Manifestoberta Performance Report

Context Version 2026-1-1 (56Topics)

2026-08-20

Contents

Summary

- As training parameters, we used the following setting: learning rate: 1e-5, weight decay: 0.01, epochs: 1, warm-up ratio: 0.6, per_device_train_batch_size=16, per_device_eval_batch_size=64, gradient accumulation steps: 2.
- Performance for the 2026-1-1 model 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 64.41%. In 80.99% of the cases, the true category is among the two most confident predictions of the model, and in 87.83% 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 48.18% in Slovenia and highest accuracy value of 76.45% in Spain (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 17.40% of all quasi-sentences in our test set) it was, in fact, right in 94.09% 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

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, 43.41% of the sentences to be coded could be automatically classified with an accuracy of at least 72.80%.

- 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 87.83% for the top 3 suggestions of the model and 93.65% 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 2026-1-1

Accuracy	0.64
Top2_Acc	0.81
Top3_Acc	0.88
Weighted Precision	0.63
Macro Precision	0.55
Weighted Recall	0.64
Macro Recall	0.51
Weighted F1	0.64
Macro F1	0.53
MCC	0.63
Cross-Entropy	1.17

Table 1: Classification Results - Overall

Category	Precision	Recall	F1	n(%)	n_predicted(%)
101	0.50	0.44	0.47	0.42%	0.37%
102	0.40	0.37	0.38	0.07%	0.07%
103	0.58	0.50	0.53	0.47%	0.40%
104	0.74	0.81	0.77	1.55%	1.69%
105	0.70	0.66	0.68	0.36%	0.34%
106	0.64	0.63	0.63	0.41%	0.41%
107	0.64	0.66	0.65	2.11%	2.17%
108	0.62	0.70	0.66	0.96%	1.10%
109	0.50	0.26	0.34	0.18%	0.10%
110	0.61	0.67	0.64	0.26%	0.29%
201	0.61	0.60	0.61	2.05%	2.01%
202	0.66	0.62	0.64	3.66%	3.45%
203	0.54	0.36	0.43	0.31%	0.20%
204	0.60	0.52	0.56	0.31%	0.27%
301	0.64	0.64	0.64	2.97%	2.98%
302	0.26	0.17	0.20	0.17%	0.11%
303	0.62	0.52	0.56	4.23%	3.55%
304	0.68	0.71	0.70	1.36%	1.42%
305	0.45	0.50	0.48	1.80%	1.98%
401	0.46	0.38	0.42	1.14%	0.92%
402	0.54	0.58	0.56	2.52%	2.69%
403	0.52	0.50	0.51	2.93%	2.81%
404	0.32	0.29	0.31	0.54%	0.49%
405	0.46	0.34	0.40	0.25%	0.18%
406	0.49	0.26	0.34	0.50%	0.26%
407	0.47	0.48	0.47	0.57%	0.58%
408	0.36	0.15	0.22	1.59%	0.68%
409	0.26	0.10	0.15	0.30%	0.11%
410	0.50	0.48	0.49	2.17%	2.07%
411	0.70	0.77	0.73	9.09%	10.07%
412	0.41	0.26	0.32	0.47%	0.30%
413	0.53	0.43	0.47	0.37%	0.30%
414	0.55	0.61	0.58	1.27%	1.43%
415	0.43	0.42	0.43	0.15%	0.15%
416	0.58	0.55	0.56	3.06%	2.89%
501	0.70	0.75	0.72	4.29%	4.63%
502	0.76	0.84	0.80	3.38%	3.75%
503	0.64	0.61	0.62	6.16%	5.86%
504	0.74	0.77	0.76	11.35%	11.95%
505	0.53	0.29	0.38	0.41%	0.23%
506	0.77	0.81	0.79	5.24%	5.46%
507	0.59	0.26	0.36	0.08%	0.03%
601	0.54	0.50	0.52	1.51%	1.39%
602	0.37	0.27	0.31	0.43%	0.32%
603	0.69	0.67	0.68	1.18%	1.14%
604	0.65	0.63	0.64	0.41%	0.39%
605	0.68	0.80	0.73	3.59%	4.19%
606	0.51	0.48	0.49	1.36%	1.29%
607	0.61	0.64	0.62	1.23%	1.28%
608	0.55	0.53	0.54	0.24%	0.22%
701	0.61	0.63	0.62	3.08%	3.17%
702	0.40	0.50	0.45	0.05%	0.06%
703	0.74	0.89	0.81	3.40%	4.09%
704	0.37	0.29	0.33	0.25%	0.20%
705	0.43	0.32	0.36	0.65%	0.49%
706	0.45	0.39	0.42	1.15%	0.99%

Table 2: Classification Results - Categories

Country	Accuracy	Weighted_Precision	Weighted_Recall	Weighted_F1	Macro_F1	N
Albania	75.59%	0.78	0.76	0.76	0.60	340
Argentina	62.51%	0.63	0.63	0.61	0.47	4,767
Armenia	63.07%	0.67	0.63	0.66	0.60	241
Australia	70.08%	0.71	0.70	0.70	0.50	12,281
Austria	63.14%	0.65	0.63	0.62	0.53	3,833
Belgium	51.88%	0.53	0.52	0.51	0.41	16,710
Bosnia-Herzegovina	66.47%	0.71	0.66	0.67	0.51	2,526
Brazil	55.52%	0.56	0.56	0.54	0.41	20,650
Bulgaria	58.85%	0.66	0.59	0.58	0.46	1,028
Canada	56.49%	0.58	0.56	0.55	0.41	10,767
Chile	58.09%	0.63	0.58	0.57	0.54	1,267
Colombia	71.93%	0.77	0.72	0.74	0.48	7,735
Costa Rica	71.44%	0.73	0.71	0.71	0.59	8,708
Croatia	68.02%	0.72	0.68	0.71	0.75	172
Cyprus	61.15%	0.68	0.61	0.64	0.52	471
Czech Republic	57.46%	0.63	0.57	0.57	0.53	724
Denmark	62.34%	0.73	0.62	0.68	0.62	316
Dominican Republic	72.60%	0.75	0.73	0.72	0.63	2,551
Ecuador	58.86%	0.60	0.59	0.58	0.51	4,071
Estonia	73.03%	0.73	0.73	0.72	0.56	4,780
Finland	65.22%	0.67	0.65	0.65	0.51	2,565
France	60.48%	0.63	0.60	0.60	0.42	1,966
Germany	69.23%	0.70	0.69	0.68	0.57	3,477
Greece	72.58%	0.74	0.73	0.73	0.52	3,847
Hungary	65.84%	0.68	0.66	0.65	0.54	2,447
Iceland	68.85%	0.70	0.69	0.69	0.58	902
Ireland	63.97%	0.66	0.64	0.63	0.43	4,355
Israel	63.71%	0.64	0.64	0.63	0.45	5,136
Italy	67.62%	0.71	0.68	0.71	0.63	210
Japan	69.07%	0.71	0.69	0.69	0.54	1,413
Latvia	62.79%	0.70	0.63	0.66	0.62	129
Lithuania	66.54%	0.69	0.67	0.65	0.52	2,292
Luxembourg	64.99%	0.66	0.65	0.65	0.48	6,552
Mexico	58.81%	0.64	0.59	0.59	0.43	2,668
Netherlands	60.53%	0.63	0.61	0.61	0.44	6,175
New Zealand	71.21%	0.73	0.71	0.71	0.63	844
North Macedonia	75.33%	0.78	0.75	0.76	0.65	2,517
Norway	64.87%	0.68	0.65	0.65	0.46	9,577
Panama	74.95%	0.79	0.75	0.76	0.58	1,541
Peru	69.88%	0.71	0.70	0.69	0.53	5,992
Poland	69.41%	0.73	0.69	0.70	0.54	2,632
Portugal	67.19%	0.78	0.67	0.69	0.65	128
Romania	64.86%	0.64	0.65	0.64	0.58	925
Russia	56.15%	0.66	0.56	0.57	0.53	390
Serbia	70.00%	0.77	0.70	0.71	0.65	500
Slovakia	65.35%	0.68	0.65	0.65	0.49	3,166
Slovenia	48.18%	0.52	0.48	0.48	0.39	2,368
South Africa	69.97%	0.72	0.70	0.70	0.53	1,675
South Korea	75.12%	0.76	0.75	0.74	0.57	2,472
Spain	76.45%	0.77	0.76	0.76	0.62	16,120
Sweden	63.07%	0.66	0.63	0.63	0.51	3,853
Switzerland	63.03%	0.65	0.63	0.62	0.52	1,412
Turkey	67.35%	0.67	0.67	0.66	0.56	8,047
Ukraine	56.97%	0.60	0.57	0.59	0.50	423
United Kingdom	64.39%	0.65	0.64	0.64	0.50	3,923
United States	60.40%	0.65	0.60	0.60	0.55	3,639
Uruguay	53.69%	0.55	0.54	0.52	0.37	3,582

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

Table 3: Classification Results per Country

Parfam	Accuracy	Weighted_Precision	Weighted_Recall	Weighted_F1	Macro_F1	N
10	64.83%	0.65	0.65	0.64	0.47	7,783
20	60.98%	0.61	0.61	0.60	0.47	29,385
30	66.60%	0.66	0.67	0.66	0.53	45,014
40	67.67%	0.67	0.68	0.67	0.51	20,429
50	61.18%	0.61	0.61	0.60	0.48	29,921
60	63.90%	0.63	0.64	0.63	0.48	42,426
70	65.25%	0.66	0.65	0.65	0.49	10,843
80	69.67%	0.71	0.70	0.69	0.49	9,887
90	63.77%	0.64	0.64	0.63	0.51	22,481
95	63.00%	0.64	0.63	0.63	0.51	4,013
98	70.67%	0.71	0.71	0.70	0.57	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%	94.09%	17.40%	17.40%
90%-95%	84.99%	10.77%	28.17%
85%-90%	78.26%	8.20%	36.38%
80%-85%	72.80%	7.03%	43.41%
75%-80%	66.88%	6.35%	49.76%
70%-75%	63.18%	6.00%	55.76%
65%-70%	58.71%	5.79%	61.55%
60%-65%	54.31%	5.76%	67.31%
55%-60%	49.73%	5.96%	73.26%
50%-55%	46.52%	6.16%	79.42%
45%-50%	41.56%	5.82%	85.24%
40%-45%	36.71%	4.73%	89.97%
35%-40%	32.34%	3.94%	93.91%
30%-35%	28.24%	2.97%	96.88%
25%-30%	23.94%	1.90%	98.79%
20%-25%	20.40%	0.93%	99.71%
15%-20%	15.02%	0.27%	99.99%
10%-15%	6.06%	0.01%	100.00%

Table 5: Model Calibration - Probability Groups

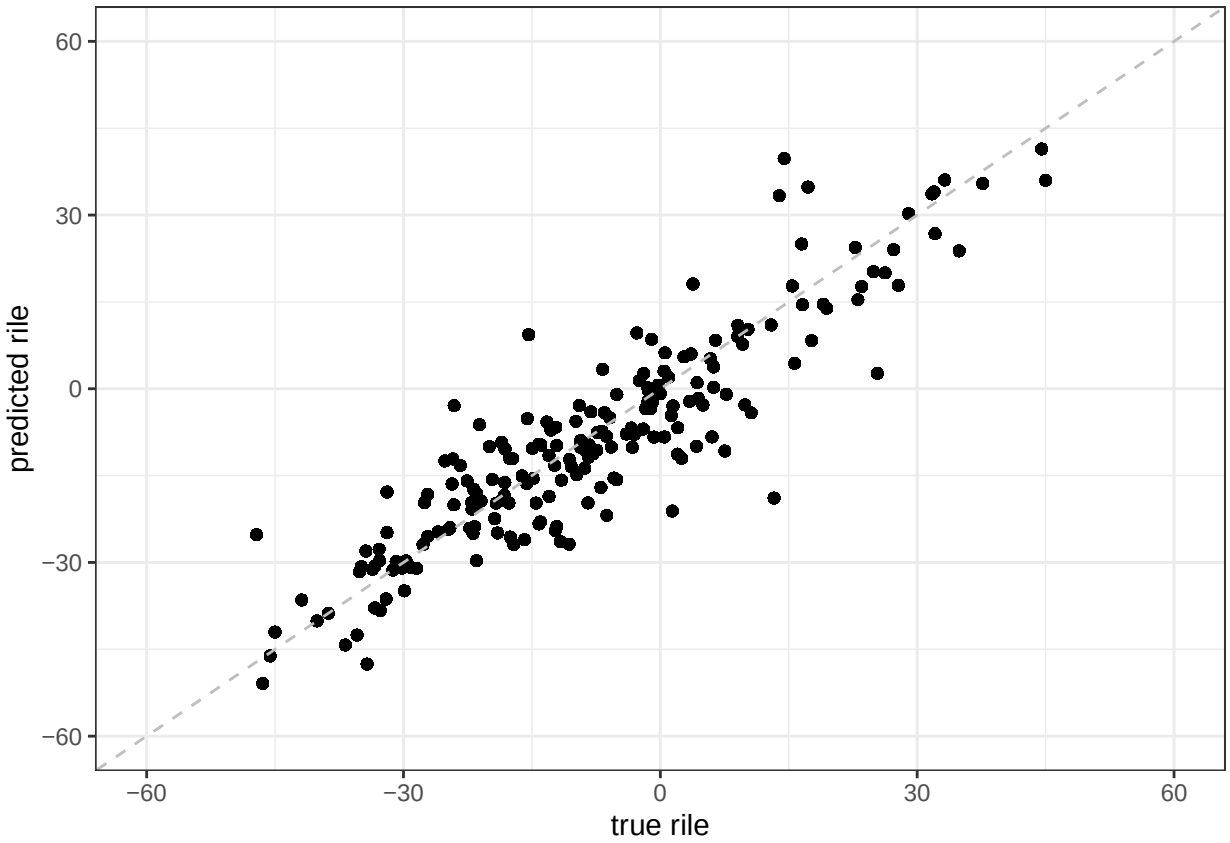

Figure 1: True rile values vs. predicted rile values

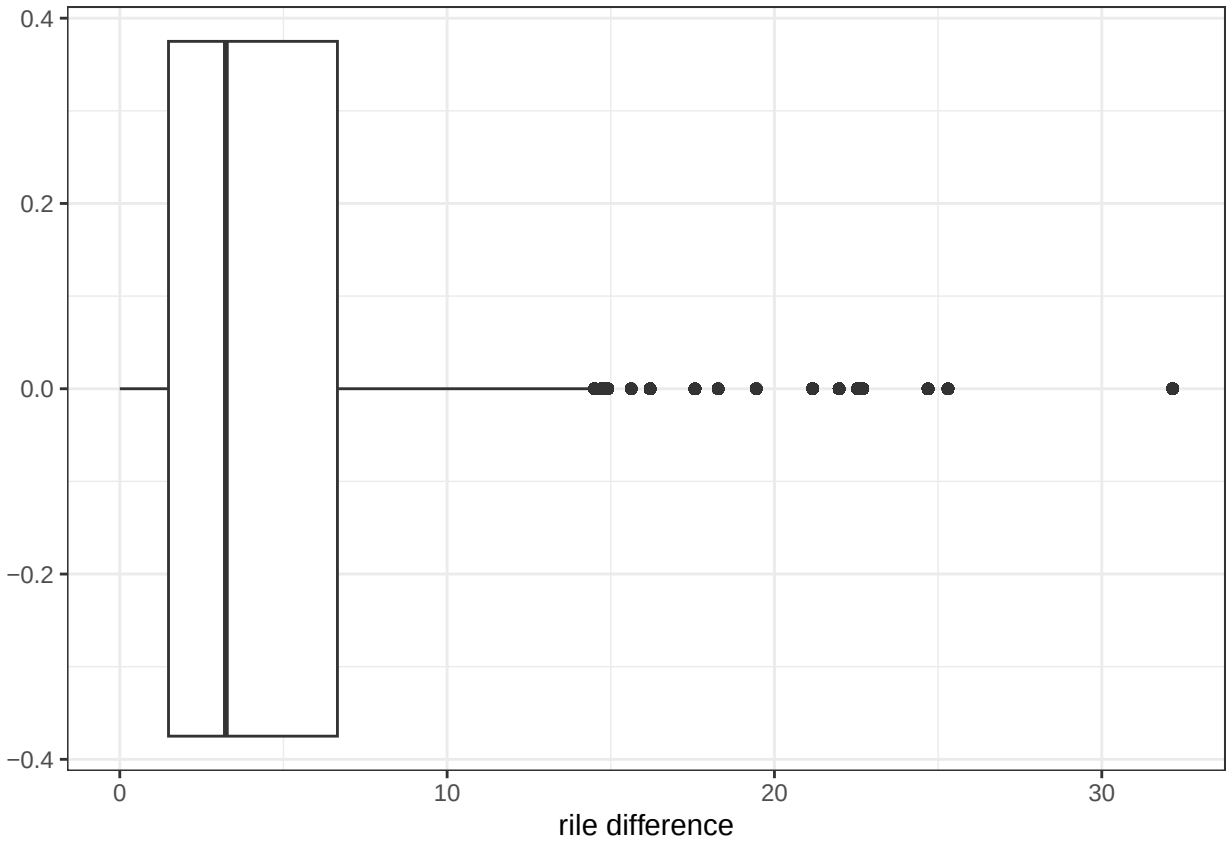

Figure 2: Absolute rile differences between true and predicted cmp codes

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	F1 weighted	F1 macro	MCC	Cross Entropy
2026-1-1	0.619	0.792	0.864	0.61	0.53	0.62	0.49	0.61	0.5	0.6	1.26
2025-1-1	0.620	0.792	0.865	0.61	0.53	0.62	0.48	0.61	0.5	0.6	1.25