Do AI models learn human biases?

A Romanian Language Case Study comparing biases reflected by WEAT and CA-WEAT tests

Table of contents

01

Problem Statement

AI models don’t just *mirror* **human biases**—they can also *reshape*, *weaken*, or even *invert* them, especially in multilingual settings.

Problem statement

What is the problem?

AI language models learn from human-generated text, inheriting and reflecting existing biases. For example, if "doctor" appears more frequently with male references and "nurse" with female ones, the model may reinforce these stereotypes in its predictions.

Why does this matter?

AI bias can result in unfair decisions (e.g., hiring, content moderation). While most research focuses on English, lesser-studied languages like Romanian may reveal unique bias patterns.

What we investigate?

We explore how AI models reflect bias beyond social factors, focusing on cultural and linguistic influences. Do monolingual models retain stronger biases? Do multilingual ones reduce them? How do biases appear in Romanian?

02

Related Work & Research Questions

Prior research using WEAT and CA-WEAT tests found that ***monolingual embeddings retain stronger biases***, while **multilingual embeddings tend to attenuate** them.

—Our goal is to examine whether this holds true for Romanian

Hypotheses

Hypothesis 1

Do monolingual embeddings retain stronger biases than multilingual ones?

Hypothesis 2

Does multilinguality attenuate bias, and if so, to what extent?

Hypothesis 3

Can CA-WEAT better capture Romanian-specific cultural biases compared to WEAT?

03

WEAT & CA-WEAT Explained

WEAT

WEAT (Word Embedding Association Test) is a predefined test that measures bias in AI language models by analyzing how different word categories relate to each other.

It consists of **preset word lists** grouped into **target categories** (e.g., "flowers" vs. "insects") and **attribute categories** (e.g., "pleasant" vs. "unpleasant").

The test checks whether a model associates certain words more strongly with one category over another, revealing patterns of bias in word embeddings.

CA-WEAT

CA-WEAT (Culturally Adapted WEAT) builds on WEAT by replacing the predefined word lists with culturally relevant terms specific to a language.

Since direct translations don’t always capture the same associations, CA-WEAT uses **words chosen by native speakers** to reflect cultural context more accurately.

For example, while WEAT might include "aster" for flowers, a Romanian version might use "ghiocel" (spring snowflake), which holds stronger cultural significance.

This allows us to measure biases that are more representative of a specific language and culture.

Aster or Ochiul Boului de Munte

Ghiocel Symbol of Mărțișor

04

Data Collection
Experimental Setup

Data used

WEAT

WEAT word lists were translated into Romanian

CA-WEAT

CA-WEAT lists were created from scratch by asking Romanian native speakers for culturally relevant words

TEAMWORK MAKES THE DREAM WORK, OUR FRIENDS LOVED THIS CHALLENGE!!

fata am innebunit de tot 2:15 PM

melcu e insecta? 2:15 PM

molusca 2:16 PM

am inteles:)))))) 2:16 PM

concept placut: alcool 2:20 PM

🐛 2:13 PM

cum se chema asta 2:13 PM

https://forms.gle/yQrXwp6GARPYfjfS8 1:01 PM

ba stiu ca e un werid request dar credeti ca ati putea sa ma ajutati si pe mine cu niste raspunsuri la formularul asta ASAP 1:01 PM

Mamișor❤️ 9m ago
Am completat formularul, doamne bătaie de cap😁

care are ddl azi 1:02 PM

am reusit 1:47 PM

mi am stors creierul 1:47 PM

te pwppppp 2:36 PM

sa stii ca am pus la lucruri placute bere

😂

nu m am putut abtine 2:36 PM

14:47 68%

Search...

insecta cu coarne

insecta maica domnului
European firebug

insecta care suge sange

melc

miriapod

insecta cu multe picioare

insecta taratoare

pian de suflat

delie
Images

centipede

01

FastText (Monolingual, Static)

Only trained on Romanian. Expected to retain stronger cultural biases

02

mBERT (Multilingual, Contextual)

Trained on 104 language including Romanian. Expected to reduce biases through cross-lingual learning.

03

XLM-R (Multilingual, Contextual)

More advanced multilingual model. Expected to attenuate biases the most.

Embedding Models Tested

05

Statistical Analysis

How did we measure bias?

1. **Cosine Similarity:** Measures how "close" words are in the embedding space.
2. **Association Score:** Determines how much a target word (e.g., "flower") is associated with a list of attributes (e.g., "pleasant").
3. **Statistical s Score:** Compares association scores across categories.
4. **Effect Size d (Cohen’s ddd)**:
    a. Measures how strong the bias is.
    b. Scale:
        - **$d<0.2$** → Small bias
        - **$d>0.8$** → Large bias

06

Results

Results

Findings by model

FastText (Monolingual)

Exhibited the **strongest biases**, confirming that **monolingual embeddings** preserve cultural biases.

mBERT (Multilingual)

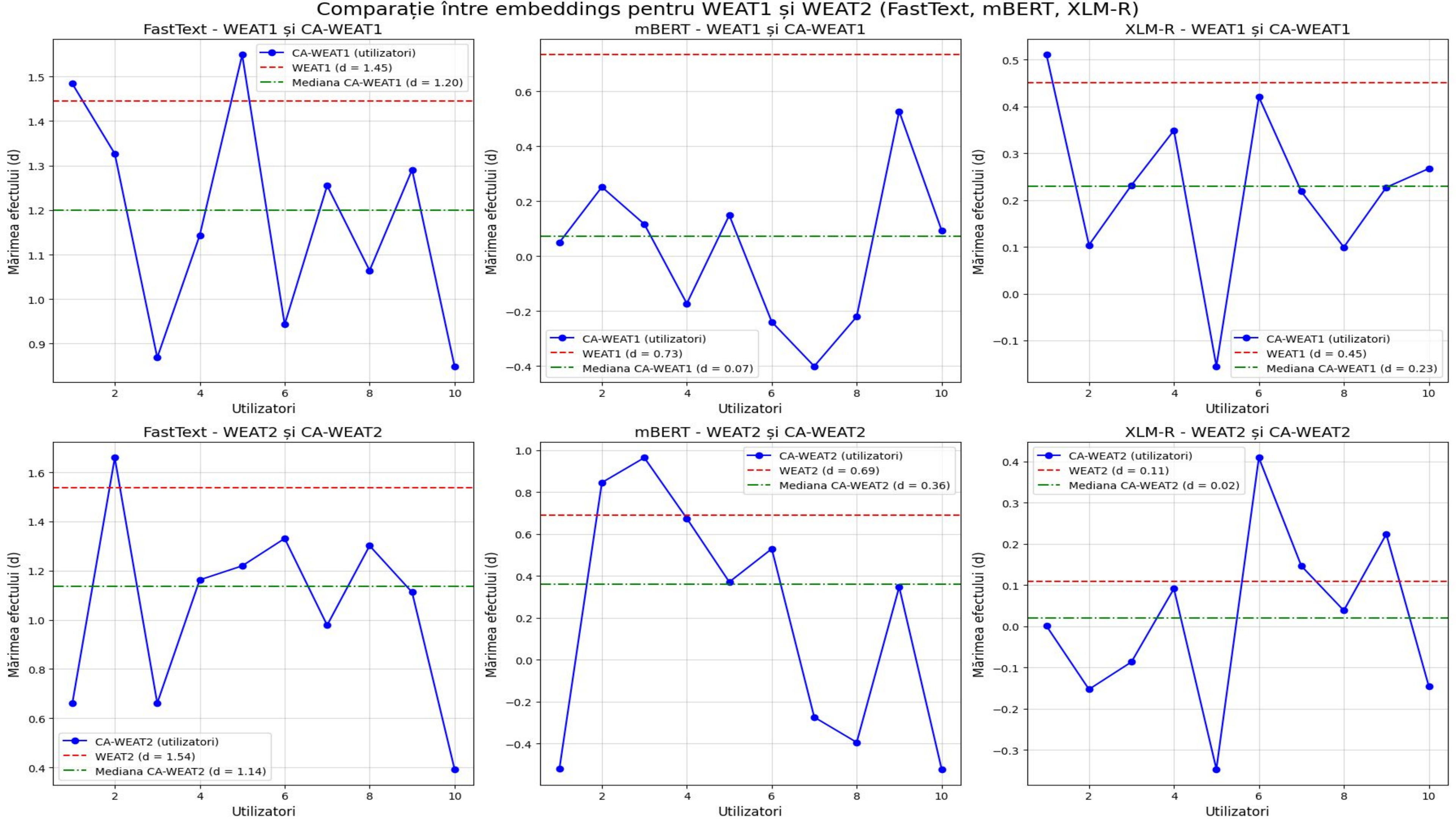
Bias was reduced compared to FastText. Bias was still detectable, suggesting multilinguality helps but does not eliminate bias.

XLM-R (Multilingual)

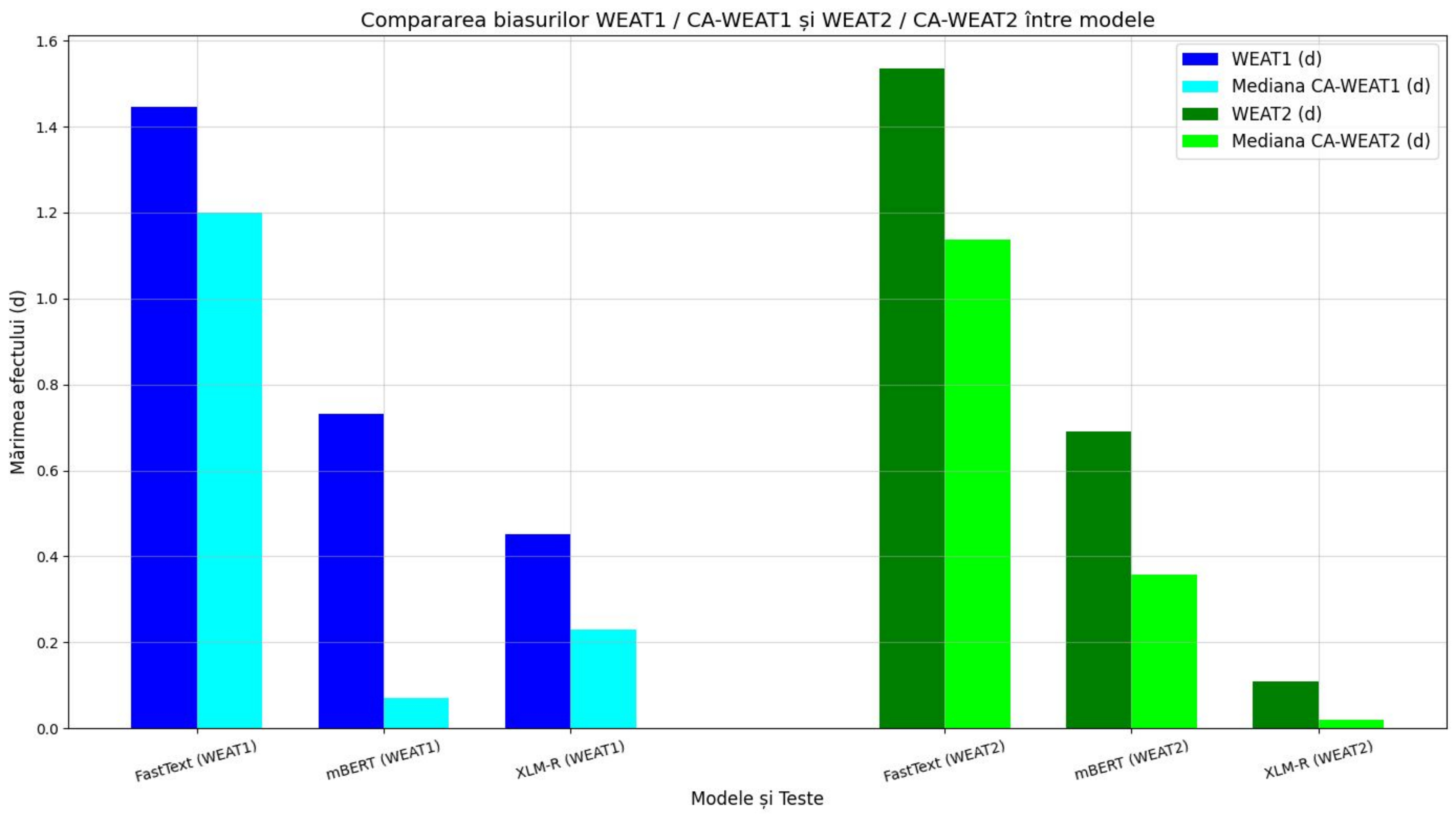
Bias was weakest, supporting the idea that cross-lingual learning helps attenuate bias

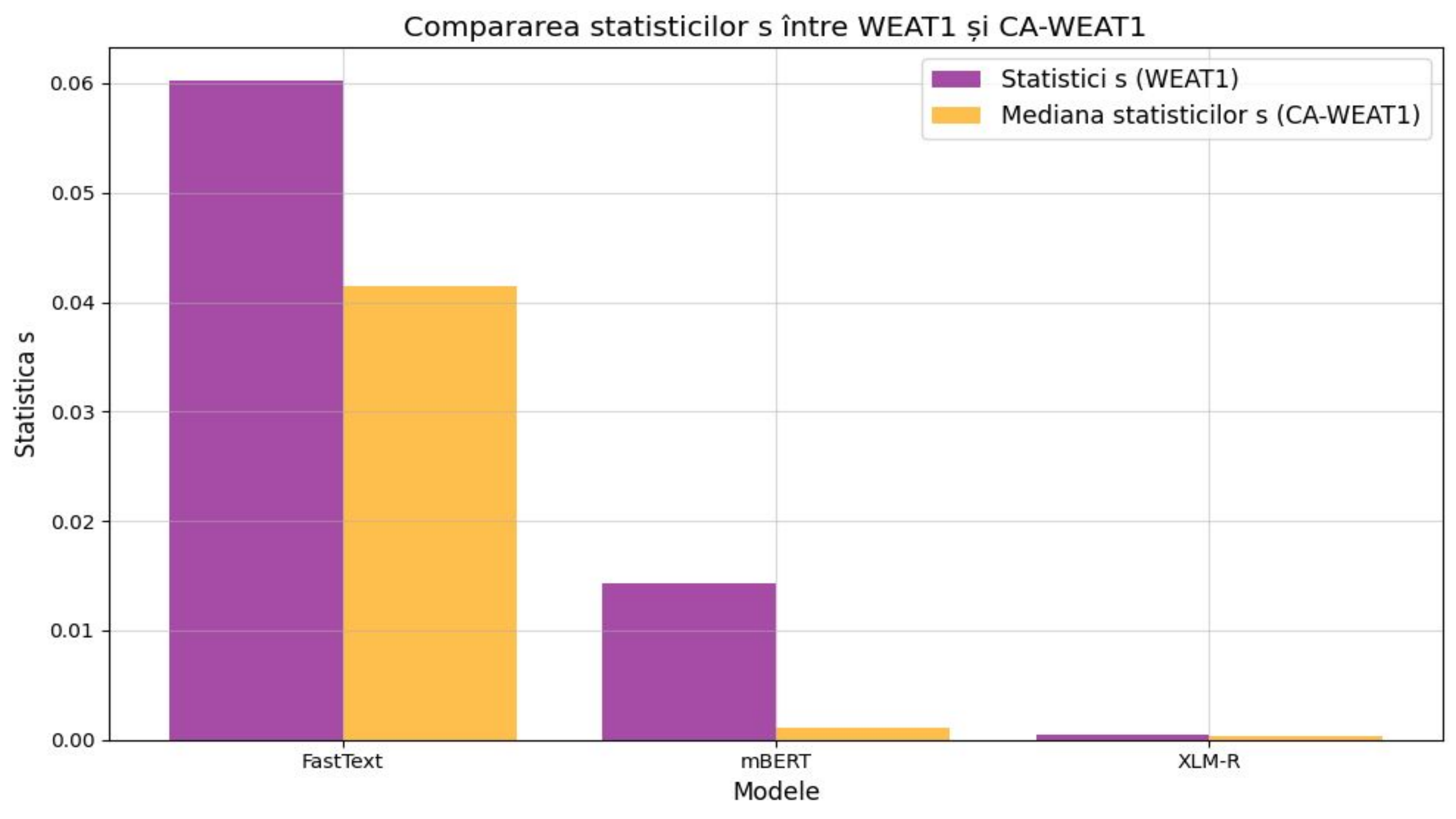
CA-WEAT vs WEAT

CA-WEAT revealed cultural nuances not captured by WEAT.



Comparație între embeddings pentru WEAT1 și WEAT2 (FastText, mBERT, XLM-R)
FastText - WEAT1 și CA-WEAT1
mBERT - WEAT1 și CA-WEAT1
XLM-R - WEAT1 și CA-WEAT1
FastText - WEAT2 și CA-WEAT2
mBERT - WEAT2 și CA-WEAT2
XLM-R - WEAT2 și CA-WEAT2
CA-WEAT1 (utilizatori)
WEAT1 (d = 1.45)
Mediana CA-WEAT1 (d = 1.20)
WEAT1 (d = 0.73)
Mediana CA-WEAT1 (d = 0.07)
WEAT1 (d = 0.45)
Mediana CA-WEAT1 (d = 0.23)
CA-WEAT2 (utilizatori)
WEAT2 (d = 1.54)
Mediana CA-WEAT2 (d = 1.14)
WEAT2 (d = 0.69)
Mediana CA-WEAT2 (d = 0.36)
WEAT2 (d = 0.11)
Mediana CA-WEAT2 (d = 0.02)
Mărimea efectului (d)
Utilizatori


Compararea biasurilor WEAT1 / CA-WEAT1 și WEAT2 / CA-WEAT2 între modele
WEAT1 (d)
Mediana CA-WEAT1 (d)
WEAT2 (d)
Mediana CA-WEAT2 (d)
Mărimea efectului (d)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
FastText (WEAT1)
mBERT (WEAT1)
XLM-R (WEAT1)
FastText (WEAT2)
mBERT (WEAT2)
XLM-R (WEAT2)
Modele și Teste



Compararea statisticilor s între WEAT1 și CA-WEAT1
Statistici s (WEAT1)
Mediana statisticilor s (CA-WEAT1)
Statistica s
0.06
0.05
0.04
0.03
0.02
0.01
0.00
FastText
mBERT
XLM-R
Modele