

Table 5: Model Calibration - Probability Groups