Precog Recruitment Task

Language Representations — Bonus Task: Harmful Associations

Nilanjan Sarkar (MS by Research, CSE) — ID 2025701014

Abstract

This report evaluates harmful associations in both *static* and *contextual* representations. For static embeddings, we use pre-trained fastText and quantify associations with the **Word Embedding Association Test (WEAT)**. We further project occupations on a learned **gender direction** to visualize male/female bias. For contextual models, we analyze DistilBERT using the **Sentence Encoder Association Test (SEAT)** and a minimal-pair evaluation on the **CrowS-Pairs** dataset . Across settings, we find sizable male→science bias and evidence of race→valence bias, and we observe that the contextual model prefers stereotypical sentences more often than their anti-stereotypical counterparts on a held-out subset.

1 Data and Models Used (as implemented)

Static embedding. Pre-trained English **fastText** vectors were loaded. Nearest neighbors for probe words (e.g., *nurse*, *doctor*, *engineer*, *receptionist*, *king*, *queen*) showed linguistically plausible neighbors and some gendered associations (e.g., *receptionist* → *waitress* among neighbors).

Contextual model. DistilBERT (a distilled variant of Bidirectional Encoder Representations from Transformers) was used in two ways: (i) sentence embeddings for SEAT via simple templates ("This is {word}.", "That is {word}.") with mean pooling, and (ii) masked-language-model scoring for CrowS-Pairs using **Pseudo Log-Likelihood (PLL)**.

2 Results

2.1 Static embeddings (fastText)

Nearest neighbors (qualitative). Examples indicated plausible semantic neighborhoods and some gendered associations (e.g., *receptionist* with *waitress*).

WEAT. Using standard wordlists:

Test	Effect size d
Gender–Career/Family	+0.386
Gender–Science/Arts	+1.602
Race–Valence	+0.809

The strongest association observed was male→science and female→arts; race→valence bias was also significant in this run. Gender career/family showed a weaker, non-significant effect.

Occupation projections. Projecting occupations onto the learned gender axis yielded a ranked list with intuitive tendencies (female-coded vs. male-coded professions), consistent with WEAT trends.

2.2 Contextual model (DistilBERT)

SEAT (template sentences, mean-pooled embeddings).

Test	Effect size d	p-value
Gender–Career/Family	+0.658	
Gender–Science/Arts	+0.549	
Race–Valence	+0.717	

In this configuration, only Race–Valence reached conventional significance. Template choice and pooling are known to affect sensitivity, so we interpret these magnitudes directionally.

CrowS-Pairs (PLL) on a 50-pair sample. Overall stereotypical preference: **64.0%** of pairs were scored higher for the stereotypical sentence than the anti-stereotypical alternative. For n=50, a 95% confidence interval (Wilson) is [**50.1%, 75.9%**]. Per-category rates (as plotted in the notebook) varied, but the aggregate preference aligns with the dataset's intent as a bias stress test.

3 How the contextual evaluation differs from static

- **Unit of analysis.** Static embeddings map each word to a single vector, so WEAT uses word-level cosine similarities. Contextual models produce token/sentence representations conditioned on surrounding words; SEAT and CrowS-Pairs operate on *sentences*.
- **Scoring.** Static WEAT uses cosine-based association scores; contextual PLL compares model likelihoods of minimally-different sentences.
- **Sensitivity to phrasing.** Contextual evaluations depend on templates (SEAT) and surface form (CrowS-Pairs); small wording changes can affect results, which is both a feature (captures context) and a caveat (introduces variance).

4 Discussion and caveats

Overall, the static fastText embedding exhibits strong male→science and race→valence effects. DistilBERT shows a significant race→valence signal under SEAT and prefers stereotypes in CrowS-Pairs on a random 50-pair subset. Two implementation choices likely dampen some contextual signals: (i) mean pooling for sentence vectors and (ii) a small number of simple templates. Nonetheless, the qualitative direction is consistent across settings.

Limitations. Results may vary with different wordlists (WEAT), template families (SEAT), PLL variants, or larger samples from CrowS-Pairs.