A Dataset Statement

A.1 Data Example

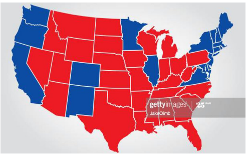
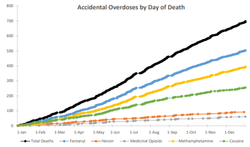
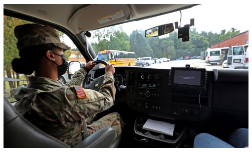
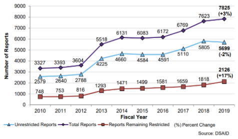
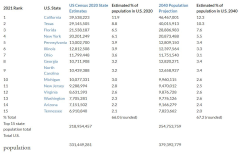
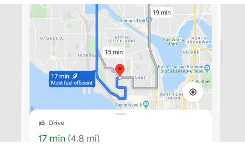
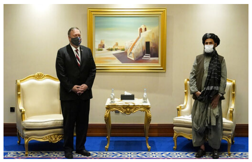
<table>
<tr><th>Claim</th><th>Text Evidence</th><th>Image Evidence</th><th>Truthfulness</th></tr>
<tr><td><b>#1</b>: A Boeing B-17E bomber from World War II was found in the jungle</td><td>The four-engine B-17E Flying Fortress was built by Boeing in November 1941, flew from California to Hawaii days after the Japanese attack on Pearl Harbor, and then island-hopped to Australia.</td><td> </td><td><i><b>Supported</b></i></td></tr>
<tr><td><b>#2</b>: If you just count all the deaths in the red states, we are number two in the world in deaths, just behind Brazil</td><td>If a red state isdefined as one that voted for Trump in 2016, he's spot-on. If it's a state that currently has a Republican governor, those red state death tolls would rank third in the world, not second.</td><td>
</td><td><i><b>Supported</b></i></td></tr>
<tr><td><b>#3</b>: San Francisco had twice as many drug overdose deaths as COVID deaths last year</td><td>That's more than twice San Francisco's 257 deaths due to COVID-19</td><td>
</td><td><i><b>Supported</b></i></td></tr>
<tr><td><b>#4</b>: 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</td><td>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</td><td></td><td><i><b>Supported</b></i></td></tr>
<tr><td><b>#5</b>: A photograph shows actor Tom Cruise sitting on top of the Burj Khalifa skyscraper without a harness</td><td>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</td><td> </td><td><i><b>Supported</b></i></td></tr>
<tr><td><b>#6</b>: 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</td><td>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</td><td>
</td><td><i><b>Supported</b></i></td></tr>
<tr><td><b>#7</b>: By 2040, 70\% of the population is expected to live in just 15 states</td><td>Recent census data from the University of Virginia’s Cooper Center shows that about 70 percent of the U.S. population will live in the 15 largest states in 2040.</td><td>
<table>
<tr><th>2021 Rank</th><th>U.S. State</th><th>US Census 2020 State Estimates</th><th>Estimated % of population in U.S. 2020</th><th>2040 Population Projection</th><th>Estimated % of population in U.S. 2040</th></tr>
<tr><td>1</td><td>California</td><td>39,538,223</td><td>11.9</td><td>46,467,001</td><td>12.3</td></tr>
<tr><td>2</td><td>Texas</td><td>29,145,505</td><td>8.8</td><td>40,015,913</td><td>10.3</td></tr>
<tr><td>3</td><td>Florida</td><td>21,538,187</td><td>6.5</td><td>28,886,983</td><td>7.6</td></tr>
<tr><td>4</td><td>New York</td><td>20,201,249</td><td>6.1</td><td>20,873,488</td><td>5.5</td></tr>
<tr><td>5</td><td>Pennsylvania</td><td>13,002,700</td><td>3.9</td><td>12,809,150</td><td>3.4</td></tr>
<tr><td>6</td><td>Illinois</td><td>12,812,508</td><td>3.9</td><td>12,397,564</td><td>3.3</td></tr>
<tr><td>7</td><td>Ohio</td><td>11,799,448</td><td>3.6</td><td>11,751,540</td><td>3.1</td></tr>
<tr><td>8</td><td>Georgia</td><td>10,711,908</td><td>3.2</td><td>12,820,271</td><td>3.4</td></tr>
<tr><td>9</td><td>North Carolina</td><td>10,439,388</td><td>3.2</td><td>12,658,927</td><td>3.4</td></tr>
<tr><td>10</td><td>Michigan</td><td>10,077,331</td><td>3.0</td><td>9,960,115</td><td>2.6</td></tr>
<tr><td>11</td><td>New Jersey</td><td>9,288,994</td><td>2.8</td><td>9,470,012</td><td>2.5</td></tr>
<tr><td>12</td><td>Virginia</td><td>8,631,393</td><td>2.6</td><td>9,876,728</td><td>2.6</td></tr>
<tr><td>13</td><td>Washington</td><td>7,705,281</td><td>2.3</td><td>9,776,126</td><td>2.6</td></tr>
<tr><td>14</td><td>Arizona</td><td>7,151,502</td><td>2.2</td><td>9,166,279</td><td>2.4</td></tr>
<tr><td>15</td><td>Tennessee</td><td>6,910,840</td><td>2.1</td><td>7,823,662</td><td>2.0</td></tr>
<tr><td>% Total</td><td></td><td></td><td>66.0 (rounded)</td><td></td><td>67.2 (rounded)</td></tr>
<tr><td>Top 15 state population total</td><td></td><td>218,954,457</td><td></td><td>254,753,759</td><td></td></tr>
<tr><td>Total U.S. population</td><td></td><td>331,449,281</td><td></td><td>379,392,779</td><td></td></tr>
</table></td><td><i><b>Supported</b></i></td></tr>
<tr><td><b>#8</b>: A planned update for Google Maps will change the app to no longer show the fastest routes by default.</td><td>Soon, Google Maps will default to the route with the lowest carbon footprint when it has approximately the same ETA as the fastest route.</td><td>
</td><td><i><b>Supported</b></i></td></tr>
<tr><td><b>#9</b>: 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</td><td>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</td><td></td><td><i><b>Supported</b></i></td></tr>
</table>

Figure 1: Examples of Multimodal Fact Checking

A.2 Access to the dataset

The MOCHEG dataset can be accessed from `https://doi.org/10.5281/zenodo.6653771`. Elementary code to process the data and run baseline experiments is 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:

[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`

            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

    (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.

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

    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].

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 15,601 claims where each claim is annotated with a truthfulness label and ruling statement, with 33,880 text evidence and 12,112 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 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.

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

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, and we have obtained permission from both Snopes and Politifact to publish the data for the research purpose. Our data crawler scripts are based on the conll2019-snopes-crawling repository [8], which is under Apache License 2.0. In our experiments, we applied information retrieval models[9] and text generation model[10], which are under Apache License 2.0. We referred to the controllable generation model[11] Lai et al. (2021), which is under MIT License[12].

A.7 Ethics Statement

We carefully follow the ethics guidelines [13] 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 possibility of offensive content.

On the other hand, our dataset contains 2,916 tweets, and we will share their Twitter IDs and scripts to crawl tweets based on Twitter API. In the case that Twitter delete some tweets in the future, dataset users can retrieve them from archive websites (e.g., Wayback machine[14]) or request them from us (according to Twitter policy[15], they permit sharing up to 50,000 hydrated Twitter content per recipient via non-automated means).

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.

[6] `https://creativecommons.org/licenses/by/4.0/`
[7] `https://www.apache.org/licenses/LICENSE-2.0`
[8] `https://github.com/UKPLab/conll2019-snopes-crawling`
[9] `https://github.com/UKPLab/sentence-transformers`
[10] `https://github.com/huggingface/transformers`
[11] `https://github.com/laihuiyuan/pre-trained-formality-transfer`
[12] `https://github.com/laihuiyuan/pre-trained-formality-transfer/blob/main/LICENSE`
[13] `https://neurips.cc/public/EthicsGuidelines`
[14] `https://archive.org/web`
[15] `https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases`