WEAT Analysis of Human and Language Model Biases in Romanian

2025-2026

Topliceanu Maria-Adina
University of Bucharest
maria-adina.topliceanu@s.unibuc.ro

Chera Gabriel-Alexandru
University of Bucharest
gabriel-alexandru.chera@s.unibuc.ro

***Abstract*—This paper presents the way human language reflects cultural and social biases, which can be encoded in both human judgments and artificial intelligence (AI) language models. While biases have already been studied in English using the Word Embedding Association Test (WEAT) and its multilingual variant, CA-WEAT, there is little work exploring such biases in Romanian. In this study, we present analysis of biases in Romanian.**

I. Introduction

Understanding how cultural and social biases appear in language is an important step in analyzing both human reasoning and the behavior of artificial intelligence models. Previous studies have shown that word embeddings and language models trained on human-written texts often inherit, amplify, or even invert human biases. However, most existing work evaluates these phenomena in English, leaving a gap in understanding how such biases manifest in other languages. In this project, we aim to examine whether the Word Embedding Association Test (WEAT), and its multilingual variant CA-WEAT, reveal measurable biases in Romanian, both in humans and in AI models.

A. Recent Studies

Studies have investigated linguistic bias and cultural variation in word embeddings across different languages. The *Corpus del Español REAL* [1] provides a large, dialect-aware Spanish corpus used as a foundation for recent multilingual bias studies. Building on this corpus, España-Bonet and Barrón-Cedeño [2] analyzed how multilingual language models tend to attenuate or even reverse human biases when compared to language models. Beyond Spanish, Jiao [4] investigated gender bias in Chinese word embeddings, demonstrating that even typologically distant languages exhibit measurable associations similar to those found in English WEAT tests.

B. Motivation

We chose this project because Romanian is rarely included in multilingual bias studies, despite being widely spoken. Most WEAT and CA-WEAT research focuses on English [3], Spanish [2], or Chinese [4], leaving a gap in evaluating biases embedded in Romanian culture language. We believe that studying Romanian biases contributes to a more inclusive understanding of cultural variability in AI behavior.

C. Aim

The main problem we address is whether linguistic biases that appear in English WEAT studies also appear in Romanian, and whether AI language models replicate biases shown by native Romanian speakers. Specifically, we investigate:

- whether Romanian speakers associate certain target words with specific concepts,
- whether AI models exhibit similar associations,
- and how Romanian bias patterns compare with those found in other languages.

D. Personal Learning Outcomes

- **Maria-Adina:** Through this project, I learned how language models trained on human-written texts can exhibit social and cultural biases, and how these biases can be measured using WEAT. I also gained experience in designing experiments to compare human judgments with AI outputs, processing and analyzing survey and model data. In the future, I would like to explore how biases vary across dialects of Romanian.
- **Gabriel-Alexandru:** Working on this project helped me understand how language models reflect patterns and biases that appear in human language, and how these can be measured using WEAT. I learned to run experiments, and analyze both model and human evaluation data. This experience also improved my skills in interpreting results and understanding their implications. Going forward, I would like to investigate how different training data or embedding techniques might influence the presence of biases.

E. Approach

Our approach is based on four main steps:

1) We selected word categories inspired by CA-WEAT and translated them into Romanian.
2) We created a survey in which participants rated words from 1–7 based on perceived associations.
3) We collected equivalent ratings from AI models.
4) We computed WEAT-style bias scores for Romanian and compared human–AI similarities as well as differences across languages.

II. APPROACH

As we previously stated, we took a different approach for this study we conducted.

A. Data

We began our approach by selecting the dictionary of target found in the data provided on Github by Corpus del Español REAL [5] and translating the entire word list into Romanian preserving meaning and context.

B. Datasets

After preparing the Romanian version of the word sets, we designed a Google Forms survey in which participants rated each word on a Likert scale from 1 to 7. The survey was completed by 40 Romanian speakers of diverse ages, ensuring a broad range of perspectives.

To evaluate the same set of words using AI language models, we created a standardized prompt and submitted it to publicly available free APIs such as ChatGPT, Gemini, NotebookLM, Meta AI, DeepSeek, and Perplexity AI. Each model provided numerical ratings for the same Romanian word lists, and these outputs were collected and stored in a CSV file for further analysis.

We then cleaned and aligned the human survey CSV with the AI CSV, ensuring both datasets shared the same structure, word order, and rating format. This preprocessing step allowed us to load both files into Python seamlessly.

C. Use of WEAT

WEAT works by measuring how closely different groups of words are associated with particular attributes inside a word-embedding space.Each word is represented as a vector, and WEAT computes the cosine similarity between vectors to estimate how semantically related two words are. In our project, we apply WEAT directly to the dataset of human ratings and AI-generated ratings to quantify potential cultural or linguistic biases. After collecting all similarity judgments (human Likert scores and model-produced scores), we treat each set of target and attribute words following the standard WEAT structure. Instead of word-embedding vectors, our "vectors" are the numerical ratings assigned by participants and language models. We compute the cosine similarity between target–attribute pairs to evaluate how strongly each group associates certain words with specific attributes.

D. Resource Requirements

No GPU resources were required for this project, as all computations involved only the calculation of cosine similarities and WEAT-based statistical evaluations. The analyses, including processing of human survey data and AI model outputs, were performed efficiently on a standard CPU with 4–8 GB of RAM.

E. Source

All code, survey data and intermediate files are available at the following repository: https://github.com/mariaxadina/WEAT-Analysis-of-Human-and-Language-Model-Biases-in-Romanian.

III. LIMITATIONS

IV. RESULTS

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

V. CONCLUSIONS AND FUTURE WORK

REFERENCES

[1] Corpus del Español REAL. Available at: https://cereal-es.github.io/CEREAL/#data.

[2] The (Undesired) Attenuation of Human Biases by Multilinguality. Cristina España-Bonet, Alberto Barrón-Cedeño. Available at: https://aclanthology.org/2022.emnlp-main.133/.

[3] Elote, Choclo and Mazorca: on the Varieties of Spanish. Alberto Barrón-Cedeño and Cristina España-Bonet. Available at: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://aclanthology.org/2024.naacl-long.204.pdf.

[4] Investigating Gender Bias in Word Embeddings for Chinese by Meichun Jiao. Available at: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.diva-portal.org/smash/get/diva2:1621875/FULLTEXT01.pdf.

[5] CA-WEAT Data. Available at: https://github.com/cristinae/CA-WEAT/blob/main/data/CA-WEATv1.tsv.