Manifestoberta Performance Report

Sentence Version 2023a (56Topics)

2023-11-01

Summary

- Performance was measured on 198 manifestos, which represent 199046 annotated quasi-sentences
- Overall the model manages to assign the correct category to quasi-sentences in the test data set with an accuracy of 56.87%. In 73.03% of the cases, the true category is among the two most confident predictions of the model, and in 80.65% 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 isn't as closely aligned with the true distribution of categories as they are in our context model. The model is over and under predicting some codes (Table 2)
- The model performs in a acceptable range across all countries/languages present in the test data set, with the lowest accuracy value of 43% in Argentina (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 9.11% of all quasi-sentences in our test set) it was, in fact, right in 94,99% of those cases. (Table 4)
- True rile values and rile values calculated based on model predictions are rather strongly correlated (Plot 1).

Usage Recommendations

As the model performs worse than the context model version yet is easier to apply, it is primarily intended for small ad-hoc analyses and prototyping. Generally, we recommend using the more capable context model for more robust and reliable results.

Results

accuracy	0.57
top2_acc	0.73
top3_acc	0.81
precision	0.49
recall	0.43
f1macro	0.45
mcc	0.55
cross-entropy	1.50

Table 1: Classification Results - Overall

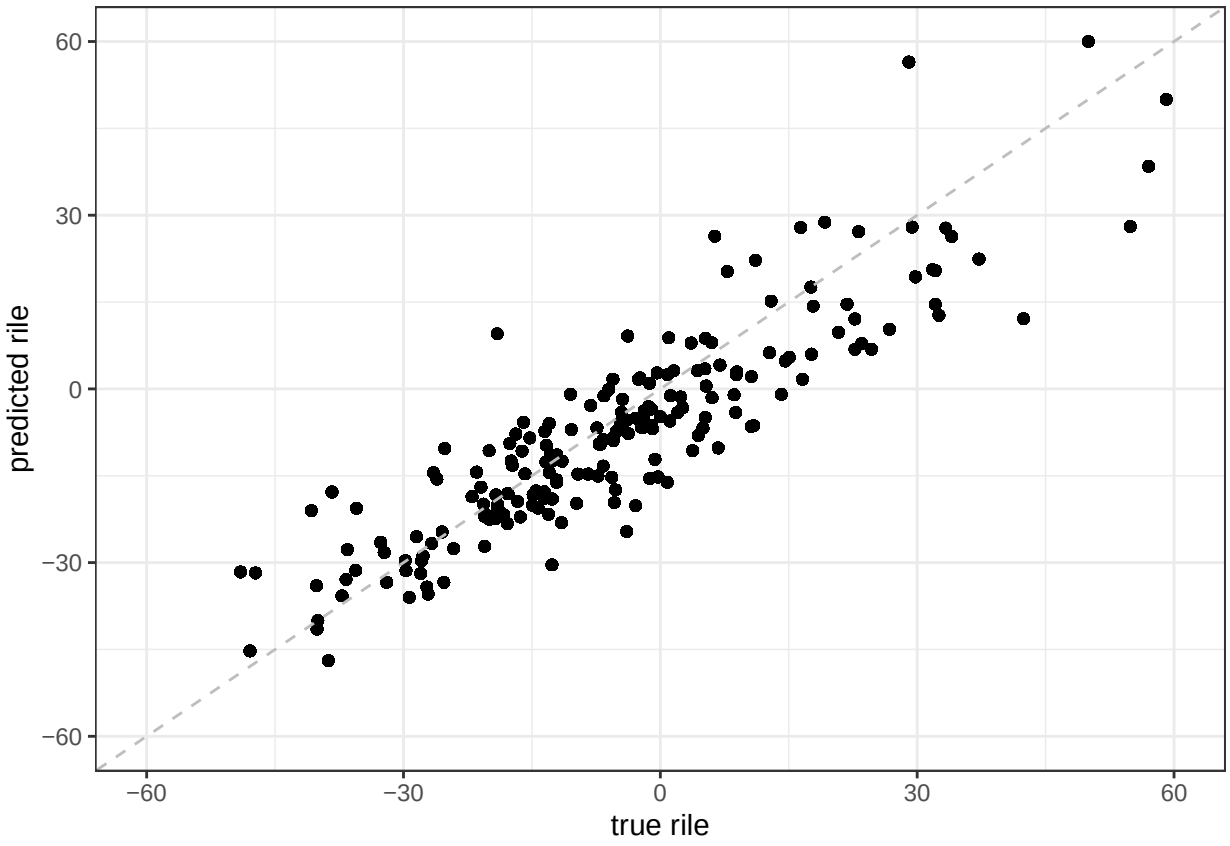

Figure 1: True rile values vs. predicted rile values

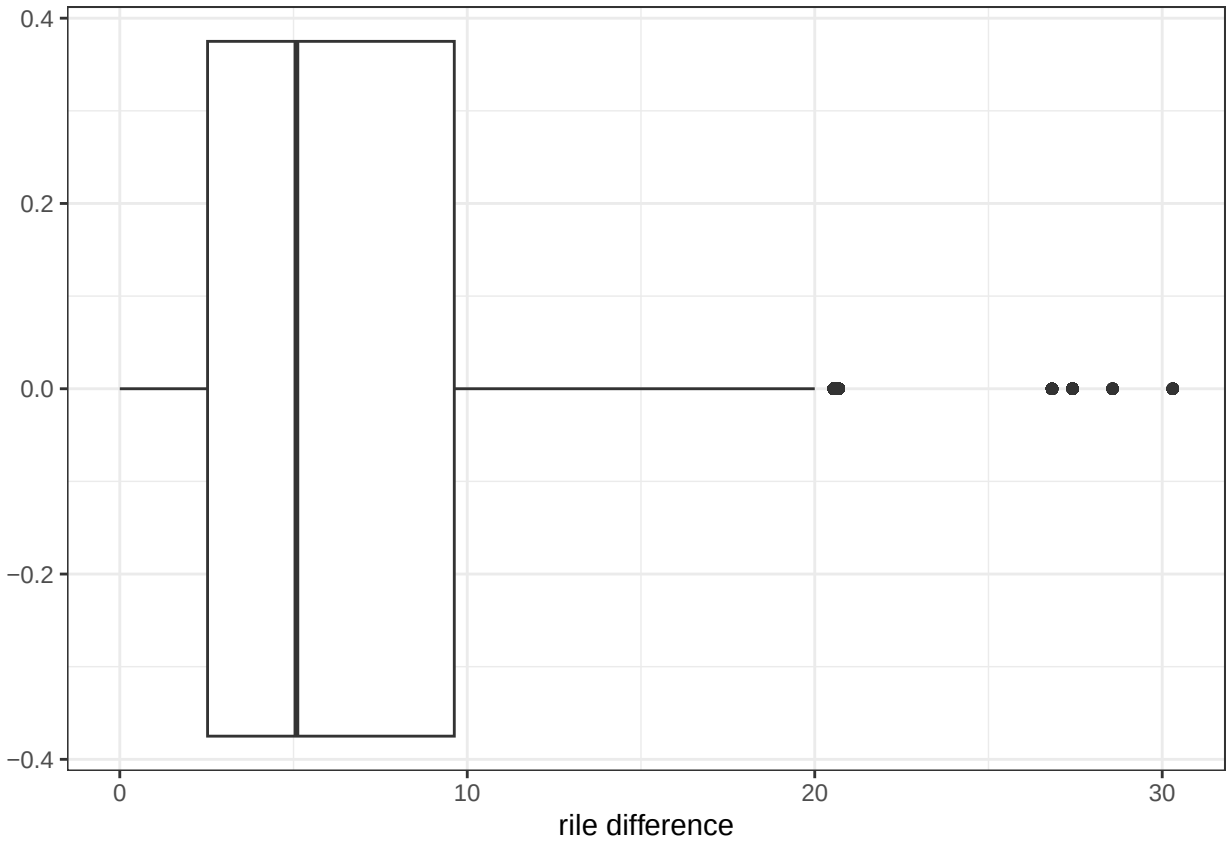

Figure 2: Absolute rile differences between true and predicted cmp codes

Category	Precision	Recall	F1	n(%)	n_predicted(%)
101	0.51	0.33	0.40	0.36%	0.24%
102	0.32	0.21	0.25	0.08%	0.05%
103	0.36	0.26	0.30	0.34%	0.24%
104	0.72	0.72	0.72	1.36%	1.34%
105	0.63	0.65	0.64	0.31%	0.33%
106	0.59	0.52	0.55	0.28%	0.25%
107	0.56	0.60	0.58	2.01%	2.15%
108	0.52	0.63	0.57	0.96%	1.17%
109	0.24	0.21	0.22	0.12%	0.10%
110	0.55	0.52	0.53	0.42%	0.39%
201	0.59	0.52	0.55	2.44%	2.15%
202	0.56	0.58	0.57	3.34%	3.45%
203	0.50	0.38	0.43	0.23%	0.18%
204	0.54	0.33	0.41	0.23%	0.14%
301	0.59	0.59	0.59	2.53%	2.53%
302	0.36	0.16	0.23	0.21%	0.09%
303	0.46	0.52	0.49	4.31%	4.87%
304	0.59	0.56	0.58	1.46%	1.38%
305	0.45	0.48	0.47	2.32%	2.48%
401	0.36	0.27	0.31	1.16%	0.86%
402	0.52	0.50	0.51	2.73%	2.63%
403	0.48	0.49	0.48	3.40%	3.49%
404	0.28	0.17	0.21	0.66%	0.39%
405	0.31	0.31	0.31	0.18%	0.18%
406	0.31	0.33	0.32	0.31%	0.33%
407	0.31	0.35	0.33	0.31%	0.34%
408	0.32	0.18	0.23	1.52%	0.84%
409	0.40	0.15	0.22	0.36%	0.13%
410	0.43	0.48	0.46	1.82%	2.00%
411	0.62	0.66	0.64	7.93%	8.35%
412	0.41	0.20	0.27	0.61%	0.30%
413	0.50	0.47	0.48	0.45%	0.42%
414	0.45	0.52	0.48	1.15%	1.33%
415	0.37	0.16	0.22	0.24%	0.10%
416	0.53	0.45	0.49	2.91%	2.47%
501	0.63	0.72	0.67	5.14%	5.88%
502	0.70	0.76	0.73	3.02%	3.27%
503	0.53	0.57	0.55	5.63%	6.01%
504	0.61	0.73	0.66	9.97%	11.83%
505	0.39	0.27	0.32	0.59%	0.40%
506	0.69	0.75	0.72	5.27%	5.67%
507	0.47	0.12	0.19	0.11%	0.03%
601	0.52	0.40	0.45	1.92%	1.46%
602	0.20	0.22	0.21	0.22%	0.24%
603	0.58	0.58	0.58	1.32%	1.32%
604	0.57	0.47	0.52	0.57%	0.47%
605	0.62	0.65	0.63	4.01%	4.21%
606	0.51	0.42	0.46	1.54%	1.27%
607	0.51	0.55	0.53	1.03%	1.12%
608	0.48	0.37	0.42	0.46%	0.36%
701	0.57	0.62	0.59	3.30%	3.63%
702	0.61	0.18	0.28	0.10%	0.03%
703	0.74	0.72	0.73	3.48%	3.37%
704	0.50	0.20	0.29	0.52%	0.21%
705	0.43	0.22	0.29	1.04%	0.52%
706	0.47	0.27	0.35	1.73%	0.99%

Table 2: Classification Results - Categories

country	accuracy	precision	recall	f1	n
Argentina	43.67%	0.34	0.45	0.38	1,699
Armenia	57.52%	0.44	0.52	0.53	113
Australia	70.29%	0.49	0.52	0.50	4,659
Austria	54.43%	0.40	0.44	0.42	5,412
Belgium	46.62%	0.42	0.34	0.36	29,445
Bolivia	47.90%	0.40	0.42	0.38	1,334
Brazil	51.44%	0.40	0.41	0.50	313
Bulgaria	50.89%	0.43	0.43	0.40	896
Canada	51.51%	0.40	0.39	0.38	5,791
Chile	58.41%	0.42	0.49	0.47	2,188
Colombia	58.93%	0.51	0.48	0.52	577
Costa Rica	62.09%	0.51	0.52	0.50	5,125
Croatia	67.09%	0.51	0.62	0.58	2,932
Cyprus	52.27%	0.42	0.44	0.46	991
Czech Republic	57.81%	0.51	0.52	0.52	1,088
Denmark	50.57%	0.42	0.39	0.42	967
Dominican Republic	66.71%	0.53	0.62	0.55	3,220
Ecuador	50.22%	0.46	0.41	0.44	448
Estonia	63.94%	0.53	0.50	0.50	2,166
Finland	69.99%	0.63	0.55	0.59	1,743
France	56.06%	0.38	0.46	0.43	858
Georgia	67.19%	0.65	0.63	0.74	64
Germany	59.29%	0.55	0.46	0.49	6,738
Greece	62.15%	0.43	0.43	0.43	5,839
Hungary	62.84%	0.55	0.46	0.49	4,987
Ireland	54.35%	0.42	0.46	0.41	2,414
Israel	61.64%	0.55	0.50	0.51	2,333
Italy	50.32%	0.40	0.48	0.46	787
Japan	65.04%	0.50	0.61	0.55	655
Latvia	66.06%	0.49	0.55	0.57	218
Lithuania	61.55%	0.50	0.48	0.48	1,675
Luxembourg	55.36%	0.42	0.50	0.45	3,486
Mexico	51.45%	0.43	0.39	0.39	9,410
Moldova	60.59%	0.55	0.48	0.54	1,426
Montenegro	58.25%	0.54	0.55	0.54	946
Netherlands	54.56%	0.44	0.39	0.40	14,565
New Zealand	59.33%	0.46	0.47	0.44	7,706
North Macedonia	66.05%	0.52	0.51	0.52	10,976
Norway	59.17%	0.49	0.43	0.42	8,960
Panama	59.26%	0.33	0.47	0.44	1,593
Peru	59.76%	0.47	0.53	0.46	5,080
Poland	59.16%	0.46	0.47	0.47	2,370
Portugal	64.21%	0.50	0.51	0.48	6,276
Romania	64.22%	0.61	0.62	0.60	735
Russia	60.39%	0.56	0.53	0.54	818
Serbia	61.02%	0.33	0.53	0.45	862
Slovakia	59.81%	0.46	0.45	0.45	3,847
Slovenia	50.16%	0.45	0.38	0.38	2,205
South Africa	65.36%	0.49	0.50	0.50	1,195
South Korea	63.43%	0.46	0.49	0.48	1,936
Spain	58.87%	0.44	0.45	0.44	7,143
Sweden	61.27%	0.43	0.54	0.52	883
Switzerland	48.47%	0.36	0.41	0.39	2,575
Turkey	48.05%	0.52	0.46	0.47	4,208
Ukraine	53.21%	0.50	0.53	0.52	327
United Kingdom	56.27%	0.48	0.46	0.46	1,843

Table 3: Classification Results - Countries

parfam	accuracy	precision	recall	f1	n
10	51.58%	0.42	0.39	0.39	16,804
20	56.55%	0.44	0.43	0.41	21,206
30	58.82%	0.48	0.43	0.44	38,018
40	54.62%	0.45	0.42	0.42	32,358
50	57.44%	0.47	0.43	0.44	23,732
60	59.74%	0.47	0.43	0.44	26,234
70	57.31%	0.45	0.43	0.42	11,284
80	58.32%	0.50	0.44	0.47	1,694
90	56.19%	0.46	0.47	0.46	14,951
95	57.49%	0.46	0.43	0.44	12,394
98	61.11%	0.64	0.64	0.60	324
999	61.70%	0.65	0.62	0.75	47

Table 4: Classification Results - Parfam

prob_estimates	accuracy	n(%)	cum_n(%)
> 95%	94.99%	9.11%	9.11%
90%-95%	86.75%	8.24%	17.35%
85%-90%	80.10%	7.06%	24.41%
80%-85%	73.92%	6.20%	30.61%
75%-80%	68.96%	5.84%	36.44%
70%-75%	63.54%	5.89%	42.33%
65%-70%	58.55%	5.82%	48.15%
60%-65%	54.70%	5.93%	54.08%
55%-60%	49.91%	6.16%	60.24%
50%-55%	46.16%	6.63%	66.87%
45%-50%	41.65%	6.77%	73.64%
40%-45%	36.85%	6.36%	80.00%
35%-40%	32.35%	5.77%	85.77%
30%-35%	28.18%	4.96%	90.73%
25%-30%	24.19%	4.05%	94.78%
20%-25%	19.12%	2.77%	97.55%
15%-20%	16.02%	1.63%	99.19%
10%-15%	11.14%	0.71%	99.90%
5%-10%	6.40%	0.10%	100.00%

Table 5: Model Calibration - Probability Groups