Manifestoberta Performance Report

Sentence Version 2026a (56Topics)

2026-08-20

Contents

Summary 1

Usage Recommendations 1

Results 2026-1-1 2

Model History 8

Summary

- Performance was measured on 204 manifestos, which represent 223798 annotated quasi-sentences.
- Overall the model manages to assign the correct category to quasi-sentences in the test data set with an accuracy of 56.76%. In 72.79% of the cases, the true category is among the two most confident predictions of the model, and in 80.25% 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 45.58% in Belgium and highest accuracy value of 68.77% in Estonia (Table 4).
- Probability estimates of the model are well calibrated and properly reflect the likelihood of a correct prediction. If the model reports a confidence of 95% or higher (which happened or 7.32% of all quasi-sentences in our test set) it was, in fact, right in 95.09% 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 2026-1-1

Accuracy	0.57
Top2_Acc	0.73
Top3_Acc	0.80
Weighted Precision	0.56
Macro Precision	0.48
Weighted Recall	0.57
Macro Recall	0.42
Weighted F1	0.56
Macro F1	0.44
MCC	0.55
Cross-Entropy	1.51

Table 1: Classification Results - Overall

Category	Precision	Recall	F1	n(%)	n_predicted(%)
101	0.46	0.30	0.36	0.42%	0.28%
102	0.34	0.33	0.33	0.07%	0.07%
103	0.50	0.28	0.36	0.47%	0.26%
104	0.64	0.71	0.67	1.55%	1.73%
105	0.63	0.56	0.59	0.36%	0.33%
106	0.57	0.56	0.56	0.41%	0.41%
107	0.54	0.60	0.57	2.11%	2.35%
108	0.54	0.61	0.57	0.96%	1.07%
109	0.46	0.21	0.29	0.18%	0.08%
110	0.52	0.52	0.52	0.26%	0.26%
201	0.53	0.53	0.53	2.05%	2.04%
202	0.59	0.56	0.58	3.66%	3.45%
203	0.46	0.31	0.37	0.31%	0.21%
204	0.58	0.38	0.46	0.31%	0.21%
301	0.56	0.58	0.57	2.97%	3.05%
302	0.25	0.07	0.11	0.17%	0.05%
303	0.49	0.46	0.47	4.23%	3.91%
304	0.63	0.58	0.61	1.36%	1.25%
305	0.36	0.47	0.41	1.80%	2.33%
401	0.36	0.27	0.31	1.14%	0.85%
402	0.47	0.52	0.50	2.52%	2.77%
403	0.43	0.44	0.44	2.93%	2.98%
404	0.29	0.22	0.25	0.54%	0.41%
405	0.40	0.28	0.33	0.25%	0.17%
406	0.40	0.21	0.28	0.50%	0.27%
407	0.42	0.37	0.39	0.57%	0.49%
408	0.29	0.08	0.12	1.59%	0.44%
409	0.16	0.07	0.09	0.30%	0.13%
410	0.41	0.42	0.42	2.17%	2.21%
411	0.63	0.70	0.66	9.09%	10.25%
412	0.43	0.19	0.26	0.47%	0.21%
413	0.43	0.39	0.41	0.37%	0.34%
414	0.45	0.52	0.48	1.27%	1.48%
415	0.38	0.30	0.34	0.15%	0.12%
416	0.52	0.43	0.47	3.06%	2.51%
501	0.60	0.71	0.65	4.29%	5.06%
502	0.73	0.74	0.74	3.38%	3.45%
503	0.58	0.55	0.57	6.16%	5.91%
504	0.64	0.71	0.68	11.35%	12.54%
505	0.32	0.15	0.20	0.41%	0.19%
506	0.72	0.75	0.73	5.24%	5.49%
507	0.50	0.12	0.20	0.08%	0.02%
601	0.45	0.42	0.44	1.51%	1.41%
602	0.25	0.19	0.21	0.43%	0.32%
603	0.61	0.53	0.56	1.18%	1.03%
604	0.57	0.48	0.52	0.41%	0.35%
605	0.58	0.71	0.64	3.59%	4.41%
606	0.44	0.41	0.43	1.36%	1.27%
607	0.53	0.51	0.52	1.23%	1.19%
608	0.41	0.36	0.38	0.24%	0.21%
701	0.54	0.60	0.57	3.08%	3.36%
702	0.54	0.36	0.43	0.05%	0.03%
703	0.67	0.74	0.70	3.40%	3.75%
704	0.31	0.16	0.21	0.25%	0.13%
705	0.31	0.12	0.17	0.65%	0.24%
706	0.40	0.23	0.29	1.15%	0.68%

Table 2: Classification Results - Categories

Country	Accuracy	Weighted_Precision	Weighted_Recall	Weighted_F1	Macro_F1	N
Albania	62.35%	0.67	0.62	0.63	0.46	340
Argentina	57.56%	0.60	0.58	0.57	0.44	4,767
Armenia	60.58%	0.65	0.61	0.63	0.53	241
Australia	61.14%	0.62	0.61	0.61	0.41	12,281
Austria	54.40%	0.56	0.54	0.53	0.43	3,833
Belgium	45.58%	0.47	0.46	0.44	0.34	16,710
Bosnia-Herzegovina	57.64%	0.64	0.58	0.59	0.45	2,526
Brazil	48.64%	0.50	0.49	0.47	0.33	20,650
Bulgaria	57.10%	0.64	0.57	0.57	0.45	1,028
Canada	50.91%	0.53	0.51	0.50	0.37	10,767
Chile	51.46%	0.59	0.51	0.51	0.47	1,267
Colombia	59.60%	0.66	0.60	0.62	0.38	7,735
Costa Rica	61.35%	0.65	0.61	0.61	0.48	8,708
Croatia	57.56%	0.62	0.58	0.58	0.52	172
Cyprus	58.17%	0.69	0.58	0.61	0.50	471
Czech Republic	52.90%	0.60	0.53	0.53	0.46	724
Denmark	52.53%	0.61	0.53	0.53	0.54	316
Dominican Republic	65.74%	0.69	0.66	0.64	0.54	2,551
Ecuador	51.02%	0.55	0.51	0.50	0.42	4,071
Estonia	68.77%	0.69	0.69	0.68	0.52	4,780
Finland	57.66%	0.60	0.58	0.57	0.42	2,565
France	53.87%	0.58	0.54	0.53	0.38	1,966
Germany	62.09%	0.64	0.62	0.61	0.50	3,477
Greece	64.44%	0.68	0.64	0.65	0.47	3,847
Hungary	56.60%	0.61	0.57	0.56	0.45	2,447
Iceland	59.65%	0.63	0.60	0.60	0.48	902
Ireland	56.79%	0.59	0.57	0.56	0.38	4,355
Israel	58.53%	0.59	0.59	0.58	0.40	5,136
Italy	61.90%	0.64	0.62	0.62	0.55	210
Japan	65.46%	0.68	0.65	0.66	0.51	1,413
Latvia	59.69%	0.66	0.60	0.62	0.63	129
Lithuania	59.95%	0.63	0.60	0.59	0.47	2,292
Luxembourg	56.52%	0.58	0.57	0.57	0.43	6,552
Mexico	49.74%	0.57	0.50	0.49	0.37	2,668
Netherlands	49.36%	0.54	0.49	0.50	0.36	6,175
New Zealand	66.94%	0.69	0.67	0.68	0.58	844
North Macedonia	63.93%	0.68	0.64	0.64	0.54	2,517
Norway	57.75%	0.60	0.58	0.57	0.37	9,577
Panama	64.11%	0.72	0.64	0.66	0.50	1,541
Peru	59.31%	0.63	0.59	0.59	0.46	5,992
Poland	60.03%	0.64	0.60	0.60	0.48	2,632
Portugal	64.06%	0.71	0.64	0.66	0.60	128
Romania	60.00%	0.60	0.60	0.59	0.53	925
Russia	47.44%	0.57	0.47	0.48	0.43	390
Serbia	63.80%	0.72	0.64	0.66	0.59	500
Slovakia	59.41%	0.61	0.59	0.59	0.44	3,166
Slovenia	46.96%	0.51	0.47	0.46	0.37	2,368
South Africa	65.07%	0.68	0.65	0.65	0.46	1,675
South Korea	65.78%	0.67	0.66	0.65	0.48	2,472
Spain	67.87%	0.68	0.68	0.67	0.53	16,120
Sweden	55.39%	0.57	0.55	0.54	0.40	3,853
Switzerland	53.68%	0.58	0.54	0.53	0.42	1,412
Turkey	59.85%	0.59	0.60	0.59	0.47	8,047
Ukraine	54.37%	0.58	0.54	0.56	0.48	423
United Kingdom	57.46%	0.59	0.57	0.57	0.43	3,923
United States	51.88%	0.57	0.52	0.52	0.45	3,639
Uruguay	48.16%	0.51	0.48	0.48	0.33	3,582

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

Parfam	Accuracy	Weighted_Precision	Weighted_Recall	Weighted_F1	Macro_F1	N
10	57.78%	0.59	0.58	0.57	0.57	7,783
20	53.47%	0.54	0.53	0.52	0.52	29,385
30	58.50%	0.58	0.59	0.58	0.58	45,014
40	61.53%	0.61	0.62	0.61	0.61	20,429
50	53.75%	0.54	0.54	0.53	0.53	29,921
60	55.61%	0.55	0.56	0.55	0.55	42,426
70	57.47%	0.58	0.57	0.57	0.57	10,843
80	60.22%	0.62	0.60	0.60	0.60	9,887
90	57.63%	0.58	0.58	0.57	0.57	22,481
95	52.43%	0.55	0.52	0.53	0.53	4,013
98	61.82%	0.62	0.62	0.61	0.61	1,616

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%	95.09%	7.32%	7.32%
90%-95%	88.14%	8.06%	15.38%
85%-90%	80.90%	6.97%	22.35%
80%-85%	75.33%	6.38%	28.73%
75%-80%	70.96%	6.06%	34.79%
70%-75%	65.57%	5.88%	40.67%
65%-70%	61.37%	5.80%	46.46%
60%-65%	55.89%	5.82%	52.28%
55%-60%	51.98%	6.20%	58.49%
50%-55%	46.82%	6.37%	64.86%
45%-50%	43.17%	6.63%	71.49%
40%-45%	38.50%	6.26%	77.75%
35%-40%	33.45%	5.76%	83.51%
30%-35%	29.79%	5.14%	88.65%
25%-30%	25.11%	4.30%	92.95%
20%-25%	20.67%	3.51%	96.46%
15%-20%	17.33%	2.18%	98.64%
10%-15%	13.12%	1.19%	99.83%
5%-10%	7.05%	0.17%	100.00%

Table 5: Model Calibration - Probability Groups

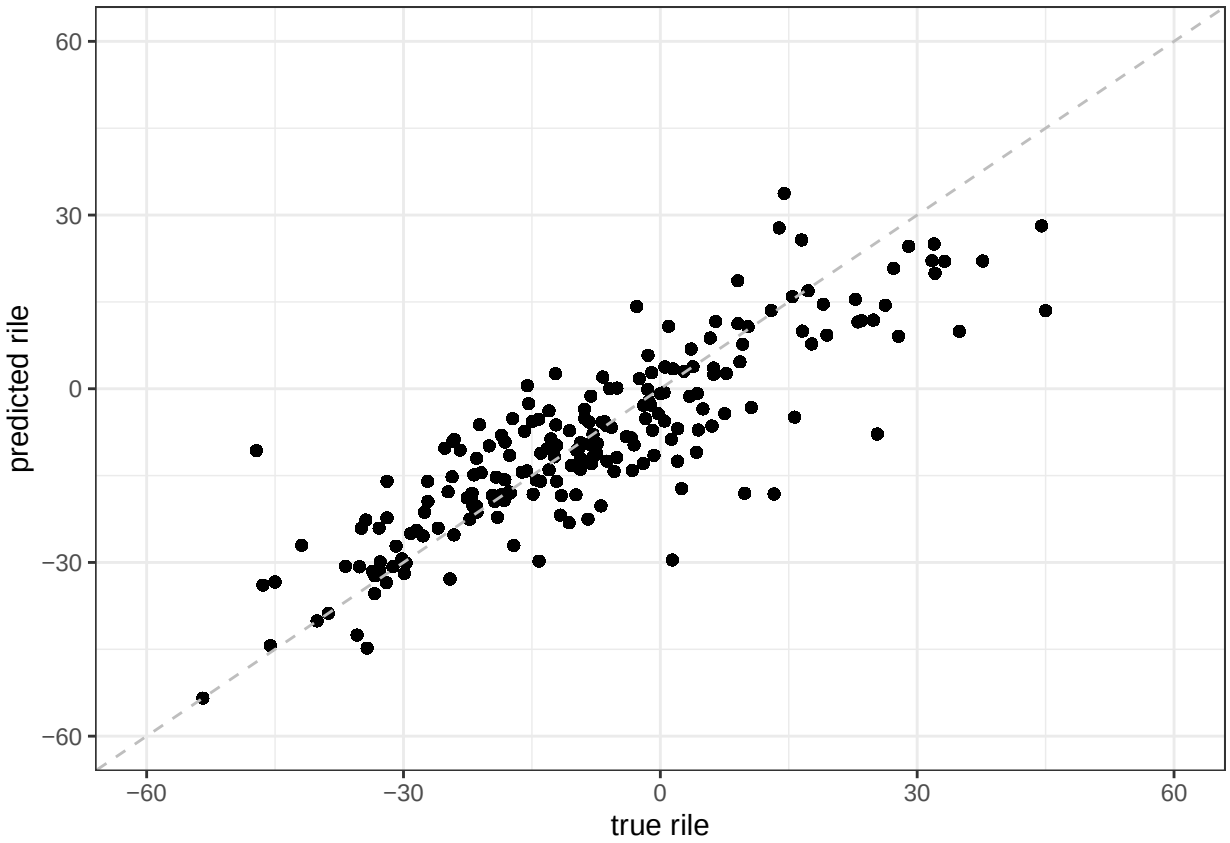

Figure 1: True rile values vs. predicted rile values

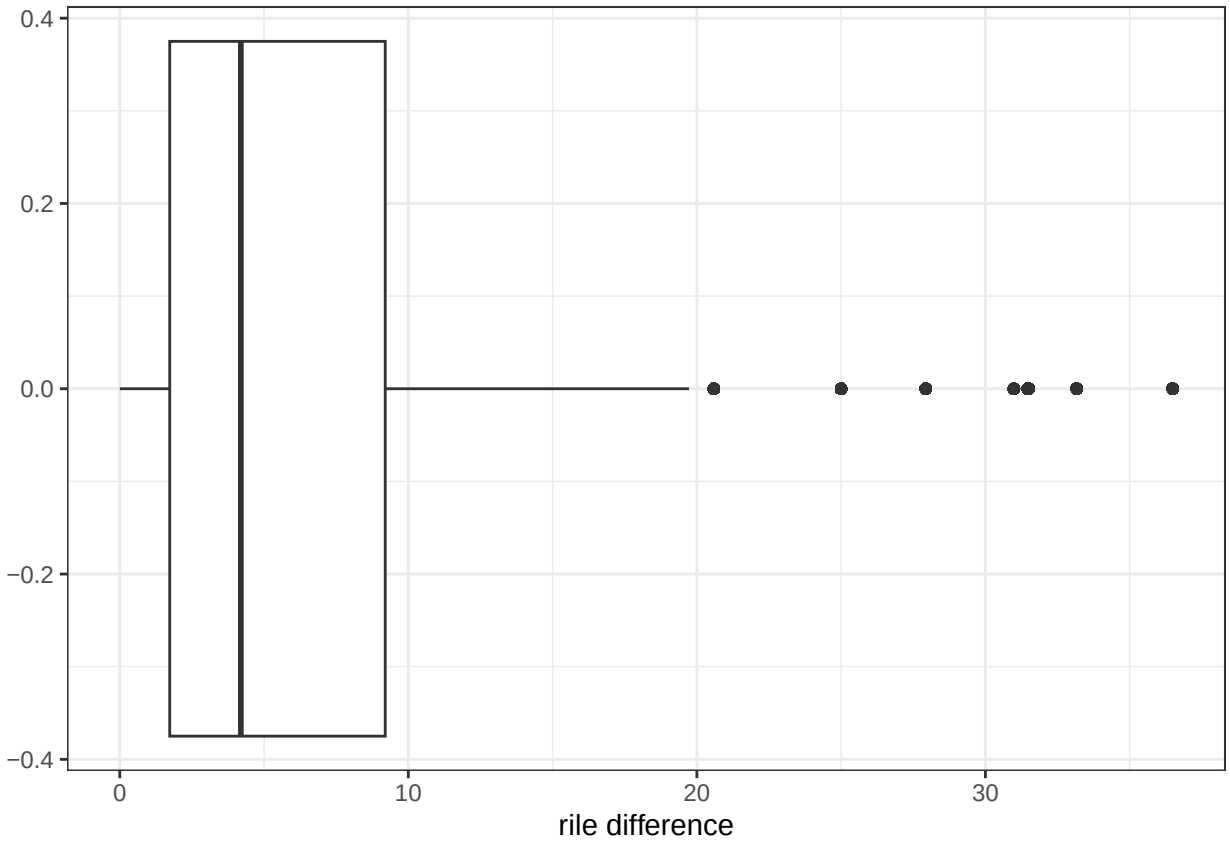

Figure 2: Absolute rile differences between true and predicted cmp codes

Model History

The model is regularly updated. All versions are available on Hugging Face (https://huggingface.co/manifesto-project). In order to make models comparable, we created a test data set of quasi-sentences that are unseen by all models. This test data set consists of documents that were not sampled into the training data for all models. This amounts to 98 documents, which contain 94680 quasi-sentences.

There is no significant difference in performance between the 2026-1-1 model and the previous 2025-1-1 model in the metrics reported below. However, a closer look at individual languages - for example Macedonian, Icelandic, and Czech - shows improved performance for the 2026-1-1 model. These gains are likely related to the fact that new election manifestos in these languages were added in the current update. This illustrates that model improvements are not always fully reflected in the overall metrics, but may become visible when looking more closely at specific languages.

A comparison with the 2023-1-1 and 2024-1-1 models is no longer included, as no fixed test set had been defined for these versions. Additionally, the 2023-1-1 model was retrained on the full corpus before publication, a practice we did not pursue for the subsequent models. However, in the report for the 2025-1-1 model, you can find a comparison of all models from 2023-1-1 to 2025-1-1, based on the 2026 update data, which were unseen by all models. This comparison showed that the 2025-1-1 model performs best. It is therefore reasonable to assume that the 2026-1-1 model also performs better than 2023-1-1 and 2024-1-1.

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
2026-1-1	0.540	0.713	0.797	0.52	0.47	0.54	0.38	0.52	0.39	0.52	1.59
2025-1-1	0.544	0.717	0.802	0.53	0.47	0.54	0.38	0.53	0.40	0.52	1.57