Summer 2022

An Assessment of Scientific Claim Verification Frameworks: Final Presentation

Ethan Landers
Old Dominion University, eland007@odu.edu

Jian Wu (Mentor)
Old Dominion University, jwu@cs.odu.edu

An Assessment of Scientific Claim Verification Frameworks: Final Presentation

NSF CS REU @ ODU, Summer 2022

Ethan Landers
Twitter @ethanlanders_

Mentor: Jian Wu
Twitter @fanchyna

Introduction

- The spread of scientific disinformation is on the rise
- Automate scientific claim verification
- The state-of-the-art model: MultiVerS (Wadden et al. 2021) trained and evaluated on SciFact (Wadden et al. 2020)
- Scientific questions:
  - How does the model trained on SciFact do with detecting open-domain scientific disinformation?
  - How does the model trained on other datasets do with detecting open-domain scientific disinformation?

Research Goals

1. Replicate results of the SOTA scientific claim verification model listed within SciFact leaderboard: MultiVerS
2. Create an open-domain scientific claims dataset for testing the model generalizability
    a. Model is good at detecting COVID-19 misinformation
3. New dataset will be tested by MultiVerS
    a. Labeling: whether the model can verify whether the claim is true or not
    b. Rationales: whether the model can identify sentences supporting or conflicting the given claim
4. Analyze results (prediction, recall, F1)

How MultiVerS Works

- Use the Long-former model (Beltagy et al., 2020) to encode both claim and its context
- Multitask rationale selection and label prediction: doing them both to more effectively use training data
- Use Vert5Erini (Pradeep et al., 2021) to select candidate abstract

How MultiVerS Works

- Training datasets

Dataset	Domain	Train claims
SciFact (Wadden 2020)	Biomedical	1109
COVIDFact (Saakyan et al. 2021)	Covid	903
HealthVer (Sarrouti et al., 2021)	Covid	1622
Fever (Thorne et al. 2018)	Wikipedia	130,644

- Multitask rationale selection and label prediction (Wadden, 2022)
  - Softmax Score
    - > 0.5: SUPPORTS/REFUTES label
    - < 0.5: NEI predicted label

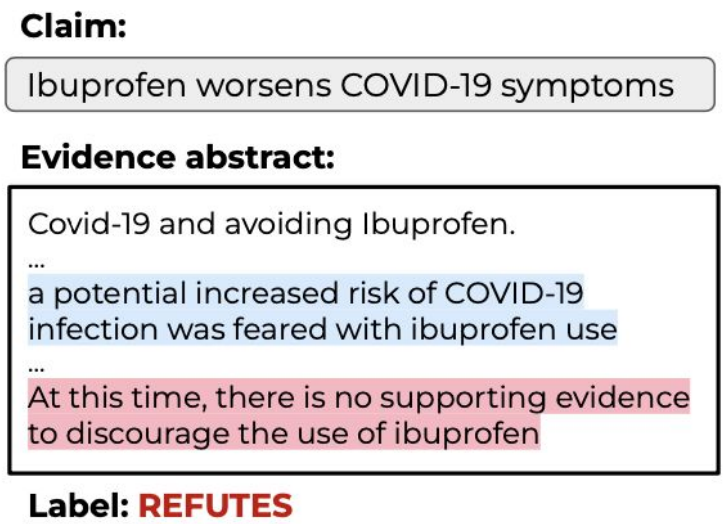


Figure 1: A claim from the HealthVer dataset, refuted by a research abstract. The sentence in red is a *rationale* reporting a finding that REFUTES the claim. However, this finding cannot be interpreted properly without the context in blue, which specifies that the finding applies to Ibuprofen as a treatment for COVID. MULTIVERS incorporates the full context of the evidence-containing abstract when predicting fact-checking labels.

Twitter @ethanlanders_ 08/04/2022

Procedures

- Create a new dataset using claims from Snopes.com
  - 61 instances, each containing a claim, at least one relevant paper (DOI)
  - Open domain, 18 true claims, 40 false claims, 3 mixture claims
- Preprocessing
  - Tokenize sentences in paper abstracts
    - NLTK
  - Generate compatible files in JSON format
- Label tokenized abstract sentences
  - Double independent labeling (with another undergraduate student: Dominik Soos)
  - Consensus rate: 75%
- Run MultiVerS model with the new dataset

claimid	claim	doi	link	abstitle	abssentid	abssent	labels
1	A photograph shows	10.1177/174387211	https://www.snopes.	Law, Scale, Anti-zoc	1	This article uses “Cc	N
1	A photograph shows	10.1177/174387211	https://www.snopes.	Law, Scale, Anti-zoc	2	It investigates what	N
1	A photograph shows	10.1177/174387211	https://www.snopes.	Law, Scale, Anti-zoc	3	The article contests	N
1	A photograph shows	10.1177/174387211	https://www.snopes.	Law, Scale, Anti-zoc	4	It is only after an ex	N
1	A photograph shows	10.1177/174387211	https://www.snopes.	Law, Scale, Anti-zoc	5	A way forward is im	N
2	Populations of wild	10.1128/JVI.00083-2	https://www.snopes.	Susceptibility of Whi	1	The origin of severe	N
2	Populations of wild	10.1128/JVI.00083-2	https://www.snopes.	Susceptibility of Whi	2	Current evidence su	N
2	Populations of wild	10.1128/JVI.00083-2	https://www.snopes.	Susceptibility of Whi	3	Understanding the h	N
2	Populations of wild	10.1128/JVI.00083-2	https://www.snopes.	Susceptibility of Whi	4	Here, we demonstra	S
2	Populations of wild	10.1128/JVI.00083-2	https://www.snopes.	Susceptibility of Whi	5	Intranasal inoculatio	S
2	Populations of wild	10.1128/JVI.00083-2	https://www.snopes.	Susceptibility of Whi	6	Notably, infected an	S
2	Populations of wild	10.1128/JVI.00083-2	https://www.snopes.	Susceptibility of Whi	7	Viral RNA was detec	S
2	Populations of wild	10.1128/JVI.00083-2	https://www.snopes.	Susceptibility of Whi	8	All inoculated and in	S
2	Populations of wild	10.1128/JVI.00083-2	https://www.snopes.	Susceptibility of Whi	9	The work provides in	S
6	Sniffing rosemary in	10.1177/204512531	https://www.snopes.	Plasma 1,8-cineole	1	The mode of influen	R
6	Sniffing rosemary in	10.1177/204512531	https://www.snopes.	Plasma 1,8-cineole	2	This study was desig	N
7	"Rocks falling into th	10.1126/science.115	https://www.snopes.	Long-Term Sea-Lev	1	Earth's long-term se	N
7	"Rocks falling into th	10.1126/science.115	https://www.snopes.	Long-Term Sea-Lev	2	However, published	N

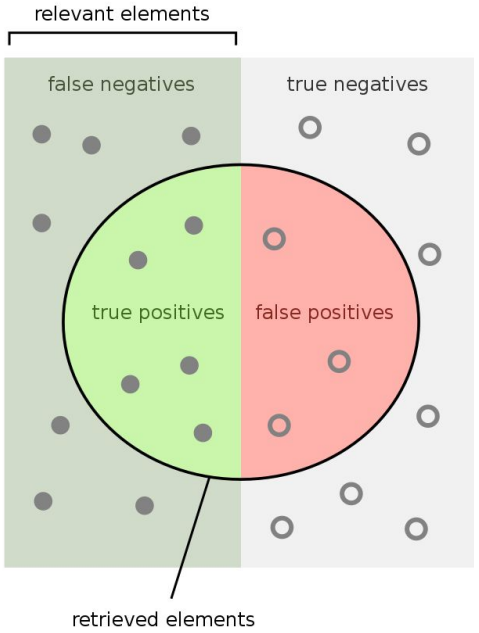
Evaluation Metric Meanings

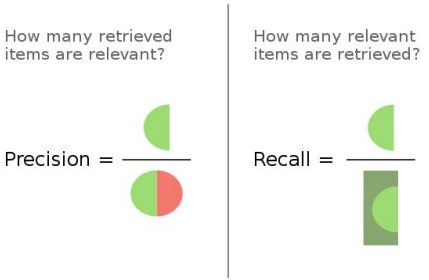
Precision is a calculation describing the number of true positives divided by all positives (false and true positives).

Recall is a calculation describing the number of true positives divided by both false negatives and true positives.

F-measure is calculated from both precision and recall, being a great accuracy measure for a test.

Abst means how accurate the predicted label was on the abstract level
Sent means how accurate the predicted label was on the abstract sentences level





(Wikipedia, 2022)

Twitter @ethanlanders_ 08/04/2022

MultiVerS Metrics on Different Checkpoints on My Data

	Fever	SciFact	Fever_Sci	HealthVer	CovidFact	SciFact Leaderboard
Abst (P)	0.0	1.0	1.0	0.22	0.43	0.74
Abst (R)	0.0	0.03	0.03	0.06	0.08	0.71
Abst (F1)	0.0	0.05	0.05	0.08	0.14	0.72
Sent (P)	0.0	0.0	0.0	0.0	0.29	0.75
Sent (R)	0.0	0.0	0.0	0.0	0.02	0.74
Sent (F1)	0.0	0.0	0.0	0.0	0.03	0.74

Twitter @ethanlanders_ 08/04/2022

Results

- All versions of MultiVerS overall did a poor job detecting disinformation within the new dataset
- Checkpoint trained on CovidFact achieves best abstract_label_only
- The MultiVerS model is sensitive to data domains
- Future work:
  - MultiVerS is still based on encoding the semantics of the text
  - It lacks background knowledge and inference

Works Cited

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." *arXiv preprint arXiv:2004.05150* (2020).

Pradeep, Ronak, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. "Scientific claim verification with VerT5erini." *arXiv preprint arXiv:2010.11930* (2020).

Saakyan, A., Chakrabarty, T., & Muresan, S. (2021). COVID-Fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Sarrouti, M., Abacha, A. B., M'rabet, Y., & Demner-Fushman, D. (2021, November). Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3499-3512).

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Wadden, D., Lo, K., Wang, L., Cohan, A., Beltagy, I., & Hajishirzi, H. (2022, July). MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 61-76).

Wikimedia Foundation. (2022, July 18). *F-score*. Wikipedia. Retrieved August 4, 2022, from https://en.wikipedia.org/wiki/F-score