End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models

Barry Menglong Yao
University at Buffalo
myao2@buffalo.edu

Aditya Shah
Virginia Tech
aditya31@vt.edu

Lichao Sun
Lehigh University
lis221@lehigh.edu

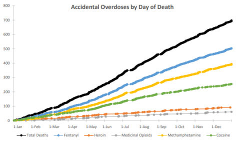
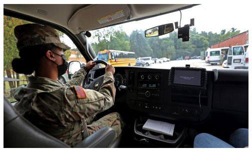
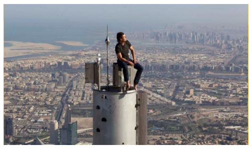
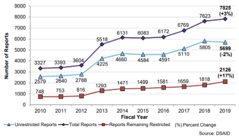
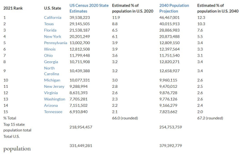
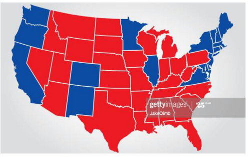
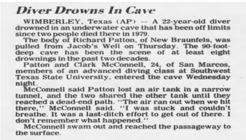
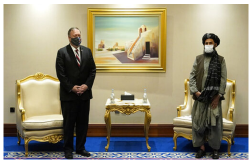
Jin-Hee Cho
Virginia Tech
jicho@vt.edu

Lifu Huang
Virginia Tech
lifuh@vt.edu

Submitted to the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks. Do not distribute.

A Supplementary Material

A.1 Experiment Analysis

Claim	Text Evidence	Image Evidence	Truthfulness
#1: San Francisco had twice as many drug overdose deaths as COVID deaths last year	That's more than twice San Francisco's 257 deaths due to COVID-19		***Supported***
#2: To address a shortage of school bus drivers in September 2021, Massachusetts Gov. Charlie Baker directed National Guard troops to help transport K-12 students to school	Governor Charlie Baker today will activate the Massachusetts National Guard in response to requests from local communities for assistance with school transportation as the 2021-2022 school year gets underway in the Commonwealth. Beginning with training on Tuesday, 90 Guard members will prepare for service in Chelsea, Lawrence, Lowell, and Lynn		***Supported***
#3: A photograph shows actor Tom Cruise sitting on top of the Burj Khalifa skyscraper without a harness	Special mounts had to be made for the 65 millmeter Imax cameras, special safety had to be put in place, because in a building that's 800 meteres tall [it's 2,723 feet] you couldn't run the risk of anything falling		***Supported***
#4: We had the highest number of (military) sexual assaults ever reported in the last year' and 'we had the lowest conviction rate and the lowest prosecution rate	The number of reported military sexual assaults increased in all but one year between 2010 and 2019, and the number reached a record in 2019		***Supported***
#5: By 2040, 70\% of the population is expected to live in just 15 states	That's more than twice San Francisco's 257 deaths due to COVID-19		***Supported***
#6: If you just count all the deaths in the red states, we are number two in the world in deaths, just behind Brazil	If it's a state that currently has a Republican governor, he's mostly accurate		***Supported***
#7: No One Realizes How Dangerous This Popular Vacation Spot in California Actually Is	Jacob's Well Natural Area remains a popular recreational destination today		***NEI***
#8: The man next to Mike Pompeo in a November 2020 photo 'is the guy the Trump administration helped get out of jail in 2018 and who is now the 'president' of Afghanistan	The U.S. envoy chosen by President Donald Trump, Zalmay Khalilzad, has publicly confirmed that he requested and secured the release of senior Taliban official Abdul Ghani Baradar from prison in Pakistan ahead of negotiations to end the war in Afghanistan		***Supported***

2021 Rank	U.S. State	US Census 2020 State Estimates	Estimated % of population in U.S. 2020	2040 Population Projection	Estimated % of population in U.S. 2040
1	California	39,538,223	11.9	46,467,001	12.3
2	Texas	29,145,505	8.8	40,015,913	10.3
3	Florida	21,538,187	6.5	28,886,983	7.6
4	New York	20,201,249	6.1	20,873,488	5.5
5	Pennsylvania	13,002,700	3.9	12,809,150	3.4
6	Illinois	12,812,508	3.9	12,397,564	3.3
7	Ohio	11,799,448	3.6	11,751,540	3.1
8	Georgia	10,711,908	3.2	12,820,271	3.4
9	North Carolina	10,439,388	3.2	12,658,927	3.4
10	Michigan	10,077,331	3.0	9,960,115	2.6
11	New Jersey	9,288,994	2.8	9,470,012	2.5
12	Virginia	8,631,393	2.6	9,876,728	2.6
13	Washington	7,705,281	2.3	9,776,126	2.6
14	Arizona	7,151,502	2.2	9,166,279	2.4
15	Tennessee	6,910,840	2.1	7,823,662	2.0
% Total			66.0 (rounded)		67.2 (rounded)
Top 15 state population total		218,954,457		254,753,759	
Total U.S. population		331,449,281		379,392,779	

Diver Drowns In Cave

WIMBERLEY, Texas (AP) — A 22-year-old diver drowned in an underwater cave that has been off limits since two people died there in 1979.

The body of Richard Patton, of New Braunfels, was pulled from Jacob's Well on Thursday. The 90-foot-deep cave has been the scene of at least eight drownings in the past two decades.

Patton and Clark McConnell, 24, of San Marcos, members of an advanced diving class at Southwest Texas State University, entered the cave Wednesday night.

McConnell said Patton lost an air tank in a narrow tunnel, and the two shared the other tank until they reached a dead-end path. "The air ran out when we hit there," McConnell said. "I was stuck and couldn't breathe. It was a last-ditch effort to get out of there. I don't remember what happened."

McConnell swam out and reached the passageway to the surface.

Figure 1: Examples of Multimodal Fact Checking

A.2 Access to the dataset

The MOCHEG dataset can be accessed from `http://nlplab1.cs.vt.edu/~menglong/project/multimodal/fact_checking/MOCHEG/dataset/`. Elementary code to process the data and run baseline experiments will be publicly available on the Github repository `https://github.com/VT-NLP/Mocheg`. The new version of our dataset will also be notified in this Github repository. The authors of this paper will ensure proper long-term maintenance and access to the dataset. The DOI is 10.5281/zenodo.6653771[1]. Structured metadata in the *schema.org*[2] format is accessed from our server [3]

A.3 Dataset Format

Our dataset is split into training, development, test subsets, and a collection of documents and images as the source of evidence:

1. Text collection named “Corpus3.csv”, which contains the articles as the sources for the text evidence retrieval task; Each entry stands for one document and consists of three key fields:
   - (a) relevant_document_id: The ID of the document in the text collection
   - (b) claim_id: The ID of the claim which is relevant to this document
   - (c) Origin Document: The document content. Its usage is as follows:
     - i. Input (collection) for the text evidence retrieval task
2. The image collection is saved in the “images” folder, which contains all images as the sources for the image evidence retrieval task. Each image is named in the format “@claim_id-@relevant_document_id-@img_id-@description”. Its usage is as follows:
   - (a) Input (collection) for the image evidence retrieval task
3. Training subset, saved in the “train” folder, which contains the following items:
   - (a) “Corpus2.csv”, which contains the claim, text evidence, truthfulness label for the claim verification task, and ruling outline, which explains the reasoning and ruling process and is used for the explanation generation task. Each entry stands for one piece of evidence. If there are multiple pieces of evidence for one claim, there will be multiple rows for this claim. In detail, it contains the following key fields:
     - i. Claim: The claim content we need to check the truthfulness. Its usage is as follows:
       - A. Input (query) for the evidence retrieval task
       - B. Input for the claim verification task
       - C. Input for the explanation generation task
     - ii. claim_id
     - iii. Evidence: One piece of text evidence that is relevant to this claim. It records the ground truth text evidence in the text evidence retrieval task. It can be retrieved from the text collection. Its usage is as follows:
       - A. Ground truth text for the text evidence retrieval task
       - B. Input for the claim verification task
       - C. Input for the explanation generation task
     - iv. evidence_id: The ID of the evidence
     - v. cleaned_truthfulness: The truthfulness label (i.e., *support*, *refute* and *not enough information*). Its usage is as follows:
       - A. Ground truth for the claim verification task
       - B. Input for the explanation generation task
     - vi. ruling_outline: It is a short paragraph to explain the reasoning and ruling process. Its usage is as follows:
       - A. Ground truth for the explanation generation task
     - vii. Origin: It is the ruling article on the fact-checking websites. The ruling_outline can be seen as the summarization of the Origin.
     - viii. Snopes URL: The url for the corresponding fact-checking article

[1] https://doi.org/10.5281/zenodo.6653771

[2] http://schema.org/

[3] http://nlplab1.cs.vt.edu/~menglong/project/multimodal/fact_checking/MOCHEG/homepage.html

   (b) "images" folder, which contains the image evidence that is relevant to the claims in the training subset. It records the ground truth image evidence in the image evidence retrieval task. They can be retrieved from the image collection. Each image is named in the format "@claim_id-proof-@img_id-@description". Its usage is as follows:
      i. Ground truth images for the image evidence retrieval task
      ii. Input for the claim verification task
      iii. Input for the explanation generation task
   (c) "text_evidence_qrels_sentence_level.csv". It records the ID of the ground truth sentence in the text evidence retrieval task. It is in the trec qrel[4] format with four fields:
      i. TOPIC: In our case, it is the claim id
      ii. ITERATION: Constant 0, no special meaning
      iii. DOCUMENT#: In our case, it is the corpus id which is in the format "@claim_id-@relevant_document_id-@sentence_id"
      iv. RELEVANCY: 1 for relevant and 0 for irrelevant
   (d) "text_evidence_qrels_article_level.csv". It records the ID of the ground truth article in the text evidence retrieval task. Its format is similar to the trec qrel format, and it has five fields:
      i. TOPIC: In our case, it is the claim id
      ii. ITERATION
      iii. DOCUMENT#: In our case, it is the relevant_document_id
      iv. RELEVANCY: 1 for relevant and 0 for irrelevant
      v. evidence_id: Since we have saved the ground truth text evidence in the "Corpus2.csv" in the training, development, and test datasets, we add the corresponding evidence_id here.
   (e) "img_evidence_qrels.csv". It records the ID of the ground truth image in the image evidence retrieval task. Its format is similar to the trec qrel format, and it has five fields:
      i. TOPIC: In our case, it is the claim id
      ii. ITERATION
      iii. DOCUMENT#: In our case, it is the image name in the image collection
      iv. RELEVANCY: 1 for relevant and 0 for irrelevant
      v. evidence_id: Since we have saved the ground truth image evidence in the "images" folder in the training, development, and test datasets, we add the corresponding image name here.

4. Development subset, saved in the "val" folder. The format is same with Training subset
5. Test subset, saved in the "test" folder. The format is same with Training subset
6. supplementary folder. This folder contains some objects which are optional for the dataset. All supplementary objects can be generated by the scripts in our Github repository, but the generation may take several hours. To make the process smooth, we include these side products in the dataset.
   (a) Corpus3_sentence_level.csv: We split the documents in the "Corpus3.csv" into sentence level and store them in this file. It has five fields:
      i. claim_id
      ii. relevant_document_id
      iii. paragraph_id: The ID for this sentence. Although this field is for just one sentence currently, it is called "paragraph_id" to support the future work where we can merge several sentences into one paragraph for our experiments.
      iv. corpus_id: It is in the format "@claim_id-relevant_document_id-@paragraph_id"
      v. paragraph: The sentence content.
   (b) img_corpus_emb.pkl: The embedding for the image collection, encoded by "clip-ViT-B-32" checkpoint [5].

[4] https://trec.nist.gov/data/qrels_eng/

[5] https://www.sbert.net/docs/pretrained_models.html

A.4 Intended use

The dataset can be used for end-to-end multimodal fact-checking and explanation generation task, where the system needs to sequentially or jointly perform all three sub-tasks, including *multimodal evidence retrieval*, *multimodal claim verification*, and *multimodal explanation generation*.

The dataset can also be used directly for these three sub-tasks separately.

The dataset can also be used in the unimodal setting, like text-only explanation generation.

A.5 Data Statement

We follow the data statement structure of Bender and Friedman (2018) to give additional insights into the dataset. The MOCHEG consists of 21,184 claims where each claim is annotated with a truthfulness label and ruling statement, with 43,148 text evidence and 15,373 image evidence. We describe the dataset construction process in Section 3 in our paper.

A.5.1 Curation Rationale

PolitiFact and Snopes are two widely used websites to fight against the spreading of misinformation, where journalists are asked to manually check and verify each claim and write a ruling article to share their judgment. Considering this, we use these two websites as the data sources and crawl all claims from these websites. We then remove some claims which do not contain evidence.

A.5.2 Language Variety

The content in our dataset is in US (en-US) mainstream Englishes.

A.5.3 Speaker Demographic

It is expected that most of the speakers speak English as a native language. Our data source focuses on political topics.

A.5.4 Annotator Demographic

The journalists in Politifact and Snopes provide the annotations. However, their personal information, like gender, and age, is not directly available on the websites.

A.5.5 Speech Situation

Generally, the claims are from online speeches, public statements, news articles, and social media platforms, such as Facebook, Twitter, Instagram, TikTok, and so on.

A.5.6 Content Characteristics

Our dataset is a multi-modal dataset with text and images.

A.6 Author Statement and Licensing

We bear all responsibility in case of violation of rights. Our dataset is licensed under the CC BY 4.0[6]. The associated codes to MOCHEG for data crawler and baseline are licensed under Apache License 2.0[7].

These data annotations incorporate material from Politifact and Snopes, which is licensed pursuant to the Politifact Copyright Policy[8] and Snopes Copyright Policy [9]. Our data crawler scripts are based on

[6] `https://creativecommons.org/licenses/by/4.0/`

[7] `https://www.apache.org/licenses/LICENSE-2.0`

[8] https://www.politifact.com/copyright/

[9] https://www.snopes.com/terms-and-conditions/

the conll2019-snopes-crawling repository [10], which is under Apache License 2.0. In our experiments, we applied information retrieval models[11] and text generation model[12], which are under Apache License 2.0. We referred to the controllable generation model[13] Lai et al. (2021), which is under MIT License[14].

A.7 Ethics Statement

We carefully follow the ethics guidelines [15] and have not found potential societal impacts so far. Our work can be used to fact-check and stop the spread of misinformation. Our dataset does not use features or label information about sensitive personally identifiable information, like individual names.

Since our dataset contains internet claims, some claims may be offensive. However, we crawl the articles from some reputational fact-checking websites, like Politifact and Snopes, to decrease the possibilities for offensive content.

A.8 Reproducible Result

All checkpoints are publicly available in the checkpoint folder [16]. The results in the paper can be reproduced with these checkpoints.

A.9 Experiment Details

The claim verification model is trained with 1 Quadro RTX 8000 for 7 hours. The explanation generation model is trained with 4 Quadro RTX 8000 for 8 hours.

References

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Thank you BART! Rewarding Pre-Trained Models Improves Formality Style Transfer. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2:484–494.

[10] https://github.com/UKPLab/conll2019-snopes-crawling

[11] https://github.com/UKPLab/sentence-transformers

[12] https://github.com/huggingface/transformers

[13] https://github.com/laihuiyuan/pre-trained-formality-transfer

[14] https://github.com/laihuiyuan/pre-trained-formality-transfer/blob/main/LICENSE

[15] https://neurips.cc/public/EthicsGuidelines

[16] http://nlplab1.cs.vt.edu/~menglong/project/multimodal/fact_checking/MOCHEG/checkpoint