MSVEC: A Multidomain Testing Dataset for Scientific Claim Verification

ACM MobiHoc 2023 REUNS Workshop

Michael Evans, Dominik Soós, Ethan Landers, Jian Wu

Michael Evans | mevan028@odu.edu | @mevansci

Acknowledgements

- This project is partially supported by National Science Foundation REU Site Award #2149607 and the Virginia Commonwealth Cyber Initiative Grant. Travel to the conference was supported by the NSF REU travel grant program.

- We acknowledge partial support from the Computer Science Department of Old Dominion University and the Old Dominion University Virginia Modeling, Analysis & Simulation Center.

Motivation

- Scientific news claims do not always faithfully report what is found in research papers.
- Available methods of scientific claim verification (SCV) are accompanied by limitations in domain adaptability and scalability.
- Language models (such as BERT) perform well on text classification tasks but poorly on scientific news verification (Soos, Landers & Wu, 2023).

Research Questions and Methods

- Can large language models verify scientific claims? If so, how well does it do in multiple domains using GPT-3.5 as a case study?
  - **Task 1:** Stance labeling. Determine whether an abstract supports/refutes a given claim.
  - **Task 2:** Sentence rationale. Identify sentence rationales as evidence of the stances.

- Future work: How well does GPT perform on SCV compared with humans?

Claim + Research Paper Pairs

Scientific News Claim

Use of Hand Sanitiser Can Seriously Mess With Breath Alcohol Test Results

HUMANS 22 November 2017 By SIGNE DEAN

(heller/Shutterstock)

Research Paper Abstract

This study was to determine if alcohol-based hand sanitizers (ABHSs) to the hands of a breath test operator will affect the breath alcohol instruments (EBTs)…A small, but significant, number (10%) resulted in positive breath alcohol concentrations, while (31.5%) resulted in a status code…EBT operators should forego the use of ABHS in the 15 min preceding subject testing.

Building the Dataset

Data Acquisition:

- MSVEC news claims were scraped from credible scientific news outlets or fact-checking websites, including Snopes.com, ScienceAlert.com, and Reuters.com. Webpages posted from 2014 to 2022 were crawled.

Stance Labeling:

- Scientific news containing URLs to scientific papers to back up the justification of the labels were manually selected.

Sentence Rationales:

- Scientific paper sentences were indexed and manually annotated by a computer science student.

Data Acquisition

***ScienceAlert** (Title)*

Specimens of Earth's Oldest Known Life Forms Have Been Discovered in Tasmania

NATURE 14 November 2017 By MICHELLE STARR

***Reuters** (First Paragraph)*

REUTERS FACT CHECK JULY 6, 2022 / 1:13 PM / A YEAR AGO

Fact Check-No evidence that U.S. schoolchildren are self-identifying as animals and disrupting classrooms

***Snopes** (Explicit HTML Claim)*

Claim:

> 3D-printed Sarco capsules for assisted suicides, which can be activated from the inside by the person intending to die, have been approved under Swiss law.

Rating:

False

About this rating

Overview of Scientific Claim Verification Datasets

Table 1: Comparison of MSVEC with existing SCV datasets.

Dataset	# Claims	Domains	Source	
SciFact-open	279	Biomedical	Research Papers	Wadden et al. (2022 ACL)
HealthVer	230	Covid	Web	Sarrouti et al. (2021)
Covid-Fact	46	Covid	Web + Generated	Saakyan et al. (2021)
MSVEC	200	Multiple	Fact-checking websites + Research Papers	

Our work

MSVEC Dataset Domain Diversity

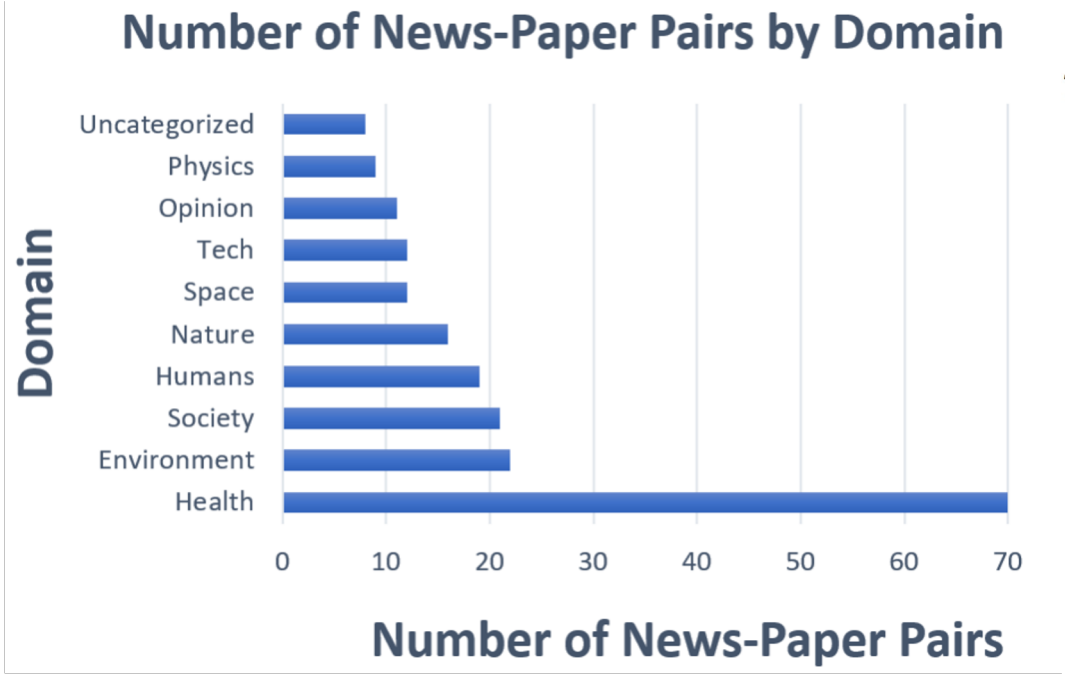

Table 3: The top 10 web domains of URLs linking to scientific papers.

Web Domain	# Papers	Percentage
onlinelibrary.wiley.com	49	24.5%
ncbi.nlm.nih.gov	23	11.5%
jamanetwork.com	18	9.00%
sciencedirect.com	18	9.00%
pnas.org	13	6.50%
pubs.acs.org	12	6.00%
tandfonline.com	9	4.50%
bmj.com	6	3.00%
link.springer.com	5	2.50%
science.org	3	1.50%

Experimental Setup

- **Model:** GPT-3.5-Turbo (Trained up to September 2021)
    - **Parameters / Tokens:** 175 billion / 4,096
- **Input:** News claim + research paper abstract
- **Hyperparameter:** Temperature for adjusting creativity in the model's responses from (0 to 1). Higher temperature = more creative.
- **Consensus:** Majority voting of 3 identical queries.

```python
def get_completion(prompt, model="gpt-3.5-turbo"):
    messages = [{"role": "user", "content": prompt}]
    response = openai.ChatCompletion.create(
        model=model,
        messages=messages,
        temperature=temp, # this is the degree of randomness of the model's output
    )
    return response.choices[0].message["content"]

for index, row in df.iterrows():
    claimIDs.append(row['id'])
    claims.append(row['claim'])
    abstracts.append(row['published_paper_abstract'])

print("Done with CSV read")

for index, row in df.iterrows():

    prompt = "Claim: " + claims[index] +"\nAbstract: " + abstracts[index] + "\n Ques

    if requests>0 and requests%20 == 0:
        print("Sleeping for 70 seconds...") # Can only make 20 calls a minute, sleep
        time.sleep(70)
        print("Continuing...currently on request " + str(requests))

    requests += 1

    responses.append(get_completion(prompt))

    df.at[index, 'GPT_Response_0.25'] = responses[index] # Name of column in output
```

Steps:

- Read CSV to dataframe
- Construct each prompt at current index
- Call get_completion function
- Sleep every 20 requests
- Emit responses (ex: SUPPORT, 900)

Stance Labeling Prompt and Response

Figure 1: An example stance labeling prompt and response.

Claim: Use of Hand Sanitiser Can Seriously Mess With Breath Alcohol Test Results

Abstract: This study was undertaken to determine if the application of alcohol-based hand sanitizers (ABHSs) to the hands of a breath test operator will affect the results obtained on evidential breath alcohol instruments (EBTs)…A small, but significant, number of initial analyses (13 of 130, 10%) resulted in positive breath alcohol concentrations…EBT operators should forego the use of ABHS in the 15 min preceding subject testing.

Question: Is the abstract relevant to the claim? Answer with one word and a number: SUPPORT if the abstract supports the claim, CONTRADICT if the abstract contradicts the claim or NEI if the abstract does not provide enough information about the claim to decide and a number on a scale of 0-1000 rate how relevant the abstract is to the claim.

Answer: SUPPORT, 900

GPT Query Responses

Task 1: Use majority voting to decide GPT's stances with respect to a news article

Temperature	Query 1	Query 2	Query 3	Voting Results
0.25	SUPPORT	SUPPORT	SUPPORT	SUPPORT
0.50	SUPPORT	SUPPORT	SUPPORT	SUPPORT
0.75	SUPPORT	SUPPORT	*CONTRADICT*	SUPPORT

Task 2: Identifying rationales that justify the stances

7240	Use of Hand Sanitiser	https://ww	Humans	22-Nov-17	TRUE	https://(	The Effe	Ellen Str	0. This study was undert	Journal of	2017	1	0
7240	Use of Hand Sanitiser	https://ww	Humans	22-Nov-17	TRUE	https://(	The Effe	Ellen Str	1. This study obtained b	Journal of	2017	1	1
7240	Use of Hand Sanitiser	https://ww	Humans	22-Nov-17	TRUE	https://(	The Effe	Ellen Str	2. A small, but significan	Journal of	2017	1	0
7240	Use of Hand Sanitiser	https://ww	Humans	22-Nov-17	TRUE	https://(	The Effe	Ellen Str	3. These status codes we	Journal of	2017	1	1
7240	Use of Hand Sanitiser	https://ww	Humans	22-Nov-17	TRUE	https://(	The Effe	Ellen Str	4. Replicate subject sam	Journal of	2017	1	0
7240	Use of Hand Sanitiser	https://ww	Humans	22-Nov-17	TRUE	https://(	The Effe	Ellen Str	5. As ABHS application c	Journal of	2017	1	1

GPT-3.5 Stance Labeling Consistency

Table 5: GPT-3.5 consistency of responses for each temperature out of 3 identical queries.

Temperature	Consistency	# Queries	Percentage
0.25	3/3	163	81.5%
0.25	2/3	37	18.5%
0.50	3/3	129	64.5%
0.50	2/3	71	35.5%
0.75	3/3	115	57.5%
0.75	2/3	85	42.5%

Evaluation Metrics

- **Precision:** The fraction of positives that were *actually* true.
- **Recall:** The fraction of true samples that are identified positive.
- **F1 Score:** Hamonic mean of precision and recall.
  $F1=2PR/(P+R)$

GPT-3.5 Stance Labeling Results

Table 4: GPT-3.5 stance labeling results. The best performance is highlighted in bold.

Temperature	Class	Precision	Recall	F1
0.25	**SUPPORT**	**0.902**	**0.490**	**0.635**
0.50	SUPPORT	0.900	0.477	0.623
0.75	SUPPORT	0.848	0.444	0.583
0.25	**CONTRADICT**	**0.347**	**0.837**	**0.491**
0.50	CONTRADICT	0.342	0.837	0.485
0.75	CONTRADICT	0.306	0.755	0.435

Support Class Results by Domain

Table 6: GPT-3.5 *SUPPORT* class stance labeling results by subject.

Domain	Size	Temperature	Precision	Recall	F1
Health	70	0.25	0.875	0.447	0.592
Environment	22	0.25 / 0.75	0.714	0.385	0.500
Society	21	0.50	0.917	0.647	0.759
Humans	19	0.75	0.857	0.462	0.600
Nature	16	0.50	1.000	0.563	0.720
Space	12	0.25	1.000	0.667	0.800
Tech	12	0.25	1.000	0.667	0.800
Opinion	11	0.25 / 0.50 / 0.75	1.000	0.333	0.500
Physics	9	0.75	1.000	0.429	0.600
Uncategorized	8	0.25 / 0.50	1.000	0.750	0.857

GPT-3.5 Stance Labeling Summary

- GPT performed the stance labeling task best at lower temperatures, achieving an F1 score of 0.635 and 0.491 for the support and contradict classes respectively.

- Domain-specific F1 scores ranged from 0.500 to 0.857 for the support class.

- GPT was more likely to answer “support” for a false news claim than to answer “contradict” when given a true news claim.

Sentence Rationale Prompt

Claim: In a Surprise Discovery, Engineers Have Turned a Laser Beam Into a Liquid Stream

Abstract:

0. Transforming a laser beam into a mass flow has been a challenge both scientifically and technologically.

1. We report the discovery of a new optofluidic principle and demonstrate the generation of a steady-state water flow by a pulsed laser beam through a glass window.

2. To generate a flow or stream in the same path as the refracted laser beam in pure water from an arbitrary spot on the window, we first fill a glass cuvette with an aqueous solution of Au nanoparticles.

[3 - 5]

6. The principle of this light-driven flow via ultrasound, that is, photoacoustic streaming by coupling photoacoustics to acoustic streaming, is general and can be applied to any liquid, opening up new research and applications in optofluidics as well as traditional photoacoustics and acoustic streaming.

Question: Which of the numbered sentences support the claim? Answer with only a list of numbers.

Answer: 1, 3, 4, 5, 6

Sentence Rationales Results for Support Class

Table 8: GPT-3.5 sentence rationale results. The best results are highlighted in bold.

Temperature	Class	Precision	Recall	F1
0.25	Rationale	0.792	0.444	0.569
0.50	Rationale	0.825	0.462	0.593
0.75	**Rationale**	**0.829**	**0.483**	**0.610**
0.25	Non-Rationale	0.282	0.652	0.394
0.50	Non-Rationale	0.306	0.708	0.428
0.75	**Non-Rationale**	**0.313**	**0.704**	**0.433**

***Results assume correctly identified rationales**

GPT-3.5 Sentence Rationales Summary

- GPT performed the sentence rationale task best at higher temperatures, achieving an F1 score of 0.610 for the rationale class and 0.433 for the non-rationale class.

Summary and Future Work

- Developed a multidomain testing dataset containing scientific claims from news articles with evidence papers and human-annotated rationales. Our dataset contains news claims from 10 domains and consists of 151 true and 49 false claim-paper pairings.

- Evaluated the performance of a zero-shot method with GPT-3.5 against the MSVEC dataset on two sub-tasks: stance labeling and identifying sentence rationales.

- Future work: Compare LLMs with humans; compare state-of-the-art LMs with LLMs; determine the model's bias towards supporting false claims.

References

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).

Wadden, D., Lo, K., Kuehl, B., Cohan, A., Beltagy, I., Wang, L. L., & Hajishirzi, H. (2022). SciFact-open: Towards open-domain scientific claim verification. arXiv preprint arXiv:2210.13777.