Archaeology of Intelligent Machines Comparing Romanian from Romania with Romanian from Moldova

1st Semester of 2024-2025

Authors

cosmina-anamaria.barbu@s.unibuc.ro
anca-ioana.oprea@s.unibuc.ro
mara-andreea.spataru@s.unibuc.ro

Abstract

This study presents a comprehensive computational analysis of stylistic differences between Romanian and Moldovan written texts using machine learning techniques. We employed logistic regression with TF-IDF features on a balanced corpus to classify and analyze texts from both variants. The model achieved a mean accuracy of 82.495% (σ = 2.605%), demonstrating significant and consistent stylistic differences. Through SHAP analysis and feature importance evaluation, we identified distinctive patterns in pronoun usage, grammatical structures and vocabulary choices that characterize each variant. Our findings reveal systematic differences in formality and syntactic preferences between Romanian and Moldovan texts.

1 Introduction

The relationship between **Romanian and Moldovan written texts** presents an interesting case study in linguistic variation. While sharing the same basic language structure, these variants have developed distinct characteristics influenced by different socio-political and cultural contexts. Our study aims to quantify and characterize these differences using computational methods, focusing on:

- Identifying distinctive linguistic features
- Measuring the extent of differentiation
- Understanding stylistic preferences in each variant
- Quantifying the reliability of computational distinction between variants

All contributions to the project have been thoroughly researched, analysed, and discussed collaboratively by all three team members.

We implemented a comprehensive machine learning pipeline consisting of:

- Text preprocessing and normalization
- Feature extraction using TF-IDF vectorization
- Classification using logistic regression
- Feature importance analysis using SHAP values
- Statistical analysis of linguistic patterns

We chose to analyze the stylistic differences between Romanian and Moldovan texts because the project combines traditional linguistics with modern computational methods, offering new insights into language variation. Additionally, the work has practical applications for machine translation and text processing tools. At the same time, we discovered that while Romanian and Moldovan have been studied linguistically, computational analysis of their stylistic differences remains relatively unexplored and we had access to sufficient data from both variants to perform a meaningful analysis.

Previous research on Romanian-Moldovan language variation has primarily focused on traditional linguistic approaches. Silviu Berejan analyzed the sociolinguistic situation and status of Romanian language in the Republic of Moldova, examining official language policies and their impact. Similar computational methods have been applied to other closely related language variants - for example (Berejan, 2002).
Jörg Tiedemann and Nikola Ljubešić used machine learning approaches to distinguish between Croatian, Serbian, and Bosnian texts, achieving high accuracy rates in language identification tasks (Tiedemann and Ljubešić, 2012).
In terms of methodological approach, our work builds on Alina Ciobanu and Liviu Dinu who used computational methods to analyze Romanian dialectal variations, though they focused on historical language changes rather than contemporary Romanian-Moldovan differences (Ciobanu

and Dinu, 2016).

Each author's perspective on the project:

Anca: "The biggest eye-opener was how messy real-world text processing is. I spent days battling processing the entire dataset of files, but it taught me that careful preprocessing is as crucial as the fancy algorithms. Next time, I'd like to dive into dialectal variations within each country - I bet there's a fascinating north-south pattern we could uncover."

Cosmina: "What fascinated me most about this project was that numbers can tell cultural stories. Each feature importance score revealed something about how language reflects society. For future work, I'm curious about applying this to social media text - I think the differences would be even more pronounced in casual communication."

Mara: "My main learning came from the statistical analysis of the results and cross-validation implementation. I learned how to properly evaluate classification models with imbalanced data and how to interpret feature importance in linguistic contexts. In the future, I would like to explore more sophisticated feature engineering techniques and perhaps extend this analysis to spoken language differences between Romanian and Moldovan."

2 Approach

Project Implementation Details

All code and data are available at: GitHub repo. Documentation and instructions for reproduction are included in the README file.

Software Tools Used

- Python 3.8 for implementation
- Scikit-learn for machine learning components
- SHAP library for feature analysis
- Pandas for data manipulation
- NumPy for numerical operations
- Matplotlib and Seaborn for visualizations

Processing Time

Data preprocessing: 4 minutes for the entire corpus

Model training: 30 seconds per fold in cross-validation

SHAP analysis: 2 minutes

Total pipeline execution: 10 minutes

Machine Learning Architecture

Model: Logistic Regression

Vectorization: TF-IDF with custom vocabulary

Cross-validation: 5-fold stratified shuffle split

Feature selection: built-in L2 regularization

Implementation Tricks

During development, we implemented several key techniques to improve our model's performance. We developed a **custom text cleaning function** specifically designed for Romanian diacritics to ensure proper character normalization. To reduce noise in our analysis, we implemented a **proper name removal** system and applied balanced sampling to handle differences in corpus sizes between variants. We normalized all texts to fall within a 300-1000 word limit to ensure comparable document lengths. A significant innovation in our approach was using a **custom vocabulary for TF-IDF** that combined both collocations and stopwords. This allowed us to capture both multi-word expressions and functional words that could signal stylistic differences. We also implemented **sentence shuffling and reconstitution techniques** to maintain text coherence while anonymizing content. These combined approaches helped focus our model on the most relevant linguistic features while reducing potential noise from irrelevant variations.

Evaluation Report

Classification accuracy: 82.53% Highest accuracy: 86.15% Lowest accuracy: 79.89% Mean accuracy: 82.53% Standard deviation: 2.27%

Visualizations and Evidence

The SHAP summary plot displayed here illustrates **the impact of different words** on our model's predictions.

Each word on the Y-axis is shown with its corresponding SHAP values on the X-axis, where positive values (dots to the right of the vertical line) indicate Romanian classification and negative values (dots to the left) indicate Moldovan classification. Words like 'ale', 'ce', and 'il' show predominantly positive SHAP values, suggesting they are strong indicators of Romanian texts. Conversely, words

like 'ul', 'noi', and 'te' lean towards negative values, marking them as characteristic of Moldovan texts. The density of pink dots at various positions shows the frequency and strength of each word's impact, while the blue vertical lines represent baseline values.

Particularly interesting is the distribution of words like 'in' and 'de', which show a spread across both positive and negative values, suggesting their usage patterns differ between the two variants. This visualization effectively captures the linguistic features our model uses to distinguish between Romanian and Moldovan texts.

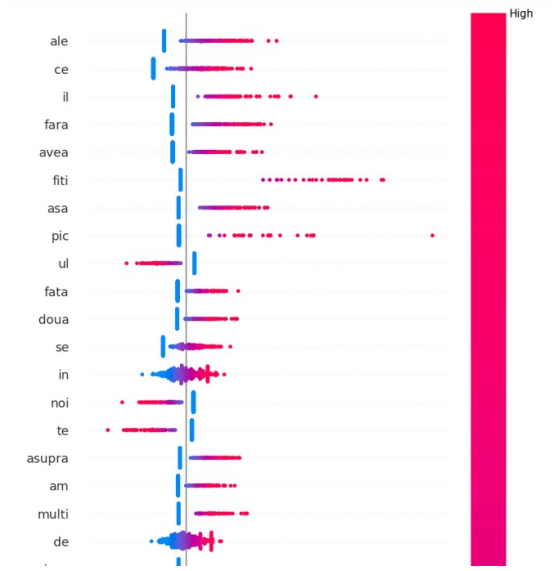

Figure 1: SHAP visualisation for words distribution.

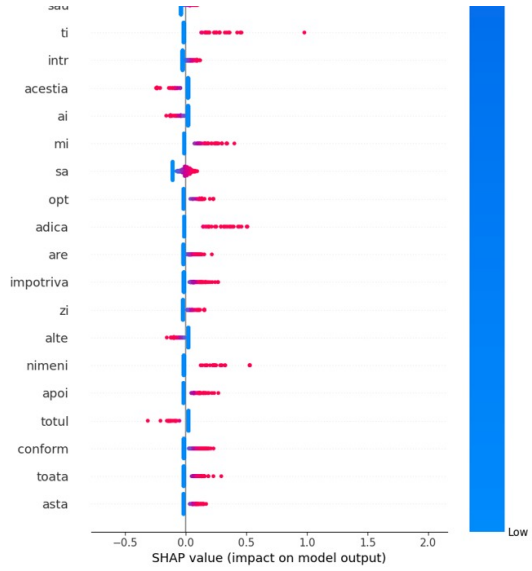

Figure 2: SHAP visualisation for words distribution.

Moldovan Texts:

The SHAP analysis revealed characteristic patterns in Moldovan texts, particularly in pronominal usage. Direct pronouns showed significant negative SHAP values (-2.63), indicating their strong association with Moldovan style. These pronouns, combined with informal linguistic markers and conversational elements, formed a consistent pattern that helped distinguish Moldovan texts from their Romanian counterparts.

Figure 3: SHAP visualisation for a Moldovan sample.

Romanian Texts:

The SHAP analysis of Romanian texts revealed distinct stylistic patterns, with formal language markers showing substantial positive SHAP values (3.65). These texts consistently exhibited a preference for complex grammatical constructions and literary vocabulary choices, distinguishing them from their Moldovan counterparts. This formal pattern manifested through the use of literary expressions and academic language constructions, which were key predictors for identifying Romanian texts.

Figure 4: SHAP visualisation for a Romanian sample.

Data Preparation

Corpus Statistics:

The Romanian texts have an average length of 588.94 words, with a standard deviation of 230.03 and lengths ranging between 300 and 1,000 words. While the Moldovan texts have an average length of 655.77 words, with a standard deviation of 261.91 and lengths ranging between 300 and 1,000 words.

Text Preprocessing

Our preprocessing relied on several text normalization functions: **no_diac()** for standardizing Romanian diacritics to ASCII characters, **remove_proper_names()** for filtering out proper nouns, and **tokenize_regex()** for word-level tokenization using custom boundary detection. These

functions worked together to ensure consistent text representation across both language variants.

Methodology

We used **StratifiedShuffleSplit** with 5 splits (90% training, 10% testing for each split) to evaluate our logistic regression classifier, ensuring class proportions were maintained across splits. Each training set processed approximately 3000 text samples, with the model achieving **consistent balanced accuracy** scores across different data partitions. The entire pipeline typically completed in under 4 minutes on standard hardware.

3 Limitations

Several important limitations should be noted in our approach.

First, our analysis relies heavily on word-level features, potentially **missing important syntactic and semantic patterns** that could differentiate the two variants.

Second, our current preprocessing pipeline requires **significant computational resources** for large-scale text processing, particularly during the proper name removal and sentence shuffling stages.

Third, the **model's performance might be biased** by our specific corpus composition - while we balanced the samples between variants, we cannot guarantee they represent the full spectrum of language use in either country.

Additionally, our TF-IDF vectorization approach with custom vocabulary, while effective, **limits the model's ability to adapt** to new vocabulary or evolving language patterns without retraining. Future work should explore more sophisticated language models that can capture deeper linguistic structures while maintaining computational efficiency.

4 Conclusions and Future Work

Reflecting on our project journey, we believe **our approach** to Romanian-Moldovan text classification was both challenging and rewarding. While we're satisfied with achieving **over 82% accuracy**, we could have explored more sophisticated approaches. Specifically, we could have investigated transformer models like BERT, which might have captured more nuanced linguistic differences. **The most enjoyable aspect** was discovering how machine learning could quantify subtle language variations that native speakers intuitively recognize. In terms of improvements, we'd love to explore social media text analysis, where we suspect the differences between variants might be even more pronounced. **The most challenging part** was finding appropriate texts and loading them to match the existing structure - ensuring each text satisfied our criteria while maintaining consistency with previously processed data. However, these challenges taught us valuable lessons about real-world data collection and preprocessing in NLP applications. **This project was particularly engaging** as it offered a refreshing departure from typical computer science assignments - having a machine learning model uncover subtle linguistic variations in our own language was both novel and fascinating. We learned that NLP techniques we study can be meaningfully applied to culturally relevant analyses, making theoretical concepts come alive in ways that were both intellectually stimulating and personally enjoyable to explore.

References

Silviu Berejan. 2002. Aspecte ale standardizării limbii române în republica moldova. *Limba Română*, XII(10):4–14.

Alina Maria Ciobanu and Liviu P. Dinu. 2016. A computational perspective on romanian dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4702–4705, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634. The COLING 2012 Organizing Committee.