Sentimental RO-BERT

Ruxandra Maria Gonțescu, Eduard Marian Florin Marin, Vlad Mihai Olăeriu, Florin Brad*, Ioana Pintilie*, Marius Drăgoi**

University of Bucharest, Romania

*Bitdefender, Romania

UNIVERSITY OF BUCHAREST
VIRTUTE ET SAPIENTIA

gontescuruxandra@yahoo.com, eduardmarin233@gmail.com, vladolaeriu@gmail.com, fbrad@bitdefender.com*,mdragoi@bitdefender.com*,ioana.pintilie@s.unibuc.ro*

1. Introduction

At the moment, current multilingual models fail to obtain optimal results in comparison with models trained exclusively on Romanian language data.
The need for better models in our native language inspired our research question: how much does data from at face value opposite domains matter in the performance of SOTA Romanian language models?

2. Dataset and Preprocessing

Datasets:
One of the datasets used in this study is **LaRoSeDa** (Large Romanian Sentiment Data), which consists of 15,000 reviews, balanced between positive and negative sentiments.
Texts from two distinct sources were utilized to enrich language model pretraining. Legal documents from **MARCELL** (non-emotional) provided structured language examples, while **Romanian literature** (emotional) introduced expressive language patterns.
With the models enhanced, we employed the **REDv2** dataset to refine the model's performance specifically for sentiment analysis tasks within twits from different areas of interest: sport, news, entertainment, socializing.

Dataset stats	
LaRoSeDa-average words: positive text	31.59
LaRoSeDa-average words: negative text	40.80
REDv2-average words:	22.86
MARCELL-average words	104.2
Literature-average words	115.1

Data preprocessing:
Experiments:

- Conducted training with and without preprocessing.
- Underwent training on the data split given by the *authors of the dataset*, but also on random split.

Preprocessing Steps:

- *Stopwords Removal*: Removal of common, less informative words.
- *Diacritics Stripping*: Standardization by removing diacritical marks.
- *Lowercasing*: Uniform text case for consistent processing.
- *Stemming*: Reducing words to their stem for better generalization.

5. Conclusion

Our results suggest that literature texts aren't superior compared to juridic documents, even though they differ in emotional tone and word choice. Interestingly, in the LaROSeDa dataset, feeding the models the title, which is very short, significantly improved the results.

3. Models

RoBERT-base pre-trained BERT-base model on Romanian corpus: RoWiki, OSCAR, RoTex.

RoBERT-small smaller version of RoBERT-base, 12 layers, 256 hidden size, 8 attention heads (*19M weights*)

RomanianBERT-cased: pre-trained BERT-base model on Romanian corpus; we upgrated it to fit the classification task by adding some fully connected layers after it; the fully connected architectures used:

- **NN1**: Dropout 0.2 + Linear
- **NN2**: Dropout 0.2 + Linear + GELU + Linear
- **NN3**: Dropout 0.2 + LSTM + Linear classifier

4. Results

LaRoSeDa: as shown below, the influence of the title is *significant* in the classification task. Morever, the model RoBERT-small kept up in performance with the bigger RoBERT-base and RomanianBERT-cased models.

Results of the models run on the *content*:

Architecture	Accuracy	Precision	Recall	F1
RoBERT-small raw	0.935	0.932	0.937	0.935
RoBERT-base raw	**0.956**	**0.960**	**0.952**	**0.956**
RomanianBERT-cased raw	0.901	0.928	0.931	0.93

Results of the models run on the *title and content*:

Architecture	Accuracy	Precision	Recall	F1
RoBERT-small raw	0.976	0.972	0.979	0.975
RoBERT-small raw + random split	**0.983**	**0.987**	**0.979**	**0.983**
RoBERT-base	0.970	0.960	0.981	0.970
RomanianBERT-cased preprocessed	0.901	0.928	0.931	0.93
RomanianBERT-cased preprocessed + NN1	0.947	0.957	0.906	0.93
RomanianBERT-cased raw + NN1	0.967	0.957	0.978	0.967
RomanianBERT-cased raw + NN2	0.982	0.982	0.982	0.982
RomanianBERT-cased raw + NN3	0.971	0.965	0.977	0.971

REDv2: from the results below, the best-performing RoBERT was the one that was pre-trained on a smaller dataset with law samples. Another point worth mentioning, is for the literature dataset, the increase of the dataset size resulted in a better accuracy of the model.

Architecture	Dataset	Samples	Accuracy	Hamming Loss
RoBERT-small	literature	3000	0.539	0.111
RoBERT-small	literature	6000	0.546	0.112
RoBERT-small	**law**	**3000**	**0.562**	**0.107**
RoBERT-small	law	6000	0.546	0.112