Manifestoberta Performance Report

Sentence Version 2025a (56Topics)

2026-04-29

Contents

Summary

- Performance was measured on 201 manifestos, which represent 192131 annotated quasi-sentences.
- Overall the model manages to assign the correct category to quasi-sentences in the test data set with an accuracy of 55.62%. In 71.90% of the cases, the true category is among the two most confident predictions of the model, and in 79.63% among the top three (Table 2).
- Lower macro F1 scores (as well as macro precision and recall scores) reflect problems with single categories, especially rare/exotic categories like 102, 409, or 702 (Table 2 and Table 3). That is why we indicate weighted F1 scores here, to account for class imbalance.
- 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 3).
- The model performs considerably well in all countries/languages present in the test data set, with the lowest accuracy value of 44.4% in Brazil (Table 4).
- 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 6.90% of all quasi-sentences in our test set) it was, in fact, right in 95.22% of those cases (Table 5).
- 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

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 2025-1-1

Accuracy	0.56
Top2_Acc	0.72
Top3_Acc	0.80
Weighted Precision	0.55
Macro Precision	0.48
Weighted Recall	0.56
Macro Recall	0.41
Weighted F1	0.55
Macro F1	0.43
MCC	0.54
Cross-Entropy	1.54

Table 1: Classification Results - Overall

Category	Precision	Recall	F1	n(%)	n_predicted(%)
101	0.48	0.24	0.32	0.40%	0.20%
102	0.30	0.18	0.22	0.08%	0.05%
103	0.43	0.30	0.36	0.45%	0.32%
104	0.69	0.67	0.68	1.70%	1.66%
105	0.54	0.57	0.55	0.35%	0.37%
106	0.57	0.53	0.55	0.34%	0.31%
107	0.54	0.60	0.57	2.10%	2.32%
108	0.54	0.58	0.56	1.10%	1.17%
109	0.34	0.17	0.23	0.18%	0.09%
110	0.52	0.47	0.49	0.35%	0.31%
201	0.54	0.52	0.53	2.36%	2.25%
202	0.56	0.55	0.56	3.54%	3.48%
203	0.37	0.27	0.31	0.29%	0.21%
204	0.57	0.34	0.42	0.37%	0.22%
301	0.59	0.58	0.59	3.13%	3.08%
302	0.30	0.08	0.13	0.20%	0.05%
303	0.42	0.51	0.46	3.64%	4.41%
304	0.61	0.50	0.55	1.36%	1.12%
305	0.42	0.52	0.46	2.39%	2.97%
401	0.37	0.26	0.31	1.14%	0.80%
402	0.50	0.51	0.51	2.65%	2.68%
403	0.42	0.45	0.43	3.03%	3.27%
404	0.24	0.17	0.20	0.42%	0.30%
405	0.39	0.25	0.30	0.26%	0.17%
406	0.36	0.21	0.27	0.52%	0.30%
407	0.44	0.37	0.40	0.48%	0.40%
408	0.29	0.09	0.14	1.41%	0.45%
409	0.41	0.17	0.24	0.41%	0.17%
410	0.47	0.43	0.45	2.26%	2.06%
411	0.62	0.67	0.65	7.72%	8.28%
412	0.41	0.14	0.21	0.61%	0.21%
413	0.38	0.36	0.37	0.36%	0.35%
414	0.43	0.52	0.47	1.27%	1.55%
415	0.35	0.21	0.26	0.17%	0.10%
416	0.52	0.41	0.46	2.91%	2.34%
501	0.60	0.72	0.66	4.38%	5.20%
502	0.71	0.74	0.72	3.10%	3.25%
503	0.54	0.59	0.56	6.11%	6.69%
504	0.62	0.69	0.66	10.69%	11.91%
505	0.48	0.18	0.26	0.72%	0.27%
506	0.68	0.76	0.71	4.87%	5.44%
507	0.52	0.07	0.12	0.11%	0.02%
601	0.42	0.40	0.41	1.69%	1.61%
602	0.32	0.20	0.25	0.48%	0.30%
603	0.56	0.56	0.56	1.13%	1.13%
604	0.55	0.40	0.47	0.45%	0.33%
605	0.63	0.67	0.65	4.23%	4.45%
606	0.48	0.37	0.42	1.53%	1.17%
607	0.51	0.51	0.51	1.27%	1.27%
608	0.40	0.37	0.38	0.32%	0.30%
701	0.55	0.62	0.58	3.42%	3.85%
702	0.54	0.20	0.29	0.07%	0.03%
703	0.66	0.72	0.69	3.09%	3.37%
704	0.26	0.23	0.24	0.26%	0.22%
705	0.35	0.16	0.22	0.81%	0.36%
706	0.38	0.23	0.29	1.37%	0.82%

Table 2: Classification Results - Categories

Country	Accuracy	Weighted_Precision	Weighted_Recall	Weighted_F1	Macro_F1	N
Argentina	58.03%	0.63	0.58	0.57	0.47	1,749
Australia	63.92%	0.64	0.64	0.63	0.46	14,488
Austria	53.35%	0.58	0.53	0.53	0.39	4,206
Belgium	46.77%	0.48	0.47	0.46	0.33	18,925
Bosnia-Herzegovina	56.38%	0.66	0.56	0.58	0.47	1,678
Brazil	44.40%	0.47	0.44	0.44	0.32	8,033
Bulgaria	56.42%	0.63	0.56	0.56	0.44	1,028
Canada	50.70%	0.52	0.51	0.50	0.36	10,838
Chile	52.04%	0.54	0.52	0.52	0.39	3,213
Colombia	61.82%	0.70	0.62	0.65	0.60	571
Costa Rica	62.71%	0.65	0.63	0.62	0.51	4,813
Croatia	67.40%	0.72	0.67	0.69	0.64	770
Cyprus	67.53%	0.75	0.68	0.70	0.72	77
Czech Republic	45.83%	0.62	0.46	0.51	0.46	216
Denmark	54.85%	0.60	0.55	0.55	0.44	423
Dominican Republic	65.31%	0.69	0.65	0.64	0.54	2,551
Ecuador	51.25%	0.54	0.51	0.51	0.45	3,774
Estonia	67.57%	0.69	0.68	0.67	0.55	1,952
Finland	53.73%	0.56	0.54	0.53	0.40	3,508
France	47.28%	0.56	0.47	0.48	0.35	3,513
Germany	60.16%	0.62	0.60	0.60	0.47	5,274
Greece	52.65%	0.54	0.53	0.52	0.38	5,745
Hungary	54.09%	0.58	0.54	0.53	0.38	8,279
Iceland	61.72%	0.63	0.62	0.63	0.53	802
Ireland	56.83%	0.59	0.57	0.56	0.40	4,355
Israel	59.59%	0.64	0.60	0.60	0.53	735
Italy	52.14%	0.57	0.52	0.51	0.36	1,565
Lithuania	60.34%	0.62	0.60	0.59	0.42	7,156
Luxembourg	56.10%	0.60	0.56	0.56	0.41	4,720
Mexico	49.70%	0.55	0.50	0.49	0.40	2,318
Moldova	62.96%	0.67	0.63	0.63	0.58	791
Montenegro	60.37%	0.63	0.60	0.60	0.52	381
Netherlands	50.16%	0.53	0.50	0.50	0.37	8,187
New Zealand	53.88%	0.58	0.54	0.53	0.45	2,253
North Macedonia	64.04%	0.66	0.64	0.64	0.51	4,928
Norway	59.82%	0.66	0.60	0.62	0.44	1,665
Panama	66.19%	0.73	0.66	0.68	0.52	1,541
Peru	58.79%	0.63	0.59	0.59	0.45	7,268
Poland	58.40%	0.61	0.58	0.58	0.46	1,959
Portugal	67.02%	0.69	0.67	0.67	0.55	1,804
Romania	58.66%	0.62	0.59	0.60	0.48	1,616
Russia	46.67%	0.54	0.47	0.48	0.33	255
Serbia	51.41%	0.59	0.51	0.52	0.49	284
Slovakia	59.70%	0.60	0.60	0.58	0.43	3,367
Slovenia	50.26%	0.55	0.50	0.50	0.44	1,363
South Africa	61.97%	0.66	0.62	0.62	0.54	1,065
South Korea	53.39%	0.60	0.53	0.54	0.44	2,171
Spain	69.60%	0.71	0.70	0.69	0.54	5,875
Sweden	55.62%	0.59	0.56	0.54	0.40	4,297
Switzerland	50.15%	0.51	0.50	0.49	0.43	1,936
Turkey	53.59%	0.60	0.54	0.55	0.46	3,527
Ukraine	54.18%	0.57	0.54	0.55	0.49	419
United Kingdom	61.85%	0.63	0.62	0.62	0.49	2,860
United States	56.50%	0.62	0.56	0.57	0.49	1,462
Uruguay	48.49%	0.51	0.48	0.48	0.34	3,582

Note: N relates to the number of quasi-sentences in the test dataset.

Table 3: Classification Results - Countries

Parfam	country_name	Accuracy	Weighted_Precision	Weighted_Recall	Weighted_F1	Macro_F1	N
10	Sweden	54.11%	0.55	0.54	0.53	0.53	15,064
20	Denmark	51.42%	0.52	0.51	0.51	0.51	20,136
30	Costa Rica	57.47%	0.57	0.57	0.57	0.57	41,201
40	South Korea	56.02%	0.56	0.56	0.55	0.55	24,905
50	Sweden	55.98%	0.56	0.56	0.55	0.55	16,786
60	South Korea	58.90%	0.58	0.59	0.58	0.58	33,746
70	Sweden	52.95%	0.55	0.53	0.52	0.52	11,327
80	Sweden	57.42%	0.58	0.57	0.55	0.55	3,483
90	Belgium	52.22%	0.54	0.52	0.52	0.52	21,748
95	Iceland	51.58%	0.58	0.52	0.53	0.53	2,443
98	Colombia	65.02%	0.67	0.65	0.65	0.65	1,292

Note: N relates to the number of quasi-sentences in the test dataset.

Table 4: Classification Results - Parfam

prob_estimates	accuracy	n(%)	cum_n(%)
> 95%	95.22%	6.90%	6.90%
90%-95%	87.89%	7.52%	14.42%
85%-90%	80.26%	6.77%	21.19%
80%-85%	75.24%	6.18%	27.37%
75%-80%	69.83%	5.89%	33.26%
70%-75%	64.84%	5.78%	39.05%
65%-70%	60.42%	5.81%	44.86%
60%-65%	54.81%	5.85%	50.71%
55%-60%	51.64%	6.16%	56.87%
50%-55%	46.31%	6.63%	63.50%
45%-50%	43.46%	6.85%	70.35%
40%-45%	38.74%	6.46%	76.81%
35%-40%	33.64%	6.03%	82.84%
30%-35%	29.81%	5.52%	88.36%
25%-30%	24.70%	4.50%	92.86%
20%-25%	20.61%	3.45%	96.32%
15%-20%	16.37%	2.34%	98.66%
10%-15%	12.15%	1.20%	99.85%
5%-10%	8.57%	0.15%	100.00%

Table 5: Model Calibration - Probability Groups

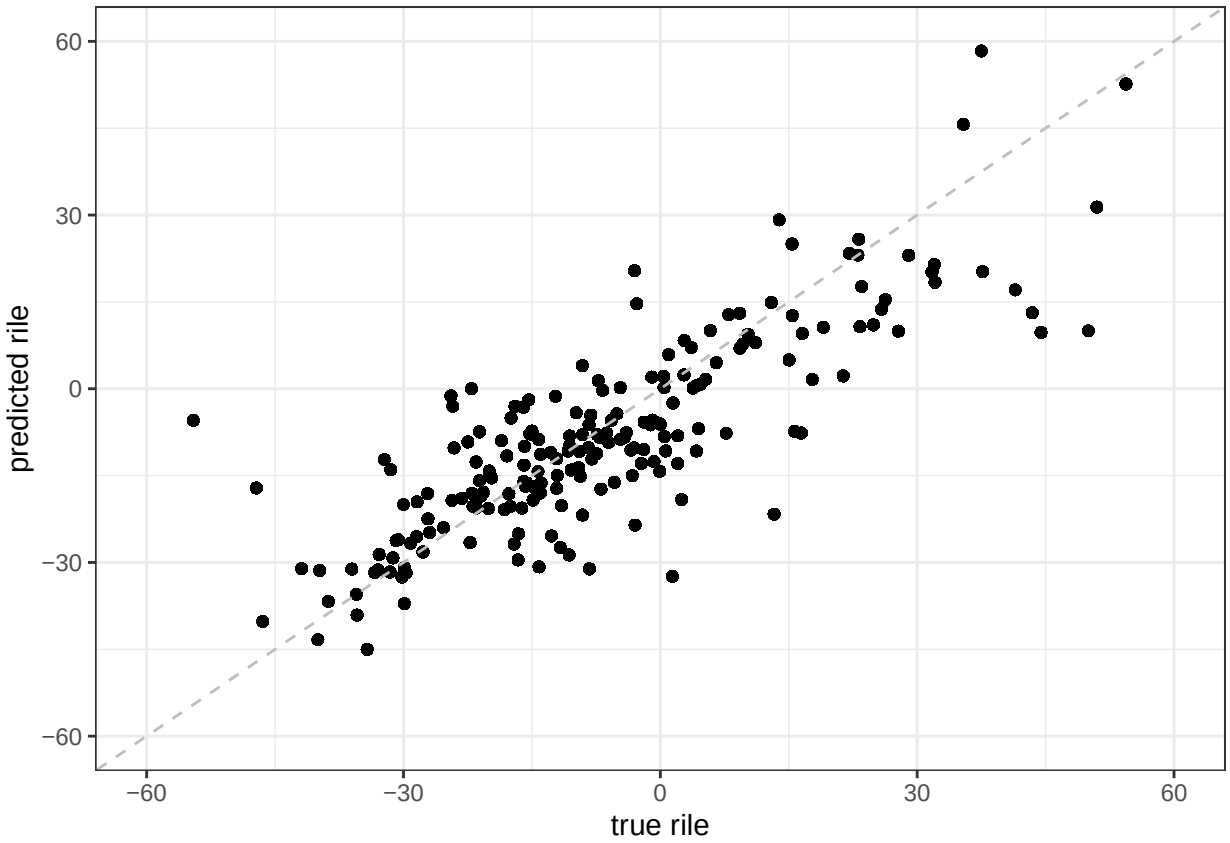

Figure 1: True rile values vs. predicted rile values

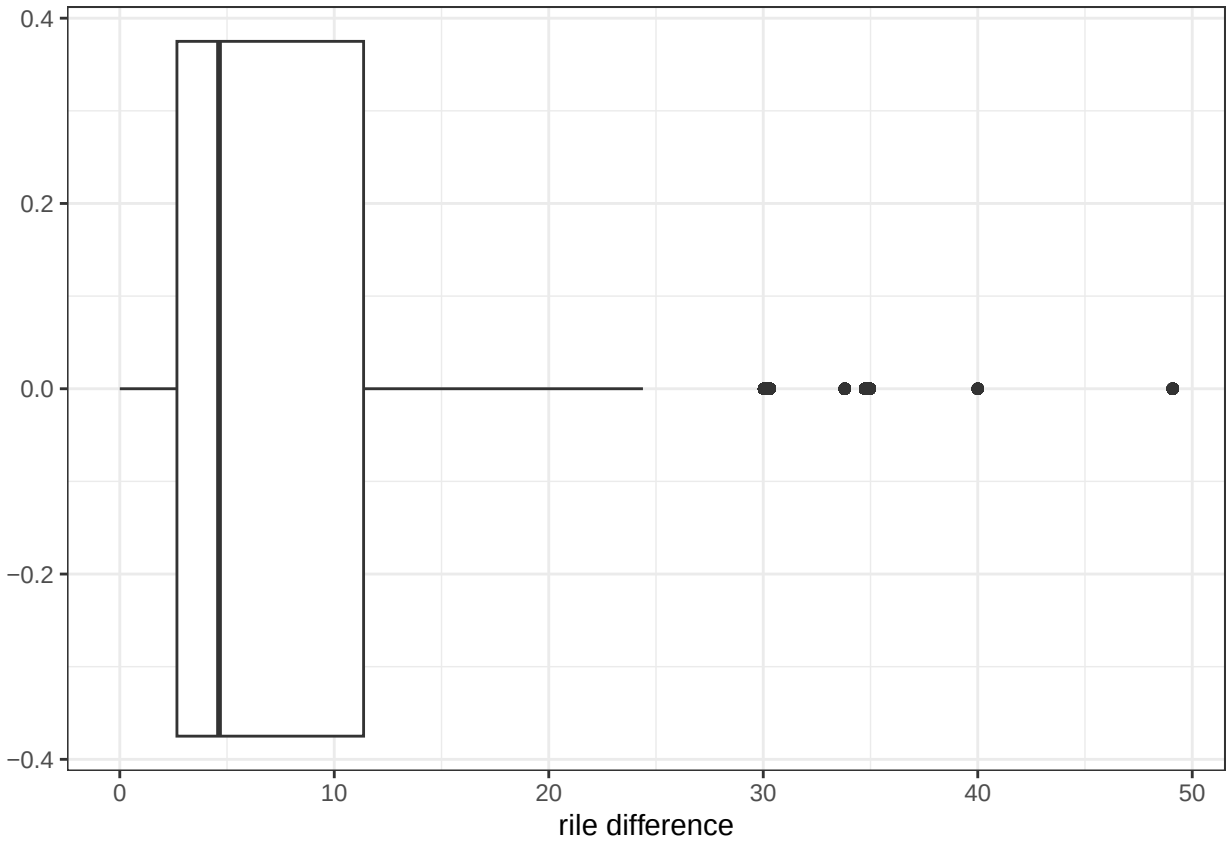

Figure 2: Absolute rile differences between true and predicted cmp codes

Model History

The model is regularly updated. All versions are available on Hugging Face (https://huggingface.co/manifesto-project). The performance of the 2025-1-1 model is slightly better than previous models (The macro F1 score of the 2023-1-1 model is slightly better, as this model was retrained on the full corpus before publication, a practice we did not pursue for the subsequent models). In order to make models comparable, we created a test data set of quasi-sentences that are unseen by ALL three models. This test dataset consists of documents that were not sampled into the training data for all three models plus the quasi-sentences that enter our corpus with the 2026-1 update. This amounts to 37 documents, which contain 59994 quasi-sentences.

Table 6: Classification Results - Comparison between Models

Model	Accuracy	Top2 Acc	Top3 Acc	Weighted Precision	Macro Precision	Weighted Recall	Macro Recall	F1weighted	F1macro	MCC	Cross Entropy
2025-1-1	0.603	0.766	0.838	0.63	0.45	0.6	0.42	0.60	0.41	0.58	1.34
2024-1-1	0.595	0.762	0.835	0.63	0.43	0.6	0.42	0.59	0.39	0.57	1.37
2023-1-1	0.599	0.761	0.837	0.63	0.45	0.6	0.42	0.60	0.41	0.58	1.37