

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

SENTIROM

Andra-Gabriela Ursa

Faculty of Mathematics and Computer Science, Babeş-Bolyai University Cluj-Napoca, Romania,

[andra.ursa@stud.ubbcluj.ro](mailto:andra.ursa@stud.ubbcluj.ro)

Laura Dioşan

Faculty of Mathematics and Computer Science, Babeş-Bolyai University Cluj-Napoca, Romania,

[laura.diosan@stud.ubbcluj.ro](mailto:laura.diosan@stud.ubbcluj.ro)

## Extended Abstract

Sentiment analysis is a vital technology for understanding opinions, emotions, and attitudes expressed in textual data, especially on digital platforms where traditional communication cues such as tone or facial expressions are absent. While considerable progress has been made for major languages like English, many less-resourced languages—including Romanian—still face significant challenges due to a scarcity of annotated corpora, linguistic tools, and tailored computational models. This research addresses this gap by presenting SENTIROM, a comprehensive system designed to advance Romanian sentiment analysis through the integration of modern natural language processing (NLP) techniques and machine learning approaches.

To accurately classify sentiment, SENTIROM employs a multi-strategy experimental framework investigating four approaches combining word embeddings and clustering: (1) a baseline using **Word2Vec** embeddings trained from scratch on Romanian reviews coupled with **K-Means** clustering; (2) pre-trained **BERT-ro** models and Romanian-specific tokenizers, combined with K-Means clustering; (3) a fine-tuned English BERT model applied to an English translation of the Romanian dataset to evaluate cross-lingual transferability and model generalization; and (4) a custom BERT-ro model trained from scratch on Romanian datasets, paired with K-Means clustering to leverage language-specific nuances.

Our experiments, conducted on multiple datasets including **LaRoSeDa**, the Romanian Twitter sentiment dataset **SART**, and a subset of **Amazon** reviews, demonstrate that the custom-trained BERT-ro model paired with K-Means clustering significantly outperforms the other methods, achieving an accuracy of **96.8%** on LaRoSeDa and an even higher accuracy of **98.3%** on the SART dataset. These results markedly surpass previous benchmarks, including the **90.9%** accuracy achieved by earlier BERT-ro models. The Word2Vec baseline model achieved a modest 57.86% accuracy, highlighting the superiority of transformer-based language models. Meanwhile, the English BERT model's performance on translated datasets was notably lower, indicating the crucial importance of language-specific training and tokenization in sentiment analysis.

The methodology involved careful preprocessing steps, including punctuation removal, normalization, tokenization adapted to Romanian linguistic characteristics, and dataset balancing to reduce labeling bias. SENTIROM's use of K-Means clustering to group text embeddings allows an unsupervised final layer of sentiment assignment, demonstrating that clustering combined with contextual embeddings can effectively capture sentiment polarity in Romanian texts.

Looking forward, SENTIROM opens promising avenues for extending this approach to other low-resource languages and more diverse datasets, such as multilingual corpora and social media content.

Key contributions include:

- the first comprehensive evaluation of BERT-based architectures for Romanian sentiment classification across multiple datasets;
- a novel fine-tuned and trained-from-scratch BERT-KMeans pipeline tailored to Romanian text;
- a critical assessment of translation-based transfer learning;

Future work could focus on integrating sentiment analysis with emotion detection, sarcasm recognition, and domain adaptation to improve robustness and real-world applicability. The demonstrated effectiveness of custom BERT models also suggests potential for industry applications in customer feedback analysis, social media monitoring, and market research, where accurate sentiment understanding is crucial.

In summary, SENTIROM significantly advances the field of Romanian sentiment analysis by delivering a high-accuracy, language-aware system that overcomes resource limitations. It establishes a new state of the art and lays a solid foundation for future research and practical NLP applications in Romanian and other underrepresented languages.

CCS CONCEPTS • Information systems → Information retrieval → Retrieval tasks and goals → Sentiment analysis

• Computing methodologies → Artificial intelligence → Natural language processing • Information systems → Information retrieval → Retrieval tasks and goals → Clustering and classification

**Additional Keywords and Phrases:** Sentiment analysis, Romanian language, BERT, Word2Vec, K-Means, LaRoSeDa, NLP, low-resource language.

## REFERENCES

- [1] Tache, A. M., Gaman, M., & Ionescu, R. T. 2021. Clustering word embeddings with self-organizing maps. application on laroseda--a large romanian sentiment data set. *arXiv preprint arXiv:2101.04197*.
- [2] D. C. Neagu, A. B. Rus, M. Grec, M. A. Boroianu, N. Bogdan, A. Gal, 2022 "Towards sentiment analysis for romanian twitter content ", Algorithms 15 (10) 357
- [3] S. D. Dumitrescu, A.-M. Avram, S. Pyysalo. 2020 "The birth of romanian bert" , arXiv preprint arXiv:2009.08712
- [4] A. Barila, M. Danubianu, B. Gradinaru, 2022 "Romanian-lexicon-based sentiment analysis for assesing teachers activity, International journal of computer science and network security: IJCSNS 22 (10) 43-5
- [5] Ciobotaru, Alexandra and Dinu, Liviu P. 2023 Proceedings of the 27th International Conference on Knowledge-Based and Intelligent Infor mation & Engineering Systems (KES 2023)
- [6] Haque, T. U., Saber, N. N., Shah, F. M. 2018, May. Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE international conference on innovative research and development (ICIRD) (pp. 1-6). IEEE.
- [7] Ionescu, R. T., & Butnaru, A. M. 2018. Improving the results of string kernels in sentiment analysis and Arabic dialect identification by adapting them to your test set. *arXiv preprint arXiv:1808.08409*.
- [8] Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- [9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova 2018 "Bert: Pre-training of deep bidirectional transformers for language understanding" arXiv preprint arXiv:1810.04805
- [10] Rahman, W., Hasan, M. K., Zadeh, A., Morency, L. P., Hoque, M. E. 2019. M-bert: Injecting multimodal information in the BERT structure. arXiv preprint arXiv:1908.05787.
- [11] Jorg Tiedemann. (2012). Parallel data, tools and interfaces in opus. In LREC.
- [12] Pedro Javier Ortiz Suarez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Cardiff, United Kingdom.
- [13] Ciobotaru, A., Constantinescu, M. V., Dinu, L. P., & Dumitrescu, S. 2022, June. RED v2: enhancing red dataset for multi-label emotion detection. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 1392-1399)