The Impact of Romanian Dialects on Language Models

Florin-Silviu Dinu
florin-silviu.dinu@s.unibuc.ro

Andrei-Virgil Ilie
andrei-virgil.ilie@s.unibuc.ro

Mihai Dilirici
mihai.dilirici@s.unibuc.ro

Abstract

Large Language Models (LLMs) have achieved remarkable proficiency in understanding and generating text across many languages. However, speficific language differences such as various dialects might pose challenges for them. This paper analyzes the impact of language varieties on LLMs by comparing romanian dialects from different parts of Romania, Moldova and from variations in other global Romanian dialects. We will assess the models' ability to distinguish, adapt, and generate responses across these linguistic varieties. We will also present some text statistics on the different data we have used and try to see if other Machine Learning approaches can detect the dialect differences.

1 Introduction

Dialectical differences arise naturally in a language and usually indicate a persons' geographical area of living and sometimes their background or occupation but whether a certain difference constitutes a dialect is often subject to interpretation (Encyclopaedia Britannica). Regarding the dialects present in the Romanian language, different studies often contradict each other, starting from considering Daco-Romanian, Aromanian, Megleno-Romanian and Istro-Romanian as dialects of a single language to dividing Daco-Romanian in 2 to even 20 dialects, starting from Wallachian and Moldavian but then distinguishing between different regions as well (Wikipedia contributors, 2025).

An important part of Romanian language dialectical variations that is often neglected is the large diaspora of Romanian speakers. The number of people living in each country is difficult to estimate, having Romanian citizens, Moldovan citizens and naturalized citizens of the respective country but an estimate of the numbers given national censuses pertaining to the countries analyzed in this study is (Wikipedia contributors, c):

1. **Italy** - 1,081,836 Romanians and 122,667 Moldovans
2. **Spain** - 1,079,726 Romanians and 17,868 Moldovans
3. **Germany** - 909,795 Romanians
4. **United Kingdom (England & Wales)** - 539,000 Romanians and 18,000 Moldovans
5. **Canada (by mother tongue)** - 93,160 Romanians

In addition to the recent diaspora, there is a historical diaspora as well, in this study being analyzed Serbia with 23,044 Romanians, out of which 327 are Aromanians and Ukraine with 150,989 Romanians and 258,619 Moldovans (Wikipedia contributors, c).

This paper aims to analyze using natural language processing the differences between the written Romanian language in newspapers from different regions with the aim to differentiate between the dialects and ascertain the impact they have on LLMs.

2 Related Work

2.1 Dialect identification

There are multiple reasons for dialect identification, some of the most important factors being enumerated in a survey by Joshi et al. (Joshi et al., 2025):

1. Performance of NLP models per-capita GDP - especially when it comes to machine translation and automatic speech recognition
2. Healthcare monitoring - existing a disparity between people based on dialect
3. Racial biases in hate speech detection - in the English language, hate-speech classifiers may have a bias towards African-American English

4. Prejudice in the prediction of employability and criminality - employability may be affected when language models are used and confronted with dialects

Even though multilingual LLMs are trained on a large corpus, when it comes to comparing the US English dialect with Indian English dialect, Maldivian English dialect and Turkish English dialect, the LLMs usually perform better on the US English dialect (Srirag et al., 2024). This study also shows that fine-tuning for a specific dialect, Maldivian English in this case greatly improves performance (Srirag et al., 2024).

A toxicity identification study performed on Arabic, Bengali, Chinese, Common Turkic, English, Finnish, High German, Kurdish, Norwegian and Sotho-Tswana shows that in high-resource languages the models employed for evaluating toxicity perform better than in the low-resource languages (Faisal et al., 2024). This can represent a significant factor of inconsistent behavior when a specific LLM based toxicity detection system is used in a region (Faisal et al., 2024).

For the Romanian language, a dataset for distinguishing between the Romanian/Wallachian and Moldovan dialect has been created by Butnaru & Ionescu (Butnaru and Ionescu, 2019). As with the current study, the data was gathered from newspapers, the criteria being the popularity of the newspapers with 15,403 samples from Moldovan newspapers and 18,161 samples from Romanian newspapers (Butnaru and Ionescu, 2019). After named entity recognition and replacement, binary classifiers were used for dialects, the best performer being Kernel Ridge Regression with an accuracy of 94.13 on the test dataset proving that the dialects can be separated (Butnaru and Ionescu, 2019).

2.2 RoLlama

Masala et al. have trained Llama2 8b using CulturaX dataset and translated datasets then evaluated the project on academic benchmarks, MT-bench, on a series of Romanian language downstream tasks, and on RoCulturaBench (Masala et al., 2024).

What the authors have noted is that there is a need for translated datasets since the Romanian language does not yet have high quality datasets (Masala et al., 2024) which is a problem that the current dataset made for this paper tries to alleviate.

2.3 RoQLlama

Dima et al. have trained Llama 2 7b using the RoWiki dataset, RoTex, and a new dataset the authors introduced, RoMedQA which is a dataset of 4,127 single-choice medical questions (Dima et al., 2024). The tasks that the model was evaluated on include Moroco for dialect classification (Dima et al., 2024; Butnaru and Ionescu, 2019).

For the dialect classification, the model reached 46.48% accuracy but the NFI score shows that the model does not follow instructions in 11.71% of the cases (Dima et al., 2024).

2.4 RoBERT

Dumitrescu et al. have trained RoBERT, a Romanian-specific BERT model, using the Romanian Wikipedia, OSCAR, and news corpora (Dumitrescu et al., 2021). The model was evaluated on Moroco for dialect classification (Dumitrescu et al., 2021; Butnaru and Ionescu, 2019), SiMoNERo for named entity recognition, and RO-STSB for semantic similarity (Dumitrescu et al., 2021).

In sentiment analysis, RoBERT-base achieved a macro F1 score of 71.61%, surpassing mBERT's 69.57% (Dumitrescu et al., 2021). For dialect identification, it obtained a macro F1 score of 93.40%, outperforming mBERT's 91.66% (Dumitrescu et al., 2021). Unlike RoLlama and RoQLlama, RoBERT is bidirectional, making it more suitable for masked language modeling.

3 Method

3.1 Dataset

The dataset was obtained by scraping Romanian language newspapers and then extracting the title and contents using Goose3 (Goose3 Developers).

For Romanian newspapers the county specific lists published on e-Ziare (e-ziare.ro) were used and then filtered by availability and county only newspapers, excluding all websites that were either defunct or country or region specific. A number of articles were extracted from each newspaper, usually targeting local news but in cases of no such category, a mix of other categories were used excluding politics.

For Moldovan newspapers e-Ziare (e-ziare.ro) was also used but given the fact that they weren't usually local and a large portion were defunct, alternative methods have been employed. One such method was to use the newspapers of the members of The Association of Independent press, which is

a Moldovan press association (API.md). Another way of identifying newspapers that issued a low result was to go through the Wikipedia page that listed newspapers from the Republic of Moldova (Wikipedia contributors, a). This method resulted in a lot of defunct newspapers or those that had a paper only version. Alternatively, Google search engine (Google) was used with "Raion_Name ziar" or similar prompt that issued some results as well. In total 11 raions were identified as having Romanian language newspapers and 1 municiplaity from the Gagauz Autonomous Region which is a significantly lower than the 32 raions and 2 autonomous regions in the country. This may happen for a number of reasons, one hypothesis being the two national language system, papers probably existing but in the Russian language and another being the fact that in 2003 the country changed its administrative division from counties to raions (Wikipedia contributors, b).

For the Ukrainian newspapers, the Romanian General Consulate in Cernauti was consulted as it publishes a list of mostly functioning Romanian language websites in the region (Ministry of Foreign Affairs of Romania).

For the Serbian newspapers, Google was queried and issued 2 results (Google).

For the rest of the diaspora, e-Ziare (e-ziare.ro) and Google (Google) were employed in finding newspapers.

Finally the dataset was grouped on regions. For Random Forest preprocessing, the diacritics were modified to use the Romanian standard using a method developed by Tindeche & Dinu which was further improved outside the repository by the authors (Tindeche and Dinu, 2025)

In total we have gathered 31570 usable examples. The distribution by regions can be seen in Table 1. Some regions such as "Italia" and "Hincesti" have really small number of examples. This is due to the fact that our cleaning script was not able to get the text from the HTML pages for much of the data.

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**

Table 1: Number of Examples by Region

3.2 Preprocessing

One important step before using the data was preprocessing. When collecting the data we have saved raw *.html pages. Those include much meaningless data such as HTML elements, advertisements, navigation menus or embedded media. For this task of cleaning the data we have used the Goose3 library (Goose3 Developers) to parse and

extract article titles, and cleaned text. This was important for isolating the core textual content from the irrelevant noise. This helped us to achieve better results in our tasks. The cleaned data was then structured in a JSON format with the following keys:

- Title - the title of the article
- Content - the actual content of the article
- Medatadata - the original html file name for traceability

The pipeline processed articles from multiple directories. Each directory represented a region of interest for our project. After the preprocessing, the script replicated this folder structure, while saving the cleaned output. The final step was creating the actual dataset by combining all of the json files for a region in a single json file which contained all of the data for that region.

The preprocessing was the one developed by Tindeche & Dinu (Tindeche and Dinu, 2025) with additional steps produced by the authors outside of the repository. The texts was lemmatized. Lemmatization reduces words to their root form called lemma. This is important because it links similar meaning words as one word.

3.3 Method

The dataset produced in this paper will be used to compare RoBERT (Dumitrescu et al., 2021) with RoLlama 2 (Masala et al., 2024) and RoQLlama (Dima et al., 2024) and do a dialect classification.

3.3.1 Random Forest

Random Forest (Breiman, 2001) is an ensemble supervised learning method that constructs multiple decision trees during training and combines their outputs to improve accuracy and robustness. A key aspect of this method is bootstrapping, where random subsets of the training data are selected to train each tree, helping to reduce correlation between them. For classification tasks, such as ours, the final prediction is determined by majority voting among the trees. Due to its ability to capture complex relationships and its resilience to noisy data, Random Forest has been widely applied in various domains, including text classification, financial modeling, and medical diagnosis.

3.3.2 Language Models

The following language models that have been described in 2 have been used: RoLlama (Masala et al., 2024), RoQLlama (Dima et al., 2024) and RoBert (Dumitrescu et al., 2021).

Each article was given to the model and then the negative log likelihood was outputted. Perplexity was computed for each (Hugging Face) as can be seen in Equation 1 where the power of e is the negative log likelihood. The tokenizer was instructed to truncate at 512 tokens.

$$PPL(X) = e^{\left\{-\frac{1}{t}\sum_i^t \log p_\theta(x_i|x_{<i})\right\}} \quad (1)$$

Equation 1: Perplexity equation (Hugging Face)

4 Statistics

We have conducted a quantitative analysis of the content of the collected news articles to evaluate the impact of language varieties. Understanding the data was important. We assessed how linguistic variations manifest in data and how they might influence the performance of LLMs.

The key linguistic metrics we have used in our quantitative analysis were:

- 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.

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

Table 2: Vocabulary Statistics for Romanian Regions

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

Table 3: Sentence and Word Length Statistics for Romanian Regions

These linguistic metrics can be seen by region in the following tables:

- Romania: Vocabulary statistics in Table 2 and sentence level statistics in Table 3
- Moldova: Vocabulary statistics in Table 4 and sentence level statistics in Table 5
- Other Countries: Vocabulary statistics in Table 6 and sentence level statistics in Table 7

4.1 Random Forest

We have conducted 2 experiments for the random forest method. One of them was using the whole text of news articles, and the other using 1 sentence as 1 training example. In our trials we have only used stopwords from the texts/sentences. This is important, because stop words exhibit distinct usage patterns across dialects, even when the overall content remains similar. Stop words are function words that reflect syntactic and grammatical structures unique to each dialect. Differences in frequency and choise make stop words indicators of

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

Table 4: Vocabulary Statistics for Moldova Regions

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

Table 5: Sentence and Word Length Statistics for Moldova Regions

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

Table 6: Vocabulary Statistics for Other Countries

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

Table 7: Sentence and Word Length Statistics for Other Countries

dialectal variation. Another reason of why we have chose to train the classifier only on stop words is because the actual content might give away cues to the classifier. For example, news articles from the UK may contain words such as "UK", "London". This could influence the model into thinking that articles which contain those words might be from UK dialect.

4.1.1 Word Embeddings

The first step after the preprocessing is to create the numerical representations of words, namely word embeddings. For this, we have tried 2 different approaches: TF-IDF and Word2Vec with different hyperparameters. We have also conducted a hyperparameter grid search for finding the best ones in our case.

4.2 News Texts

Because the dataset is imabalanced, we have decided to use a maximum of 1000 articles per region, and regions with a minimum of 200 articles.

4.2.1 TF-IDF

The model identifies different dialects really well, given the fact that we have achieved an F1 score of 0.8190 in the iteration with 2500 features and 200 estimators. This suggests that there might actually be big differences between romanian dialects. The results can be seen in Table 8.

4.2.2 Word2Vec

Word2Vec has gotten worse results than TF-IDF. We have also conducted a parameter grid search, but we were not able to achieve an F1 score higher than 0.5593, again for 2500 features and 200 estimators. The results can be checked in Table 9.

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

Table 8: Random Forest TF-IDF on Articles

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

Table 9: Random Forest Word2Vec on Articles

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

Table 10: Random Forest TF-IDF on Sentences

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

Table 11: Random Forest Word2Vec on Sentences

4.3 News Sentences

For the news section we have used a maximum of 10000 sentences per region.

4.3.1 TF-IDF

This iteration was worse than on the articles. This may be because sentences are too short for random forest to understand the nuances of different dialects. The results are found in Table 10

4.3.2 Word2Vec

This were the wors results as we can see in Table 11

4.4 RoLlama

As per Table 12 the mean perplexity with a deviation of 2.2864 but fluctuations can be seen especially in the Quebec region of Canada where the articles represent an anomaly since a lot of them are written in whether English or French without proper labeling. The overall mean perplexity is 5.4099 which means that in general the model is well trained and no clear difference being made between values from Romanian geographical regions, Moldova's regions or any other diaspora.

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

Titles on the other hand being shorter and having to summarize more information and raise interest in the article have a mean perplexity of 10.065 with a deviation of 2.7154 as can be seen in Table 13

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

One important aspect is that Moldova, followed by Bucovina and then Criuleni have on average the titles with most perplexity, while Calarasi and Canada's Quebec region have the least perplexity. This can be a local indicator of strategic newspaper communication style.

4.5 RoQLlama

Regarding content perplexity, the mean is 7.1868 with a deviation of 2.8268 as seen in Table 14. The top is led by Maramures with a mean of 17.4622 but what's interesting comparing to the previous case is that the Quebec region of Canada has a lower perplexity meaning that the model is not as surprised to see texts in foreign languages than in the other cases most likely due to its training. The bottom mean perplexity is in Ukraine followed by Crisana meaning that the dialects don't seem to impact the model as much.

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

The maximum mean perplexity for titles is seen in Ardeal which is due to an anomaly. The mean is 52.476 for all regions with a deviation of 58.7014 and it does not seem to be the impact of an anomaly as seen in Table 15. The highest mean perplexity is in Moldova which corresponds with the previous model. It looks like this model issues a higher and more unstable perplexity when given the titles.

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

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

4.6 RoBERT

As shown in Table 16, RoBERT achieves a mean perplexity of 0.5789, demonstrating stable performance across regions. Unlike RoLlama, which exhibited larger fluctuations, RoBERT maintains consistent values, indicating a well-trained model with reliable generalization across Romanian geographical regions and the diaspora.

The title perplexity, seen in Table 17, suggests that RoBERT maintains a stable representation of shorter, more condensed text. The mean title perplexity of 0.6327 is consistent with content perplexity trends, indicating the model does not struggle significantly with summarization tasks.

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

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

The sentence-level perplexity in Table 18 reveals that most perplexity scores remain close to the overall dataset mean, reinforcing the model's robustness in handling different text structures. While some regions, such as Comrat and Balti, show slightly higher perplexity, there is no drastic deviation compared to RoLlama.

Overall, RoBERT presents lower perplexity values than RoLlama and RoQLlama, meaning it produces more predictable and coherent Romanian text. However, its ability to generalize beyond training data remains an open question that requires further benchmarking.

5 Conclusion

The dataset is large enough to be considered relevant but different cleaning techniques must be employed. Also, language identification has to be employed for filtering incorrectly labeled articles from websites. Even so, those incorrectly labeled articles can be used as adversarial examples as shown in this paper.

Dialect classification can be done successfully proving that there are clear differences between the dialects. Still, better classification techniques must

be used in order to establish a benchmark.

Variations in perplexity may appear which is a good way of hinting towards the dialect of a specific text but not enough to prove that the dialect exists.

Ethical Statement

The dataset that we presented contains scraped Romanian language newspaper data and can be used for training or benchmarking language models. Given the fact that dialect identification may lead to discrimination in employment, criminality and other areas (Joshi et al., 2025) this dataset must only be used to fight against this behaviour not to reinforce it.

References

API.md. Membrii api. https://api.md/ro/membrii-api/. Accessed: 2025-02-04.

L Breiman. 2001. Random forests. *Machine Learning*, 45:5–32.

Andrei M. Butnaru and Radu Tudor Ionescu. 2019. Moroco: The moldavian and romanian dialectal corpus.

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Crăciun, and Dumitru-Clementin Cercel. 2024. Roqllama: A lightweight romanian adapted language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Stefan Daniel Dumitrescu, Andrei-Marius Avram, and Dumitru-Clementin Cercel. 2021. Robert - a romanian bert model.

e-ziare.ro. e-ziare.ro. https://e-ziare.ro/. Accessed: 2025-02-04.

Encyclopaedia Britannica. Dialect. https://www.britannica.com/topic/dialect. Accessed: 2025-02-04.

Fahim Faisal, Md Mushfiqur Rahman, and Antonios Anastasopoulos. 2024. Dialectal toxicity detection: Evaluating llm-as-a-judge consistency across language varieties.

Google. Google. https://google.com. Accessed: 2025-02-04.

Goose3 Developers. goose3/goose3: A python library for web content extraction. https://github.com/goose3/goose3/tree/master. Accessed: 2025-02-04.

Hugging Face. Perplexity in transformers. https://huggingface.co/docs/transformers/perplexity. Accessed: 2025-02-04.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Comput. Surv.* Just Accepted.

Mihai Masala, Denis C. Ilie-Ablachim, Alexandru Dima, Dragos Corlatescu, Miruna Zavelca, Ovio Olaru, Simina Terian-Dan, Andrei Terian-Dan, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "vorbeşti româneşte?" a recipe to train powerful romanian llms with english instructions.

Ministry of Foreign Affairs of Romania. Cernăuți. https://cernauti.mae.ro/node/392. Accessed: 2025-02-04.

Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2024. Evaluating dialect robustness of language models via conversation understanding.

Alexandru Tindeche and Florin Dinu. 2025. Romanian-language-study_romania-moldova. https://github.com/AlexTindeche/Romanian-Language-Study_Romania-Moldova. Accessed: 2025-02-04.

Wikipedia contributors. a. Listă de ziare din republica moldova. https://ro.wikipedia.org/wiki/List%C4%83_de_ziare_din_Republica_Moldova. Accessed: 2025-02-04.

Wikipedia contributors. b. Raioanele republicii moldova. https://ro.wikipedia.org/wiki/Raioanele_Republicii_Moldova. Accessed: 2025-02-04.

Wikipedia contributors. c. Romanian diaspora. https://en.wikipedia.org/wiki/Romanian_diaspora. Accessed: 2025-02-04.

Wikipedia contributors. 2025. Romanian dialects. https://en.wikipedia.org/wiki/Romanian_dialects. Accessed: 2025-02-04.