Towards Debiasing Fact Verification Models

Tal Schuster*[,1], **Darsh J Shah***[,1], **Yun Jie Serene Yeo**[2],
Daniel Filizzola[1], **Enrico Santus**[1], **Regina Barzilay**[1]

[1]Computer Science and Artificial Intelligence Lab, MIT
[2]DSO National Laboratories, Singapore
{tals, darsh, yeoyjs, danifili, esantus, regina}@csail.mit.edu

Abstract

Fact verification requires validating a claim in the context of evidence. We show, however, that in the popular FEVER dataset this might not necessarily be the case. Claim-only classifiers perform competitively with top evidence-aware models. In this paper, we investigate the cause of this phenomenon, identifying strong cues for predicting labels solely based on the claim, without considering any evidence. We create an evaluation set that avoids those idiosyncrasies. The performance of FEVER-trained models significantly drops when evaluated on this test set. Therefore, we introduce a regularization method which alleviates the effect of bias in the training data, obtaining improvements on the newly created test set. This work is a step towards a more sound evaluation of reasoning capabilities in fact verification models.[1]

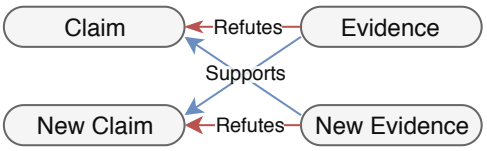
1 Introduction

Creating quality datasets is essential for expanding NLP functionalities to new tasks. Today, such datasets are often constructed using crowdsourcing mechanisms. Prior research has demonstrated that artifacts of this data collection method often introduce idiosyncratic biases that impact performance in unexpected ways (Poliak et al., 2018; Gururangan et al., 2018). In this paper, we explore this issue using the FEVER dataset, designed for fact verification (Thorne et al., 2018).

The task of fact verification involves assessing claim validity in the context of evidence, which can either support, refute or contain not enough information. Figure 1(A) shows an example of a FEVER claim and evidence. While validity of some claims may be asserted in isolation (e.g. through common sense knowledge), contextual verification is key for a fact-checking task (Alhindi et al., 2018). Datasets should ideally evaluate this ability. To assess whether this is the case for FEVER, we train a claim-only BERT (Devlin et al., 2019) model that classifies each claim on its own, without associated evidence. The resulting system achieves 61.7%, far above the majority baseline (33.3%).

Our analysis of the data demonstrates that this unexpectedly high performance is due to idiosyncrasies of the dataset construction. For instance, in §2 we show that the presence of negation phrasing highly correlates with the REFUTES label, independently of provided evidence.

To address this concern, we propose a mecha-



(A) ORIGINAL pair from the FEVER dataset

Claim:
Stanley Williams stayed in Cuba his whole life.
Evidence:
Stanley [...] was part of the West Side Crips, a street gang which has its roots in South Central Los Angeles.

(B) Manually GENERATED pair

Claim:
Stanley Williams moved from Cuba to California when he was 15 years old.
Evidence:
Stanley [...] was born in Havana and didn't leave the country until he died.

Figure 1: An illustration of a REFUTES claim-evidence pair from the FEVER dataset (A) that is used to generate a new pair (B). From the combination of the ORIGINAL and manually GENERATED pairs, we obtain a total of four pairs creating symmetry.

Asterisk (*) denotes equal contribution.
[1]Data and code: https://github.com/TalSchuster/FeverSymmetric

nism for avoiding bias in the test set construction. We create a SYMMETRIC TEST SET where, for each claim-evidence pair, we manually generate a synthetic pair that holds the same relation (e.g. SUPPORTS or REFUTES) but expressing a different, contrary, fact. In addition, we ensure that in the new pair, each sentence satisfies the inverse relation with the original pair's sentence. This process is illustrated in Figure 1, where an original REFUTES pair is extended with a synthetic REFUTES pair. The new evidence is constrained to support the original claim, and the new claim is supported by the original evidence. In this way, we arrive at three new pairs that complete the symmetry.

Determining veracity with the claim alone in this setting would be equivalent to a random guess. Unsurprisingly, the performance of FEVER-trained models drop significantly on this test set, despite having complete vocabulary overlap with the original dataset. For instance, the leading evidence-aware system in the FEVER Shared Task, the NSMN classifier by Nie et al. (2019)[2], achieves only 58.7% accuracy on the symmetric test set compared to 81.8% on the original dataset.

While this new test set highlights the aforementioned problem, other studies have shown that FEVER is not the only biased dataset (Poliak et al., 2018; Gururangan et al., 2018). A potential solution which may be applied also in other tasks is therefore to develop an algorithm that alleviates such bias in the training data. We introduce a new regularization procedure to downweigh the giveaway phrases that cause the bias.
The contributions of this paper are threefold:

- We show that inherent bias in FEVER dataset interferes with context-based fact-checking.
- We introduce a method for constructing an evaluation set that explicitly tests a model's ability to validate claims in context.
- We propose a new regularization mechanism that improves generalization in the presence of the aforementioned bias.

2 Motivation and Analysis

In this section, we quantify the observed bias and explore the factors causing it.

[2] https://github.com/easonnie/combine-FEVER-NSMN

	Train		Development	
Bigram	**LMI**·10^{-6}	$p(l\|w)$	**LMI**·10^{-6}	$p(l\|w)$
did not	1478	0.83	1038	0.90
yet to	721	0.90	743	0.96
does not	680	0.78	243	0.68
refused to	638	0.87	679	0.97
failed to	613	0.88	220	0.96
only ever	526	0.86	350	0.82
incapable being	511	0.89	732	0.96
to be	438	0.50	454	0.65
unable to	369	0.88	346	0.95
not have	352	0.78	211	0.92

Table 1: Top 10 LMI-ranked bigrams in the train set of FEVER for REFUTES with its $p(l|w)$. The corresponding figures for the development set are also provided. Statistics for other labels are in Appendix B.2.

Claim-only Classification Claim-only aware classifiers can significantly outperform all baselines described by Thorne et al. (2018).[3] BERT, for instance, attains an accuracy of 61.7%, which is just 8% behind NSMN. We hypothesize that these results are due to two factors: (1) idiosyncrasies distorting performance and (2) word embeddings revealing world knowledge.

Idiosyncrasies Distorting Performance We investigate the correlation between phrases in the claims and the labels. In particular, we look at the n-gram distribution in the training set. We use Local Mutual Information (LMI) (Evert, 2005) to capture high frequency n-grams that are highly correlated with a particular label, as opposed to $p(l|w)$ that is biased towards low frequency n-grams. LMI between w and l is defined as follows:

$$LMI(w,l) = p(w,l) \cdot \log\left(\frac{p(l|w)}{p(l)}\right), \qquad (1)$$

where $p(l|w)$ is estimated by $\frac{\text{count}(w,l)}{\text{count}(w)}$, $p(l)$ by $\frac{\text{count}(l)}{|D|}$, $p(w,l)$ by $\frac{\text{count}(w,l)}{|D|}$ and $|D|$ is the number of occurrences of all n-grams in the dataset.

Table 1 shows that the top LMI-ranked n-grams that are highly correlated with the REFUTES class in the training set exhibit a similar correlation in the development set. Most of the n-grams express strong negations, which, in hindsight, is not surprising as these idiosyncrasies are induced by the way annotators altered the original claims to generate fake claims.

[3] We evaluate on the development set as the test set is hidden. Hyper-parameter fine-tuning is performed on a 20% split of the training set, which is finally joined to the remaining 80% for training the best setting. See Appendix B.1.

World Knowledge Word embeddings encompass world knowledge, which might augment the performance of claim-only classifiers. To factor out the contribution of world knowledge, we trained two versions of claim-only InferSent (Poliak et al., 2018) on the FEVER claims: one with GloVe embeddings (Pennington et al., 2014) and the other with random embeddings.[4] The performance with random embeddings was 54.1%, compared to 57.3% with GloVe, which is still far above the majority baseline (33.3%). We conjecture that world knowledge is not the main reason for the success of the claim-only classifier.

3 Towards Unbiased Evaluation

Based on the analysis above, we conclude that an unbiased verification dataset should exclude 'give-away' phrases in one of its inputs and also not allow the system to solely rely on world knowledge. The dataset should enforce models to validate the claim with respect to the retrieved evidence. Particularly, the truth of some claims might change as the evidence varies over time.

For example, the claim *"Halep failed to ever win a Wimbledon title"* was correct until July 19. A fact-checking system that retrieves information from Halep's Wikipedia page should modify its answer to "false" after the update that includes information about her 2019 win.

Towards this goal, we create a SYMMETRIC TEST SET. For an original claim-evidence pair, we manually generate a synthetic pair that holds the same relation (i.e. SUPPORTS or REFUTES) while expressing a fact that contradicts the original sentences. Combining the ORIGINAL and GENERATED pairs, we obtain two new cross pairs that hold the inverse relations (see Figure 1). Examples of generated sentences are provided in Table 2.

This new test set completely eliminates the ability of models to rely on cues from claims. Considering the two labels of this test set[5], the probability of a label given the existence of any n-gram in the claim or in the evidence is $p(l|w) = 0.5$, by construction.

Also, as the example in Figure 1 demonstrates, in order to perform well on this dataset, a fact verification classifier may still take advantage of world knowledge (e.g. geographical locations), but reasoning should only be with respect to the context.

[4]We use InferSent because BERT, being pretrained on Wikipedia, comprises world knowledge (Talmor et al., 2019).

[5]NOT ENOUGH INFO cases are easy to generate so we focus on the two other labels.

4 Towards Unbiased Training

Creating a large symmetric dataset for training is outside the scope of this paper as it would be too expensive. Instead, we propose an algorithmic solution to alleviate the bias introduced by 'give-away' n-grams present in the claims. We reweight the instances in the dataset to flatten the correlation of claim n-grams with respect to the labels. Specifically, for 'give-away' phrases of a particular label, we increase the importance of claims with different labels containing those phrases.

We assign an additional (positive) balancing weight $\alpha^{(i)}$ to each training example $\{x^{(i)}, y^{(i)}\}$, determined by the words in the claim.

Bias in the Re-Weighted Dataset For each n-gram w_j in the vocabulary V of the claims, we define the bias towards class c to be of the form:

$$b_j^c = \frac{\sum_{i=1}^{n} I_{[w_j^{(i)}]}(1+\alpha^{(i)}) I_{[y^{(i)}=c]}}{\sum_{i=1}^{n} I_{[w_j^{(i)}]}(1+\alpha^{(i)})}, \quad (2)$$

where $I_{[w_j^{(i)}]}$ and $I_{[y^{(i)}=c]}$ are the indicators for w_j being present in the claim from $x^{(i)}$ and label $y^{(i)}$ being of class c, respectively.

Optimization of the Overall Bias Finding the α values which minimize the bias leads us to solving the following objective:

$$\min \left(\sum_{j=1}^{|V|} \max_c (b_j^c) + \lambda \|\vec{\alpha}\|_2 \right). \quad (3)$$

Re-Weighted Training Objective We calculate the α values separately from the model optimization, as a pre-processing step, by optimizing Eq. 3. Using these values, the training objective is reweighted from the standard $\sum_{i=1}^{n} L(x^{(i)}, y^{(i)})$ to

$$\sum_{i=1}^{n} (1+\alpha^{(i)}) L(x^{(i)}, y^{(i)}). \quad (4)$$

This re-weighting is independent of the model architecture and can be easily added to any objective, similar to Jiang and Nachum (2019) where they learn instance weights to address labeling bias in datasets.

Source	Claim	Evidence	Label
ORIGINAL	Tim Roth is an English actor.	Timothy Simon Roth (born 14 May 1961) is an English actor and director.	SUPPORTS
GENERATED	Tim Roth is an American actor.	Timothy Simon Roth (born 14 May 1961) is an American actor and director.	SUPPORTS
ORIGINAL	Aristotle spent time in Athens.	At seventeen or eighteen years of age, he joined Plato's Academy in Athens and remained there until the age of thirty-seven (c. 347 BC).	SUPPORTS
GENERATED	Aristotle did not visit Athens.	At seventeen or eighteen years of age, he missed the opportunity to join Plato's Academy in Athens and never visited the place.	SUPPORTS
ORIGINAL	Telemundo is a English-language television network.	Telemundo (telemundo) is an American Spanish-language terrestrial television network owned by Comcast through the NBCUniversal division NBCUniversal Telemundo Enterprises.	REFUTES
GENERATED	Telemundo is a Spanish-language television network.	Telemundo (telemundo) is an American English-language terrestrial television network owned by Comcast through the NBCUniversal division NBCUniversal Telemundo Enterprises.	REFUTES
ORIGINAL	Magic Johnson did not play for the Lakers.	He played point guard for the Lakers for 13 seasons.	REFUTES
GENERATED	Magic Johnson played for the Lakers.	He played for the Giants and no other team.	REFUTES

Table 2: Examples of pairs from the Symmetric Dataset. Each generated claim-evidence pair holds the relation described in the right column. Crossing the generated sentences with the original ones creates two additional cases with an opposite label (see Figure 1).

5 Experiments

We use the SYMMETRIC TEST SET to (1) investigate whether top performing sequence classification models trained on the FEVER dataset are actually verifying claims in the context of evidence; and (2) measure the impact of the re-weighting method described in §4 over a classifier.

To achieve the first goal, we use three classifiers. The first is a pre-trained, current FEVER state-of-the-art classifier, **NSMN** (Nie et al., 2019) which is a variation of the ESIM (Chen et al., 2017) model, with a number of additional features, such as contextual word embeddings (Peters et al., 2018). In addition, we train our own **ESIM** model with GloVe embeddings, using the available code from Gardner et al. (2017). The third is a **BERT** classifier[6] that we fine-tune for 3 epochs to classify the relation based on the concatenation of the claim and evidence (with a delimiter token). To measure the impact of our regularization method, we also train the ESIM and BERT models with the re-weighting method.

[6] https://github.com/huggingface/pytorch-pretrained-BERT

Symmetric Test Set The full SYMMETRIC TEST SET consists of 956 claim-evidence pairs, created following the procedure described in §3. The new pairs originated from 99 SUPPORTS and 140 REFUTES pairs that were randomly picked from the cases which NSMN correctly predicts.[7] After its generation, we asked two subjects to annotate randomly sampled 285 claim-evidence pairs (i.e. 30% of the total pairs in SYMMETRIC TEST SET) with one label among SUPPORTS, REFUTES or NOT ENOUGH INFO, flagging non-grammatical cases. They agreed with the dataset labels in 94% of cases, attaining a Cohen κ of 0.88 (Cohen, 1960). Typos and small grammatical errors were reported in 2% of the cases. Given the small size of this dataset, we only use it as a test set.

Results Table 3 summarizes the performance of the three models on the SUPPORTS and REFUTES pairs from the FEVER DEV set and on the created SYMMETRIC TEST SET pairs. All models perform relatively well on FEVER DEV but achieve less than 60% accuracy on the synthetic ones. We

[7] Due to our focus on the performance drop with respect to the newly generated pairs rather than on the intention of multiplying the difficulties for the top performing model.

Model	FEVER DEV		GENERATED	
	BASE	R.W	BASE	R.W
NSMN	81.8	-	58.7	-
ESIM	80.8	76.0	55.9	59.3
BERT	**86.2**	84.6	58.3	**61.6**

Table 3: Classifiers' accuracy on the SUPPORTS and REFUTES cases from the FEVER DEV set and on the GENERATED pairs for the SYMMETRIC TEST SET in the setting of without (BASE) and with (R.W) re-weight.

conjecture that the drop in performance is due to training data bias that is also observed in the development set (see §2) but not in the generated symmetric cases.

Our re-weighting method (§4) helps to reduce the bias in the claims. In Table 4, we revisit the give-away bigrams from Table 1. Applying the weights obtained by optimizing Eq. 3, the weighted distribution of these phrases being associated with a specific label in the training set is now roughly uniform.

The re-weighting method increases the accuracy of the ESIM and BERT models by an absolute 3.4% and 3.3% respectively. One can notice that this improvement comes at a cost in the accuracy over the FEVER DEV pairs. Again, this can be explained by the bias in the training data that translates to the development set, allowing FEVER-trained models to leverage it. Applying the regularization method, using the same training data, helps to train a more robust model that performs better on our test set, where verification in context is a key requirement.

6 Related Work

Large scale datasets are fraught with give-away phrases (McCoy et al., 2019; Niven and Kao, 2019). Crowd workers tend to adopt heuristics when creating examples, introducing bias in the dataset. In SNLI (Stanford Natural Language Inference) (Bowman et al., 2015), entailment based solely on the hypothesis forms a very strong baseline (Poliak et al., 2018; Gururangan et al., 2018).

Similarly, as shown by Kaushik and Lipton (2018), reading comprehension models that rely only on the question (or only on the passage referred to by the question) perform exceedingly well on several popular datasets (Weston et al., 2016; Onishi et al., 2016; Hill et al., 2016). To address deficiencies in the SQuAD dataset (Jia and Liang, 2017), researchers have proposed approaches for augmenting the existing dataset (Rajpurkar et al., 2018). In most cases, these augmentations are done manually, and involve constructing challenging examples for existing systems.

Bigram	R.W **LMI**$\cdot 10^{-6}$	R.W $p(l\|w)$
did not	144	0.35
yet to	30	0.33
does not	67	0.35
refused to	55	0.35
failed to	31	0.33
only ever	9	0.31
incapable being	32	0.33
to be	8	0.30
unable to	10	0.32
not have	41	0.35

Table 4: Re-weighted statistics (l = REFUTES) for the bigrams from Table 1. The weights were obtained following the optimization of Eq. 3 on the training set which contains three labels.

7 Conclusion

This paper demonstrates that the FEVER dataset contains idiosyncrasies that can be easily exploited by fact-checking classifiers to obtain high classification accuracies. Evaluating the claim-evidence reasoning of these models necessitates unbiased datasets. Therefore, we suggest a way to turn the evaluation FEVER pairs into symmetric combinations for which a decision that is solely based on the claim is equivalent to a random guess. Tested on these pairs, FEVER-trained models show degraded performance. To address this problem, we propose a simple method that supports a more robust generalization in the presence of bias.

Moving forward, we suggest using our symmetric dataset in addition to the current retrieval-based FEVER evaluation pipeline. This way, models could be tested both for their evidence retrieval and classification accuracy and for performing the reasoning with respect to the evidence.

8 Acknowledgments

We thank the MIT NLP group and the reviewers for their helpful discussion and comments. This work is supported by DSO grant DSOCL18002.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stefan Evert. 2005. The statistics of word cooccurrences: word pairs and collocations.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *International Conference on Learning Representations 2016*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

Heinrich Jiang and Ofir Nachum. 2019. Identifying and correcting label bias in machine learning. *arXiv preprint arXiv:1901.04966*.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence*.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David A. McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235. The Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*,

pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations*.