Manifestoberta Performance Report

Context Version 2025-1-1 (56Topics)

2025-04-29

Contents

Summary 1

Usage Recommendations 1

Results 2025-1-1 2

Model History 7

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 2025-1-1 model 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 63.32%. In 80.40% of the cases, the true category is among the two most confident predictions of the model, and in 87.54% among the top three (Table 2).
- We indicate weighted F1 scores here, to account for class imbalance and ensuing problems with some individual categories, especially rare/exotic categories like 102, 409, or 702 (Table 2 and Table 3).
- 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 49.54% in Czech Republic (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 15.66% of all quasi-sentences in our test set) it was, in fact, right in 94.68% 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, 42% of the sentences to be coded could be automatically classified with an accuracy of at least 73.6%.

- 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 89% 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 2025-1-1

Accuracy	0.63
Top2_Acc	0.80
Top3_Acc	0.88
Weighted Precision	0.63
Macro Precision	0.55
Weighted Recall	0.63
Macro Recall	0.50
Weighted F1	0.63
Macro F1	0.51
MCC	0.62
Cross-Entropy	1.19

Table 1: Classification Results - Overall

Category	Precision	Recall	F1	n(%)	n_predicted(%)
101	0.55	0.42	0.47	0.40%	0.30%
102	0.33	0.20	0.25	0.08%	0.05%
103	0.53	0.47	0.50	0.45%	0.40%
104	0.77	0.78	0.77	1.70%	1.72%
105	0.61	0.64	0.62	0.35%	0.37%
106	0.60	0.61	0.60	0.34%	0.34%
107	0.63	0.68	0.65	2.10%	2.27%
108	0.60	0.67	0.63	1.10%	1.22%
109	0.40	0.20	0.27	0.18%	0.09%
110	0.64	0.61	0.63	0.35%	0.33%
201	0.59	0.60	0.59	2.36%	2.41%
202	0.63	0.62	0.62	3.54%	3.47%
203	0.35	0.29	0.32	0.29%	0.24%
204	0.53	0.40	0.46	0.37%	0.28%
301	0.67	0.64	0.66	3.13%	2.98%
302	0.32	0.12	0.17	0.20%	0.07%
303	0.54	0.56	0.55	3.64%	3.75%
304	0.69	0.64	0.66	1.36%	1.27%
305	0.54	0.59	0.56	2.39%	2.62%
401	0.45	0.37	0.41	1.14%	0.95%
402	0.59	0.58	0.59	2.65%	2.58%
403	0.49	0.51	0.50	3.03%	3.15%
404	0.30	0.23	0.26	0.42%	0.32%
405	0.37	0.27	0.31	0.26%	0.19%
406	0.43	0.27	0.33	0.52%	0.33%
407	0.51	0.50	0.50	0.48%	0.47%
408	0.33	0.16	0.22	1.41%	0.68%
409	0.51	0.32	0.39	0.41%	0.26%
410	0.56	0.49	0.52	2.26%	1.95%
411	0.72	0.73	0.72	7.72%	7.90%
412	0.41	0.18	0.25	0.61%	0.27%
413	0.53	0.40	0.46	0.36%	0.27%
414	0.51	0.59	0.55	1.27%	1.48%
415	0.42	0.28	0.34	0.17%	0.11%
416	0.58	0.49	0.53	2.91%	2.45%
501	0.67	0.80	0.72	4.38%	5.23%
502	0.76	0.84	0.80	3.10%	3.43%
503	0.60	0.67	0.63	6.11%	6.91%
504	0.72	0.74	0.73	10.69%	10.95%
505	0.61	0.38	0.47	0.72%	0.45%
506	0.75	0.82	0.79	4.87%	5.30%
507	0.68	0.24	0.36	0.11%	0.04%
601	0.52	0.49	0.50	1.69%	1.59%
602	0.38	0.30	0.33	0.48%	0.38%
603	0.64	0.69	0.66	1.13%	1.20%
604	0.65	0.50	0.56	0.45%	0.35%
605	0.70	0.75	0.72	4.23%	4.49%
606	0.57	0.42	0.48	1.53%	1.13%
607	0.57	0.62	0.60	1.27%	1.36%
608	0.52	0.50	0.51	0.32%	0.30%
701	0.64	0.63	0.64	3.42%	3.40%
702	0.49	0.36	0.42	0.07%	0.05%
703	0.76	0.88	0.82	3.09%	3.55%
704	0.28	0.38	0.33	0.26%	0.34%
705	0.43	0.37	0.40	0.81%	0.70%
706	0.43	0.40	0.41	1.37%	1.29%

Table 2: Classification Results - Categories

Country	Accuracy	Weighted_Precision	Weighted_Recall	Weighted_F1	Macro_F1	N
Argentina	64.09%	0.66	0.64	0.63	0.55	1,749
Australia	71.74%	0.71	0.72	0.71	0.53	14,488
Austria	63.01%	0.66	0.63	0.62	0.48	4,206
Belgium	53.82%	0.55	0.54	0.53	0.39	18,925
Bosnia-Herzegovina	67.64%	0.74	0.68	0.69	0.55	1,678
Brazil	51.21%	0.53	0.51	0.51	0.40	8,033
Bulgaria	59.92%	0.67	0.60	0.59	0.48	1,028
Canada	57.10%	0.58	0.57	0.56	0.41	10,838
Chile	63.90%	0.66	0.64	0.64	0.53	3,213
Colombia	73.73%	0.78	0.74	0.74	0.61	571
Costa Rica	72.22%	0.74	0.72	0.72	0.62	4,813
Croatia	72.47%	0.75	0.72	0.72	0.66	770
Cyprus	70.13%	0.76	0.70	0.75	0.77	77
Czech Republic	49.54%	0.63	0.50	0.55	0.49	216
Denmark	61.94%	0.68	0.62	0.65	0.54	423
Dominican Republic	71.58%	0.74	0.72	0.71	0.56	2,551
Ecuador	58.51%	0.60	0.59	0.58	0.50	3,774
Estonia	73.26%	0.74	0.73	0.73	0.61	1,952
Finland	61.80%	0.65	0.62	0.61	0.50	3,508
France	54.63%	0.60	0.55	0.55	0.40	3,513
Germany	68.03%	0.69	0.68	0.68	0.55	5,274
Greece	60.19%	0.61	0.60	0.60	0.45	5,745
Hungary	66.37%	0.68	0.66	0.66	0.51	8,279
Iceland	65.46%	0.68	0.65	0.66	0.52	802
Ireland	63.88%	0.66	0.64	0.63	0.45	4,355
Israel	66.26%	0.70	0.66	0.66	0.62	735
Italy	61.98%	0.66	0.62	0.61	0.45	1,565
Lithuania	65.75%	0.68	0.66	0.65	0.48	7,156
Luxembourg	63.33%	0.68	0.63	0.64	0.47	4,720
Mexico	55.09%	0.60	0.55	0.54	0.46	2,318
Moldova	67.76%	0.72	0.68	0.68	0.62	791
Montenegro	66.40%	0.66	0.66	0.65	0.57	381
Netherlands	60.30%	0.61	0.60	0.60	0.46	8,187
New Zealand	61.56%	0.64	0.62	0.61	0.52	2,253
North Macedonia	74.41%	0.76	0.74	0.75	0.64	4,928
Norway	67.75%	0.73	0.68	0.70	0.52	1,665
Panama	76.18%	0.79	0.76	0.77	0.61	1,541
Peru	68.90%	0.70	0.69	0.68	0.53	7,268
Poland	69.47%	0.72	0.69	0.69	0.58	1,959
Portugal	78.16%	0.80	0.78	0.78	0.66	1,804
Romania	62.87%	0.67	0.63	0.64	0.50	1,616
Russia	50.20%	0.53	0.50	0.49	0.37	255
Serbia	67.96%	0.73	0.68	0.69	0.59	284
Slovakia	65.01%	0.66	0.65	0.64	0.51	3,367
Slovenia	53.71%	0.58	0.54	0.53	0.46	1,363
South Africa	69.30%	0.72	0.69	0.69	0.60	1,065
South Korea	61.91%	0.66	0.62	0.62	0.52	2,171
Spain	74.23%	0.76	0.74	0.74	0.58	5,875
Sweden	62.02%	0.65	0.62	0.62	0.48	4,297
Switzerland	53.46%	0.54	0.53	0.52	0.49	1,936
Turkey	62.97%	0.71	0.63	0.62	0.49	3,527
Ukraine	54.65%	0.58	0.55	0.55	0.49	419
United Kingdom	67.80%	0.68	0.68	0.67	0.51	2,860
United States	58.76%	0.65	0.59	0.59	0.53	1,462
Uruguay	55.30%	0.57	0.55	0.54	0.39	3,582

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

Table 3: Classification Results - Countries

Parfam	Accuracy	Weighted_Precision	Weighted_Recall	Weighted_F1	Macro_F1	N
10	60.53%	0.61	0.61	0.60	0.45	15,064
20	59.16%	0.59	0.59	0.59	0.45	20,136
30	65.11%	0.65	0.65	0.64	0.52	41,201
40	64.24%	0.64	0.64	0.63	0.50	24,905
50	63.43%	0.63	0.63	0.63	0.48	16,786
60	67.38%	0.67	0.67	0.66	0.50	33,746
70	61.65%	0.63	0.62	0.61	0.45	11,327
80	64.71%	0.65	0.65	0.63	0.53	3,483
90	58.29%	0.61	0.58	0.58	0.47	21,748
95	62.79%	0.65	0.63	0.62	0.45	2,443
98	74.77%	0.75	0.75	0.74	0.57	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%	94.68%	15.66%	15.66%
90%-95%	85.17%	10.21%	25.87%
85%-90%	77.64%	7.97%	33.84%
80%-85%	72.28%	6.94%	40.78%
75%-80%	67.89%	6.37%	47.15%
70%-75%	63.02%	6.16%	53.30%
65%-70%	57.96%	6.00%	59.30%
60%-65%	53.91%	6.04%	65.34%
55%-60%	49.22%	6.17%	71.51%
50%-55%	46.66%	6.42%	77.93%
45%-50%	42.64%	6.07%	84.00%
40%-45%	37.80%	5.17%	89.17%
35%-40%	32.89%	4.21%	93.38%
30%-35%	28.48%	3.24%	96.62%
25%-30%	25.26%	2.10%	98.72%
20%-25%	20.08%	1.00%	99.72%
15%-20%	15.88%	0.27%	99.98%
10%-15%	11.43%	0.02%	100.00%

Table 5: Model Calibration - Probability Groups

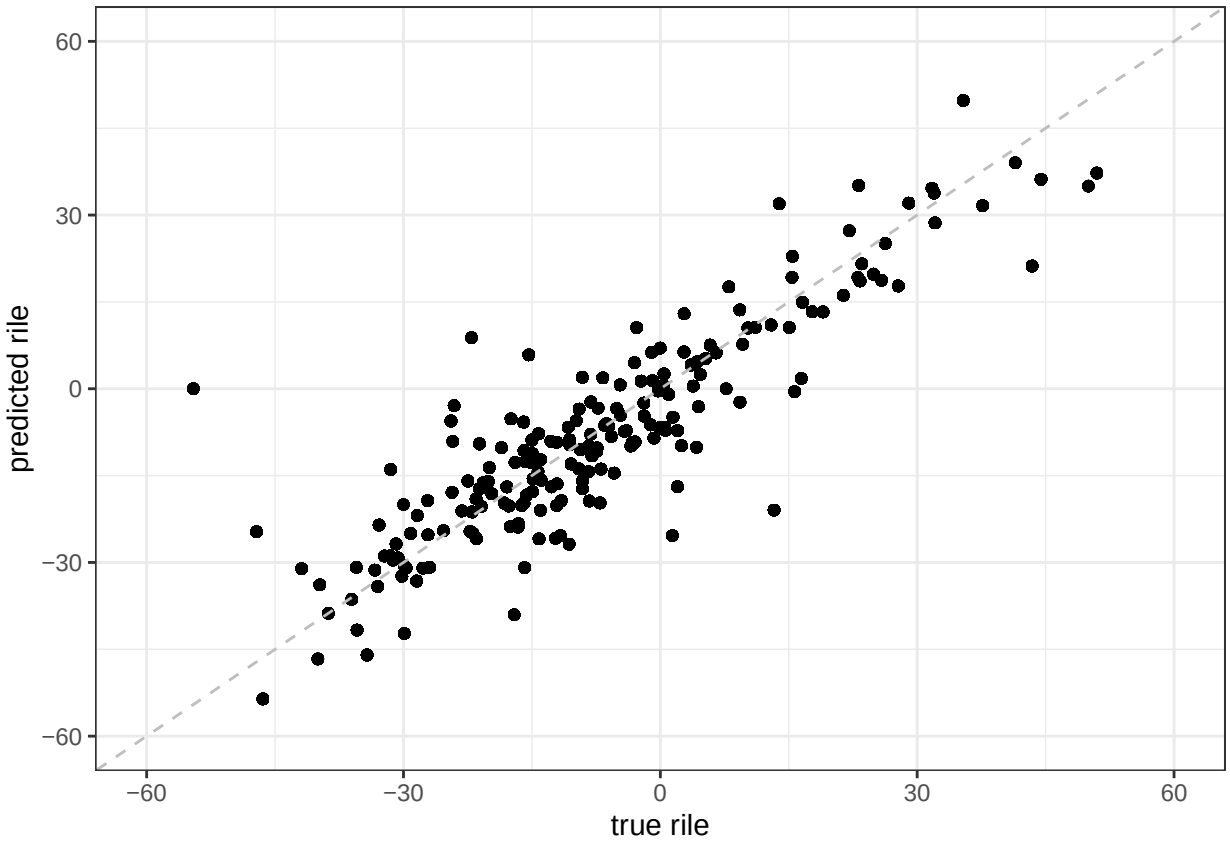

Figure 1: True rile values vs. predicted rile values

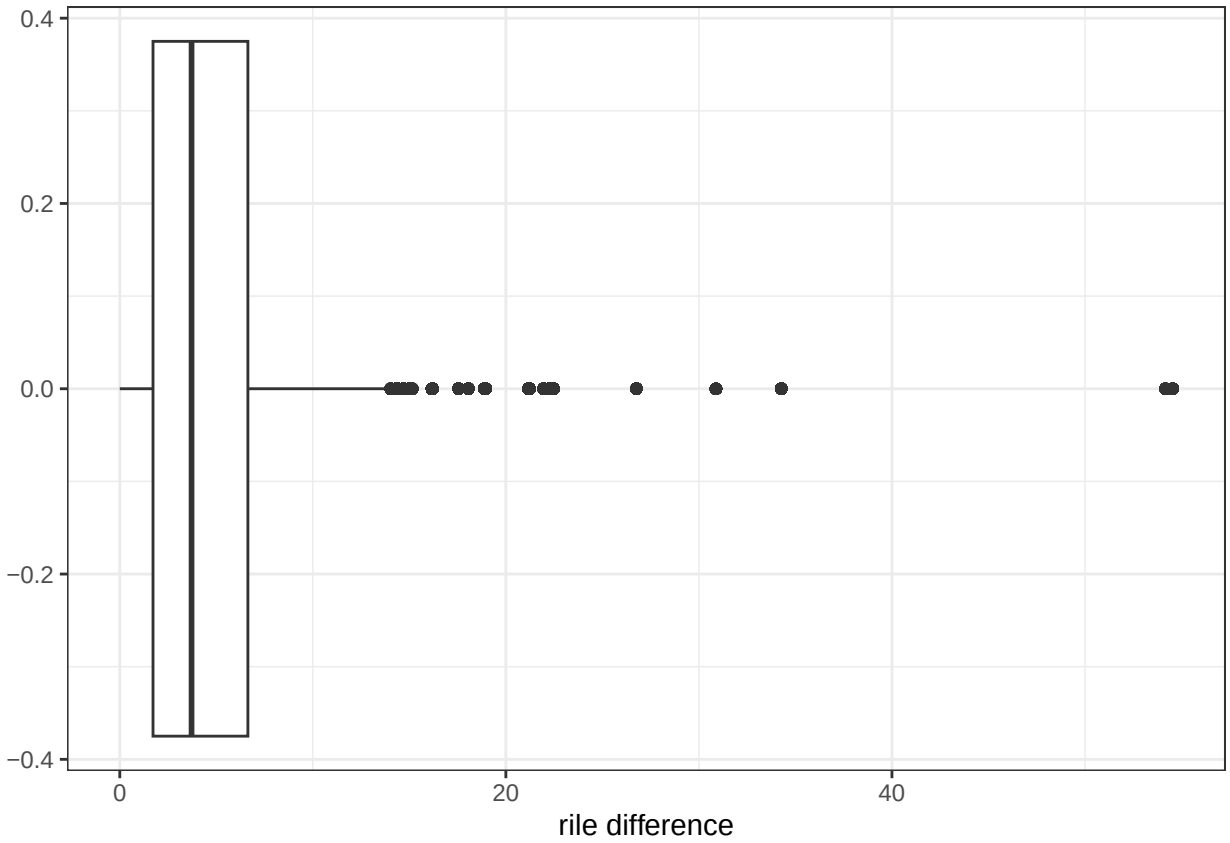

Figure 2: Absolute rile differences between true and predicted cmp codes

Model History

The model is regularly updated. All versions are available on Hugging Face (https://huggingface.co/manifesto-project). For the 2025-1-1 model, we increased the context window to 300 tokens, which increased performance. Combined with the larger training data, the performance of the 2025-1-1 model is thus slightly better than previous 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	F1 weighted	F1 macro	MCC	Cross Entropy
2025-1-1	0.638	0.802	0.876	0.67	0.46	0.64	0.43	0.64	0.42	0.62	1.23
2024-1-1	0.632	0.799	0.871	0.66	0.48	0.63	0.43	0.63	0.42	0.61	1.26
2023-1-1	0.633	0.801	0.874	0.66	0.49	0.63	0.44	0.63	0.43	0.61	1.26