Word Embedding Association Test (WEAT)
Archaeology of Intelligent Machines

1st Semester of 2024-2025

Laura-Vanesa Lungu
laura-vanesa.lungu@s.unibuc.ro

Ana-Maria Suditu
ana-maria.suditu@s.unibuc.ro

Abstract

Human biases, such as associating flowers with pleasantness and insects with unpleasantness, are often reflected in machine learning models trained on human-generated text. This project explores these biases in Romanian language models by building on the findings of (España-Bonet and Barrón-Cedeño, 2022). The study uses two methods to measure biases: the Word Embedding Association Test (WEAT), which examines universal biases, and a culturally adapted version called CA-WEAT, designed to capture cultural differences. These tests are applied to three types of language models: monolingual (FastText) and multilingual (mBERT and XLM-R). Statistical tools were used to measure and compare the strength of biases across these models. The results show that monolingual models retain stronger biases, while multilingual models reduce or attenuate biases, offering valuable insights into how language models process cultural and universal associations.

1 Introduction

Language models are systems trained to understand and process human language by analyzing large collections of text. These models often mirror the biases present in the text they are trained on. For instance, people might universally associate flowers with positive emotions and insects with negative ones. Such associations are often embedded into language models and can manifest in unintended ways. While universal biases are shared across cultures, many biases are language- or culture-specific. For example, some cultures might associate certain animals with sacredness or danger, which may not be true in others. This variability poses a challenge for multilingual language models that are designed to handle multiple languages simultaneously. The paper by (España-Bonet and Barrón-Cedeño, 2022) introduced CA-WEAT, an extension of the Word Embedding Association Test (WEAT), to evaluate biases in multilingual models while accounting for cultural differences. Their findings suggested that multilingual models tend to attenuate biases compared to monolingual ones.

Problem Statement

Bias in neural models is a significant concern in artificial intelligence. When these models are applied to real-world tasks—such as hiring decisions, loan approvals, or content moderation—they can perpetuate stereotypes and reinforce systemic inequities. For instance, a language model trained on biased data might associate certain professions predominantly with one gender, or it could amplify cultural stereotypes in its outputs.

Identifying and quantifying these biases is an essential step toward building fair and ethical AI systems. While much work has been done to study biases in major languages like English, less attention has been given to languages spoken in smaller communities or those with unique cultural characteristics. This gap is particularly concerning because models trained on multilingual corpora may generalize poorly, potentially attenuating or even reversing biases when moving between languages.

In this project, we do not aim to study harmful or discriminatory biases (e.g., racial or gender biases). Instead, we focus on cultural biases—associations and preferences rooted in the traditions, history, and cultural identity of a language. Specifically, we aim to understand how biases manifest in Romanian, a less-studied language, and how these biases differ between monolingual and multilingual models.

Related Work

Our research builds on foundational work by (Caliskan et al., 2017), who introduced WEAT to quantify biases in English word embeddings. Further studies such as (Gonen and Goldberg, 2019), investigated methods to detect and mitigate gender biases in embeddings. In the multilingual context, (Lauscher and Glavaš, 2019) introduced X-WEAT, extending bias analysis to multiple languages. More recently, (España-Bonet and Barrón-Cedeño, 2022) introduced CA-WEAT, emphasizing culturally aware evaluations, which demonstrated that biases vary significantly across languages and that multilingual embeddings often attenuate or reverse these biases. Their findings directly inspired our study, focusing on Romanian-specific adaptations of WEAT and CA-WEAT.

Reproducing Existing Work

We extended the methodology of (España-Bonet and Barrón-Cedeño, 2022) by creating culturally adapted WEAT lists for Romanian. Unlike the original multilingual approach, our focus on Romanian required addressing challenges specific to monolingual settings, such as handling linguistic nuances and ensuring cultural relevance. While (España-Bonet and Barrón-Cedeño, 2022) highlighted the attenuation effects of multilingual embeddings, our study examined how these effects interact with Romanian-specific biases, particularly in contextual embeddings like mBERT and XLM-R. This work serves as a bridge between monolingual and multilingual evaluations, shedding light on the complexities of bias attenuation across linguistic contexts.

Contributions

1. **Ana-Maria Suditu:** Managed data collection, cleaning, and structuring for the project. Translated and adapted WEAT tests into Romanian and developed culturally specific CA-WEAT lists based on input from native speakers. Ensured data readiness for analysis and validated the quality of embeddings used in testing.

2. **Laura-Vanesa Lungu:** Led the statistical analysis of biases in Romanian embeddings, focusing on both static (FastText) and contextual models (mBERT, XLM-R). Conducted tests using WEAT and CA-WEAT to quantify bias strength and compared results across monolingual and multilingual embeddings.

Learning Outcomes and Future Directions

1. **Laura-Vanesa Lungu:** Strengthened expertise in comparative analysis of embedding models. Plans to explore how statistical insights and tools can be used to uncover the propagation of subtle cultural nuances in language models. In addition, the goal is to expand her knowledge in statistical analysis and data visualization techniques, recognizing the need for further experience in advanced data analysis workflows.

2. **Ana-Maria Suditu:** Through this project, I deepened my understanding of data preprocessing techniques and gained experience in adapting and translating existing methodologies to culturally specific contexts, enhancing my skills in bridging technical and linguistic domains. In the future, I want to deepen my understanding of NLP techniques.

Summary of Approach

We adapted WEAT to Romanian by designing culturally relevant word sets that reflect the unique characteristics of the language. We evaluated biases in models such as cc.ro.300, mBERT and XLM-R, comparing their performance across training corpora. Multilinguality and its influence on bias attenuation were also studied, offering a nuanced understanding of how embeddings behave in diverse linguistic scenarios.

2 Approach

All the code and data for this project can be found in the following repository: Project

What are WEAT and CA-WEAT?

WEAT (Word Embedding Association Test): The Word Embedding Association Test (WEAT) is a statistical method developed to measure biases in word embeddings. WEAT evaluates how strongly two categories of target words (e.g., *flowers* vs. *insects*) are associated with two sets of attribute words (e.g., *pleasant* vs. *unpleasant*). The associations are measured using a similarity metric, such as cosine similarity, which calculates how "close" words are in the embedding space.

Example: In WEAT, *flowers* might be closer to *pleasant* words like *joy* or *happiness*, while *insects* might be closer to *unpleasant* words like *fear* or *disgust*. This difference in association indicates the strength of the bias encoded in the embeddings.

CA-WEAT (Culturally Adapted WEAT): The Culturally Adapted WEAT (CA-WEAT) extends the original WEAT methodology by incorporating culturally specific word lists. This adaptation addresses the limitations of directly translating WEAT word lists into other languages, where cultural and linguistic differences may affect word associations.

Why CA-WEAT is Necessary: Some concepts or categories used in WEAT may not translate well into other languages due to differences in cultural context. For example, while English lists for flowers might include *aster*, a Romanian speaker might think of *ghiocel* (spring snowflake) as a culturally relevant example of a flower. Similarly, culturally significant terms like *Palma Maicii Domnului* (Madonna Lily) add depth to the analysis, capturing nuances that are specific to Romanian speakers.

How CA-WEAT Works: Instead of relying solely on predefined lists, CA-WEAT leverages input from native speakers to generate culturally relevant target and attribute word sets. These lists are then used to perform bias analysis, revealing how biases manifest in a specific cultural and linguistic context.

Both WEAT and CA-WEAT allow researchers to quantify biases in embeddings, but CA-WEAT adds a critical layer of cultural awareness, making it particularly useful for studying less-represented languages like Romanian.

Splitting Data into Tests

The tests were divided into two groups, following the methodology of the original paper:

WEAT1 / CA-WEAT1

- **Target categories:** Flowers and Insects.
- **Attribute categories:** Pleasant and Unpleasant.
- **Purpose:** To evaluate universal and cultural biases related to nature and aesthetics.

WEAT2 / CA-WEAT2

- **Target categories:** Musical Instruments and Weapons.
- **Attribute categories:** Pleasant and Unpleasant.
- **Purpose:** To measure biases in associations related to cultural artifacts and violence.

The results of each test were computed separately and compared across embedding models (FastText, mBERT, and XLM-R).

Embedding Models

Three types of embeddings were used to capture and compare biases:

FastText (cc.ro.300.bin)

- Monolingual static embeddings trained exclusively on Romanian text.
- Biases in this model reflect cultural and linguistic nuances specific to Romanian.

mBERT (Multilingual BERT)

- A multilingual contextual embedding model trained on 104 languages.
- Captures contextual meanings of words but may attenuate biases due to cross-lingual generalization.

XLM-R (Cross-lingual RoBERTa)

- An advanced multilingual embedding model trained on 100 languages.
- Known for its robust performance in multilingual and cross-lingual tasks.

The results of each test were analyzed to understand the differences in bias strength across the three models.

Experimental Setup

- **Translation and Adaptation:** The WEAT lists were translated into Romanian, while CA-WEAT lists were created independently by gathering culturally relevant examples from participant responses.
- **Comparison Across Tests and Models:** Each model (FastText, mBERT, XLM-R) was evaluated on both WEAT1 / CA-WEAT1 and

WEAT2 / CA-WEAT2. Results were analyzed to understand the differences in bias strength between models and between the universal (WEAT) and culturally adapted (CA-WEAT) tests.

Statistical Methods and Analysis

Quantifying and Comparing Biases

To quantify and compare biases in word embeddings, the following statistical methods were employed:

- **Statistical s:** Calculates the difference in average similarity between target categories and attribute words.
- **Effect Size (d):** A standardized measure of bias strength, where larger values indicate stronger biases.
- **Bootstrap Confidence Intervals:** Used to assess the robustness of statistical results by resampling data to estimate confidence intervals.

Defining Functions for Statistical Analysis

Cosine Similarity

The cosine similarity measures the closeness between two word vectors:

$$\text{sim}(t, a) = 1 - \cos(\mathbf{t}, \mathbf{a})$$

where $\mathbf{t}$ and $\mathbf{a}$ are the vectors for the target and attribute words, respectively.

Association of a Word with a Category

The association measures how well a target word t aligns with a category A. It is computed as the average cosine similarity between t and all elements in A:

$$\text{assoc}(t, A) = \frac{1}{|A|} \sum_{a \in A} \text{sim}(t, a)$$

This determines the strength of the relationship between a word and a semantic category.

Difference in Association Between Two Categories

To compare how a word t aligns with two categories A and B, we calculate the difference in association:

$$\Delta\text{assoc}(t, A, B) = \text{assoc}(t, A) - \text{assoc}(t, B)$$

This is useful for testing if terms from one list (e.g., *flowers/insects*) are more semantically related to another list (e.g., *pleasant/unpleasant*).

Statistical s Formula

The statistic s measures the average difference in associations for two target sets X and Y with two attribute categories A and B:

$$s = \frac{1}{n} \sum_{x \in X} \big(\text{assoc}(x, A) - \text{assoc}(x, B)\big)$$
$$- \frac{1}{n} \sum_{y \in Y} \big(\text{assoc}(y, A) - \text{assoc}(y, B)\big)$$

Where:

- X, Y: Target sets (e.g., *flowers* and *insects*).
- A, B: Attribute categories (e.g., *pleasant* and *unpleasant*).
- n: Number of elements in X or Y.

Effect Size (d) - Cohen's d

The effect size d is a standardized measure of bias magnitude:

$$d = \frac{\mu\big(\Delta\text{assoc}(x, A, B)\,\forall\, x \in X\big) - \mu\big(\Delta\text{assoc}(y, A, B)\,\forall\, y \in Y\big)}{\sigma\big(\Delta\text{assoc}(w, A, B)\,\forall\, w \in X \cup Y\big)}$$

Where:

- μ: Mean.
- σ: Standard deviation.
- Δassoc: Difference in association values.

Magnitude Scale for d:

- Very small: $d < 0.01$
- Small: $d < 0.20$
- Medium: $d < 0.50$
- Large: $d < 0.80$
- Very large: $d < 1.20$
- Huge: $d \geq 2.00$

Bootstrap Confidence Intervals

Bootstrap is a resampling method used to estimate the robustness of results. Confidence intervals (CI) are computed as:

$$CI = [P_{\text{lower}}, P_{\text{upper}}]$$

Where:

- P_{lower}: Lower percentile (e.g., 2.5% for 95% CI).

- P_{upper}: Upper percentile (e.g., 97.5% for 95% CI).

This approach generates synthetic samples to validate the stability of statistical measurements.

Training and Processing Duration

- **Embedding Model Loading:**
  - **FastText:** Approximately 2 minutes to load the `cc.ro.300` model.
  - **mBERT and XLM-R:** Approximately 10–15 seconds to load using the Hugging Face `transformers` library.

- **Processing Time:**
  - Calculating similarity scores and association metrics: Approximately 45 seconds per model.
  - Total runtime for experiments: Around 30 minutes, including pre-processing, analysis, and visualization.

Results

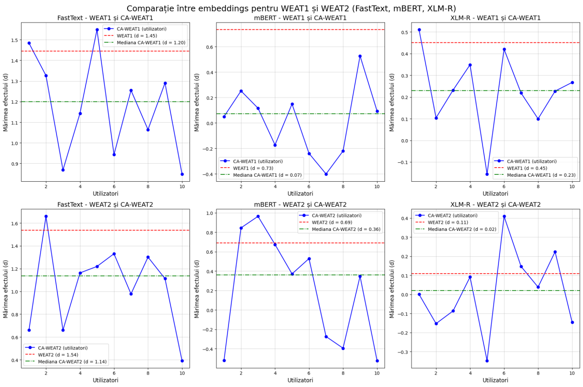

Figure 1: Results visualization showing model performance and bias evaluations.

FastText

- Exhibited the strongest biases, as expected for a monolingual model trained exclusively on Romanian text.

- Universal biases (WEAT) were more pronounced than culturally specific biases (CA-WEAT).

mBERT

- Demonstrated reduced bias strength compared to FastText, supporting the hypothesis that multilingual models attenuate biases.

- While biases were still detectable, they were notably less pronounced than in FastText.

XLM-R

- Showed the weakest biases among the three models, indicating robust bias attenuation across languages.

Key Observations

- **CA-WEAT Variability:** Individual responses introduced greater variability in CA-WEAT compared to WEAT. However, the median results for CA-WEAT aligned closely with those of WEAT.

- **Bias Attenuation:** Multilingual models (mBERT, XLM-R) consistently exhibited smaller bias effect sizes (d) compared to the monolingual model (FastText).

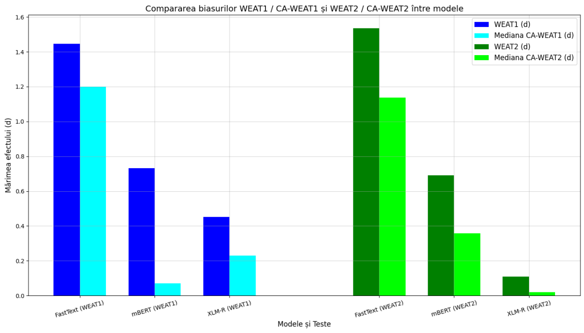

Figure 2: Bias Comparison

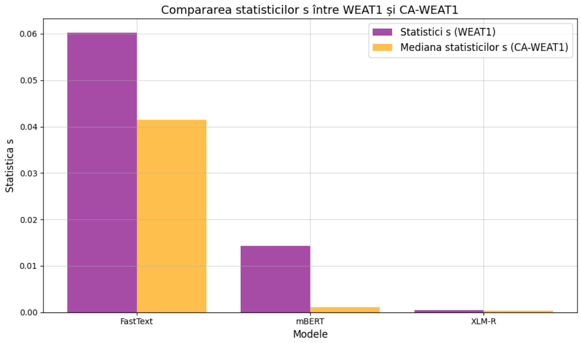

Figure 3: Statistic Visualization

3 Limitations

Translation and Adaptation Difficulties

Translating the original WEAT tests from English to Romanian presented significant challenges due to differences in word meanings and usage:

- **Restricted Lists:** Many English words have a single, unambiguous meaning, but their Romanian equivalents often have multiple meanings. This resulted in a more constrained list of words for the Romanian version of WEAT.

- **Polysemy:** Words like *corn* (which can mean both a musical instrument and a type of food) introduced ambiguity in embeddings, potentially affecting statistical results.

Cultural Nuances in CA-WEAT

The CA-WEAT test revealed unique cultural influences in Romanian responses but also highlighted certain inconsistencies:

- **Ambiguous Terms:** Certain flower names in Romanian, such as *Palma Maicii Domnului* (a culturally significant flower), carry strong cultural connotations but are less universally recognized. While these terms enriched the cultural context, they complicated direct comparisons between WEAT and CA-WEAT results.

Limited Participant Input for CA-WEAT

The CA-WEAT lists were derived from a relatively small group of native Romanian speakers. While these responses captured valuable cultural nuances, the limited sample size may not fully represent the diversity of the Romanian-speaking population.

Corpus Variability

The differences in training corpora for FastText, mBERT, and XLM-R likely influenced the biases observed. As the corpora were not standardized for this study, their varying compositions could have affected the embedding results, adding another layer of complexity to the analysis.

Resource Intensity

Running large-scale embedding evaluations, particularly with contextual models like mBERT and XLM-R, required substantial computational resources. This limitation restricted the scalability of the study and constrained the ability to perform broader experiments on larger datasets.

4 Conclusions and Future Work

Conclusions

This study investigated universal and culturally specific biases in Romanian language embeddings using WEAT and CA-WEAT tests to evaluate monolingual (FastText) and multilingual models (mBERT, XLM-R). The findings revealed that:

- Monolingual embeddings retain stronger biases, reflecting cultural and linguistic nuances of Romanian text.

- Multilingual models attenuate biases through cross-lingual generalization, with XLM-R showing the weakest biases among the evaluated models.

- CA-WEAT provided insights into Romanian-specific cultural associations, emphasizing the richness and diversity of linguistic expression.

While the project successfully demonstrated the importance of cultural adaptations in bias evaluation, incorporating a feedback loop with native speakers and validating culturally specific word lists could have improved the reliability and depth of the results. Overall, the challenges encountered highlighted the complexity of adapting linguistic tests to a specific cultural context and underscored the critical role of language and culture in shaping bias detection methodologies.

Future Work

Future research should focus on the following directions:

- Preserving Positive Cultural Biases: Beyond mitigating harmful biases, models should aim to retain culturally significant attributes that reflect a language's identity, traditions, and heritage, ensuring inclusivity and respect for linguistic diversity.

- Culturally Sensitive Model Development: Future language models should balance multilingual generalization with the ability to capture and respect the unique characteristics of individual languages. This involves training models that are both robust and culturally aware.

- Expanding CA-WEAT: The methodology should be extended to other languages, enabling comparative analyses of cultural nuances in biases across diverse linguistic contexts. Such work would provide a broader understanding of how language-specific connotations influence embedding behavior.

By prioritizing cultural inclusivity and specificity, future work can contribute to the development of AI systems that are fair, unbiased, and reflective of the richness and diversity of human language and culture.

References

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2022. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2056–2077, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Lauscher and Goran Glavaš. 2019. Are we consistently biased? multidimensional analysis of biases in distributional word vectors. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 85–91, Minneapolis, Minnesota. Association for Computational Linguistics.