Mitigating biases in fact-checking models

Joshua Cowan

Supervised by Andreas Vlachos

Background

Regular Article

Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation

Zapan Barua [a], Sajib Barua [a], Salma Aktar [a], Najma Kabir [a], Mingze Li [b]

Show more

Outline | + Add to Mendeley | Share | Cite

https://doi.org/10.1016/j.pdisas.2020.100119 Get rights and content

The Washington Post
Democracy Dies in Darkness

FACT CHECKER

How misinformation on WhatsApp led to a mob killing in India

Analysis by Elyse Samuels
Video authentication editor

February 21, 2020 at 3:00 a.m. EST

The Washington Post
Democracy Dies in Darkness

FACT CHECKER

How misinformation helped spark an attempted coup in Gabon

Analysis by Sarah Cahlan
Video reporter

February 13, 2020 at 3:00 a.m. EST

DIRECTORATE-GENERAL FOR EXTERNAL POLICIES
POLICY DEPARTMENT

European Parliament

STUDY

The impact of disinformation on democratic processes and human rights in the world

ABSTRACT

Around the world, disinformation is spreading and becoming a more complex phenomenon based on emerging techniques of deception. Disinformation undermines human rights and many elements of good quality democracy; but counter-disinformation measures can also have a prejudicial impact on human rights and democracy. COVID-19 compounds both these dynamics and has unleashed more

A solution?

- Fact checking
  - *The assessment of the truthfulness of a claim*
    (Thorne & Vlachos 2018)
- Need it done quickly
  - Diffuses over time
    (Karlova & Fisher 2013)
- Automated approaches?

FEVER (Fact extraction and verification)

- Shared tasks and dataset.
- Dataset: Collection of claims from modified Wikipedia factoids, with the original corresponding evidence.

Claim: Scandinavia does not include Greenland.
Evidence: The remote Norwegian islands of Svalbard and Jan Mayen are usually not seen as a part of Scandinavia, nor is Greenland, an overseas territory of Denmark.
Ground truth: Supports

- First shared task: Use claim to both find evidence and derive a classification judgement.
- I focused on NLI-style verification only

Biases in datasets

Artefacts in the form of superficial patterns whose exploitation facilitates impressive in-domain performance, but poor generalisation.

Claim: Scandinavia does not include Greenland.

Evidence: The remote Norwegian islands of Svalbard and Jan Mayen are usually not seen as a part of Scandinavia, nor is Greenland, an overseas territory of Denmark.

Ground truth: Supports

Hypothesis-only bias

Why the need to debias?

- Out-of-domain performance is critical here:
  Typical human speech is stylistically distinct from Wikipedia claims/evidence.

- Must engage with the actual *semantics* of the claim/hypothesis.

- Room for nefarious manipulation through pattern exploitation.

Adversarial datasets: Potency and Resilience

$$\text{Potency}(a) \stackrel{\text{def}}{=} c_a \frac{1}{|S|} \sum_{s \in S} \left(1 - f\left(\hat{Y}_{s,a}, Y_a\right)\right)$$

$$\text{Resilience}(s) \stackrel{\text{def}}{=} \frac{\sum_{a \in A} c_a \times f\left(\hat{Y}_{s,a}, Y_a\right)}{\sum_{a \in A} c_a}$$

Figure 2.2: Potency and Resilience definitions. c_a is the Correctness rate of the adversary a. S is the set of systems, with $\hat{Y}_{s,a}$ the set of the system s's prediction for the adversary a, with Y_a the ground truths and f the scoring metric used. For resilience, A is the set of adversaries.

Thorne et al. 2019

Schuster et al. 2021

Dataset-targeting debiasing: **VitaminC**

- Dataset designed for FEVER models to be trained on.
- VitaminC uses contrastive claims to induce contextual sensitivity.

> Evidence: There are 5000 confirmed cases of coronavirus in the US.
> Annotator-written claim: There are more/less than 4500 confirmed cases of coronavirus in the US.

- **Dataset targeting requires knowledge of biases in each dataset.**
- Does appear to reduce hypothesis-only bias, but they don't evaluate on any other types.
- Overuse of 'more than'/'less than' could induce new biases.

Utama et al. 2020b

Model-targeting debiasing: **Self-debias**

- Can instead detect biased examples automatically.
- Use a shallow model to act as a rapid surface learner -> overfit to shallow patterns.
- Model then predicts training set – probabilities for each class indicate (over)confidence.
- Then downweigh examples the shallow/biased model is confidently correct on.

$$\text{ExampleReweighting}(\theta_d) = -\left(1 - p_b^{(i,c)}\right) y^{(i)} \cdot \log p_d$$

$$\text{ProductofExpert}\,(\theta_d) = -y^{(i)} \cdot \log \text{softmax}\,(\log p_d + \log p_b)$$

$$\text{ConfidenceRegularisation}(\theta_d) = -\text{S}\left(p_t, p_b^{(i,c)}\right) \cdot \log p_d$$

Utama et al. 2020b

Model-targeting debiasing (2)

- All three methods delivered 2-3pp improvement on fever-symmetric, a hypothesis-only adversary.

- Also ran experiments with *annealing*, an attempt to prevent the shallow model downweighing everything due to dealing with many biases at once.
  - Had minimal impact on performance vs non-annealed results, however.

Room for improvement

- Little investigation into *other* biases in FEVER.
  - Needed for robust fact-checking systems
  - VitaminC may even introduce its own biases
- Do previous results extend to these new biases?
  - And to different, better-performing, architectures?
- Could VitaminC and self-debiases methods be combined?
  - Do they combine constructively?

Adversaries designed

Negation overlap: Only one of claim/ev is negated

- Inspired by Naik (2018).
- Implemented with spaCy + inflector
- Wordnet, Conceptnet and PPDB for (unsuccessful) antonym/alternative word resolution.

Evidence: In 2014 , she signed her first recording contract with Astralwerks and released her debut EP , titled Room 93 .

Original Claim: Halsey signed her first recording contract in 2014.

Negation-only variant Claim: Halsey did not sign her first recording contract in 2014.

Negation and Antonym variant Claim: Halsey did not sign her second recording contract in 2014.

Entity overweighting

- VitaminC *refutation* pairs often due to value of a single entity varying.
  - *Entailment* pairs often due to high entity overlap (as seen earlier).

- Adversaries are these cases, but a Not Enough Info label.
- T5 model takes named entities, generates new claim from them.

VitaminC test set claim: "The case has been going on since 2016".
Evidence for pair 1 (Ground = Refutes): "This suit sought $ 210 million in damages and was ongoing as of **2004** ."
Evidence for pair 2 (Ground = Supports): "This suit sought $ 210 million in damages and was ongoing as of **2016** .

New claim generated from pair-2 evidence: "210 million dollars were spent on the project in 2016."

Numerical mismatch

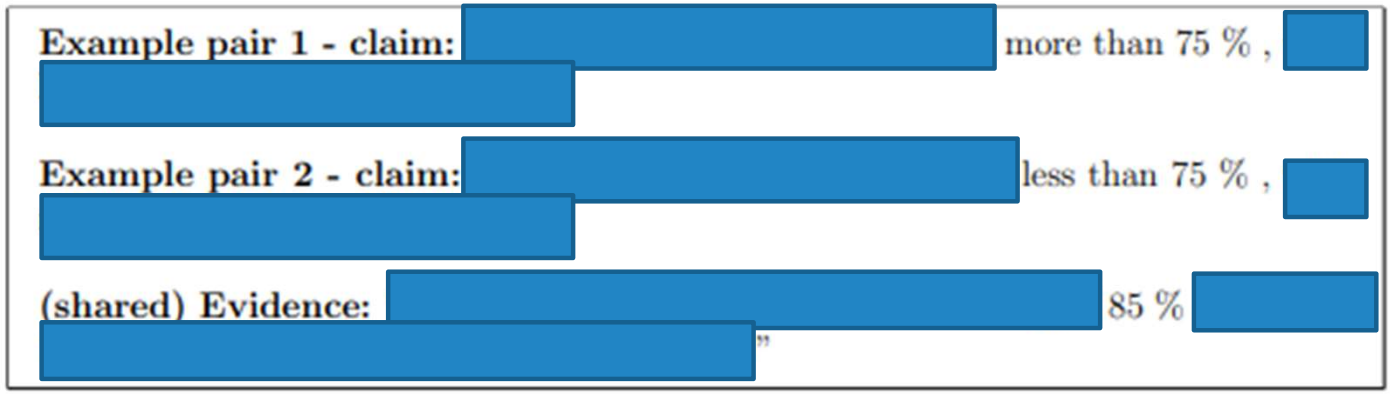
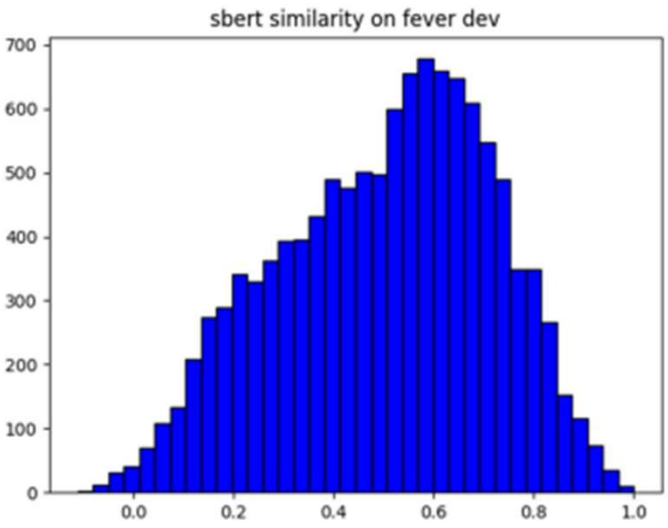


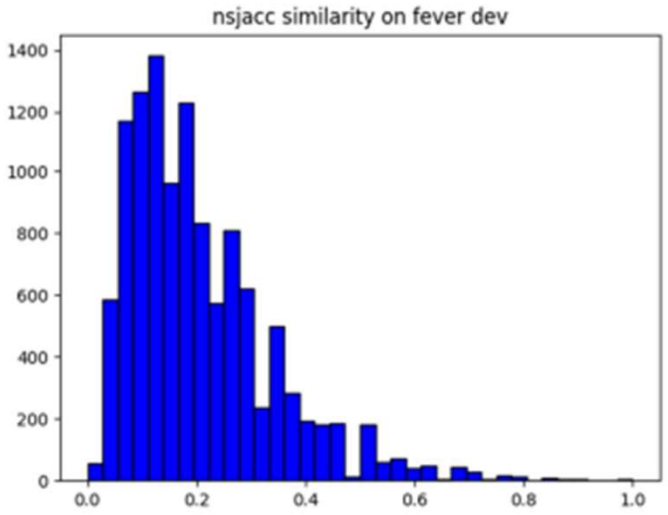
Figure 4.3: Two VitaminC examples whose claims differ only by their use of more and less than.

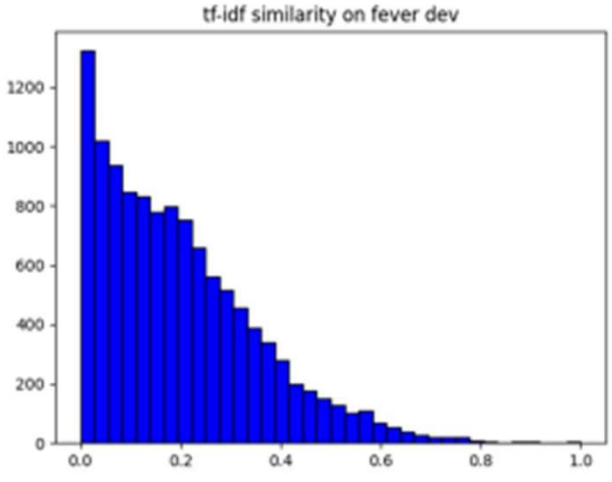
- VitaminC heavy with numbers/figures + annotators overused more/less than
- Adversaries: Cases where ranges overlap but relation is NEI/Refutes, or where they *don't* overlap but relation is supports

Textual similarity

- Inspired by Naik (2018) – *word overlap*
- Biased models may predict ‘Supports’ if high word overlap between claim and hypothesis (or ‘refutes’/’nei’ if low).







Experiments & Results

Experiments to be run

Architecture: BERT-base-uncased, ALBERT-base.

Training-sets: Fever, VitaminC only, Both datasets.

Loss-function: Plain (Standard cross-entropy loss), Example Reweighting, PoE, Confidence regularization.

For non-plain loss-functions, the following further parameterisation is possible:

Use of Annealing: True / False

Bias score source: Self-debiasing (shallow model), or a-priori/known.

Research Question A: How well do contrastive evidence-pairs inoculate fact-checking systems against different biases?

A1. Evaluating the adversarial Potency of VitaminC.

A2. Evaluating if training on vitaminC can inoculate systems against other types of bias – following the framework of Thorne et al [44].

A2.1. Finding other biases, guided by linguistic analysis and examples in [29, 44].

A2.2. Creating Potent adversarial test sets, and determining the Resilience of FEVER-only trained versus vitC+FEVER trained systems to them.

- There **are** other biases in FEVER – performance on newly derived adversarial datasets is below that of their in-domain sources
- Training models on contrastive evidence-pairs does enhance performance on non-hypothesis-only biases – higher Resilience in VitaminC+FEVER models than those trained on just one of the two.

A2.2. Creating Potent adversarial test sets, and determining the Resilience of FEVER-only trained versus vitC+FEVER trained systems to them.

Model: Vitc-albert, Metric: Accuracy		Training Set			Potency	
Correctness	Adversary Name	Both	Fever-only	Vitc-only	Raw	Standard
0.64	Entity Overweighting	0.148	0.127	0.077	0.883	0.565
1	Word overlap (Jaccard)	0.923	0.887	0.831	0.120	
1	Word overlap (Sbert)	0.920	0.883	0.741	0.152	
1	Word overlap (Tf-idf)	0.908	0.908	0.777	0.136	
1	Negation Overlap (Manual, Neg+Ant)	0.742	0.672	0.766	0.273	
0.9	Negation Overlap (Semi-auto, Neg+Ant)	0.317	0.635	0.622	0.475	0.428
0.78	Negation Overlap (Semi-auto, Neg only)	0.952	0.950	0.817	0.094	0.073
1	Numerical Mismatch	0.871	0.375	0.880	0.291	
Resilience		0.689	0.627	0.698		

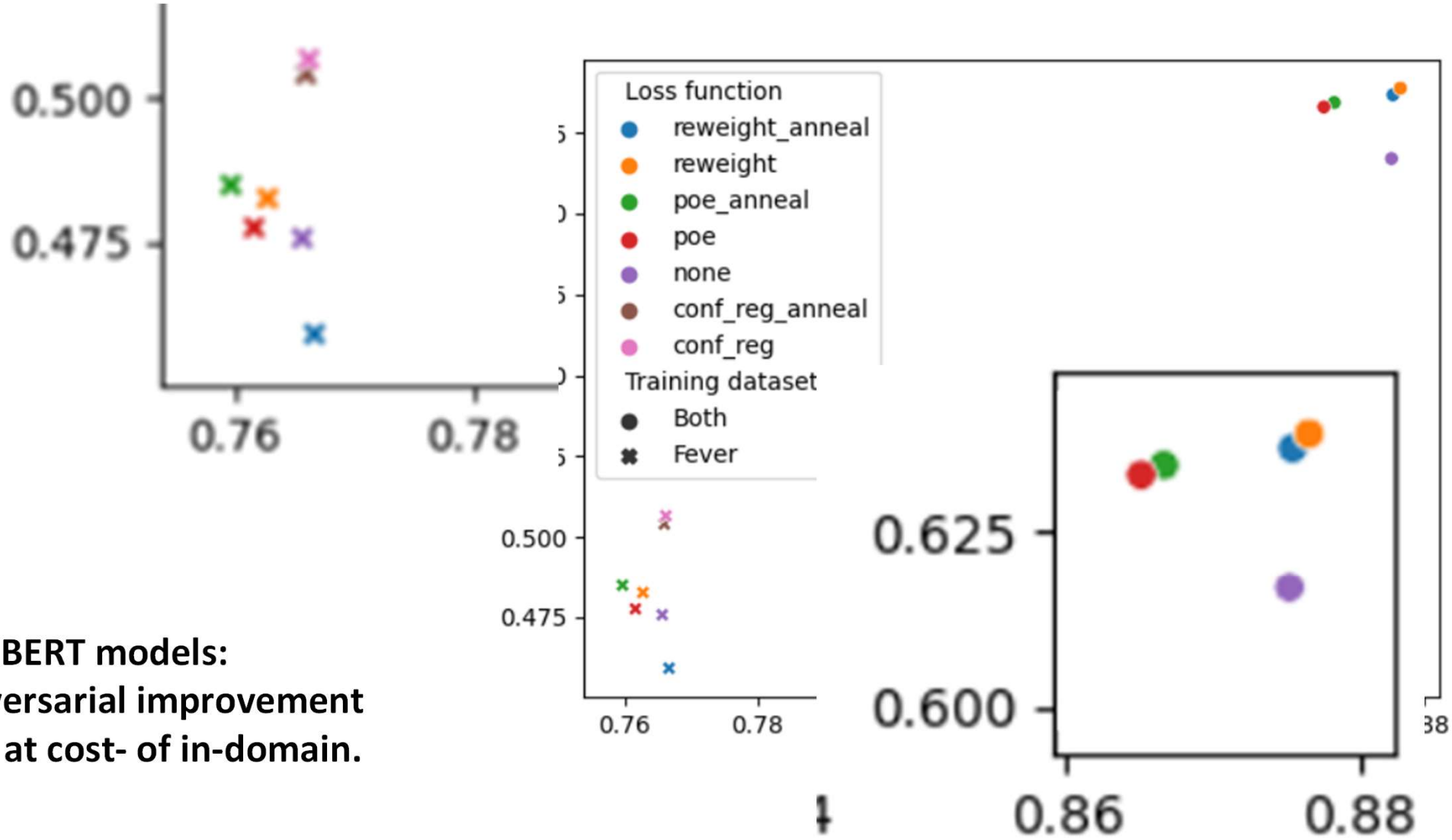
Table 5.1: Performance of Albert-base model on derived adversaries. Resilience is calculated using a single set of values for the three word overlap approaches (the average of the three).

RQB: Architecture comparison

	BERT	ALBERT
Fever-only	++	None
Both	++	- - -

Fever-only trained models				
Architecture	*ALBERT-base*		*BERT-base-uncased*	
Loss Function\Test Set	fever	fever-symmetric	fever	fever-symmetric
PoE	0.949872	0.813202	0.869428	0.733146
Reweight	0.94082	0.787921	0.870453	0.745787
(None / Cross-entropy)	0.948249	0.81882	0.869769	0.733146
Confidence Regularization	0.873783	0.65309	0.871904	0.745787
Confidence Regularization (Annealed)	0.887703	0.672753	0.872161	0.759831
Reweight (Annealed)	0.907857	0.676966	0.870367	0.758427
PoE (Annealed)	0.626473	0.5	0.869257	0.726124

Table 5.3: Comparison of baseline models across architectures, when trained solely on FEVER



For BERT models:
Adversarial improvement not at cost- of in-domain.

RQB: Comparing loss functions + Annealing

- No generally superior loss function
- Confidence regularization best performer on FEVER-trained BERT models, and Entity reweighting for those trained on both datasets.
- Annealing mostly made no difference
- Conf-reg models didn't converge for ALBERT:

Training dataset	Loss Function	Bias used	Uses Teacher?	Uses Shallow?	FEVER Test accuracy	Average Confidence
FEVER only	Conf. Reg	Ent Overweighting	✓	×	0.954569	0.206
	Conf. Reg	Shallow	✓	✓	0.873783	0.034
	(baseline)		✓*	×	0.948249	0.995
	PoE	Shallow	×	✓	0.949872	0.996

B2. Investigating if combining training with both VitC and the Self-debias framework can produce higher Resilience.

- Training on both FEVER and VitaminC is beneficial both for debiasing and for in-domain performance.
- Self-debias does not work with ALBERT-based models, on which VitaminC training delivers a better performing model.

		Correctly Predicted				
		FEVER-debias only	Vitc-base only	Both	Neither	TOTAL
Fever	Correct	1480	0	12950	0	14430
+vitc	Incorrect	731	2891	0	5368	8990
+debias	TOTAL	2211	2891	12950	5368	23420

Table 5.5: Table showing overlap between examples correctly predicted between FEVER-debias (BERT model trained on FEVER with Example Reweighting), VitC-base (BERT baseline (i.e. non-debiased) model trained on VitaminC only), and FEVER+vitc+debias (BERT model trained on both FEVER and VitC, using Example Reweighting). 23420 examples are a 50/25/25 split between FEVER-test, VitC-Real test and VitC-Synth test.

EQ1: How well do bias-agnostic approaches perform against hand-crafted known-bias ones?

- Using known rather than shallow biases yields higher resilience.
- Debiasing is also more effective in these cases.

ALBERT/ Both / Known		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	91.4%	80.2%	71.7%	81.1%
Best (Ao3) / delta	conf_reg	91.4% / 0%	80.3% / 0.1%	72.7% / 1.1%	81.5% / 0.4%
Best (NBR) / delta					

ALBERT / Fever only / Known		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	80.5%	73.3%	52.3%	68.7%
Best (Ao3) / delta	conf_reg_anneal	82.5% / 2%	76.9% /3.6%	71.6% / 19.3%	77.0% / 8.3%
Best (NBR) / delta					

BERT / Both / Known		Fact-check average	adv+sym resilience	new-biases resilience	Average
	baseline	87.5%	76.3%	67.0%	76.9%
Best (Ao3) / delta	reweight_anneal	87.8% / 0.3%	77.3% / 1%	68.1% / 1.1%	77.7% / 0.8%
Best (NBR) / delta					

Figure 5.6: Best performing adversaries from known-bias training, compared with baseline models. Ao3 is best (non-baseline) model according to average of three metrics, NBR is best according to highest new-biases resilience.

Conclusions

Key technical hurdles

- Needed to combine VitaminC with self-debias.
- Self-debias was written with Pytorch-Pretrained-Bert which doesn't support ALBERT models. Needed to migrate to Transformers.
- Self-debias was not documented and was missing nearly all code related to FEVER.
- VitaminC (obviously) had no concept of bias or teacher files, and so dataloaders and trainers needed extending to pass these through.

Future work

- Why do these methods designed for BERT not transfer to ALBERT?
- What difference is there in the dynamics of the three loss functions used in self-debias?
- Is there a better way to combine dataset-targeting and model-targeting debiasing approaches?
- How appropriate is VitaminC to train FEVER-based systems when it exhibits stylistic differences?
- Automated antonym generation needs a lot of work.

Questions?