Intersectional Bias in Word Embeddings

Morgan Wajda-Levie & Rebecca Kleinbart
Hunter College Computer Science Department
CSCI 795 ~ Professor Raja

CONTEXT

Word embeddings are **very commonly used** and **powerful**.

Word embeddings are **biased**, hurting marginalized and less powerful groups more.

Approaches to debiasing word embeddings focus mostly on gender.

PROBLEM STATEMENT AND GOALS

1. Can we *measure* **non-gender bias** of word embeddings?
2. How does the *bias metric change* when considering **intersectional bias**?
3. Stretch question: Do particularly biasing texts share features?
4. Stretch question: Can we successful *predict* which texts will contribute significantly to the bias of word embeddings?

OUR APPROACH

1. Attempt to de- bias from the source (training data) before GloVe word embedding algorithm is run.
2. Use **WEAT test** to find **bias differential**.

$$g(c, \mathcal{A}, \mathcal{B}, w) = \underset{a \in \mathcal{A}}{\text{mean}} \cos(w_c, w_a) - \underset{b \in \mathcal{B}}{\text{mean}} \cos(w_c, w_b)$$

$$B_{weat}(w) = \frac{mean_{s \in \mathcal{S}}\, g(s, \mathcal{A}, \mathcal{B}, w) - mean_{t \in \mathcal{T}}\, g(t, \mathcal{A}, \mathcal{B}, w)}{stddev_{c \in \mathcal{S} \cup \mathcal{T}}\, g(c, \mathcal{A}, \mathcal{B}, w)}$$

3. Use **perturbation algorithm** measure change in bias if a particular text is removed.

$$\Delta_p B = B(w) - B(\tilde{w})$$

Algorithm 1 Approximating Differential Bias

input *Co-occ Matrix*: X, *WEAT words*: $\{\mathcal{S}, \mathcal{T}, \mathcal{A}, \mathcal{B}\}$
  w^*, u^*, b^*, c^* = GloVe(X) *# Train embedding*
  for doc **in** corpus **do**
    $\tilde{X} = X - X^{(k)}$ *# Subtract coocs from doc k*
    for word i **in** doc $\cap (\mathcal{S} \cup \mathcal{T} \cup \mathcal{A} \cup \mathcal{B})$ **do**
      # Only need change in WEAT word vectors
      $\tilde{w}_i = w_i^* - \frac{1}{V} H_{w_i}^{-1} [\nabla_{w_i} L(\tilde{X}_i, w) - \nabla_{w_i} L(X_i, w)]$
    end for
    $\Delta_{\text{doc}} B \approx B_{\text{weat}}(w^*) - B_{\text{weat}}(\tilde{w})$
  end for

DATA SETS

1. Simple English Wikipedia
2. New York Times Annotated Corpus
3. Reduced New Yor Times Annotated Corpus

BIAS METRIC

WEAT tests substantiated through **Implicit Association tests.**

```
},
"attr1": {
    "category": "LikableNotHostile",
    "vocab": [
        "agreeable",
        "fair",
        "honest",
        "trustworthy",
        "selfless",
        "accommodating",
        "likable",
        "liked"
    ]
},
"attr2": {
    "category": "UnlikableHostile",
    "vocab": [
        "abrasive",
        "conniving",
        "manipulative",
        "dishonest",
        "selfish",
        "pushy",
        "unlikable",
        "unliked"
    ]
}
```

RESULTS (EXCERPTED)

Table 6: comparison of full and reduced NY Times corpus

index	delta_effect_sizes	delta_p_values	effect_sizes_ratio	p_values_ratio
WEAT A	0.108332	0.0137	0.102289	1.096000
WEAT B	0.193853	0.0000	0.128888	0.000000
WEAT C	0.297318	0.0084	0.285946	0.988235
WEAT D	0.383627	0.0021	0.332541	1.000000
WEAT E	0.370220	0.0067	0.334063	1.000000
WEAT F	0.183656	0.0000	0.135122	0.000000
WEAT FF	0.361930	0.0059	0.387186	1.000000
WEAT H	0.109796	0.0001	0.063377	1.000000
WEAT I	0.021619	0.0000	0.017226	0.000000
WEAT J	0.082787	0.0014	0.061180	0.823529
WEAT K	0.241868	0.0000	0.157037	0.000000
WEAT L	0.587134	0.2841	0.718016	5.727823
WEAT M	0.180225	0.0127	0.151617	0.808917
WEAT N	0.149681	0.0000	0.105932	0.000000

99th percentile accounts for an average of 48.8% of cumulativeestimated bias among our results. The 90th percentile accounts for an average of 92.2%.

Some (greater than 10%) correlation between WEAT sets in the same general area of bias (anti-Blackness and gender bias), and slight roughly 5%) correlation across some bias areas.

CONCLUSIONS

A relatively small number of documents contribute most to the overall bias.

Gender and racial bias are not strongly correlated and attempts to de-bias word embeddings should address these separately in order to be maximally effective

NEXT STEPS

> Analyzing different corpora in order to compare and evaluate bias in more recent and older texts.

> Most-biasing documents analysis in order to learn about bias and train a classifier

HUNTER
The City University of New York