The Impact of Romanian Dialects on Language Models

Florin-Silviu Dinu - 412
Andrei-Virgil Ilie - 412
Mihai Dilirici - 411

Purpose

Purpose

1. Build a dataset
2. Transform the dataset in an easy to use format
3. Compute statistics on the dataset
4. Differentiate between dialects
5. Use language models for perplexity

Dataset

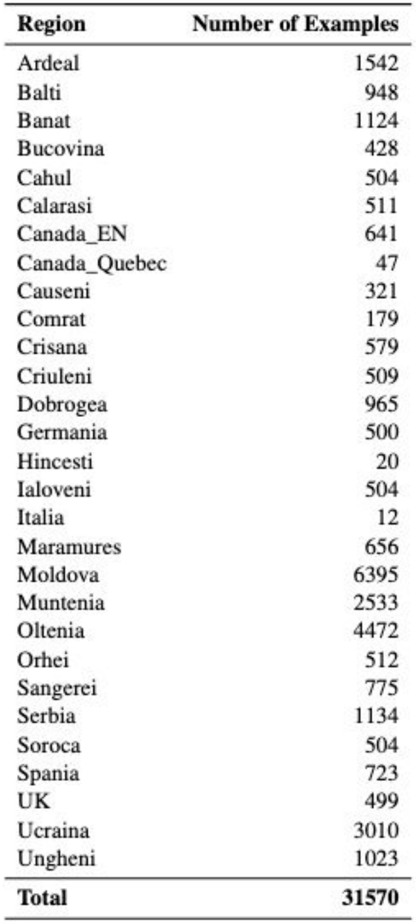
Dataset

- Romania regions: Arad, Banat, Bucovina, Crisana, Dobrogea, Maramures, Muntenia, Oltenia
- Moldova raions: Balti, Cahul, Calarasi, Causeni, Comrat, Criuleni, Hincesti, Ialoveni, Orhei, Sangerei, Soroca, Ungheni
- Historical diaspora: Serbia, Ukraine
- Diaspora: Canada (both Quebec and the rest), Germany, Italy, Spain, UK

Dataset statistics

Dataset statistics

- Dataset distribution per Region



Region	Number of Examples
Ardeal	1542
Balti	948
Banat	1124
Bucovina	428
Cahul	504
Calarasi	511
Canada_EN	641
Canada_Quebec	47
Causeni	321
Comrat	179
Crisana	579
Criuleni	509
Dobrogea	965
Germania	500
Hincesti	20
Ialoveni	504
Italia	12
Maramures	656
Moldova	6395
Muntenia	2533
Oltenia	4472
Orhei	512
Sangerei	775
Serbia	1134
Soroca	504
Spania	723
UK	499
Ucraina	3010
Ungheni	1023
Total	**31570**

Dataset statistics

- We have computed the following statistics:

  Vocabulary size(V): The total number of unique words in a given news article. A bigger vocabulary size would mean a greater lexical diversity

  Word count(N): The total number of words in the text. This helps identify text length.

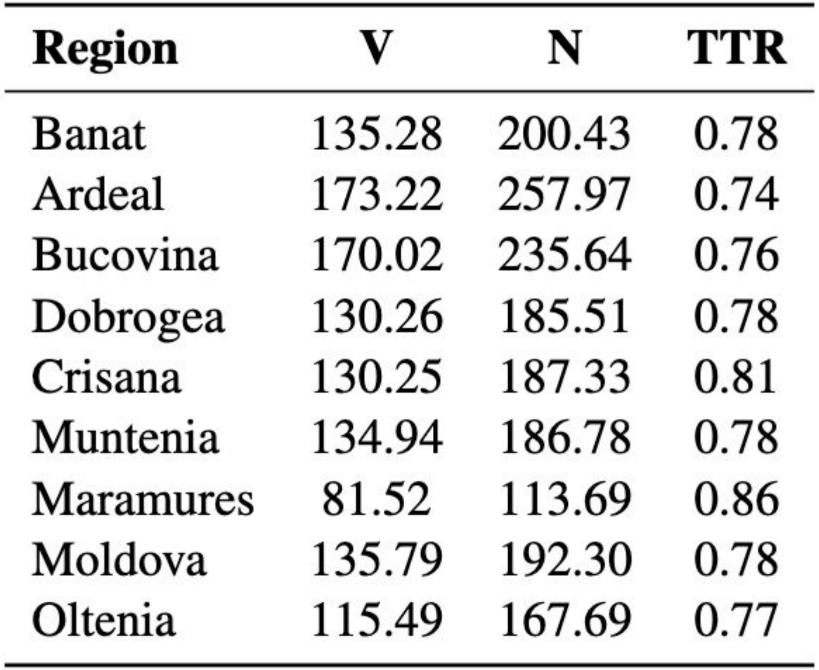
  Type-Token ratio(TTR): This metric measures lexical richness. A higher value would suggest a more varied word usage. On the other hand, lower values indicate repetition in the text.

  Average Word Length(l): The mean number of characters per word.This provides insight for morphological complexity.

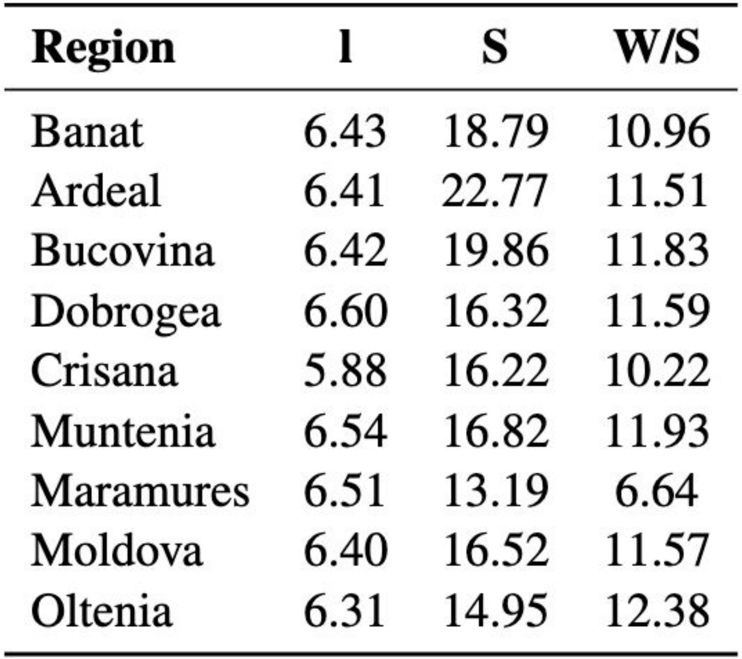
  Sentence Count(S): Total number of sentences in articles. This alows us to analyze syntactic structures and sentence segmentation patterns.

  Words per Sentence(W/S): The average number of words per sentence. It is computed by dividing the word count to the sentence count. This metric reflects syntactic complexity and sentence structuring differences among dialects.

Dataset statistics - Romania

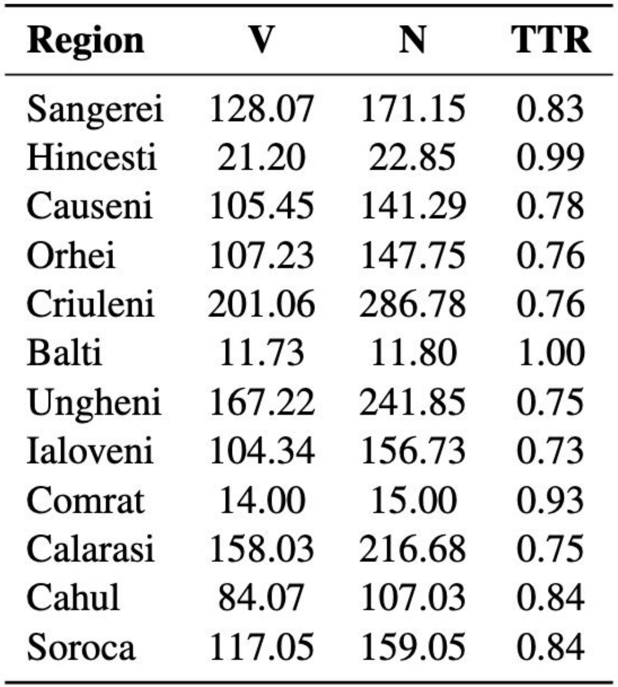


Region	V	N	TTR
Banat	135.28	200.43	0.78
Ardeal	173.22	257.97	0.74
Bucovina	170.02	235.64	0.76
Dobrogea	130.26	185.51	0.78
Crisana	130.25	187.33	0.81
Muntenia	134.94	186.78	0.78
Maramures	81.52	113.69	0.86
Moldova	135.79	192.30	0.78
Oltenia	115.49	167.69	0.77

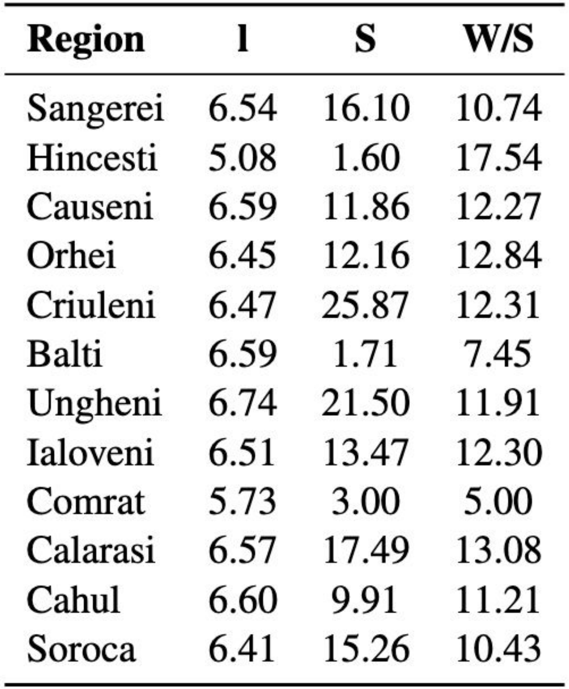


Region	l	S	W/S
Banat	6.43	18.79	10.96
Ardeal	6.41	22.77	11.51
Bucovina	6.42	19.86	11.83
Dobrogea	6.60	16.32	11.59
Crisana	5.88	16.22	10.22
Muntenia	6.54	16.82	11.93
Maramures	6.51	13.19	6.64
Moldova	6.40	16.52	11.57
Oltenia	6.31	14.95	12.38

Dataset statistics - Moldova

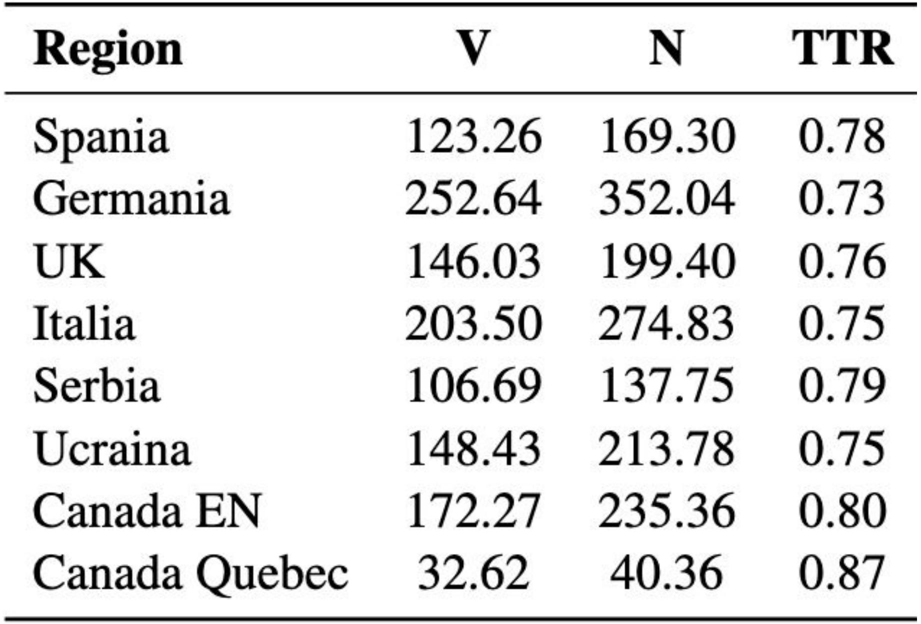


Region	V	N	TTR
Sangerei	128.07	171.15	0.83
Hincesti	21.20	22.85	0.99
Causeni	105.45	141.29	0.78
Orhei	107.23	147.75	0.76
Criuleni	201.06	286.78	0.76
Balti	11.73	11.80	1.00
Ungheni	167.22	241.85	0.75
Ialoveni	104.34	156.73	0.73
Comrat	14.00	15.00	0.93
Calarasi	158.03	216.68	0.75
Cahul	84.07	107.03	0.84
Soroca	117.05	159.05	0.84

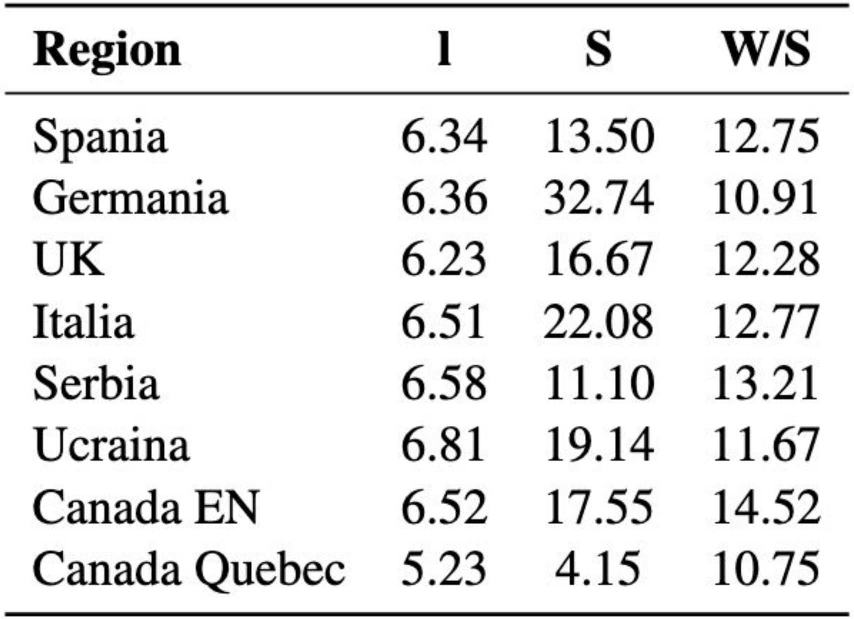


Region	l	S	W/S
Sangerei	6.54	16.10	10.74
Hincesti	5.08	1.60	17.54
Causeni	6.59	11.86	12.27
Orhei	6.45	12.16	12.84
Criuleni	6.47	25.87	12.31
Balti	6.59	1.71	7.45
Ungheni	6.74	21.50	11.91
Ialoveni	6.51	13.47	12.30
Comrat	5.73	3.00	5.00
Calarasi	6.57	17.49	13.08
Cahul	6.60	9.91	11.21
Soroca	6.41	15.26	10.43

Dataset statistics - International



Region	V	N	TTR
Spania	123.26	169.30	0.78
Germania	252.64	352.04	0.73
UK	146.03	199.40	0.76
Italia	203.50	274.83	0.75
Serbia	106.69	137.75	0.79
Ucraina	148.43	213.78	0.75
Canada EN	172.27	235.36	0.80
Canada Quebec	32.62	40.36	0.87

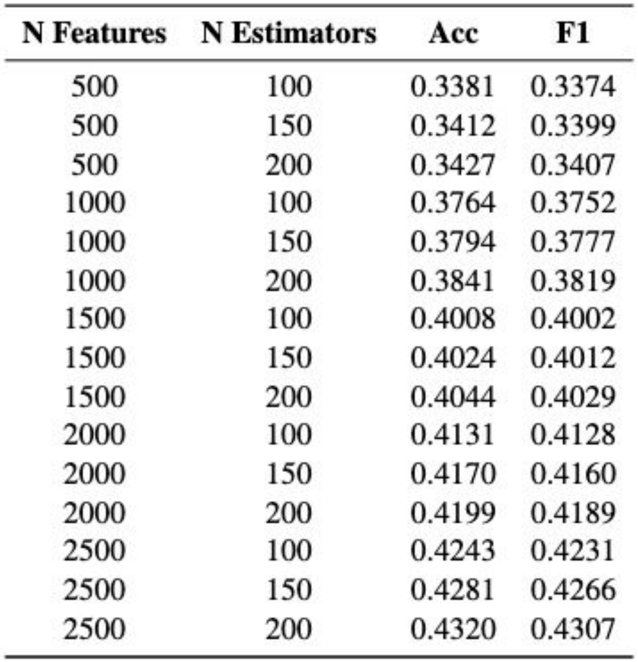


Region	l	S	W/S
Spania	6.34	13.50	12.75
Germania	6.36	32.74	10.91
UK	6.23	16.67	12.28
Italia	6.51	22.08	12.77
Serbia	6.58	11.10	13.21
Ucraina	6.81	19.14	11.67
Canada EN	6.52	17.55	14.52
Canada Quebec	5.23	4.15	10.75

Random Forest

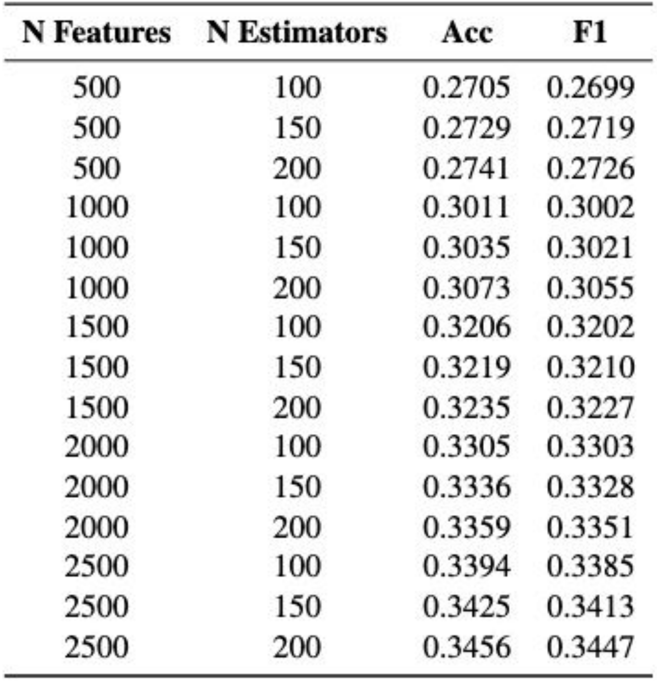
Random Forest - Texts

TF-IDF



N Features	N Estimators	Acc	F1
500	100	0.3381	0.3374
500	150	0.3412	0.3399
500	200	0.3427	0.3407
1000	100	0.3764	0.3752
1000	150	0.3794	0.3777
1000	200	0.3841	0.3819
1500	100	0.4008	0.4002
1500	150	0.4024	0.4012
1500	200	0.4044	0.4029
2000	100	0.4131	0.4128
2000	150	0.4170	0.4160
2000	200	0.4199	0.4189
2500	100	0.4243	0.4231
2500	150	0.4281	0.4266
2500	200	0.4320	0.4307

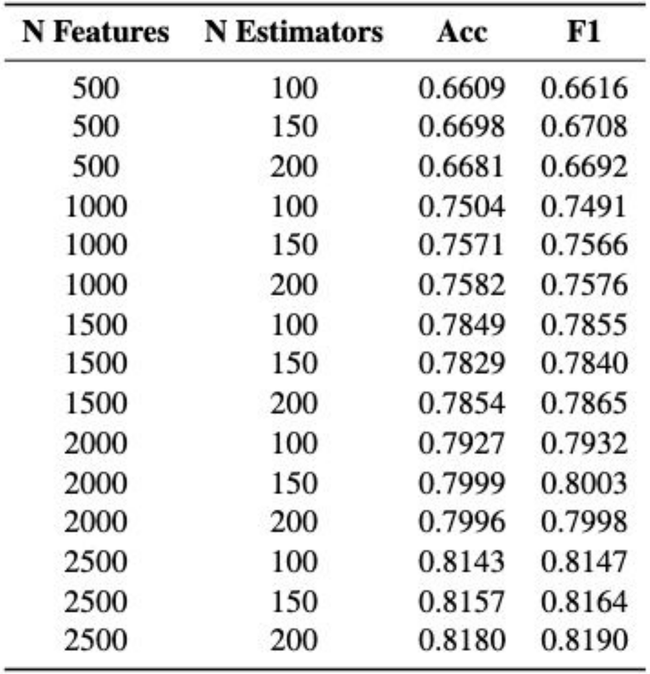
Word2Vec



N Features	N Estimators	Acc	F1
500	100	0.2705	0.2699
500	150	0.2729	0.2719
500	200	0.2741	0.2726
1000	100	0.3011	0.3002
1000	150	0.3035	0.3021
1000	200	0.3073	0.3055
1500	100	0.3206	0.3202
1500	150	0.3219	0.3210
1500	200	0.3235	0.3227
2000	100	0.3305	0.3303
2000	150	0.3336	0.3328
2000	200	0.3359	0.3351
2500	100	0.3394	0.3385
2500	150	0.3425	0.3413
2500	200	0.3456	0.3447

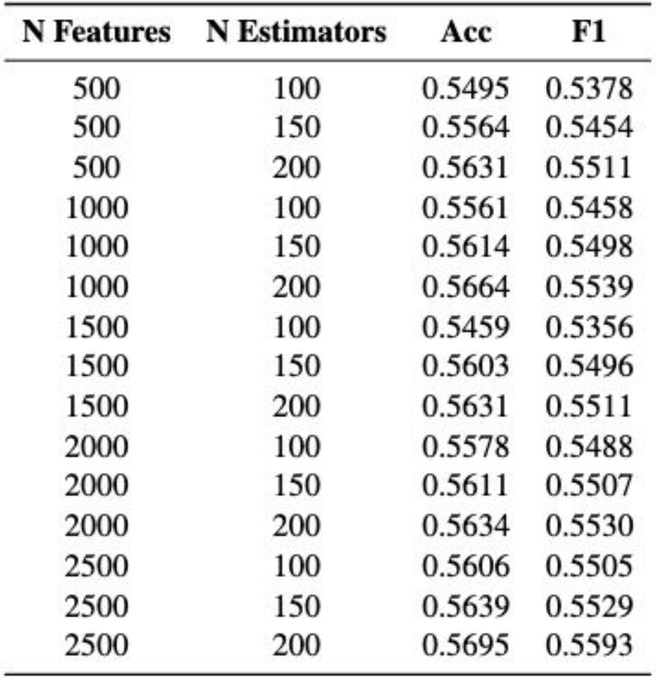
Random Forest - Texts

TF-IDF



N Features	N Estimators	Acc	F1
500	100	0.6609	0.6616
500	150	0.6698	0.6708
500	200	0.6681	0.6692
1000	100	0.7504	0.7491
1000	150	0.7571	0.7566
1000	200	0.7582	0.7576
1500	100	0.7849	0.7855
1500	150	0.7829	0.7840
1500	200	0.7854	0.7865
2000	100	0.7927	0.7932
2000	150	0.7999	0.8003
2000	200	0.7996	0.7998
2500	100	0.8143	0.8147
2500	150	0.8157	0.8164
2500	200	0.8180	0.8190

Word2Vec

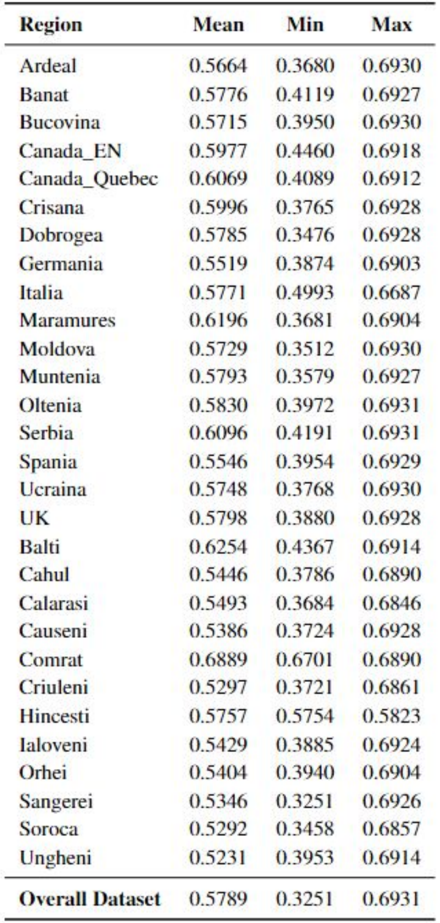


N Features	N Estimators	Acc	F1
500	100	0.5495	0.5378
500	150	0.5564	0.5454
500	200	0.5631	0.5511
1000	100	0.5561	0.5458
1000	150	0.5614	0.5498
1000	200	0.5664	0.5539
1500	100	0.5459	0.5356
1500	150	0.5603	0.5496
1500	200	0.5631	0.5511
2000	100	0.5578	0.5488
2000	150	0.5611	0.5507
2000	200	0.5634	0.5530
2500	100	0.5606	0.5505
2500	150	0.5639	0.5529
2500	200	0.5695	0.5593

RoBERT

Results

stdev: 0.035
overall mean perplexity is 0.5789

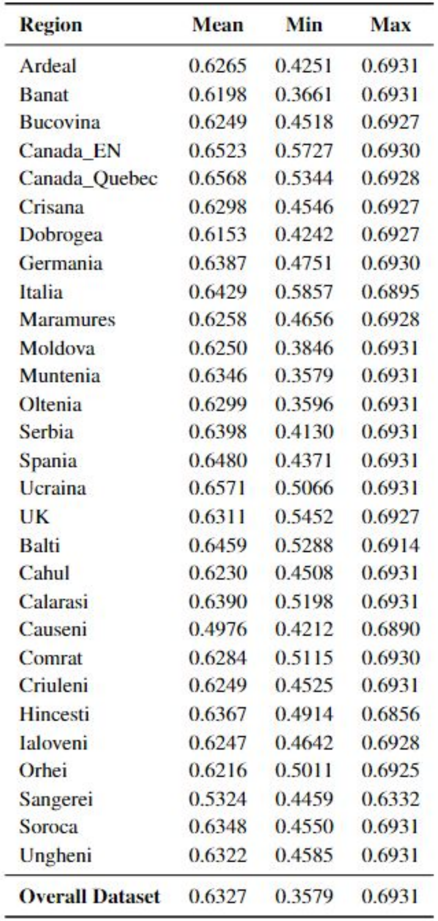


Region	Mean	Min	Max
Ardeal	0.5664	0.3680	0.6930
Banat	0.5776	0.4119	0.6927
Bucovina	0.5715	0.3950	0.6930
Canada_EN	0.5977	0.4460	0.6918
Canada_Quebec	0.6069	0.4089	0.6912
Crisana	0.5996	0.3765	0.6928
Dobrogea	0.5785	0.3476	0.6928
Germania	0.5519	0.3874	0.6903
Italia	0.5771	0.4993	0.6687
Maramures	0.6196	0.3681	0.6904
Moldova	0.5729	0.3512	0.6930
Muntenia	0.5793	0.3579	0.6927
Oltenia	0.5830	0.3972	0.6931
Serbia	0.6096	0.4191	0.6931
Spania	0.5546	0.3954	0.6929
Ucraina	0.5748	0.3768	0.6930
UK	0.5798	0.3880	0.6928
Balti	0.6254	0.4367	0.6914
Cahul	0.5446	0.3786	0.6890
Calarasi	0.5493	0.3684	0.6846
Causeni	0.5386	0.3724	0.6928
Comrat	0.6889	0.6701	0.6890
Criuleni	0.5297	0.3721	0.6861
Hincesti	0.5757	0.5754	0.5823
Ialoveni	0.5429	0.3885	0.6924
Orhei	0.5404	0.3940	0.6904
Sangerei	0.5346	0.3251	0.6926
Soroca	0.5292	0.3458	0.6857
Ungheni	0.5231	0.3953	0.6914
Overall Dataset	0.5789	0.3251	0.6931

Table 16: Content Perplexity Statistics Per Region - RoBert

Results

stdev: 0.032
overall mean perplexity is 0.6327



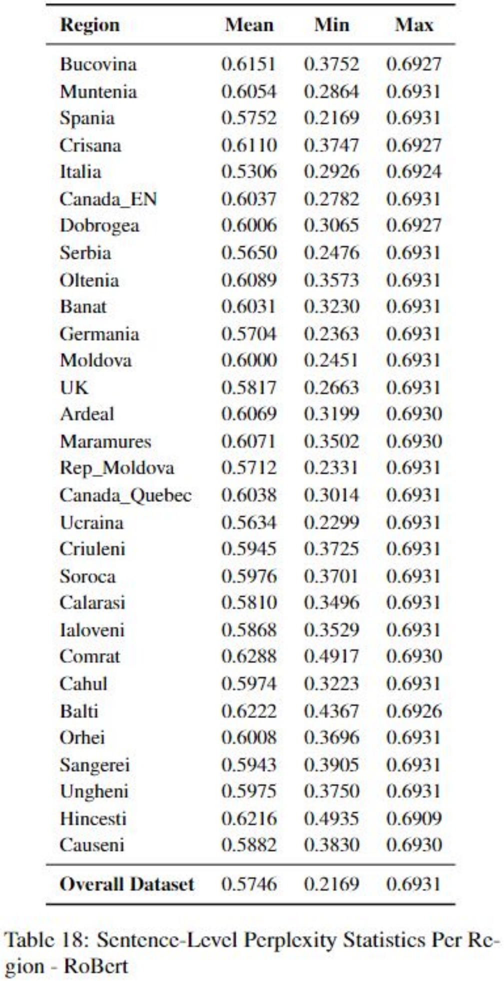
Region	Mean	Min	Max
Ardeal	0.6265	0.4251	0.6931
Banat	0.6198	0.3661	0.6931
Bucovina	0.6249	0.4518	0.6927
Canada_EN	0.6523	0.5727	0.6930
Canada_Quebec	0.6568	0.5344	0.6928
Crisana	0.6298	0.4546	0.6927
Dobrogea	0.6153	0.4242	0.6927
Germania	0.6387	0.4751	0.6930
Italia	0.6429	0.5857	0.6895
Maramures	0.6258	0.4656	0.6928
Moldova	0.6250	0.3846	0.6931
Muntenia	0.6346	0.3579	0.6931
Oltenia	0.6299	0.3596	0.6931
Serbia	0.6398	0.4130	0.6931
Spania	0.6480	0.4371	0.6931
Ucraina	0.6571	0.5066	0.6931
UK	0.6311	0.5452	0.6927
Balti	0.6459	0.5288	0.6914
Cahul	0.6230	0.4508	0.6931
Calarasi	0.6390	0.5198	0.6931
Causeni	0.4976	0.4212	0.6890
Comrat	0.6284	0.5115	0.6930
Criuleni	0.6249	0.4525	0.6931
Hincesti	0.6367	0.4914	0.6856
Ialoveni	0.6247	0.4642	0.6928
Orhei	0.6216	0.5011	0.6925
Sangerei	0.5324	0.4459	0.6332
Soroca	0.6348	0.4550	0.6931
Ungheni	0.6322	0.4585	0.6931
Overall Dataset	0.6327	0.3579	0.6931

Table 17: Titles Perplexity Statistics Per Region - RoBert

Results

stdev: 0.020

overall mean perplexity is 0.5746



Region	Mean	Min	Max
Bucovina	0.6151	0.3752	0.6927
Muntenia	0.6054	0.2864	0.6931
Spania	0.5752	0.2169	0.6931
Crisana	0.6110	0.3747	0.6927
Italia	0.5306	0.2926	0.6924
Canada_EN	0.6037	0.2782	0.6931
Dobrogea	0.6006	0.3065	0.6927
Serbia	0.5650	0.2476	0.6931
Oltenia	0.6089	0.3573	0.6931
Banat	0.6031	0.3230	0.6931
Germania	0.5704	0.2363	0.6931
Moldova	0.6000	0.2451	0.6931
UK	0.5817	0.2663	0.6931
Ardeal	0.6069	0.3199	0.6930
Maramures	0.6071	0.3502	0.6930
Rep_Moldova	0.5712	0.2331	0.6931
Canada_Quebec	0.6038	0.3014	0.6931
Ucraina	0.5634	0.2299	0.6931
Criuleni	0.5945	0.3725	0.6931
Soroca	0.5976	0.3701	0.6931
Calarasi	0.5810	0.3496	0.6931
Ialoveni	0.5868	0.3529	0.6931
Comrat	0.6288	0.4917	0.6930
Cahul	0.5974	0.3223	0.6931
Balti	0.6222	0.4367	0.6926
Orhei	0.6008	0.3696	0.6931
Sangerei	0.5943	0.3905	0.6931
Ungheni	0.5975	0.3750	0.6931
Hincesti	0.6216	0.4935	0.6909
Causeni	0.5882	0.3830	0.6930
Overall Dataset	0.5746	0.2169	0.6931

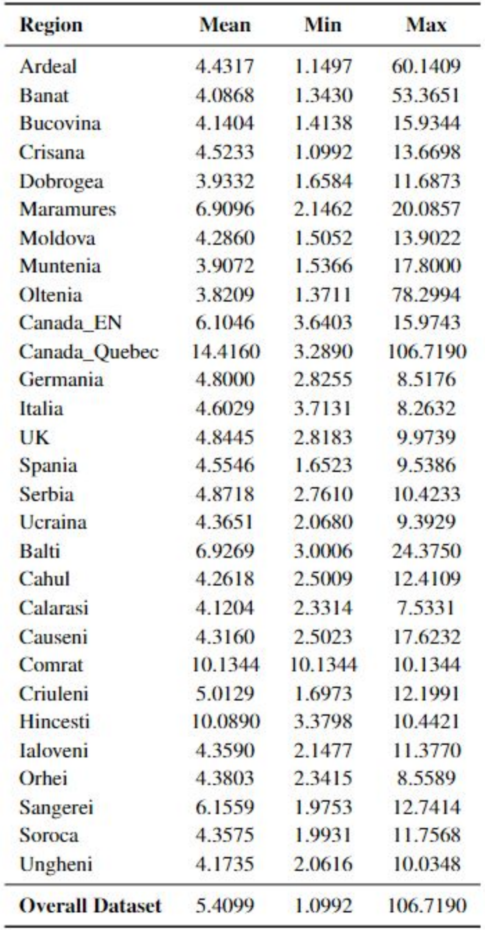
Table 18: Sentence-Level Perplexity Statistics Per Region - RoBert

RoLlama 2

Results

deviation: 2.2864

overall mean perplexity is 5.4099

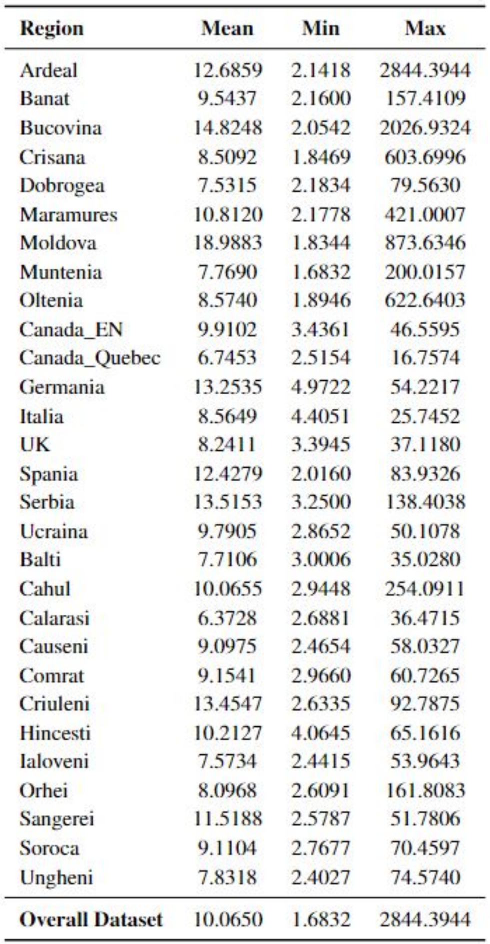


Region	Mean	Min	Max
Ardeal	4.4317	1.1497	60.1409
Banat	4.0868	1.3430	53.3651
Bucovina	4.1404	1.4138	15.9344
Crisana	4.5233	1.0992	13.6698
Dobrogea	3.9332	1.6584	11.6873
Maramures	6.9096	2.1462	20.0857
Moldova	4.2860	1.5052	13.9022
Muntenia	3.9072	1.5366	17.8000
Oltenia	3.8209	1.3711	78.2994
Canada_EN	6.1046	3.6403	15.9743
Canada_Quebec	14.4160	3.2890	106.7190
Germania	4.8000	2.8255	8.5176
Italia	4.6029	3.7131	8.2632
UK	4.8445	2.8183	9.9739
Spania	4.5546	1.6523	9.5386
Serbia	4.8718	2.7610	10.4233
Ucraina	4.3651	2.0680	9.3929
Balti	6.9269	3.0006	24.3750
Cahul	4.2618	2.5009	12.4109
Calarasi	4.1204	2.3314	7.5331
Causeni	4.3160	2.5023	17.6232
Comrat	10.1344	10.1344	10.1344
Criuleni	5.0129	1.6973	12.1991
Hincesti	10.0890	3.3798	10.4421
Ialoveni	4.3590	2.1477	11.3770
Orhei	4.3803	2.3415	8.5589
Sangerei	6.1559	1.9753	12.7414
Soroca	4.3575	1.9931	11.7568
Ungheni	4.1735	2.0616	10.0348
Overall Dataset	5.4099	1.0992	106.7190

Table 12: Content Perplexity Statistics Per Region - RoLlama

Results

deviation: 2.715
overall mean perplexity is 10.065



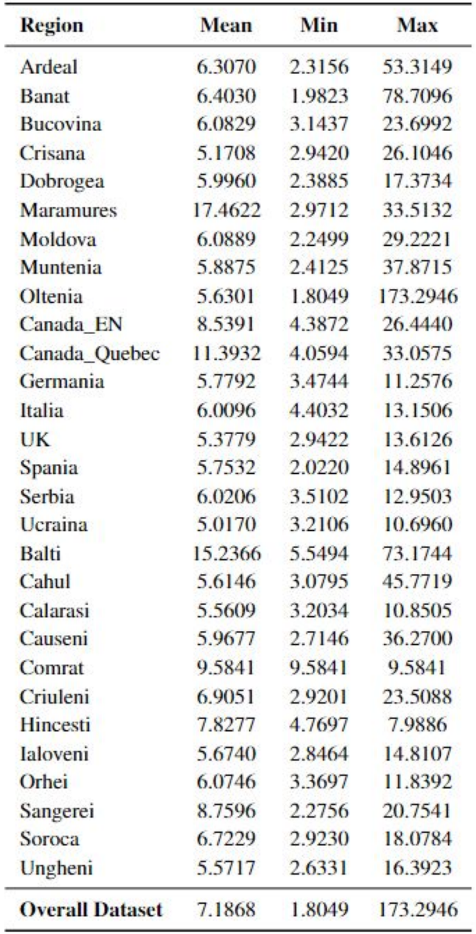
Region	Mean	Min	Max
Ardeal	12.6859	2.1418	2844.3944
Banat	9.5437	2.1600	157.4109
Bucovina	14.8248	2.0542	2026.9324
Crisana	8.5092	1.8469	603.6996
Dobrogea	7.5315	2.1834	79.5630
Maramures	10.8120	2.1778	421.0007
Moldova	18.9883	1.8344	873.6346
Muntenia	7.7690	1.6832	200.0157
Oltenia	8.5740	1.8946	622.6403
Canada_EN	9.9102	3.4361	46.5595
Canada_Quebec	6.7453	2.5154	16.7574
Germania	13.2535	4.9722	54.2217
Italia	8.5649	4.4051	25.7452
UK	8.2411	3.3945	37.1180
Spania	12.4279	2.0160	83.9326
Serbia	13.5153	3.2500	138.4038
Ucraina	9.7905	2.8652	50.1078
Balti	7.7106	3.0006	35.0280
Cahul	10.0655	2.9448	254.0911
Calarasi	6.3728	2.6881	36.4715
Causeni	9.0975	2.4654	58.0327
Comrat	9.1541	2.9660	60.7265
Criuleni	13.4547	2.6335	92.7875
Hincesti	10.2127	4.0645	65.1616
Ialoveni	7.5734	2.4415	53.9643
Orhei	8.0968	2.6091	161.8083
Sangerei	11.5188	2.5787	51.7806
Soroca	9.1104	2.7677	70.4597
Ungheni	7.8318	2.4027	74.5740
Overall Dataset	10.0650	1.6832	2844.3944

Table 13: Title Perplexity Statistics Per Region - RoLlama

RoQLlama

Results

deviation: 2.8268
overall mean perplexity is 7.1868

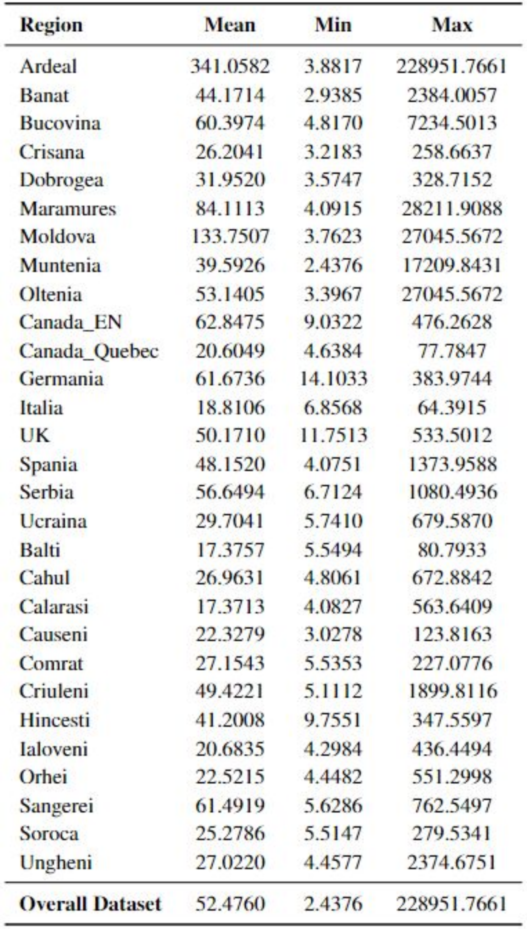


Region	Mean	Min	Max
Ardeal	6.3070	2.3156	53.3149
Banat	6.4030	1.9823	78.7096
Bucovina	6.0829	3.1437	23.6992
Crisana	5.1708	2.9420	26.1046
Dobrogea	5.9960	2.3885	17.3734
Maramures	17.4622	2.9712	33.5132
Moldova	6.0889	2.2499	29.2221
Muntenia	5.8875	2.4125	37.8715
Oltenia	5.6301	1.8049	173.2946
Canada_EN	8.5391	4.3872	26.4440
Canada_Quebec	11.3932	4.0594	33.0575
Germania	5.7792	3.4744	11.2576
Italia	6.0096	4.4032	13.1506
UK	5.3779	2.9422	13.6126
Spania	5.7532	2.0220	14.8961
Serbia	6.0206	3.5102	12.9503
Ucraina	5.0170	3.2106	10.6960
Balti	15.2366	5.5494	73.1744
Cahul	5.6146	3.0795	45.7719
Calarasi	5.5609	3.2034	10.8505
Causeni	5.9677	2.7146	36.2700
Comrat	9.5841	9.5841	9.5841
Criuleni	6.9051	2.9201	23.5088
Hincesti	7.8277	4.7697	7.9886
Ialoveni	5.6740	2.8464	14.8107
Orhei	6.0746	3.3697	11.8392
Sangerei	8.7596	2.2756	20.7541
Soroca	6.7229	2.9230	18.0784
Ungheni	5.5717	2.6331	16.3923
Overall Dataset	7.1868	1.8049	173.2946

Table 14: Content Perplexity Statistics Per Region - RoQLlama

Results

deviation: 58.7014
overall mean perplexity is 52.476



Region	Mean	Min	Max
Ardeal	341.0582	3.8817	228951.7661
Banat	44.1714	2.9385	2384.0057
Bucovina	60.3974	4.8170	7234.5013
Crisana	26.2041	3.2183	258.6637
Dobrogea	31.9520	3.5747	328.7152
Maramures	84.1113	4.0915	28211.9088
Moldova	133.7507	3.7623	27045.5672
Muntenia	39.5926	2.4376	17209.8431
Oltenia	53.1405	3.3967	27045.5672
Canada_EN	62.8475	9.0322	476.2628
Canada_Quebec	20.6049	4.6384	77.7847
Germania	61.6736	14.1033	383.9744
Italia	18.8106	6.8568	64.3915
UK	50.1710	11.7513	533.5012
Spania	48.1520	4.0751	1373.9588
Serbia	56.6494	6.7124	1080.4936
Ucraina	29.7041	5.7410	679.5870
Balti	17.3757	5.5494	80.7933
Cahul	26.9631	4.8061	672.8842
Calarasi	17.3713	4.0827	563.6409
Causeni	22.3279	3.0278	123.8163
Comrat	27.1543	5.5353	227.0776
Criuleni	49.4221	5.1112	1899.8116
Hincesti	41.2008	9.7551	347.5597
Ialoveni	20.6835	4.2984	436.4494
Orhei	22.5215	4.4482	551.2998
Sangerei	61.4919	5.6286	762.5497
Soroca	25.2786	5.5147	279.5341
Ungheni	27.0220	4.4577	2374.6751
Overall Dataset	52.4760	2.4376	228951.7661

Table 15: Title Perplexity Statistics Per Region - RoQLlama

Improvements

- Clean the data better
- Do NER on the data
- Train a model to see if the dataset is useful
- Check the reasons for variations in perplexity

Conclusions

- The dataset has a relevant size
- Dialect classification can be employed successfully
- Variations in perplexity appear but they are not enough to do dialect classification

Thanks!

Florin-Silviu Dinu
Andrei-Virgil Ilie
Mihai Dilirici