CrossAug: A Contrastive Data Augmentation Method for Debiasing Fact Verification Models

Minwoo Lee[1], Seungpil Won[1], Juae Kim[2], Hwanhee Lee[1], Cheoneum Park[2], Kyomin Jung[1]

[1] Dept. of Electrical and Computer Engineering, Seoul National University
[2] AIRS Company, Hyundai Motor Group

Fact Verification

Given a **claim** sentence and **evidence** text, predict whether the evidence:

- **SUPPORTS** the claim,
- **REFUTES** the claim,
- or has **NOT ENOUGH INFO** to support or refute the claim.

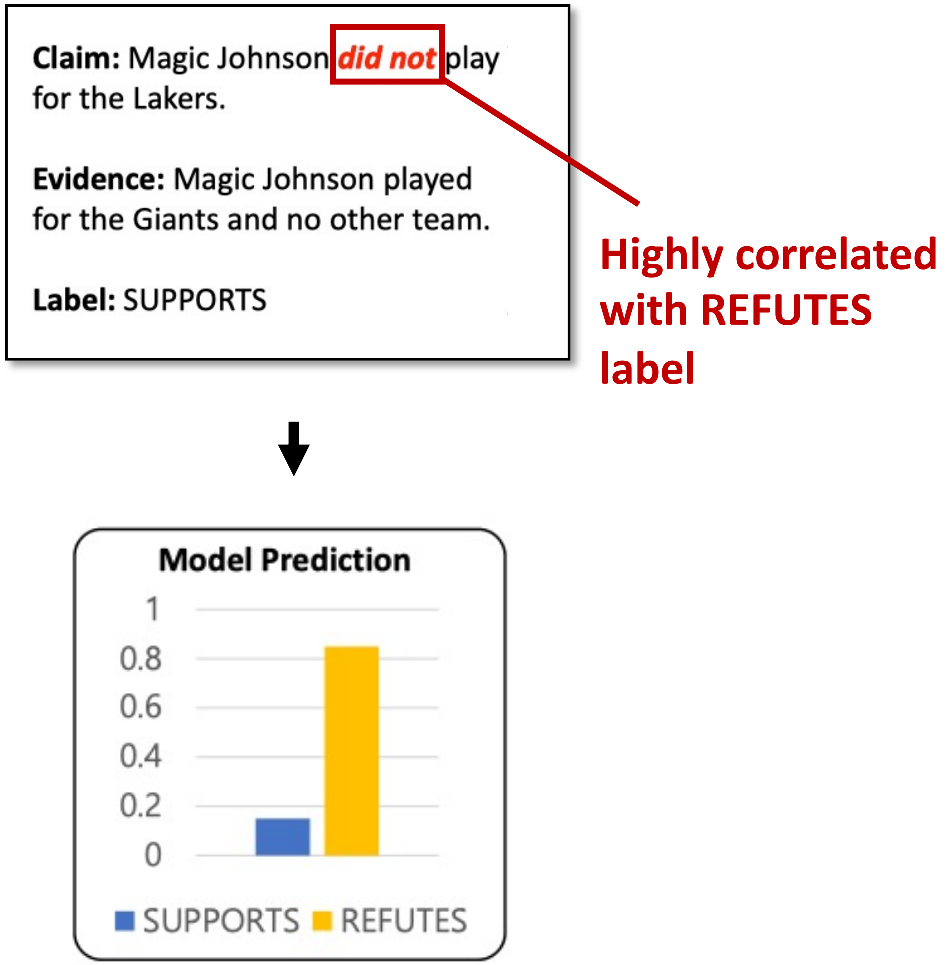
Claim: Magic Johnson ***did not*** play for the Lakers.

Evidence: Magic Johnson played for the Giants and no other team.

Label: SUPPORTS

Annotation Artifacts in Fact Verification Dataset

- Crowdsouring datasets often produce **annotation artifacts** that lead to unwanted bias in data.
- FEVER dataset [1] was also shown to contain **lexical biases** where specific phrases in claim is highly correlated with a specific label [2].
- Training leads to biased models that exploit the artifacts.



[1] 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).
[2] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards Debiasing Fact Verification Models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Debiasing Approaches

Previous Approaches

- **Regularize** learning by weighting biased samples, weighted using
  - n-gram statistics
  - biased model predictions
- Directs model to avoiding learning a specific type of bias

Debiasing Approaches

Previous Approaches

- **Regularize** learning by weighting biased samples, weighted using
  - n-gram statistics
  - biased model predictions
- Directs model to avoiding learning a specific type of bias

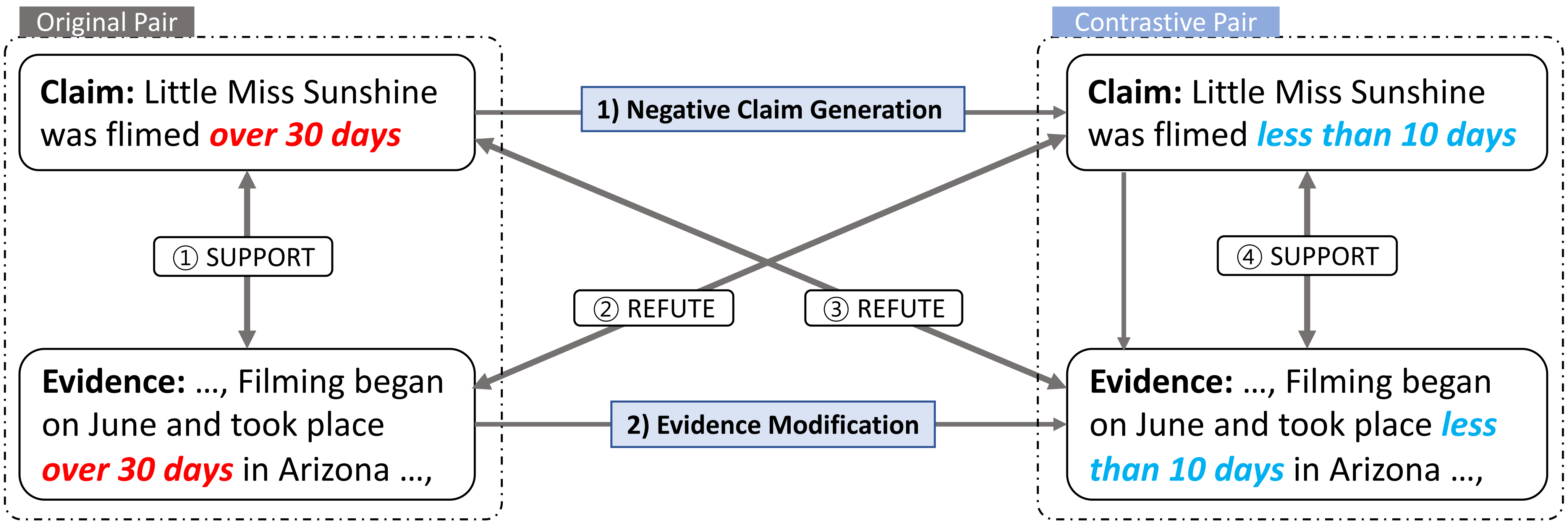
Our Approach: CrossAug

- **Augment data** with **contrastive samples** so that model can’t rely on artifacts
- Directs model to learn a more robust representation

CrossAug Overview

Two-stage data augmentation process

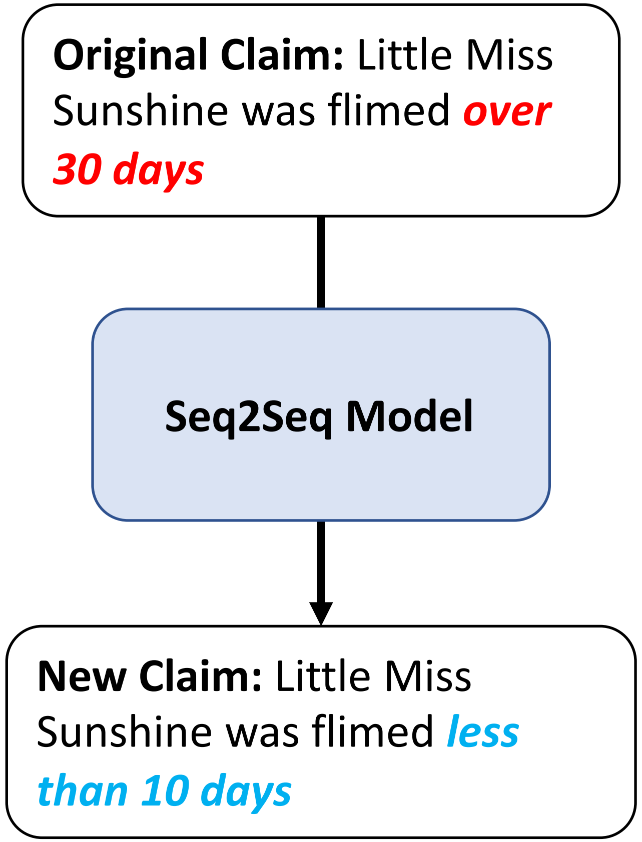
1. **Negative Claim Generation**
2. **Evidence Modification**



1. Negative Claim Generation

- **Neural seq2seq model** is used to generate negative claim c' from positive claim c.

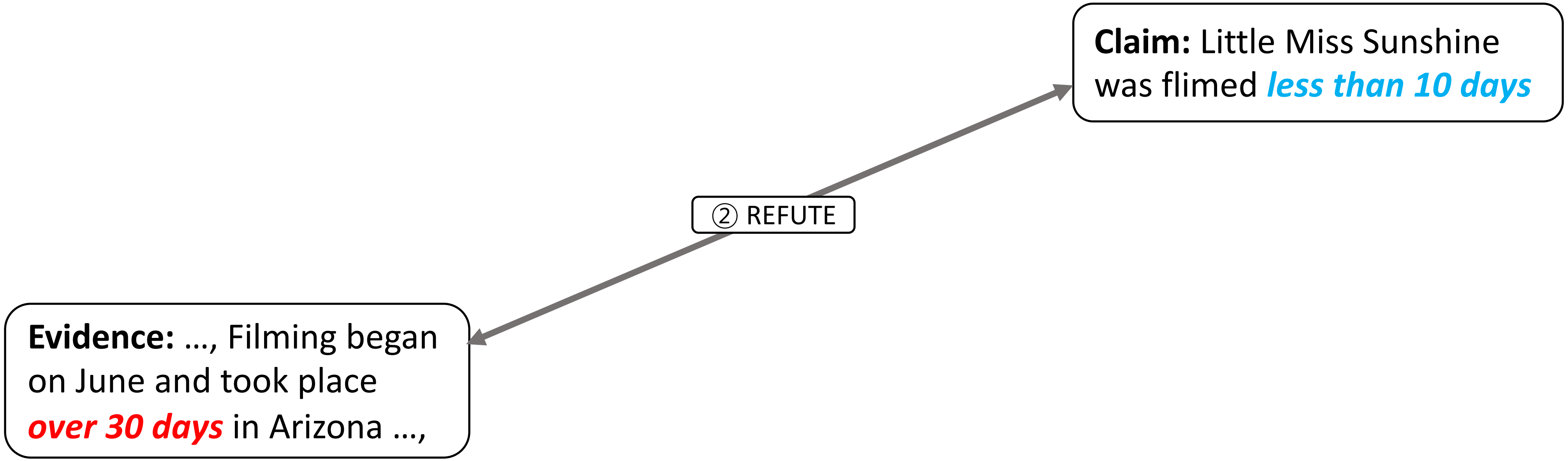
- BART model is finetuned on WikiFactCheck-English dataset [3], which provides human-written, parallel pairs of positive and negative claims.



[3] Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated Fact-Checking of Claims from Wikipedia. In Proceedings of The 12th Language Resources and Evaluation Conference. 6874–6882.

1. Negative Claim Generation

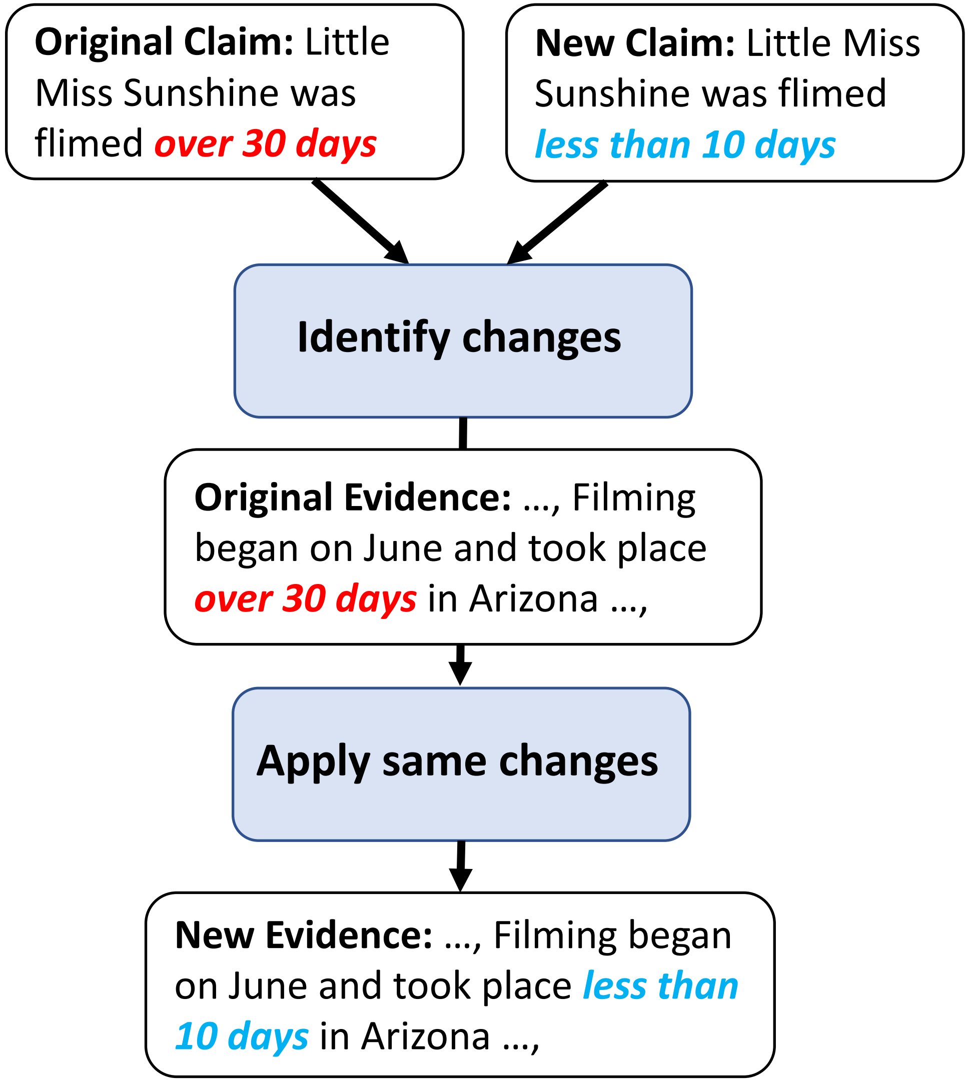
- The generated negative claim c’ is paired with original evidence with a flipped label of REFUTES to create **one additional samples**.



2. Evidence Modification

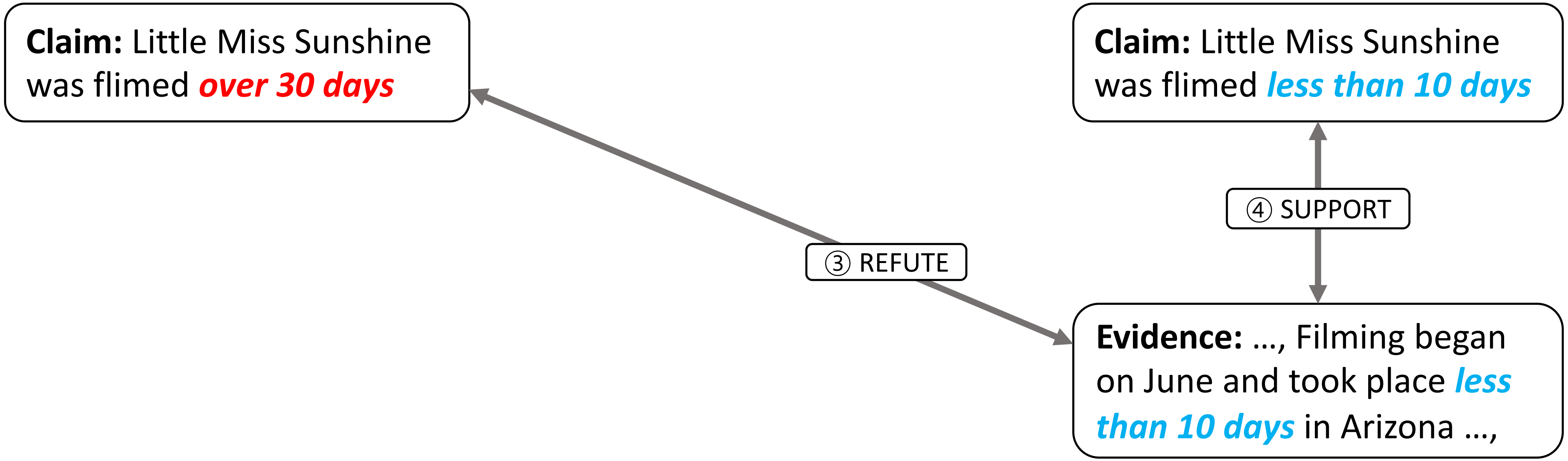
- The negative claim generated in the first stage often only differs from the positive claim by a few words, and can be seen as a **span replacement**.

- We **identify changes** in the claims and **replace the same words** in the evidence to create modified evidence.



2. Evidence Modification

- The modified evidence e is paired with both claims c and c' to generate **two additional contrastive samples**.



Results

Full Dataset Evaluation

- Our method achieved **10.13% improvement over the baseline** and **3.6% improvement over the previous SOTA debiasing technique** on the Symmetric FEVER dataset.

Train method	FEVER dev	Symmetric	Adversarial	FM2 dev	Δ sym.	Δ avg.
No augmentation (baseline)	86.15 ± 0.42	58.77 ± 1.29	49.66 ± 0.37	40.81 ± 0.43	-	-
EDA	85.09 ± 0.25	58.55 ± 1.63	51.41 ± 1.14	41.21 ± 1.11	−0.22%	+0.22%
Paraphrasing	84.33 ± 0.34	59.02 ± 1.38	**52.53** ± 1.20	40.60 ± 0.71	+0.25%	+0.27%
Re-weighting	85.56 ± 0.32	61.87 ± 1.16	49.92 ± 0.80	43.80 ± 0.46	+3.10%	+1.44%
Product of Experts (PoE)	**86.50** ± 0.35	65.30 ± 1.73	51.07 ± 1.20	**46.69 ± 1.11**	+6.53%	+3.54%
CrossAug (ours)	85.34 ± 0.68	**68.90 ± 1.68**	51.78 ± 1.02	44.17 ± 1.27	**+10.13%**	**+3.70%**
- Negative claim only augmentation	85.70 ± 0.28	61.00 ± 0.71	51.96 ± 0.90	43.06 ± 0.40	+2.23%	+1.58%
- Negative evidence only augmentation	85.87 ± 0.16	67.06 ± 0.99	51.46 ± 0.43	43.70 ± 0.97	+8.29%	+3.18%

Full Dataset Evaluation

- Our method achieved **10.13% improvement over the baseline** and **3.6% improvement over the previous SOTA debiasing technique** on the Symmetric FEVER dataset.
- Our method also led to **greatest average improvement** across various fact verification evaluation sets.

Train method	FEVER dev	Symmetric	Adversarial	FM2 dev	Δ sym.	Δ avg.
No augmentation (baseline)	86.15 ± 0.42	58.77 ± 1.29	49.66 ± 0.37	40.81 ± 0.43	-	-
EDA	85.09 ± 0.25	58.55 ± 1.63	51.41 ± 1.14	41.21 ± 1.11	−0.22%	+0.22%
Paraphrasing	84.33 ± 0.34	59.02 ± 1.38	**52.53** ± 1.20	40.60 ± 0.71	+0.25%	+0.27%
Re-weighting	85.56 ± 0.32	61.87 ± 1.16	49.92 ± 0.80	43.80 ± 0.46	+3.10%	+1.44%
Product of Experts (PoE)	**86.50** ± 0.35	65.30 ± 1.73	51.07 ± 1.20	**46.69 ± 1.11**	+6.53%	+3.54%
CrossAug (ours)	85.34 ± 0.68	**68.90 ± 1.68**	51.78 ± 1.02	44.17 ± 1.27	**+10.13%**	**+3.70%**
- Negative claim only augmentation	85.70 ± 0.28	61.00 ± 0.71	51.96 ± 0.90	43.06 ± 0.40	+2.23%	+1.58%
- Negative evidence only augmentation	85.87 ± 0.16	67.06 ± 0.99	51.46 ± 0.43	43.70 ± 0.97	+8.29%	+3.18%

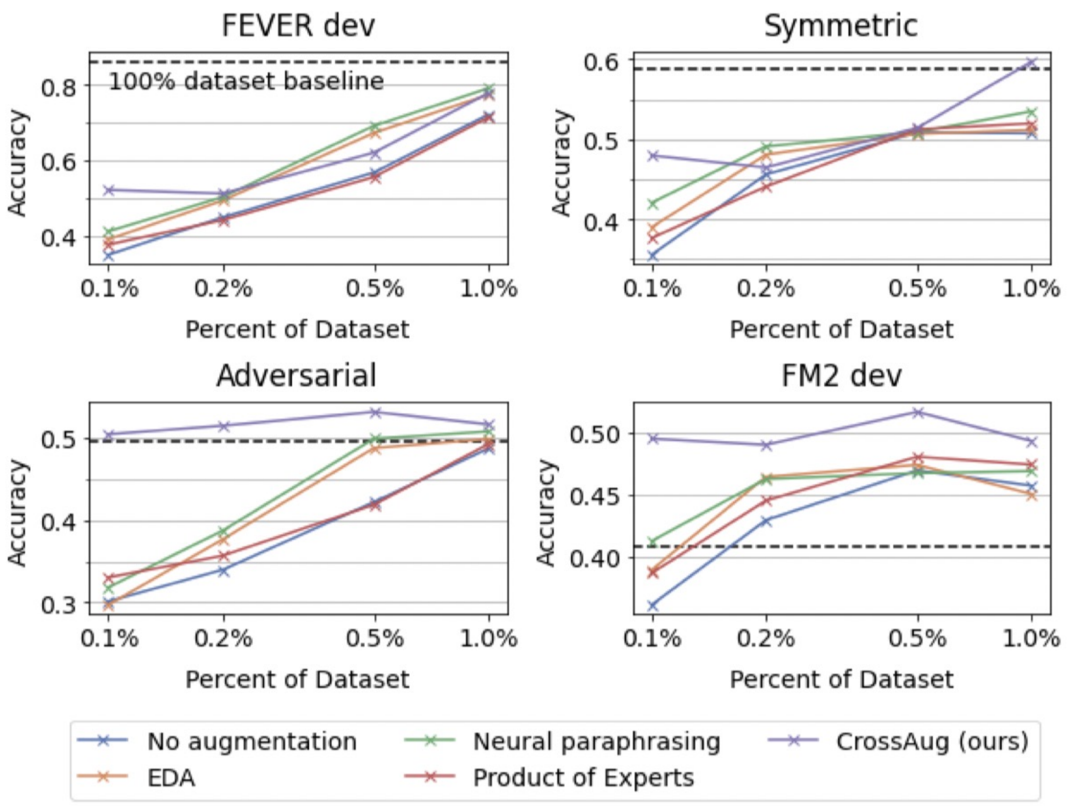
Ablation Study

- Partial data augmentation by **using only negative claims or negative evidence** still effective with moderate performance improvement.
- However, using all combination of data augmentation samples is most effective.

Train method	FEVER dev	Symmetric	Adversarial	FM2 dev	Δ sym.	Δ avg.
No augmentation (baseline)	86.15 ± 0.42	58.77 ± 1.29	49.66 ± 0.37	40.81 ± 0.43	-	-
EDA	85.09 ± 0.25	58.55 ± 1.63	51.41 ± 1.14	41.21 ± 1.11	−0.22%	+0.22%
Paraphrasing	84.33 ± 0.34	59.02 ± 1.38	**52.53** ± 1.20	40.60 ± 0.71	+0.25%	+0.27%
Re-weighting	85.56 ± 0.32	61.87 ± 1.16	49.92 ± 0.80	43.80 ± 0.46	+3.10%	+1.44%
Product of Experts (PoE)	**86.50** ± 0.35	65.30 ± 1.73	51.07 ± 1.20	**46.69 ± 1.11**	+6.53%	+3.54%
CrossAug (ours)	85.34 ± 0.68	**68.90 ± 1.68**	51.78 ± 1.02	44.17 ± 1.27	**+10.13%**	**+3.70%**
- Negative claim only augmentation	85.70 ± 0.28	61.00 ± 0.71	51.96 ± 0.90	43.06 ± 0.40	+2.23%	+1.58%
- Negative evidence only augmentation	85.87 ± 0.16	67.06 ± 0.99	51.46 ± 0.43	43.70 ± 0.97	+8.29%	+3.18%

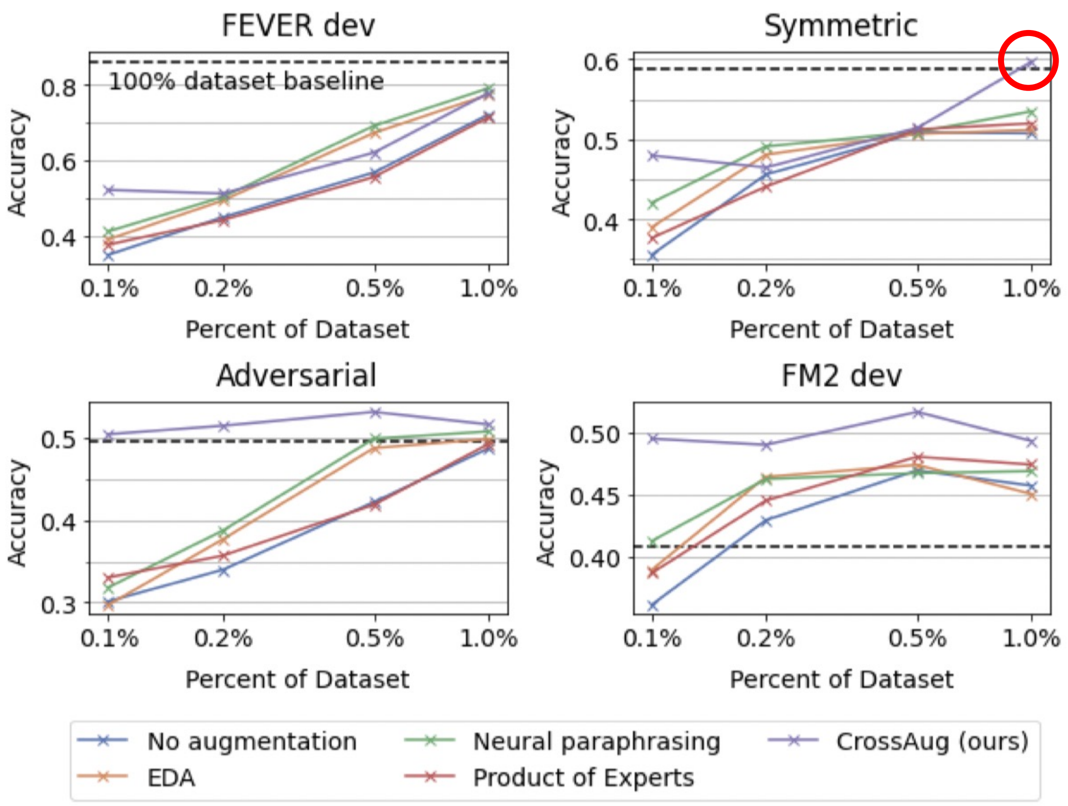
Low-resource Conditions Evaluation

- Our method showed a **consistent improvement** in the low-resource conditions with limited training samples over all evaluated datasets.



Low-resource Conditions Evaluation

- Our method also showed a **consistent improvement** in the low-resource conditions with limited training samples over all evaluated datasets.
- For the Symmetric evaluation set, we outperform the baseline trained on the full dataset **with just 1% of the original training data**.



Summary

- We propose **CrossAug**, a novel contrastive data augmentation method for debiasing fact verification models.
- Using CrossAug leads to the state-of-the-art performance on the debiased fact verification dataset.
- Our method shows consistent effectiveness even in low-resource conditions with limited training data.

Code: https://github.com/minwhoo/CrossAug