On the Automatic Detection and Classification of Linguistic Bias

Richard Diehl Martinez, Sabri Eyuboglu

West Coast NLP 2019

Dataset

Wikipedia Neutrality Corpus

50,000 biased sentences

Words responsible for bias are tagged

Schnabel himself did a **fantastic** reproduction of Basquiat's work.

No bias type labels!

Hand-labeled Data

500 biased sentences labeled as framing or epistemological

Schnabel himself did a **fantastic** reproduction of Basquiat's work. → **Framing**

Lower cost of living means a **possibly** higher standard of living → **Epistemoglical**

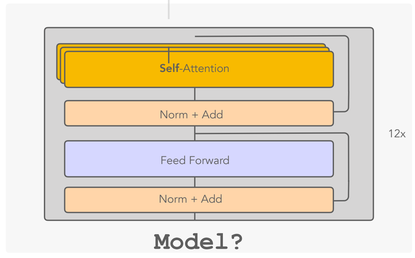
Challenge

✓ **Wikipedia Neutrality Corpus (WNC)** $n = 50{,}000$

✓ **Bias Detection Model** BERT Trained on WNC

Schnabel himself did a **fantastic** reproduction of Basquiat's work.

Biased Sentence



Model?

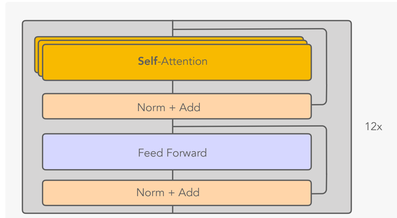
✗ **Epistemological Bias** **Framing Bias**

Principal Challenge: LACK OF LABELS

Training on Gold Labels

Bias Detection Model

Lower cost of living means a **possibly** higher standard of living



Lower cost of living means a possibly higher standard of living

BERT out of Box

Accuracy 0.573

AUROC 0.618

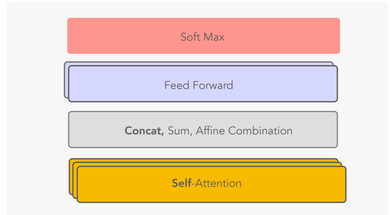
BERT

Accuracy 0.606

AUROC 0.608

Bias Classification Model

Epistemological Framing



Lower cost of living means a **possibly** higher standard of living

Attention

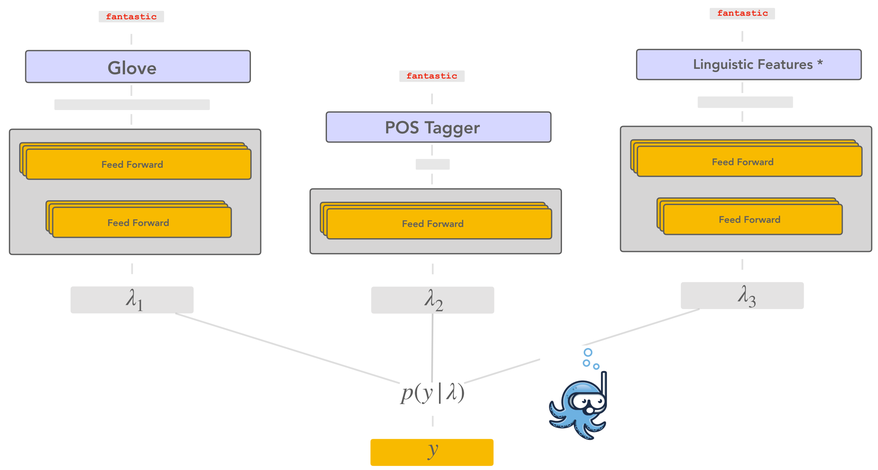
Accuracy 0.679

AUROC 0.756

Training on Weak Labels

Wikipedia Neutrality Corpus

50,000 unlabeled sentences



Wikipedia Neutrality Corpus

50,000 weakly labeled sentences

BERT

Accuracy 0.742

AUROC 0.847

Attention

Accuracy 0.704

AUROC 0.770