Towards Debiasing Fact Verification Models

Tal Schuster*, Darsh J Shah*, Serene Yeo,
Daniel Filizzola, Enrico Santus, Regina Barzilay
Task: Fact verification

Is the following claim true?

Claim:

Trevor Griffiths was born on April 4, 1935
Task: Fact verification

Bert can answer correctly (using only the claim)

Claim:
Trevor Griffiths was born on April 4, 1935
Task: Fact verification

Shouldn’t it rely on evidence?

Claim:
Trevor Griffiths was born on April 4, 1935

Evidence:
Trevor Griffiths (born 4 April 1935), is an English dramatist.

Bert
Performance on the FEVER dataset

![Bar chart showing performance on the FEVER dataset]

- Evidence aware: 69.7
- Claim only: 61.7
- Majority baseline: (33.3)
Why does the claim-only model perform so well?
Possibility 1

**World knowledge** captured in the pretraining process

1. pretraining

   *Wikipedia*

2. fact-checking

   Claim based on Wikipedia

   can leak the truthfulness of claims
Without pretrained embeddings

Random Embeddings still perform far above the baseline

GloVe embeddings: 57.3
Random embeddings: 54.1

Majority baseline (33.3)

24 gap
21 gap

claim-only InferSent model (Conneau et al., Poliak et al.)
Possibility 2

**Give-away phrases in the claims**
Give-away phrases in the claims

A claim-only model should fail:

• Magic Johnson did not play for the Lakers.

• There has been at least one windstorm in Stanley Park.

• All About Eve won an award for Best Picture.

• The New England Patriots failed to reach seven Super Bowls.

• Quinoa did not originate in South America.
Give-away phrases in the claims

But the model can "cheat":

- Magic Johnson **did not** play for the Lakers.
- There has been **at least one** windstorm in Stanley Park.
- All About Eve **won an award** for Best Picture.
- The New England Patriots **failed to** reach seven Super Bowls.
- Quinoa **did not** originate in South America.
Give-away phrases in the claims

- Probability of a claim having label $l$ if it contains phrase $w_j$:

$$ p(l|w_j) = \frac{\text{count}(l, w_j)}{\text{count}(w_j)} = \frac{\sum_{i=1}^{n} \mathbb{1}_{[w_j(i) \cdot \mathbb{1}_{[y(i)=l]}]} \cdot \sum_{i=1}^{n} \mathbb{1}_{[w_j(i)]}}{\sum_{i=1}^{n} \mathbb{1}_{[w_j(i)]}} $$

- Local Mutual Information - phrases that create the strongest bias:

$$ LMI(w_j, l) = p(w_j, l) \cdot \log \left( \frac{p(l|w_j)}{p(l)} \right) $$

Towards Debiasing Fact Verification Models
Example $p(l|w_j)$ for top LMI phrases

- **“did not”**
  - Supports: 83%
  - Refutes: 11%
  - Not enough info: 6%

- **“at least one”**
  - Supports: 90%
  - Refutes: 8%
  - Not enough info: 2%
Bias reappears in the evaluation set

Train:

“did not”

Supports 83%
Refutes 11%
Not enough info 6%

“at least one”

Supports 90%
Refutes 2%
Not enough info 8%

Evaluation:

Supports 5%
Refutes 5%
Not enough info 90%

Supports 90%
Refutes 21%
Not enough info 2%
$p(l|w_j)$ on evaluation set for top LMI phrases by training set

$p(\text{👍}|w_j)$

$p(\text{👎}|w_j)$

"did not" "yet to" "does not" "refused to" "failed to"

"least one" "at least" "person who" "stars actor" "won award"

Towards Debiasing Fact Verification Models
Give-away phrases are the main culprits

1. World knowledge captured in pretraining

2. Give-away phrases in claims
Goal: Evidence based predictions
Towards Debiasing Fact Verification Models

Claim-only should be inadequate

A claim’s truthfulness might:

- Vary with context/ perspective
  - Ross cheated on Rachel
    - They were not on a break
    - They were on a break

- Change over time
  - Bert is SOTA
    - 2019
    - 2020
Debiasing the evaluation

By creating a symmetric dataset
Creating a symmetric unbiased dataset

Magic Johnson \textcolor{red}{did not} play for the Lakers

\textcolor{red}{\leftarrow}\quad \text{thumbs down}\quad \rightarrow \quad He played point guard for the Lakers for 13 seasons
Creating a symmetric unbiased dataset

Magic Johnson **did not** play for the Lakers

He played point guard for the Lakers for 13 seasons

He spent his entire career playing for the Giants

Magic Johnson **played for** the Lakers
Creating a symmetric unbiased dataset

Magic Johnson did not play for the Lakers

He played point guard for the Lakers for 13 seasons

Magic Johnson played for the Lakers

He spent his entire career playing for the Giants
The evidence is crucial for predictions

\[ p(l|w_j) \]
Performance on the symmetric dataset

Entailment results using Bert

<table>
<thead>
<tr>
<th></th>
<th>85.9</th>
<th>57.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fever</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Symmetric</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Debiasing the training

By reweighting the training samples
Regularizing the training

• Defining the bias of phrase $j$ towards label $l$:

$$b_j^l = p(l|w_j) = \frac{\sum_{i=1}^{n} \mathbb{1}_{[w_j^{(i)}]} \cdot \mathbb{1}_{[y^{(i)}=l]}}{\sum_{i=1}^{n} \mathbb{1}_{[w_j^{(i)}]}}$$

• Setting weights ($\alpha$) for each training sample:

$$b_j^l = \frac{\sum_{i=1}^{n} \mathbb{1}_{[w_j^{(i)}]} \cdot \mathbb{1}_{[y^{(i)}=l]} \cdot (1 + \alpha^{(i)})}{\sum_{i=1}^{n} \mathbb{1}_{[w_j^{(i)}]} \cdot (1 + \alpha^{(i)})}$$
Regularizing the training

• Learning the weights by optimizing:

$$\min \left( \sum_{j=1}^{V} \max_{l} (b_j^l) + \lambda \| \tilde{\alpha} \|_2 \right)$$

• **Re-weighted** loss function:

$$\sum_{i=1}^{n} (1 + \alpha^{(i)}) \cdot \mathcal{L} (x^{(i)}, y^{(i)})$$
Statistical cues are alleviated

Original:

“did not”

- Supports: 83%
- Refutes: 11%
- Not enough info: 6%

“at least one”

- Supports: 90%
- Refutes: 2%
- Not enough info: 8%

Re-weighted:

- Supports: 37%
- Refutes: 29%
- Not enough info: 36%

- Supports: 37%
- Refutes: 35%
- Not enough info: 26%
Performance on symmetric dataset

Towards Debiasing Fact Verification Models
Towards Debiasing Fact Verification Models

• Bias in FEVER
  • give-away phrases in the claims

• Symmetric dataset

• Alleviating the bias
  • Re-weighting the training samples:
    \[ \min \left( \sum_j \max (b_j^l) \right) \]

Code and data:
https://github.com/TalSchuster/FeverSymmetric

@TalSchuster
@darshj_shah