CrowdChecked: Detecting Previously Fact-Checked Claims in Social Media

Momchil Hardalov,[1] Anton Chernyavskiy,[2]
Ivan Koychev,[1] Dmitry Ilvovsky,[2] Preslav Nakov[3]

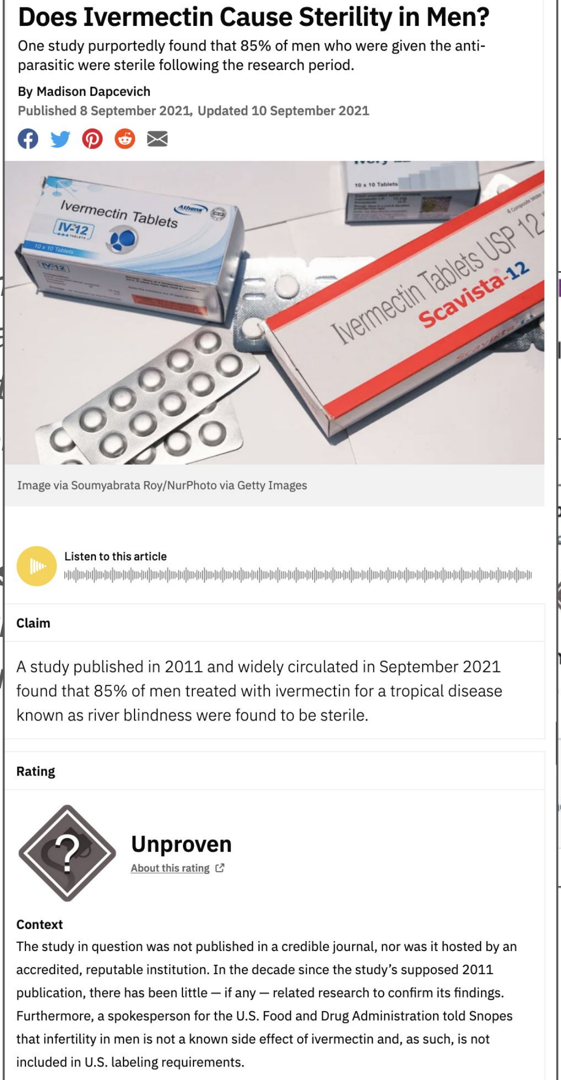
[1]Sofia University "St. Kliment Ohridski"
[2]HSE Univeristy
[3]Mohamed bin Zayed University of Artificial Intelligence

The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL-IJCNLP 2022)

Problem Definition

(Task Definition) *Given a* ***user comm***
detect whether *the* ***claim*** *it makes wa*
previously fact-checked *with respect*
collection of verified claims *and* ***the***
corresponding articles*.*

(Crowd Fact-Checker) *A* ***person on s***
media *who posts a* ***fact-checking arti***
reply *to a (potentially relevant)* ***claim i***
conversational thread.


Does Ivermectin Cause Sterility in Men?

One study purportedly found that 85% of men who were given the anti-parasitic were sterile following the research period.

By Madison Dapcevich

Published 8 September 2021, Updated 10 September 2021

Image via Soumyabrata Roy/NurPhoto via Getty Images

Listen to this article

Claim

A study published in 2011 and widely circulated in September 2021 found that 85% of men treated with ivermectin for a tropical disease known as river blindness were found to be sterile.

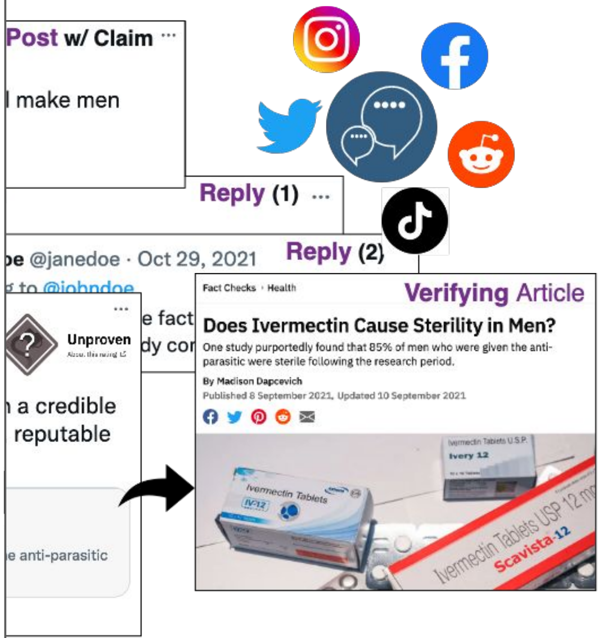
Rating

Unproven

About this rating

Context

The study in question was not published in a credible journal, nor was it hosted by an accredited, reputable institution. In the decade since the study's supposed 2011 publication, there has been little — if any — related research to confirm its findings. Furthermore, a spokesperson for the U.S. Food and Drug Administration told Snopes that infertility in men is not a known side effect of ivermectin and, as such, is not included in U.S. labeling requirements.



Motivation

- **Leverage the Knowledge of the Crowd Fact-Checkers**
  - Prior work: mostly small datasets but manually annotated
  - People can fact-check by referring to previously written “credible” fact-checks
  - Collect large-scale datasets without the need of human-in-the-loop

Motivation

- **Leverage the Knowledge of the Crowd Fact-Checkers**
  - Prior work: mostly small datasets but manually annotated
  - People can fact-check by referring to previously written “credible” fact-checks
  - Collect large-scale datasets without the need of human-in-the-loop
- **Improving the Model Learning from Noisy Data**
  - Labeling with Distant Supervision
  - Loss modifications and model self-adaptation

Motivation

- **Leverage the Knowledge of the Crowd Fact-Checkers**
  - Prior work: mostly small datasets but manually annotated
  - People can fact-check by referring to previously written “credible” fact-checks
  - Collect large-scale datasets without the need of human-in-the-loop
- **Improving the Model Learning from Noisy Data**
  - Labeling with Distant Supervision
  - Loss modifications and model self-adaptation
- **Evaluate the Model Abilities**
  - Strategy for data mixing from multiple sources (e.g., manual vs. distant labeling)
  - Measure the impact of model architecture and data selection

Contributions

- **Large-scale collection** of **330,000** pairs of **tweets–fact-checking articles**
    - Covering **diverse topics** from conversations that span **four years**.

Contributions

- **Large-scale collection** of **330,000** pairs of **tweets–fact-checking articles**
  - Covering **diverse topics** from conversations that span **four years**.
- **Two distant supervision strategies** to **label** the dataset;
  - Used techniques that **do not need human supervision**

Contributions

- **Large-scale collection** of **330,000** pairs of **tweets–fact-checking articles**
    - Covering **diverse topics** from conversations that span **four years**.
- **Two distant supervision strategies** to **label** the dataset;
    - Used techniques that **do not need human supervision**
- **Novel method** to learn from this data using **modified self-adaptive training**
    - Based on a **MNR loss**, **self-adaptive learning,** and **additional weighing**.

Contributions

- **Large-scale collection** of **330,000** pairs of **tweets–fact-checking articles**
    - Covering **diverse topics** from conversations that span **four years**.
- **Two distant supervision strategies** to **label** the dataset;
    - Used techniques that **do not need human supervision**
- **Novel method** to learn from this data using **modified self-adaptive training**
    - Based on a **MNR loss**, **self-adaptive learning,** and **additional weighing**.
- **Sizable improvements** over the **state of the art** on a standard test set.
    - **Our dataset** yields **better results** compared to **manually** annotated alternatives
    - **Proposed models** show **4% P@1, MRR, MAP@5** gains over **strong baselines**.
    - We achieve **2% improvement** over the **current state of the art**.

CrowdChecked: Newly Collected Dataset

- **Collected from Twitter**
    - All **replies** or **quote** tweets that **contain a link to a fact-check** (Snopes)
    - From October 2017 till October 2021

CrowdChecked: Newly Collected Dataset

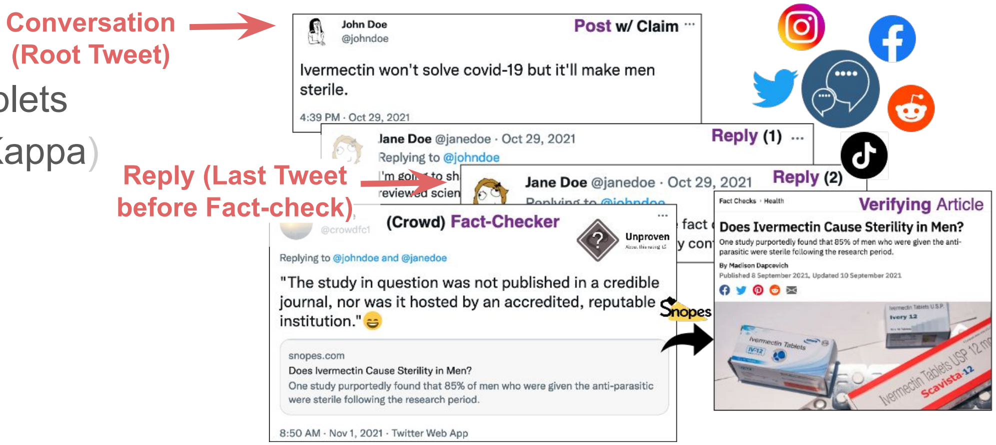
- Collected from Twitter
  - All **replies** or **quote** tweets that **contain a link to a fact-check** (Snopes)
  - From October 2017 till October 2021
- **Dataset size**
  - **330K** unique tweet–article pairs in English (collected)
    - The largest alternative contains 1.4K pairs (Shaar et al., 2021)
    - There are multimodal datasets w/ 19K pairs, 3K articles (Vo and Lee 2019)
  - **10K** unique fact-checking articles.

CrowdChecked: Newly Collected Dataset

- Collected from Twitter
  - All **replies** or **quote** tweets that **contain a link to a fact-check** (Snopes)
  - From October 2017 till October 2021
- Dataset size
  - **333K** unique tweet–article pairs in English (collected)
    - The largest alternative contains 1.4K pairs (Shaar et al., 2021)
    - There are multimodal datasets w/ 19K pairs, 3K articles (Vo and Lee 2019)
  - **10K** unique fact-checking articles.
- **Data Labeling (w/ Distant Supervision)**
  - Two labeling strategies:
    - **Jaccard Similarity** (5K–27K “correct” pairs)
    - **Semi-Supervision** (3.5K–49K “correct” pairs)
  - **Performed manual annotations** to estimate the **quality at each threshold**

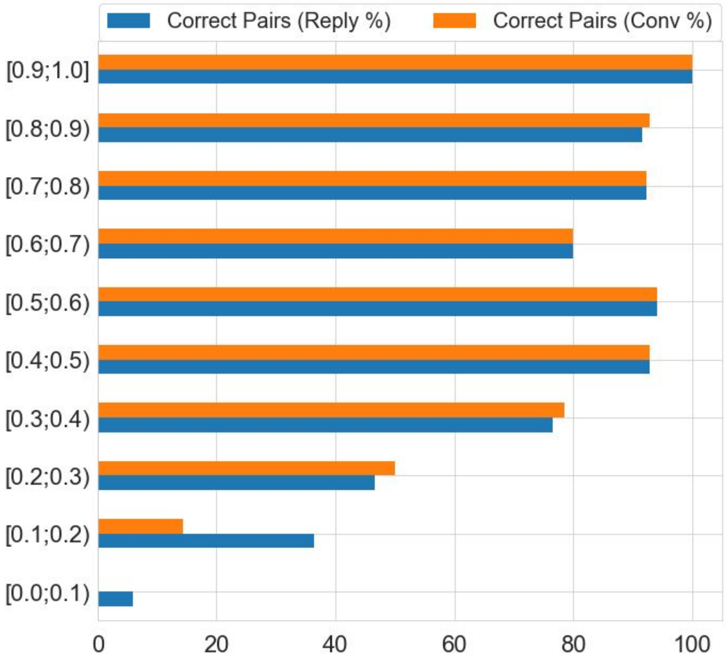
Data Labeling Quality

- **Quality Estimation**
  - 3 annotators, 150 conv–reply–tweet triplets
  - *Good* level of agreement (0.75 Fleiss Kappa)



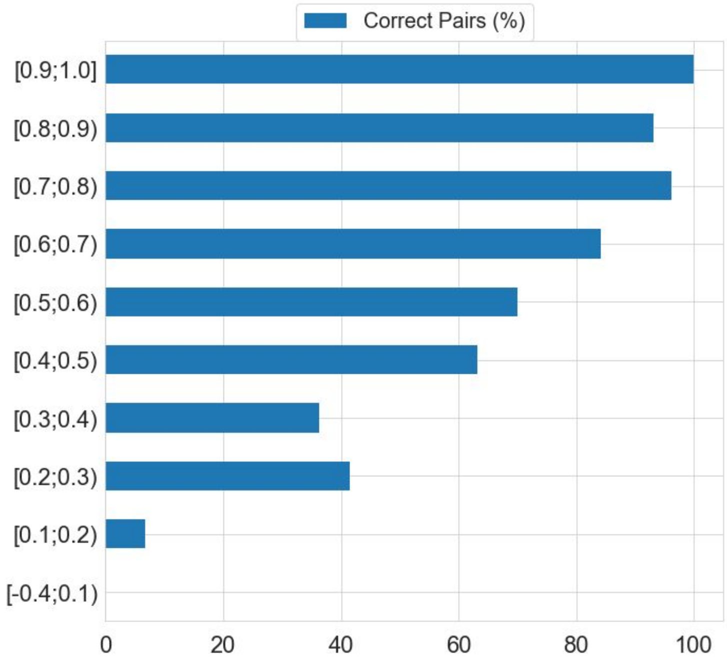
Data Labeling Quality

- Quality Estimation
  - 3 annotators, 150 conv–reply–tweet triplets
  - *Good* level of agreement (0.75 Fleiss Kappa)
- **Jaccard Similarity** (5K–27K “correct” pairs)
  - Simple, yet effective, finds diverse examples
  - Tweets and claims are pre-processed
  - Mean similarity – claim vs. article “title” and “subtitle”



Data Labeling Quality

- Quality Estimation
  - 3 annotators, 150 conv–reply–tweet triplets
  - *Good* level of agreement (0.75 Fleiss Kappa).
- Jaccard Similarity (5K–27K “correct” pairs)
  - Simple, yet effective, finds diverse examples
  - Tweets and claims are pre-processed
  - Mean similarity – claim vs. article “title” and “subtitle”
- **Semi-Supervised** (3.5K–49K “correct” pairs)
  - Based on the predictions of a Sentence-BERT
    - cosine similarity
  - Includes multiple fields in the article encoding
  - Finds examples similar to the fine-tuning dataset
    - less difficult

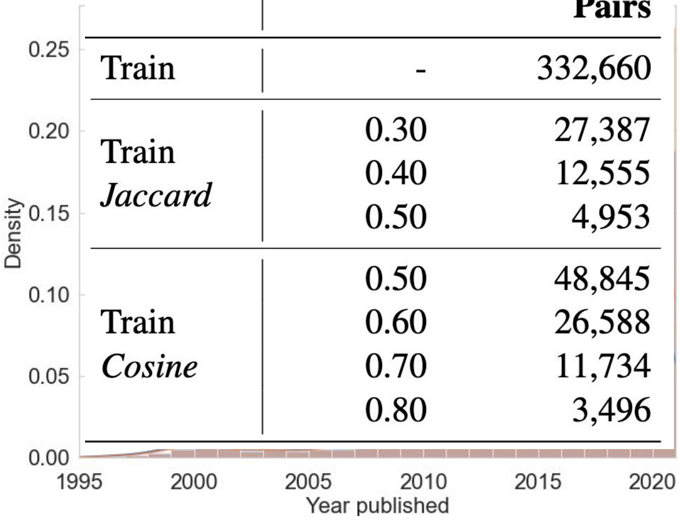


Datasets and Comparison

- CheckThat ’21 (CT) at CLEF (Shaar et al., 2021)
  - **Manually** annotated
  - Contains **1.4K English** examples (1,000 train, 200 dev/test)
  - Used **for training** and **evaluation**
  - **9K unique words** (tweets), **13.8K articles**

Datasets and Comparison

- CheckThat ’21 (CT) at CLEF (Shaar et al., 2021)
    - **Manually** annotated
    - Contains **1.4K English** examples (1,000 train, 200 dev/test)
    - Used **for training** and **evaluation**
    - **9K unique words** (tweets), **13.8K articles**
- **CrowdChecked (Ours)**
    - Labeled w/ **distant supervision**
    - 7 sets of size **3.5K–49K** (**threshold** based, **English)**
    - used **only for training**
    - **114,727 unique words** (all tweets), **10K articles**
    - claims (tweets) have **similar length** to **CT**
    - **8K common** fact-checking **articles** with **CT**

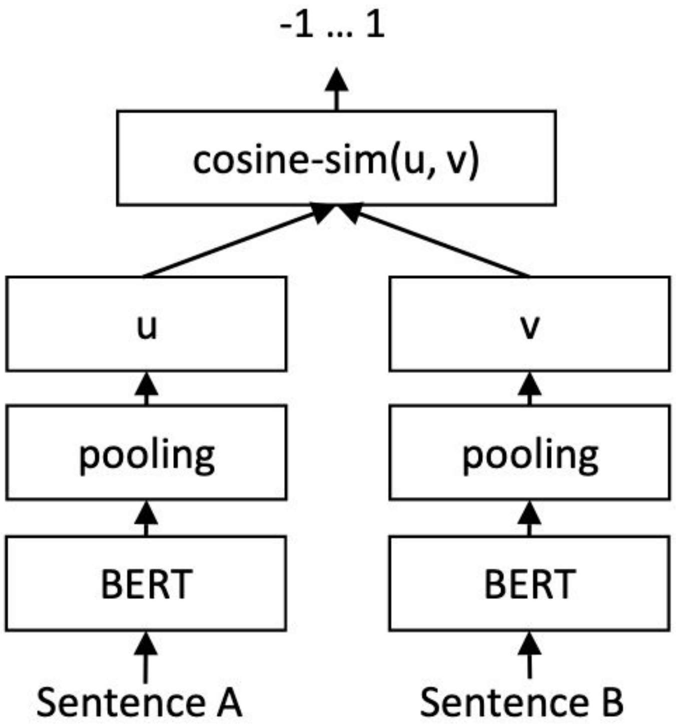


Data Split	Threshold	Tweet-Article Pairs
Train	-	332,660
Train *Jaccard*	0.30	27,387
	0.40	12,555
	0.50	4,953
Train *Cosine*	0.50	48,845
	0.60	26,588
	0.70	11,734
	0.80	3,496

Method Overview

Key Characteristics (Pipeline for Detecting Previously Fact-Checked Claims)

- **General scheme**: Sentence-BERT for semantic matching[1]
- Multiple Negatives Ranking loss[2]
  - shuffling
  - temperature
- Enriched scheme:
  - SBERT, TF.IDF, and Re-ranking[3]
- Training w/ noisy data
  - Self-adaptive training[4]
  - Loss weighting

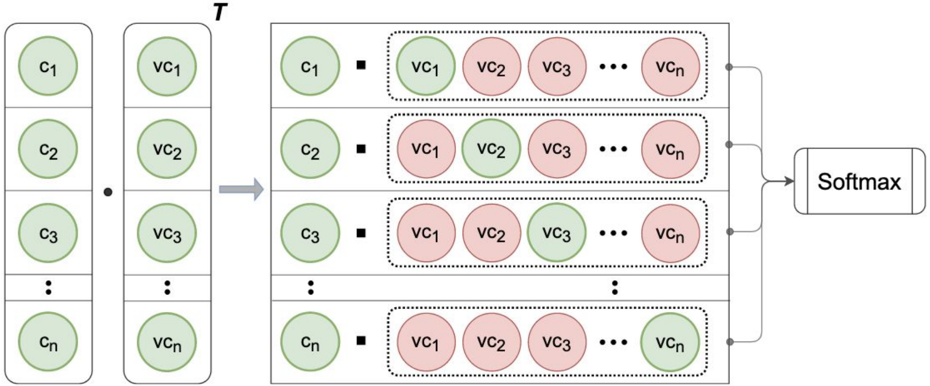


[1] https://www.aclweb.org/anthology/D19-1410.pdf
[2] https://aclanthology.org/2022.naacl-main.9.pdf
[3] http://ceur-ws.org/Vol-2936/paper-38.pdf
[4] https://proceedings.neurips.cc/paper/2020/file/e0ab531ec312161511493b002f9be2ee-Paper.pdf

Method Overview

Key Characteristics (Pipeline for Detecting Previously Fact-Checked Claims)

- General scheme: Sentence-BERT for semantic matching[1]
- **Multiple Negatives Ranking loss**[2]
  - shuffling
  - temperature
- Enriched scheme:
  - SBERT, TF.IDF, and Re-ranking[3]
- Training w/ noisy data
  - Self-adaptive training[4]
  - Loss weighting

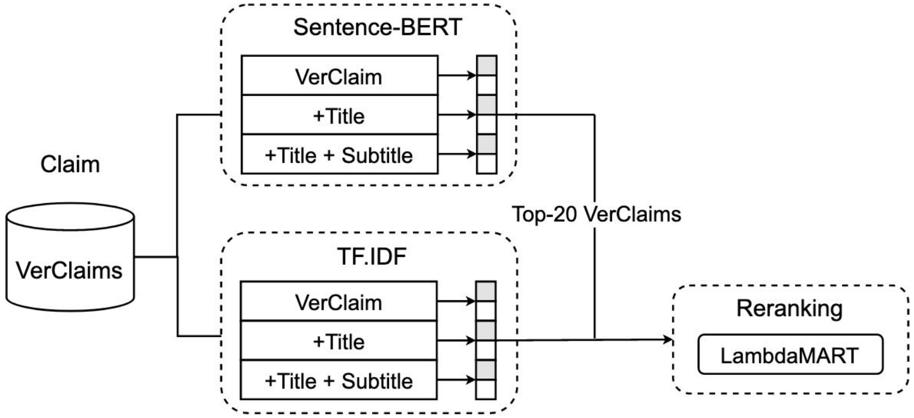


[1] https://www.aclweb.org/anthology/D19-1410.pdf
[2] https://aclanthology.org/2022.naacl-main.9.pdf
[3] http://ceur-ws.org/Vol-2936/paper-38.pdf
[4] https://proceedings.neurips.cc/paper/2020/file/e0ab531ec312161511493b002f9be2ee-Paper.pdf

Method Overview

Key Characteristics (Pipeline for Detecting Previously Fact-Checked Claims)

- General scheme: Sentence-BERT for semantic matching[1]
- Multiple Negatives Ranking loss[2]
    - shuffling
    - temperature
- **Enriched scheme**:
    - SBERT, TF.IDF, and Re-ranking[3]
- Training w/ noisy data
    - Self-adaptive training[4]
    - Loss weighting



[1] https://www.aclweb.org/anthology/D19-1410.pdf
[2] https://aclanthology.org/2022.naacl-main.9.pdf
[3] http://ceur-ws.org/Vol-2936/paper-38.pdf
[4] https://proceedings.neurips.cc/paper/2020/file/e0ab531ec312161511493b002f9be2ee-Paper.pdf

Method Overview

Key Characteristics (Pipeline for Detecting Previously Fact-Checked Claims)

- General scheme: Sentence-BERT for semantic matching[1]
- Multiple Negatives Ranking loss[2]
  - shuffling
  - temperature
- Enriched scheme:
  - SBERT, TF.IDF, and Re-ranking[3]
- **Training w/ noisy data**
  - Self-adaptive training*[4]
  - Loss weighting**

$$y^r \leftarrow \alpha \cdot y^r + (1-\alpha) \cdot \hat{y},$$

$$\mathcal{L} = -\frac{1}{m}\sum_{i=1}^{m} {y^r}_i \left(\frac{c_i^T v_i}{\tau} - \log \sum_{j=1}^{m} \exp\left(\frac{c_i^T v_j}{\tau}\right)\right)$$

*where y^r is the refined label of the r^{th} example (initialized with the original label), α is a hyper-parameter, $\hat{y}$ is the model prediction.

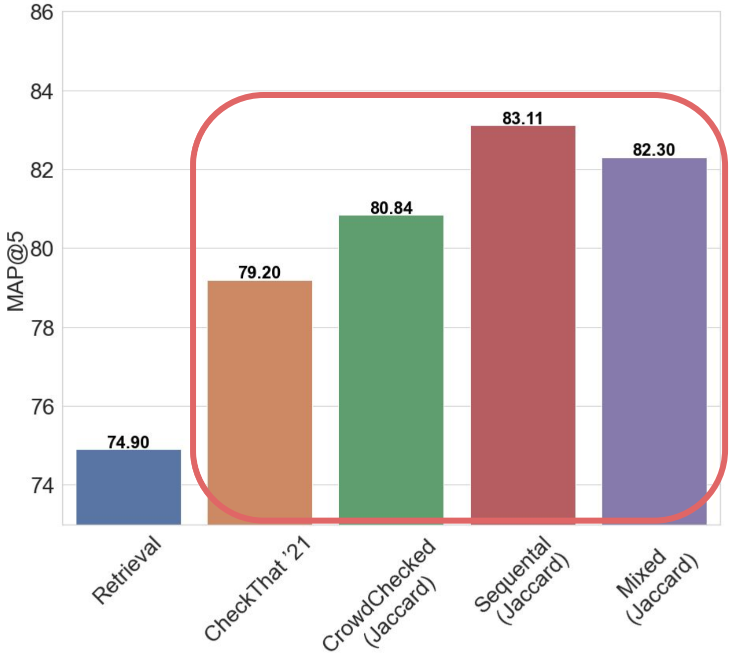
c and *v* are the claim and verifying article representations (MNR loss)

**y^r is squared (Huang et al. (2020) [4])

[1] https://www.aclweb.org/anthology/D19-1410.pdf
[2] https://aclanthology.org/2022.naacl-main.9.pdf
[3] http://ceur-ws.org/Vol-2936/paper-38.pdf
[4] https://proceedings.neurips.cc/paper/2020/file/e0ab531ec312161511493b002f9be2ee-Paper.pdf

Experimental Results

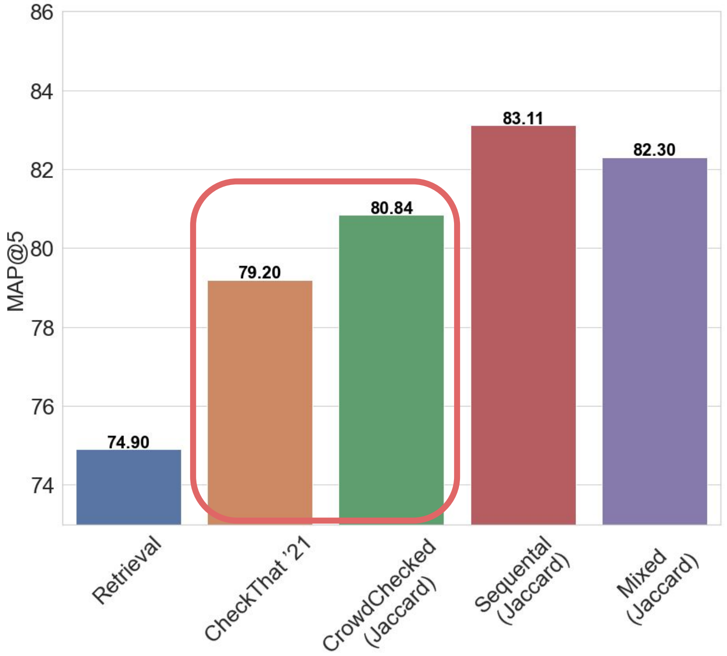
- **CrowdChecked vs. CheckThat! '21**
  - **General scheme** SBERT (better than IR)



CrowdChecked* sets are the largest from each strategy (Jaccard **27K, Cosine **49K**)

Experimental Results

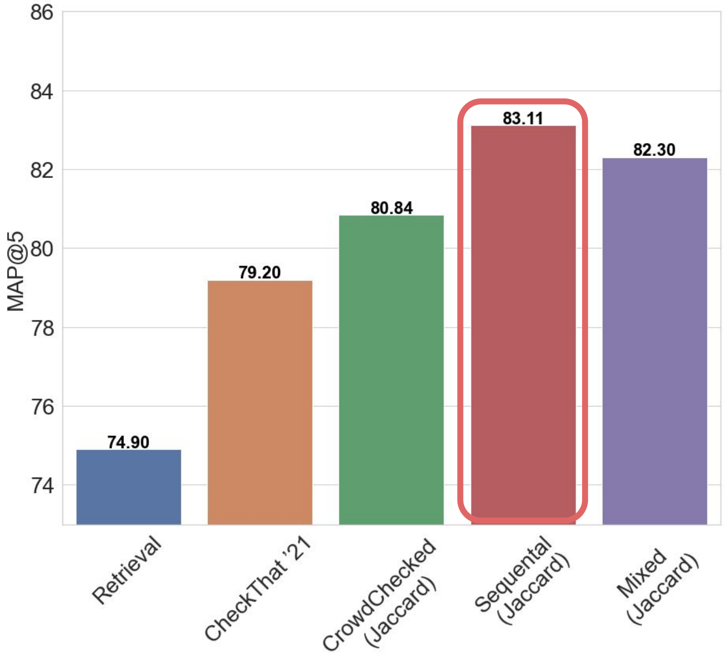
- **CrowdChecked vs. CheckThat! ‘21**
  - **General scheme** SBERT (better than IR)
  - CrowdChecked **outperforms** CheckThat



****CrowdChecked*** sets are the largest from each strategy (Jaccard **27K**, Cosine **49K**)

Experimental Results

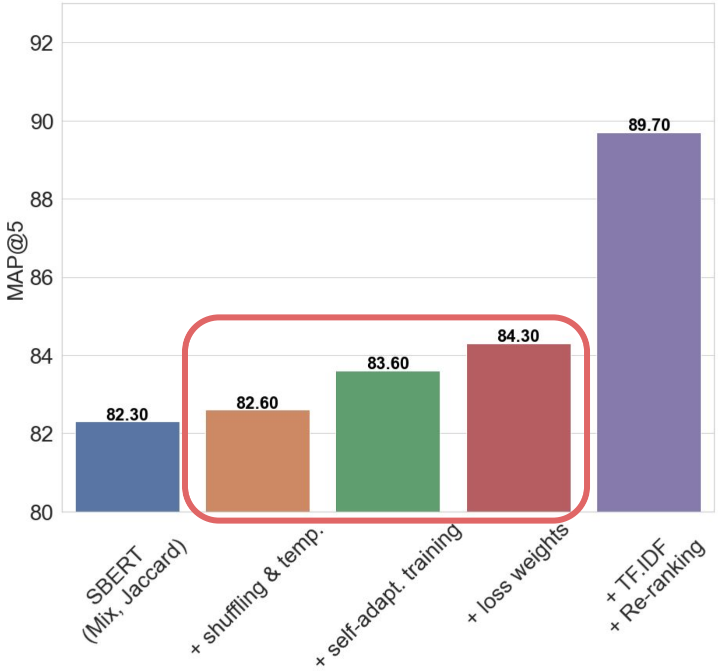
- **CrowdChecked vs. CheckThat! '21**
  - **General scheme** SBERT (better than IR)
  - CrowdChecked **outperforms** CheckThat
  - **Training sequentially** on the two datasets **yields the best results**



*****CrowdChecked*** sets are the largest from each strategy (Jaccard **27K**, Cosine **49K**)

Experimental Results

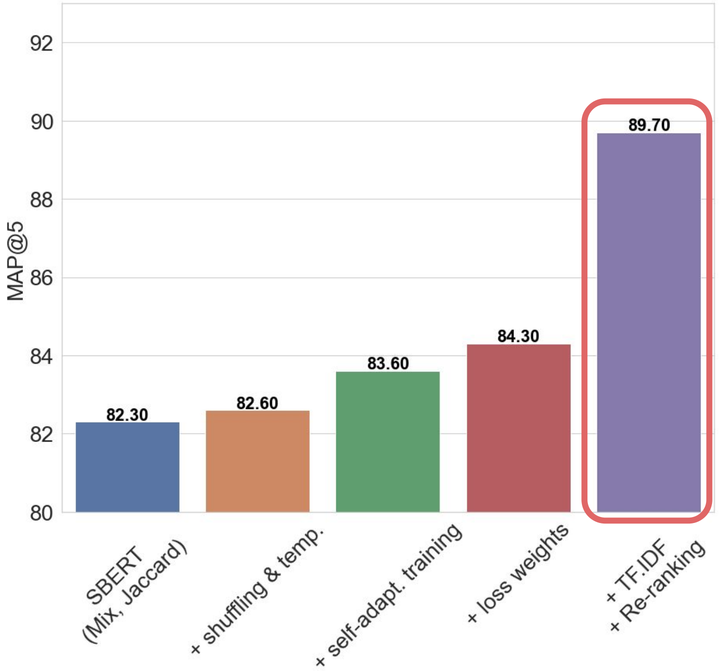
- CrowdChecked vs. CheckThat! ‘21
  - **General scheme** SBERT (better than IR)
  - CrowdChecked **outperforms** CheckThat
  - **Training sequentially** on the two datasets **yields the best results**
- **Model component analysis**
  - Pipeline components’ **contribution** (total of **2 points MAP@5**)



****CrowdChecked*** sets are the largest from each strategy (Jaccard **27K**, Cosine **49K**)

Experimental Results

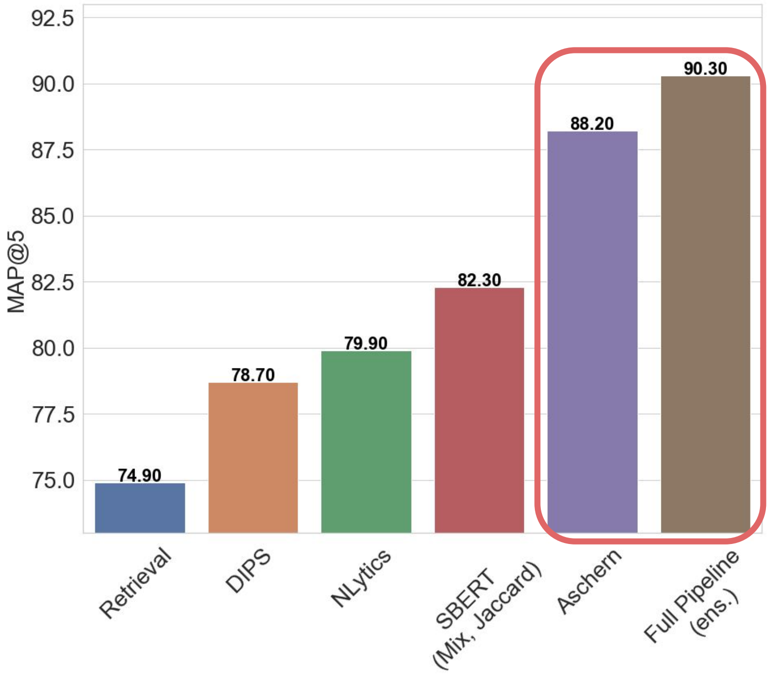
- CrowdChecked vs. CheckThat! ‘21
    - **General scheme** SBERT (better than IR)
    - CrowdChecked **outperforms** CheckThat
    - **Training sequentially** on the two datasets **yields the best results**
- **Model component analysis**
    - Pipeline components’ **contribution** (total of **2 points MAP@5**)
    - **Enriched Scheme** adds **+5 points**



****CrowdChecked*** sets are the largest from each strategy (Jaccard **27K**, Cosine **49K**)

Experimental Results

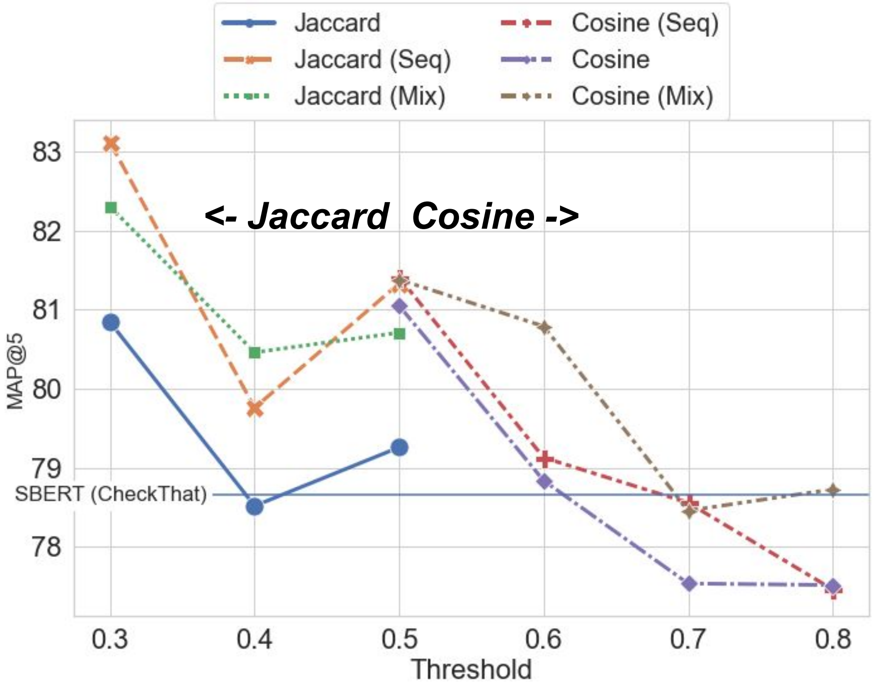
- CrowdChecked vs. CheckThat! '21
    - **General scheme** SBERT (better than IR)
    - CrowdChecked **outperforms** CheckThat
    - **Training sequentially** on the two datasets **yields the best results**
- Model component analysis
    - Pipeline components' **contribution** (total of **2 points MAP@5**)
    - **Enriched Scheme** adds **+5 points**
- **State-of-the-art comparison**
    - The **ensemble** adds **+0.6 point**
    - **SOTA** results **+2 points MAP@5**



CrowdChecked* sets are the largest from each strategy (Jaccard **27K, Cosine **49K**)

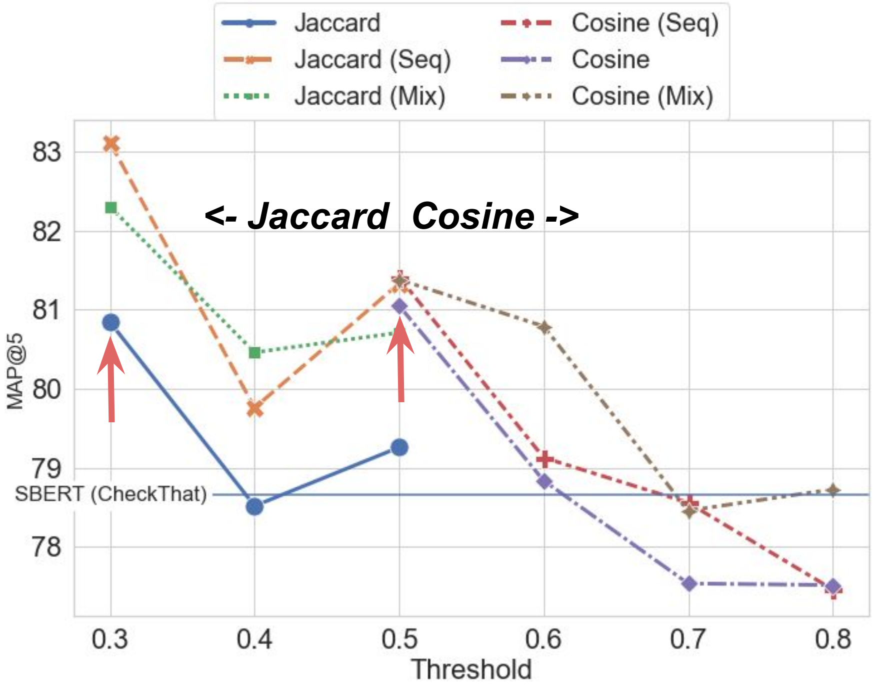
Discussion

- **Labeling function and threshold**



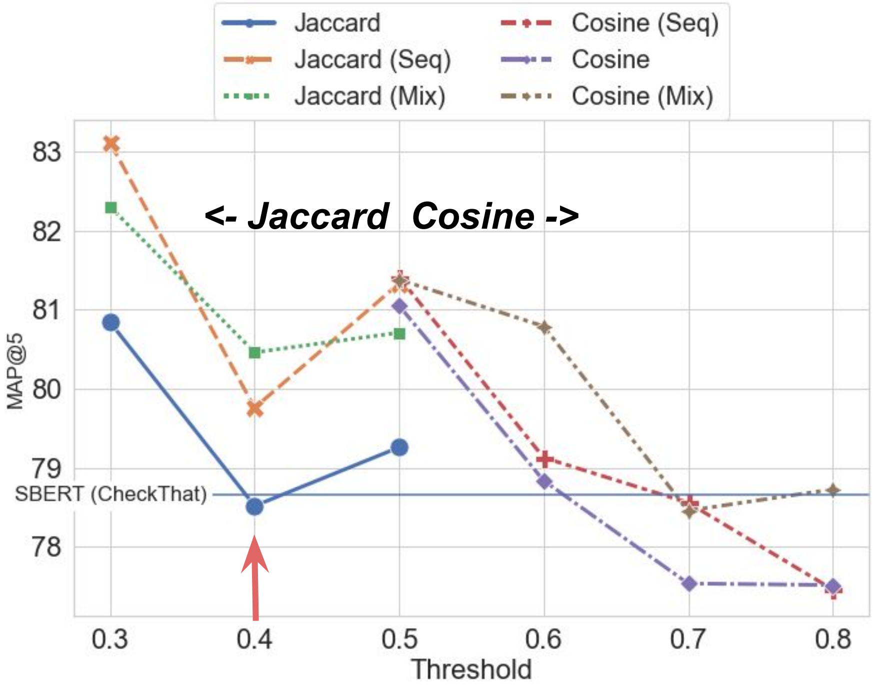
Discussion

- **Labeling function and threshold**
  - **Lower threshold** leads to **higher MAP**



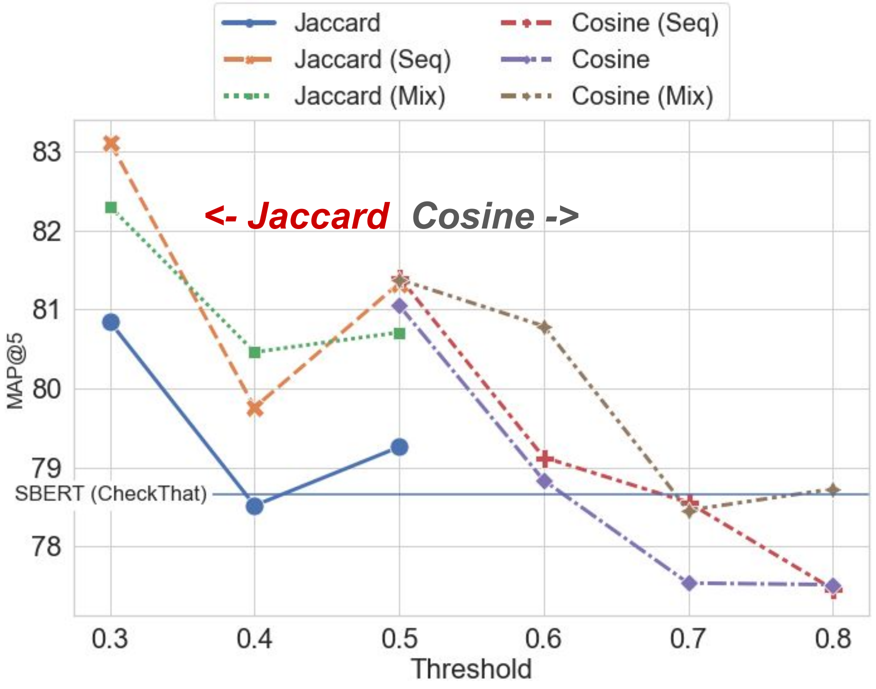
Discussion

- **Labeling function and threshold**
  - **Lower threshold** leads to **higher MAP**
  - There are **low-precision buckets**



Discussion

- **Labeling function and threshold**
  - **Lower threshold** leads to **higher MAP**
  - There are **low-precision buckets**
  - **Jaccard** outperforms Cosine



Discussion

- Labeling function and threshold
  - **Lower threshold** leads to **higher MAP**
  - There are **low-precision buckets**
  - **Jaccard** outperforms Cosine
- **Estimating the total correct pairs**

Dataset	**Data Split**	**Threshold**	**Tweet-Article Pairs** Found w/ Distant Supervision
	Train		
CrowdChecked (Our Dataset)	Train *Jaccard*	0.30	27,387
		0.40	12,555
		0.50	4,953
	Train *Cosine*	0.50	48,845
		0.60	26,588
		0.70	11,734
		0.80	3,496

Discussion

- Labeling function and threshold
  - **Lower threshold** leads to **higher MAP**
  - There are **low-precision buckets**
  - **Jaccard** outperforms Cosine
- **Estimating the total correct pairs**
  - Based on the manual annotations
    (*150 conversation–reply–tweet triplets*)

Dataset	Data Split	Threshold	Tweet-Article Pairs
CrowdChecked (Our Dataset)	Train	-	332,660
	Train *Jaccard*	0.30	27,387
		0.40	12,555
		0.50	4,953
	Train *Cosine*	0.50	48,845
		0.60	26,588
		0.70	11,734
		0.80	3,496

Discussion

- Labeling function and threshold
    - **Lower threshold** leads to **higher MAP**
    - There are **low-precision buckets**
    - **Jaccard** outperforms Cosine
- **Estimating the total correct pairs**
    - Based on the manual annotations
      (*150 conversation–reply–tweet triplets*)
    - **Jaccard: 61,500** (Expectation)
    - ***Cosine*: 90,170** (Expectation)

Dataset	Data Split	Threshold	Tweet-Article Pairs
CrowdChecked (Our Dataset)	Train	-	332,660
	Train *Jaccard*	0.30	27,387
		Estimated: 61K	[illegible]555
		0.50	4,953
	Train *Cosine*	0.50	48,845
		0.60	26,588
		Estimated: 90K	11,734
		0.80	3,496

Summary and Future Work

Summary

- We presented ***CrowdChecked***, a large dataset for detecting previously fact-checked claims
- We collected **330K pairs** of tweets and fact-checking articles form **crowd fact-checkers**
- We investigated two **techniques** for **labeling the data** using **distance supervision**
- We proposed a novel **approach for training from noisy data**
- We demonstrated that our data **yields sizable performance gains** over strong baselines
- We achieved **state-of-the-art results** using ***CrowdChecked*** and the **proposed pipeline**

Summary and Future Work

Summary

- We presented ***CrowdChecked***, a large dataset for detecting previously fact-checked claims
- We collected **330K pairs** of tweets and fact-checking articles form **crowd fact-checkers**
- We investigated two **techniques** for **labeling the data** using **distance supervision**
- We proposed a novel **approach for training from noisy data**
- We demonstrated that our data **yields sizable performance gains** over strong baselines
- We achieved **state-of-the-art results** using ***CrowdChecked*** and the **proposed pipeline**

Future Work

- Experiment with more languages
- Evaluate other distant supervision techniques, e.g., predictions from an ensemble model
- Integrate the “incorrect” pairs into the model training

Download our dataset, and train new models!

https://github.com/mhardalov/crowdchecked-claims

If you have more questions, please contact hardalov@fmi.uni-sofia.bg

Thank You for Listening!

Please check out our paper for more details:

“CrowdChecked: Detecting Previously Fact-Checked Claims in Social Media”