BABEŞ BOLYAI UNIVERSITY, CLUJ NAPOCA, ROMÂNIA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

BACHELOR'S DEGREE

SENTIROM (Sentiment Recognition in Romanian texts)

Coordonator
Prof. Laura Silvia DIOŞAN

Andra-Gabriela URSA
Computer Science
237
andra.ursa@stud.ubbcluj.ro

2024

Abstract

As our world changes, technology is reshaping how we talk and share our feelings, especially on social media. This brings up an interesting question: How well can we understand the sentiments expressed in online texts?

SENTIROM aims to contribute to Romanian sentiment analysis research focusing on finding the best approach in detecting sentiments right. With little resources available for the creation of tools for natural language processing, Romanian is one of the languages that computational linguistics has studied the least. We used the LaRoSeDa and SART dataset in our study to try and close this gap. To investigate how this dataset may be utilized in bigger projects, we improved it by adding a Word2Vec and K-Means approach, a BERT-ro and BERT with personal setup using tokenizer-ro and K-Means, and comparing it with an English dataset. Our objective is to enhance Romanian language comprehension and processing in the context of NLP activities. By applying these techniques to the LaRoSeDa dataset, we aim to demonstrate the potential for advanced NLP applications in Romanian.

The main contributions are:

- Trying out different setups for our computer models, looking at their results, and discussing each one.
- We designed and implemented an intelligent system aimed at simplifying the data collection process. This innovation is expected to significantly contribute to the writing of future corpora, ensuring that they are larger and of higher quality.
- We introduced a new approach with BERT with a Romanian tokenizer, alongside K-Means clustering. This approach has proven to be not only viable but also superior to pre-existing models in the domain being tested on LaRoSeDa, but also on SART dataset.
- We also tried to see if the dataset translated into English is having a similar result to an Amazon dataset (with the same structure as mine) run on a model similar to the ones I made before.
- The experiments in this thesis were supported by a research paper accepted by KES 2024: Ursa, A. Diosan, L. (2024). Bridging Linguistic Gaps: SENTIROM's Approach to Romanian Sentiment Analysis. International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES)