

Bridging Linguistic Gaps: SENTIROM's Approach to Romanian Sentiment Analysis

Andra-Gabriela Ursa, Laura Dioşan



Introduction

Sentiment analysis in texts

Given: A textual dataset (reviews) in Romanian.

Required: A tool capable of analyzing and interpreting sentiments and intentions in the text, providing a sentiment evaluation (positive, negative).

We need AI to solve the problem for:

1. Understanding sentiments;
2. Advancing natural language processing (NLP);
3. Developing a sentiment analysis tool

Challenges in solving the problem:

1. The complexity of natural language;
2. The lack of linguistic resources for the Romanian language;
3. Integrating advanced NLP technologies

RQ1: How effective is the novel BERT and K-Means based approach in accurately classifying sentiment in Romanian text compared to existing models?

RQ2: To what extent can the LaRoSeDa dataset be translated and compared to English sentiment analysis frameworks?

Datasets

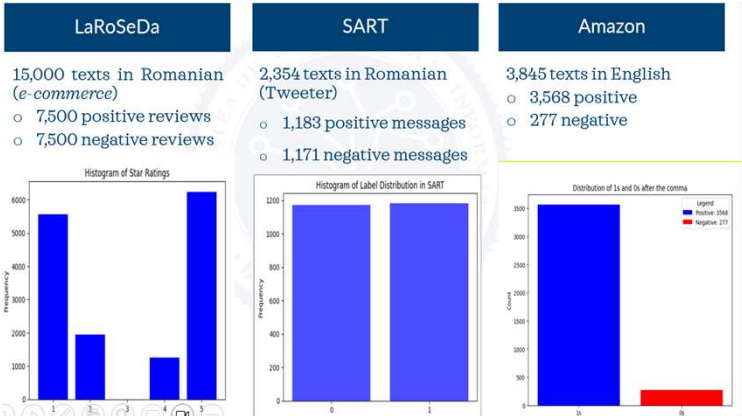


Fig.1: Datasets used with their histogram.

SENTIROM – Experimental Analysis

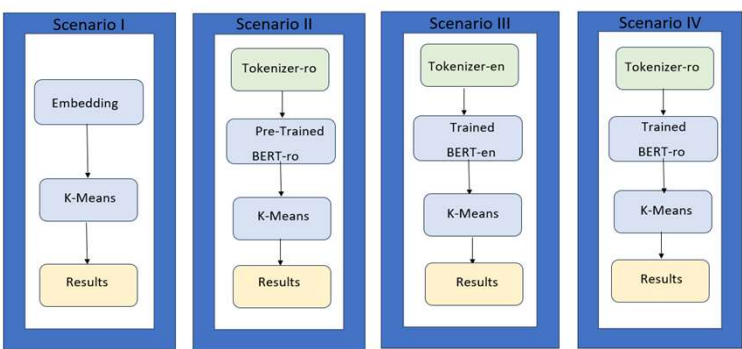


Fig. 2. Scenarios of SENTIROM

Results

Table 1. Accuracy obtained on LaRoSeDa testdata for Word2Vec and BERT-ro models

Model	Accuracy
Trained Word2Vec	57.86%
Pre-trained BERT-ro and pre-trained tokenizer-ro	71.00%
Pre-trained BERT-ro and tokenizer-ro fine-tuned on RED	79.46%

Table 2. Loss of training phase and Accuracy of testing phase obtained for our BERT-based models

Index	Model	Loss	Accuracy
1	Trained BERT-ro and tokenizer-ro with K-Means on LaRoSeDa trained on 12,000 data and tested on 3,000 data	0.0199	96.80%
2	Trained BERT-ro and tokenizer-ro with K-Means on SART trained on 1,647 data and tested on 707 data	0.0217	98.30%
3	BERT-en and tokenizer-en with K-Means on LaRoSeDa translated, trained on 700 data and tested on 300 data	0.6822	61.53%
4	BERT-en and tokenizer-en with K-Means on Amazon, trained on 700 data and tested on 300 data	0.6869	81.33%
5	BERT-en and tokenizer-en with K-Means on Amazon, trained on 2,660 data and tested on 1140	0.6602	92.72%

Example of a wrongly classified text

- **True label:** 1 (Positive) vs **Predicted label:** 0 (Negative)
- **Content:** *produsul e bun, insa nu pentru modelul cumparat de mine, in aceeaşi comanda. nu se potriveşte la toate iPad-urile de 9,7. având în vedere că în magazin am primit consultanța de la un specialist, nu mă așteptam să nu se potrivească.* (translation: *the product is good, but not for the model bought by us, in the same order. it does not fit all the iPads of 9,7. considering that I received consultancy from a specialist in the store, I didn't expect it not to fit.*)

SWOT

Strengths:

- High accuracy on target datasets
- Effectiveness of tokenizers

Weaknesses:

- Lower performance on translated datasets

Opportunities:

- Expansion to multilingual and diverse datasets
- Application in real-world scenarios

Threats:

- The labeling process can be subjective and inaccurate

Conclusions and Future Work

Briefly, the main contribution of this paper are:

- a comparison of various models for sentiment analysis in Romanian language,
- a novel BERT and K-Means based approach for sentiment analysis, and a dataset labeled (by thresholding) with binary sentiments (positive and negative).

The primary objective of this work was to develop SENTIROM, a system capable of analyzing text to determine its sentiment – categorizing it as either positive or negative. To achieve this, we conducted four experiments, all of which employed the LaRoSeDa dataset as their corpus. We consider the results to be promising and some even better than expected. Even if the training phase may have been a little long and with some problems we obtained promising results. We promote my combined approach based on a trained BERT-ro and tokenizer-ro with K-Means clustering, which proved to be an effective strategy even though I initially doubted its feasibility. It overcomes the results of the existing state of the art [3] which got a 90,90% accuracy, my work having a 96,8% accuracy.

Future work:

- In-depth analysis of incorrectly labeled sentences
- Development of a browser extension
- Addition of a new class (neutral sentiment)

References

1. R. T. Ionescu, A. M. Butnaru (2018) Improving the results of string kernels in sentiment analysis and Arabic dialect identification by adapting them to your test set, EMNLP(1084–1090)
2. S. D. Dumitrescu et al. (2020) The birth of romanian BERT, arXiv:2009.08712
3. A. M. Tache et al. (2021) Clustering word embeddings with self-organizing maps. application on LaRoSeDa—a large romanian sentiment data set, arXiv:2101.04197
4. D. C. Neagu et al. (2022) Towards sentiment analysis for romanian twitter content, Algorithms 15(10), 357
5. A. Barila et al. (2022) Romanian-lexicon-based sentiment analysis for assessing teachers activity, IJCSNS 22(10), 43–5



12th ACM Celebration of Women in Computing: womENCourage™
Braşov, Romania
17-19 September, 2025
Theme: Computer Science: a Catalyst for Educational Change

