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

Barry Menglong Yao
Virginia Tech
Blacksburg, USA
barryyao@vt.edu

Aditya Shah
Virginia Tech
Blacksburg, USA
aditya31@vt.edu

Lichao Sun
Lehigh University
Bethlehem, USA
lis221@lehigh.edu

Jin-Hee Cho
Virginia Tech
Blacksburg, USA
jicho@vt.edu

Lifu Huang
Virginia Tech
Blacksburg, USA
lifuh@vt.edu

ABSTRACT

We propose end-to-end multimodal fact-checking and explanation generation, where the input is a claim and a large collection of web sources, including articles, images, videos, and tweets, and the goal is to assess the truthfulness of the claim by retrieving relevant evidence and predicting a truthfulness label (e.g., *support*, *refute* or *not enough information*), and to generate a statement to summarize and explain the reasoning and ruling process. To support this research, we construct Mocheg, a large-scale dataset consisting of 15,601 claims where each claim is annotated with a truthfulness label and a ruling statement, and 33,880 textual paragraphs and 12,112 images in total as evidence. To establish baseline performances on Mocheg, we experiment with several state-of-the-art neural architectures on the three pipelined subtasks: multimodal evidence retrieval, claim verification, and explanation generation, and demonstrate that the performance of the state-of-the-art end-to-end multimodal fact-checking does not provide satisfactory outcomes. To the best of our knowledge, we are the first to build the benchmark dataset and solutions for end-to-end multimodal fact-checking and explanation generation. The dataset, source code and model checkpoints are available at https://github.com/VT-NLP/Mocheg.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; *Natural language generation*; Computer vision; • **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

Multimodal Fact-Checking; Evidence Retrieval; Stance Detection; Explanation Generation; Explainable Fact-Checking

ACM Reference Format:
Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591879

SIGIR '23, July 23–27, 2023, Taipei, Taiwan
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591879

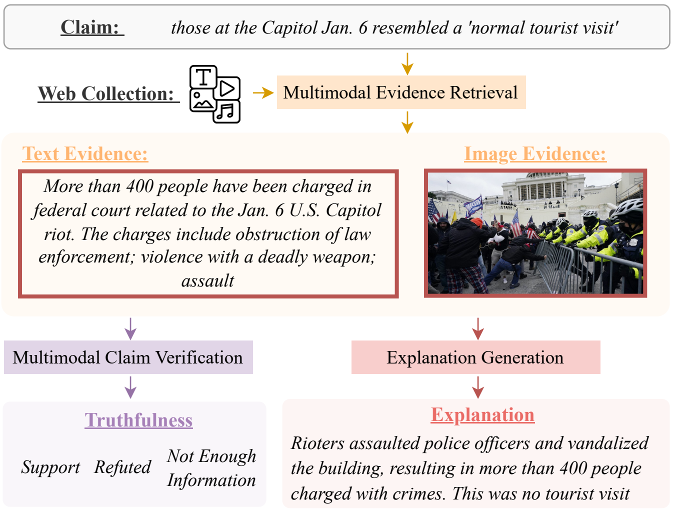

Claim: those at the Capitol Jan. 6 resembled a 'normal tourist visit'
Web Collection:
Multimodal Evidence Retrieval
Text Evidence:
More than 400 people have been charged in federal court related to the Jan. 6 U.S. Capitol riot. The charges include obstruction of law enforcement; violence with a deadly weapon; assault
Image Evidence:
Multimodal Claim Verification
Explanation Generation
Truthfulness
Support
Refuted
Not Enough Information
Explanation
Rioters assaulted police officers and vandalized the building, resulting in more than 400 people charged with crimes. This was no tourist visit

Figure 1: An example of end-to-end multimodal fact-checking and explanation generation.

1 INTRODUCTION

Misinformation has been a growing public concern in society and caused difficulty in finding reliable information online [15, 20]. For example, as Islam et al. [28] shows, the misinformation about COVID-19 has widely spread and led people to distrust medical treatment and even refuse to get vaccinated. The situation has become even more complicated with the emergence of large language models, like ChatGPT [46] since they could be intentionally misused to generate misinformation [21] or wrongly spread misinformation due to the hallucination issue [77]. To fight against misinformation, many fact-checking websites, such as *Snopes*[1] and *PolitiFact*[2], have been created where journalists manually collect thousands of claims from news and social media and verify them by referring to external reliable and relevant documents. However, it is time-consuming and hard to generalize to more broad claims.

Recently, researchers have started to investigate automatic misinformation detection and fact-checking by developing various

[1] https://www.snopes.com/
[2] https://www.politifact.com/

benchmark datasets [43, 47, 58, 64, 70] and start-of-the-art neural network architectures [38, 60, 63, 76]. However, we found the following limitations with the current fact-checking studies: (1) Most of them only consider text while ignoring the multi-media nature (e.g., images) of online articles, which are essential and useful to predict the truthfulness of claims. There are a few multimodal fact-checkinig datasets existing [1, 41, 45], however, their truthfulness labels [41] or evidence [1, 45] are automatically generated and thus cannot be guaranteed to be consistent with human judgements. (2) While current studies simply predict a truthfulness label, it is also necessary to provide a textual statement to explain the prediction. These explanations are vital to justify how the conclusion is reached step by step based on external evidence, and provide the public with rationale to analyze the reasoning process and share it with others. (3) Some prior studies [55, 70, 78] assume that a short piece of evidence text is already identified, based on which the models can directly predict the truthfulness of the target claim. However, this is not realistic in practice as the claim does not come with evidence, which should be retrieved from a knowledge base or the Internet.

To tackle these challenges, we propose end-to-end multimodal fact-checking and explanation generation, where the input consists of a claim and a large collection of web sources, including articles, images, and tweets, and the goal is to automatically retrieve information sources relevant to the claim (*Evidence Retrieval*), predict the truthfulness of the claim based on the relevant evidence (*Claim Verification*), and generate a textual explanation to explain the reasoning and ruling process (*Explanation Generation*). An example[3] is shown in Figure 1. To support this research, we introduce Mocheg, a new benchmark dataset with 15,601 claims annotated with truthfulness labels, multimodal evidence, and ruling statements, along with a large collection of web articles and images as the evidence sources. To set up the baseline performance, we explore the state-of-the-art pre-trained vision-language models for multimodal evidence retrieval, claim verification, and explanation generation. Experimental results show that there is still a huge room for further improvements in this end-to-end multimodal fact-checking and explanation generation task. Overall, the contributions of our work are as follows:

- To the best of our knowledge, this is the first study that investigates end-to-end multimodal fact-checking and explanation generation task.
- We create the first benchmark dataset for end-to-end multimodal fact-checking and explanation generation. The baseline performance of the state-of-the-art language models demonstrates that the task is still challenging, and there is a huge space to improve.

2 RELATED WORK

Multimodal Fake News Detection and Fact-checking: Most previous benchmark datasets [2, 3, 5, 25, 35, 57, 64, 70] for fake news detection and fact-checking are mainly based on text. As information is naturally in multi-modality, recent studies have started to take images [8, 18, 30, 43, 51, 55, 58, 78] and videos [40, 48, 52] into consideration. Many methods for multimodal fake news detection are based on cross-modality consistency checking [1, 56, 60, 63, 71, 76] or computing a fused representation of multimodal (textual + visual) information for final classification [31, 33, 61, 68]. [43, 55, 78] directly predict the truthfulness of multimodal claims without considering explicit evidence. [1, 41, 45] are the most related work to ours in that it considers explicit multimodal evidence. However, their labels or evidence are automatically generated without validating by humans while our label and evidence are annotated by fact-checking journalists. And we further provide journalists explanations regarding the truthfulness prediction. Compared with all these studies, our Mocheg is designed for the end-to-end multimodal fact-checking and explanation generation that requires systems to automatically retrieve multimodal evidence to predict the truthfulness of each claim and generate a ruling statement to explain the reasoning and ruling process. Table 1 compares Mocheg with mentioned datasets.

Explainable Fact-Checking: Providing explanations to the model predictions is beneficial for humans to understand the truthfulness of the claims [22–24, 65, 66]. Current explainable fact-checking studies can be divided into four categories. The first is to directly take the evidence used for claim verification as the explanation [2, 16, 25, 64]. However, the evidence usually consists of several individual sentences extracted from a large collection of documents, which are not logically connected and thus might be hard for humans to interpret. The second is to incorporate external knowledge graphs to compute a set of semantic traces starting from the claim [19]. The semantic traces can serve as explanations to justify the truthfulness of the claims. The third is to generate questions based on claims and link the claims and evidence by using these questions as a proxy [9, 11, 73]. Although these generated questions can improve the explainability, they may be similar or less relevant because, normally, the claim is short. The fourth is to apply natural language generation to generate a paragraph describing the reasoning process [4, 32, 35, 62, 75], which is the most interpretable to humans. Previous studies usually summarize fact-checking articles written by journalists in shorter paragraphs as explanations. In stark contrast, our work generates explanations based on the evidence that is automatically retrieved from the web, which is more realistic in practice. In addition, in our end-to-end multimodal setting, the system needs to sequentially or jointly perform all three sub-tasks, including *multimodal evidence retrieval*, *multimodal claim verification*, and *multimodal explanation generation.*

3 DATASET CONSTRUCTION

3.1 Data Source

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[4]. Specifically, we develop scripts based on [25] to collect all the necessary information from these two websites, including the claims that are purely based on text, truthfulness labels, text and/or image evidence that is extracted from external articles by journalists and help determine the truthfulness of claims,

[3] The example is from https://www.politifact.com/factchecks/2021/may/13/andrew-clyde/ridiculous-claim-those-capitol-jan-6-resembled-nor/

[4] We have obtained permission from both Snopes and Politifact to publish the data for the research purpose.

Feature / Dataset	Evidence Retrieval	Multimodal	Explainable Fact-checking	Annotated Label	Annotated Evidence
FEVER [64]	✓	✗	✗	✓	✓
Liar [70]	✗	✗	✗	✓	✓
Snopes[25]	✓	✗	✗	✓	✓
PUBHEALTH [35]	✓	✗	✓	✓	✓
FACTIFY [41]	✗	✓	✗	✗	✗
MuMiN [45]	✗	✓	✗	✓	✗
FakeNewsNet [58]	✓	✓	✗	✓	✗
Fauxtography [78]	✗	✓	✗	✓	✗
NewsBag [30]	✗	✓	✗	✓	✗
QProp [7]	✗	✓	✗	✓	✗
TABFACT [10]	✗	✓	✗	✓	✓
CLAIMDECOMP [9]	✓	✗	✓	✓	✓
MultiFC [5]	✓	✗	✗	✓	✗
FEVEROUS [3]	✓	✗	✗	✓	✓
Mocheg (Ours)	✓	✓	✓	✓	✓

Table 1: Comparison between Mocheg and other related datasets. The columns indicate whether the dataset requires automatic evidence retrieval, multimodal reasoning, or explanation generation and whether its label and evidence are annotated by a human.

evidence references that are linked to external articles/images containing the text and image evidence, and ruling articles that can explain and justify the truthfulness of the claims and can be viewed as a short summary of the various evidence. Note that, the claims were originally manually collected by the journalists of the two websites from many sources, e.g., online speeches, public statements, news articles, and social media platforms, such as Facebook, Twitter, Instagram, TikTok, and so on. The truthfulness labels, evidence, evidence references, and ruling articles are also manually provided by fact-checkers of the two websites[5].

Based on the evidence references, we further develop scripts to collect the articles and images that contain the evidence. Since the evidence references are linked to thousands of websites with distinct HTML templates, we utilize *Boilerpipe* [34] to extract the text and *newspaper*[6] to obtain all image links contained in the webpages and download the images based on *urllib*[7]. Some evidence references are linked to Twitter. To collect them, we first extract the Tweet IDs from the URLs of evidence references and then apply Twitter API[8] to collect the text and images from the corresponding Tweets.

3.2 Data Preprocessing

Since fact-checking websites adjust their labels over time, the initial data contains more than 75 truthfulness labels, and some labels overlap with each other, such as "*True*", "*TRUE*", and "*Status: True.*". Also, some labels have only a few instances. For example, the label "*Labeled Satire*" has only 23 instances in total. Considering these, we follow [25] and map 68 of these labels into three general categories, including *Supported*, *Refuted*, and *NEI* (*Not Enough Information*). We remove the claims that are annotated with other labels. In this way, each claim is just assigned one of the three target labels. The initial dataset also contains a lot of advertisement images. To clean the dataset, we design several rules, including (1) removing an image if its name contains any of the keywords, including "-ad-", "logo", ".gif", ".ico", "lazyload", ".cgi", "Logo", " .php", "icon", "Bubble", "svg", "rating-false", "rating-true", "banner", "-line", or its size is smaller than 400 × 400; (2) removing a claim if we can not crawl any evidence or the ruling article; (3) for each ruling article, there is usually a paragraph starting with "*Our ruling*" or "*In sum*" which summarizes the whole ruling and reasoning process to achieve the fact-checking conclusion, thus we use this paragraph as the target explanation. As a result, we collect 15,601 claims with 33,880 text evidence, where each piece of text evidence is an individual paragraph extracted from a particular evidence reference article and 12,112 image evidence[9]. Based on the evidence references, we finally collect 91,822 articles and 122,246 images which are further combined to form a constant collection of web resources for the evidence retrieval task. Within the web source collection, only 30% (27,566 out of 91,822) of articles and 10% (12,112 out of 122,246) of images contain the evidence of claims, making the evidence retrieve task realistic and challenging enough.

3.3 Task Definition

We name the dataset Mocheg and propose End-to-End[10] Multimodal Fact-Checking and Explanation Generation, with three subtasks[11]:

Task 1. Multimodal Evidence Retrieval: Given a claim and a collection of web sources containing both documents and images, the *Evidence Retrieval* task is to determine which paragraphs contained in the documents and images are related to the claim and can be further used to determine the truthfulness of the claim.

Task 2. Multimodal Claim Verification: Based on the text and image evidence retrieved in Task 1, the *Multimodal Claim Verification* task is to predict the truthfulness (*Supported*, *Refuted*, or *NEI*) of the claim. As both the input claim and retrieved evidence may contain both text and images, this task requires cross-modal reasoning.

[5] We illustrate the detailed fact-checking processing in Snopes and Politifact in Section. 8.

[6] https://newspaper.readthedocs.io/en/latest/

[7] https://docs.python.org/3/library/urllib.html

[8] https://developer.twitter.com/en/docs/api-reference-index

[9] Among the 15,601 claims, 19% of them have tweets as evidence while the remaining 81% only use other sources such as news articles or government reports as evidence. Note that the image and text evidence may be from separate sources with no clear association.

[10] The end-to-end setting in our fact-checking task means it starts with only the claim and goes through the evidence retrieval, claim verification, and explanation generation, which is almost the complete pipeline for a journalist to do fact-checking in real life.

[11] Note that we don't consider claim extraction as a subtask as all the input claims are considered worthy of being checked.

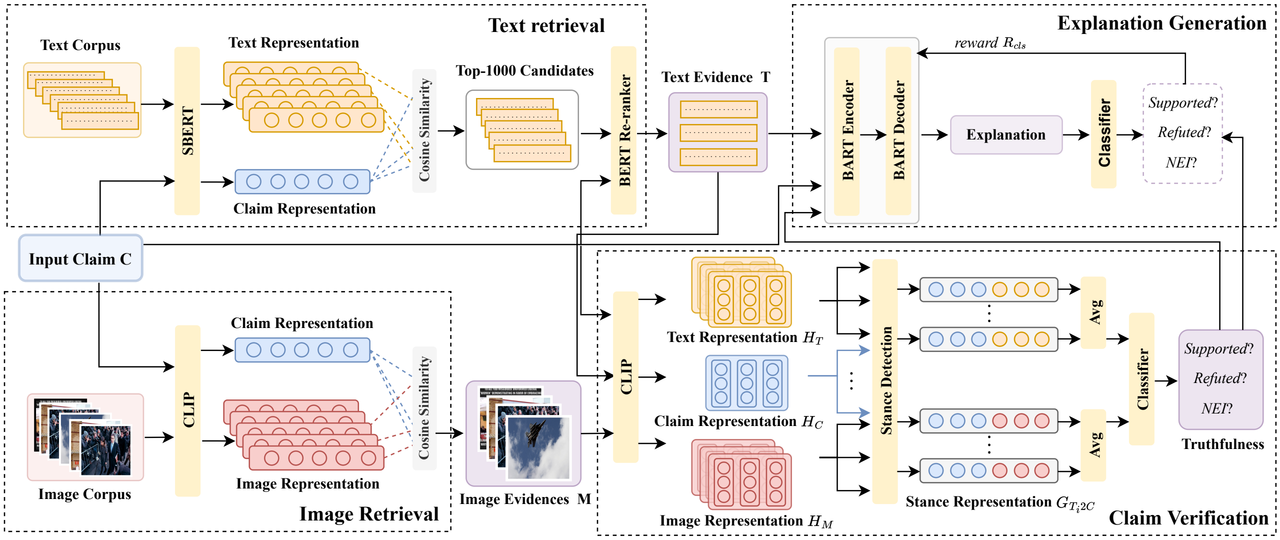


Figure 2: Overview of framework, consisting of a text evidence retrieval module (top left), an image evidence retrieval module (bottom left), a claim verification module (bottom right), and an explanation generation module (top right).

Task 3. Explanation Generation: Given an input claim, the evidence retrieved from Task 1, as well as the truthfulness predicted from Task 2, the *Explanation Generation* task aims to generate a paragraph that summarizes the evidence based on the predicted truthfulness label and explains the ruling process.

3.4 Train / Dev / Test Split

We split the whole dataset into training (`Train`), development (`Dev`), and test (`Test`) sets with the percentage of 75%, 10%, and 15%, respectively. Table 2 shows the detailed statistics for each split.

Data	`Train`	`Dev`	`Test`
# Claims	11,669	1,490	2,442
Ave. # Tokens in Claim	20	20	21
Max. # Tokens in Claim	81	77	89
# Text evidence (Paragraphs)	23.545	4,067	6,268
# Image evidence	8,927	1,178	2,007
# *Refuted* Labels	4,542	488	825
# *Supported* Labels	3,826	501	817
# *NEI* Labels	3,301	501	800
Ave. # Tokens in Explanation	132	90	105
Max. # Tokens in Explanation	600	521	600
# Document/Image in Collection	91,822 / 122,246		

Table 2: Dataset statistics of MOCHEG.

4 APPROACH

To establish the baseline performance on MOCHEG, we design a framework for *End-to-End Multimodal Fact-checking and Explanation Generation.* As illustrated in Figure 2, it consists of three components for the corresponding sub-tasks.

4.1 Evidence Retrieval

To solve this task, we apply two baseline models to retrieve text and image evidence separately.

Text Evidence Retrieval: The top left in Figure 2 illustrates the approach for text evidence retrieval. Given an input claim and a document corpus, we first split each document into sentences and then apply SBERT (Sentence-BERT) [53, 54] to take in the input claim and a sentence from the document corpus and output their contextual representations, based on which we can further compute a cosine similarity score for each pair. Based on these similarity scores, we rank all the sentences and select the top-1000 as the candidate evidence. We fine-tune the SBERT based on the following InfoNCE loss [67]:

$$\mathcal{L}(C_i, T^p, \mathcal{T}) = -\log\left(\frac{\exp(\text{cosine}(\mathbf{C}_i, \mathbf{T}^p))}{\sum_{T_j \in \mathcal{T}} \exp(\text{cosine}(\mathbf{C}_i, \mathbf{T}_j))}\right)$$

where T^p is a piece of positive evidence to a claim C_i, $\mathcal{T}$ contains T^p and a set of other negative evidence to C_i. For each claim, we use the evidence of other claims in the same batch as the negative ones[12]. $\mathbf{C}_i, \mathbf{T}^p$ and $\mathbf{T}_j$ are the sentence level representations encoded from SBERT. In this work, we use bold symbols to denote vector representations.

We further apply a re-ranking model based on BERT [12], which encodes each pair of the input claim and a candidate evidence sentence and outputs a score based on a linear classification layer. Based on these scores, we further rank all the candidate evidence and select the top-K as the text evidence. The BERT-based re-ranking model is pre-trained on the MS MARCO Passage Ranking dataset [6] which is designed for text retrieval.

Image Evidence Retrieval: As shown in the bottom left of Figure 2, given an input claim and the image corpus, we use CLIP [50] as

[12] In MOCHEG, there are 37 sentences that are labeled as positive evidence of two different claims, thus the probability of a text being positive evidence of two claims in the same batch is very low.

the encoder to learn an overall representation for the claim and a representation for each image, then compute the cosine similarity between each image and the input claim. We sort all the images in the corpus based on the cosine similarity scores and take the top-K as the candidate image evidence. We fine-tune CLIP based on the same InfoNCE loss as text evidence retrieval. Note that, during inference, we always retrieve top-K text and image evidence respectively though it's possible that there is no image or text evidence contained in the background corpus.

4.2 Claim Verification

Based on the text and image evidence, we further design a claim verification approach to predict the truthfulness of each input claim, which is shown in the bottom right of Figure 2.

Encoding with CLIP:. We formulate an input claim as $C = \{c_0, c_1, ..., c_n\}$, a piece of text evidence as $T_i = \{t_{i0}, t_{i1}, ..., t_{is}\}$, a piece of image evidence as $M_j = \{m_{j0}, m_{j1}, ..., m_{jq}\}$, where c_k denotes the k-th token of the claim, t_{ik} is the k-th token of the i-th text evidence T_i, and m_{jk} is the k-th patch of the j-th image evidence M_j. Given a claim C and its text evidence $\{T_0, T_1, ...\}$ and image evidence $\{M_0, M_1, ...\}$, we concatenate them as an overall sequence $\{C, T_0, T_1, ..., M_0, M_1 ...\}$ and feed it into CLIP to obtain their contextual representations:

$$\boldsymbol{H}_C = \{\boldsymbol{h}_{c_0}, \boldsymbol{h}_{c_1}, \ldots, \boldsymbol{h}_{c_n}\},$$
$$\boldsymbol{H}_{T_i} = \{\boldsymbol{h}_{t_{i0}}, \boldsymbol{h}_{t_{i1}}, \ldots, \boldsymbol{h}_{t_{is}}\},$$
$$\boldsymbol{H}_{M_j} = \{\boldsymbol{h}_{m_{j0}}, \boldsymbol{h}_{m_{j1}}, \ldots, \boldsymbol{h}_{m_{jq}}\}.$$

Stance detection: We then pair each piece of evidence with the input claim and detect the stance of the evidence towards the claim. As Figure 3 describes, taking text evidence as an example, we first compute an attention distribution between the claim and the evidence by using $\boldsymbol{H}_C = \{\boldsymbol{h}_{c_0}, \boldsymbol{h}_{c_1}, ..., \boldsymbol{h}_{c_n}\}$ as query, $\boldsymbol{H}_{T_i} = \{\boldsymbol{h}_{t_{i0}}, \boldsymbol{h}_{t_{i1}}, \ldots, \boldsymbol{h}_{t_{is}}\}$ as key and value to compute cross attention and obtain an updated claim representation $\boldsymbol{H}_{T_i2C} = \{\boldsymbol{h}_{\tilde{c}_0}, \boldsymbol{h}_{\tilde{c}_1}, \ldots, \boldsymbol{h}_{\tilde{c}_i}, \ldots, \boldsymbol{h}_{\tilde{c}_n}\}$ where $\boldsymbol{h}_{\tilde{c}_i}$ is defined by:

$$\boldsymbol{h}_{\tilde{c}_i} = \text{Softmax}(\boldsymbol{h}_{c_i} \cdot \boldsymbol{H}_{T_i}^{\top}) \cdot \boldsymbol{H}_{T_i}$$

We then fuse the updated claim representation $\boldsymbol{H}_{T_i2C}$ with its original representation H_C by two arithmetic operations, subtraction (-) and multiplication (*), which work best as comparison functions in [69], and obtain the stance representation $\boldsymbol{G}_{T_i2C}$ of evidence T_i towards the claim C based on max pooling.

$$\tilde{G}_{T_i2C} = \sigma([\boldsymbol{H}_{T_i2C}\boldsymbol{H}_C : \boldsymbol{H}_{T_i2C} - \boldsymbol{H}_C] \cdot \boldsymbol{W}_a + \boldsymbol{b}_a),$$
$$\boldsymbol{G}_{T_i2C} = \text{Max\_Pooling}(\tilde{\boldsymbol{G}}_{T_i2C}),$$

where [:] denotes concatenation operation, $\boldsymbol{W}_a$ and $\boldsymbol{b}_a$ are learnable parameters for aggregating the representations, and σ denotes a LeckyReLU activation function.

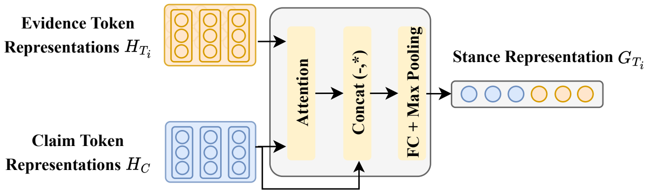

Evidence Token Representations H_{T_i}
Claim Token Representations H_C
Attention
Concat (-,*)
FC + Max Pooling
Stance Representation G_{T_i}

Figure 3: Stance Detection

Prediction: As we have multiple text and image evidence, we further compute the average of the stance representations of all text evidence and image evidence, respectively, to obtain $\boldsymbol{G}_{T2C}$ = Mean_Pooling($\boldsymbol{G}_{T_i2C}$) and $\boldsymbol{G}_{M2C}$ = Mean_Pooling($\boldsymbol{G}_{M_j2C}$). We then concatenate the overall stance representations[13] $\boldsymbol{G}_{T2C}$ and $\boldsymbol{G}_{M2C}$ obtained from both modalities to predict the truthfulness label and optimize the claim verification approach based on the cross-entropy objective:

$$\hat{\boldsymbol{y}}_{cls} = \boldsymbol{W}_h^{\top} \cdot [G_{T2C} : G_{M2C}] + \boldsymbol{b}_h,$$
$$\mathcal{L}(y_i|C) = -\log\left(\frac{\exp(\hat{\boldsymbol{y}}_{cls,i})}{\sum_{j=0}^{2}\exp(\hat{\boldsymbol{y}}_{cls,j})}\right)$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{cls}}$ denotes the probabilities over all possible labels. y_i is the truthfulness label of claim C. During training, we fix the parameters of CLIP while tuning all the other parameters.

4.3 Explanation Generation

To justify the truthfulness prediction, we further generate a ruling statement by considering the input claim, the predicted truthfulness label as well as the text evidence. The top right of Figure 2 illustrates the overall architecture for explanation generation.

Specifically, given an input claim C, its truthfulness label y_C, and text evidence $\{T_1, T_2, \ldots\}$, we concatenate them into an overall sequence X with a separator `</s>`. Then we feed this sequence as input to BART [37], which is a state-of-the-art pre-trained sequence-to-sequence model, to generate a ruling statement $S = \{s_1, s_2, \ldots, s_q\}$. During training, we use the gold truthfulness label of each claim as input. During the evaluation, we use the truthfulness label predicted by the claim verification model. The training objective is to minimize the following negative log-likelihood based on the gold ruling statement $\tilde{S} = \{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_q\}$:

$$\mathcal{L}_g = -\sum_i \log(p(\tilde{s}_i|\tilde{s}_{1:i-1}, X; \phi))$$

To ensure the generated ruling statement is consistent with the truthfulness label of the claim, we apply a truthfulness reward and optimize the generation model with reinforcement learning (RL) [36]. Specifically, we pre-train a truthfulness classification model based on BERT [13], which takes the ruling statement as input and outputs a confidence score for each truthfulness label. We use the difference between the confidence score of the correct label and the score of the wrong labels as the reward R_{cls}:

$$R_{cls} = \boldsymbol{p}(y_C|S) \quad - \sum_{y_j \neq y_C, y_j \in Y} \boldsymbol{p}(y_j|S),$$
$$\boldsymbol{p}(y|S) = \text{Softmax}(\text{BERT}_{\theta}(S)),$$

where y_C is the gold truthfulness label of C, Y is the target label set, and S is the generated ruling statement.

We then apply the reward R_{cls} for policy learning, and the policy gradient is computed as:

$$\nabla_{\phi}\mathcal{J}(\phi) = \mathbb{E}[R_{cls} \cdot \nabla_{\phi} \sum_i \log(\boldsymbol{p}(s_i|s_{1:i-1}, X; \phi))],$$

[13] Since the evidence in our corpus is annotated by journalists on Politifact and Snopes, we assume the evidence is reliable and fuse the stance of evidence to the claim to predict the truthfulness. We leave it as a future work to check the trustworthiness of evidence.

where X is the concatenated sequence of the input claim, its truthfulness label, and text evidence, and ϕ denotes the model parameters.

5 EXPERIMENTS

5.1 Evidence Retrieval

For each claim, we retrieve the top-K text and image evidence from the corresponding text and image corpus and evaluate the retrieval performance based on Precision, Recall, NDCG [29], MAP (Mean Average Precision), and S-Recall (Similarity-based Recall) scores. In S-Recall, it first computes a recall score for each gold text or image evidence based on the highest cosine similarity score between it and all retrieved text or image evidence, while each piece of evidence is represented with a vector learned from SBERT or CLIP. We use the average recall of all gold evidence as the S-Recall.

Media	K	Rec@K	Pre@K	NDCG	MAP	S-Rec
Image	5	17.01	4.71	13.81	11.93	68.22
Image	10	21.44	3.02	15.32	12.58	71.85
Text w/o Re-ranking	5	15.67	12.20	19.23	13.61	52.42
Text w/o Re-ranking	10	19.40	8.16	19.60	13.02	55.77
Text	5	19.72	14.92	23.66	14.34	54.57
Text	10	23.99	9.79	24.09	15.34	58.28

Table 3: Performance of text and image evidence retrieval. (%). *Pre* denotes *Precision* while *Rec* means *Recall*.

We show the performance of text and image evidence retrieval on the test set of Mocheg in Table 3. We can see that the performance of both image and text evidence retrieval is low, indicating the difficulty of both tasks. Taking text evidence retrieval as an example, the model needs to retrieve 2 pieces of text evidence on average for each claim from a collection of 2,792,639 sentences, which is very challenging. Also, the proposed evidence retrieval is based on semantic matching. However, in many cases, it is more important to find evidence that is relevant to the claim but describes different aspects or is against the claim, especially for refuted claims. For example, given an input claim, *"H.R. 6666 provides $100 billion to entities that perform COVID-19 testing but prohibits them from allowing any non-vaccinated persons into their facilities."* the retrieval model missed an important piece of evidence *"No provision in this bill would make testing or quarantining mandatory."*. This is against the claim and has lower similarity compared with the retrieved text *"It would provide $100 billion to organizations that do COVID-19 testing or contact tracing or that provide services to people who are isolated at home."*. In addition, for many claims, their evidence come from the comprehension of long paragraphs instead of several sentences. Although our approach successfully retrieves several relevant sentences, they are insufficient to cover all the background and indicate the truthfulness of the claims.

5.2 Claim Verification

For claim verification, we first design two common baselines: (1) *Majority Label*, which predicts the majority label (i.e., *Refuted*) in the `Training` set for all the claims in the `Test` set; and (2) *Average Similarity*, which computes average cosine similarity between the target claim and all the gold text and image evidence based on their embeddings learned from CLIP. If the average similarity is higher than $\alpha_1 \in \{0.5, 0.6, 0.7, 0.75, 0.8\}$, predict it as *Supported*; if the average similarity is lower than $\alpha_2 \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.65, 0.7\}$ and $\alpha_2 < \alpha_1$, predict it as *Refuted*; otherwise, predict it as *NEI*. We search for the best value of α_1 and α_2 on the `Development` set and then apply them to the `Test` set. We then adapt Pre-CoFactv2 [14], a multimodal fact-checking model which achieves state-of-the-art results at the Factify 2 challenge [42] at AAAI 2023[14], to be the third baseline model. As there is very little existing work on multimodal fact-checking, we further adapt SpotFakePlus [59], a multimodal fake news detection approach, to our fact-checking task, by using their model to compare the consistency of input claim and image evidence and adding a new component to check the consistency of input claim and text evidence[15]. As shown in Table 4, *Majority Label* and *Average Similarity* yield a performance score that is close to a random baseline, while Pre-CoFactv2 and SpotFakePlus underperform our approach, demonstrating that Mocheg does not contain any label distribution bias and cannot be easily solved simply by comparing the semantics between claims and evidence.

Setting	F-score (%)
Majority Label	33.78
Average Similarity (Gold Evidence)	32.72
Pre-CoFactv2 [14] (Gold Evidence)	47.17
SpotFakePlus [59] (Gold Evidence)	44.11
w/o Evidence	39.93
w/ Text Evidence (Gold)	47.54
w/ Image Evidence (Gold)	45.62
w/ Text and Image evidence (Gold)	50.78
w/ Text Evidence (System)	42.79
w/ Image Evidence (System)	40.91
w/ Text and Image evidence (System)	44.06
Human w/o Evidence	20.00
Human w/ System Evidence	62.00
Human w/ Gold Evidence	70.00

Table 4: Performance of claim verification. *Gold Evidence* denotes *gold text and image evidence* while *System Evidence* means *system-retrieved text and image evidence.*

To evaluate the impact of each type of evidence to claim verification, we design ablated models of our approach by considering the text evidence only, image evidence only, or no evidence. In addition, we compare its performance based on the system-retrieved evidence and the gold evidence to show the impact of evidence retrieval. As shown in Table 4, without considering any evidence, the model can still outperform the majority based baseline on claim verification due to the fact that some claims, such as "*Paying taxes is optional!!*," contain obvious clues or are against common sense so that the model can directly predict the truthfulness based on the claim itself. By adding text and/or image evidence, the performance of claim verification can be boosted, proving the usefulness of the evidence. The text evidence provides more significant gain than image evidence due to two reasons: (1) for about 32% of the claims (787 out of 2,442) in the `Test` set, they only have text evidence

[14] https://aiisc.ai/defactify2/

[15] Most of existing multimodal fake news detection studies aim to detect fake news by comparing the consistency between news text and news image or between the news articles and external knnowledge graphs, thus cannot be directly applied to fact-checking task.

without any associated image evidence. However, our approach always returns the top-5 most relevant images as evidence, introducing noises; (2) Texts usually carry more information than images. However, we also observe many examples that the image evidence complements the text evidence. For example, for claim #1 *A Boeing B-17E bomber from World War II was found in the jungle* in Figure 4, its image evidence plays a crucial role in confirming *the aircraft was found in the jungle.*

Finally, we also set up a human performance for claim verification by randomly sampling 50 claims and asking two annotators to label truthfulness by providing gold evidence, system evidence, or no evidence, which reach a Fleiss κ score [17] of 0.67, 0.59, and 0.42, respectively. We take a human prediction as true only if both of the two annotators provide the true label. As we can see, there is still a significant gap between machine and human performance.

5.3 Explanation Generation

We fine-tune BART based on a pre-trained `bart-large`[16] checkpoint [72] to generate the ruling statement. We use ROUGE [39], BLEU [49], and Bertscore [74] as the evaluation metrics. The BERT-based[17] classifier is pre-trained on the gold explanations and reaches an F-score of 87.59%. We fix the classifier during training of the generation model. To evaluate the impact of the evidence retrieval and claim verification on explanation generation, we compare the performance of our approach based on gold evidence and/or gold truthfulness labels with the system-based evidence and truthfulness labels. Note that we only train the model based on gold evidence and truthfulness but perform inference by taking different types of evidence or truthfulness as input. Similar as [35], we further compare our method to LEAD-3, which selects the first three sentences in evidence, and the ORACLE baseline [44], which greedily select[18] multiple evidence sentences that maximize the ROUGE-2 score. Table 5 shows the results with the following observations: (1) Without generation, the explanation is directly from the concatenation of all the text evidence. The explanation may contain all the necessary information but is not interpretable to humans as the sentences are not connected coherently or logically; (2) Evidence retrieval has a more significant impact on explanation generation than claim verification. This is reasonable because the evidence carries most of the content in the explanation and truthfulness is usually implied when comparing the evidence and the input claim. (3) The explanation in our corpus is pretty abstractive, as corroborated by the low performance of ORACLE baseline, which is the upper bound of extractive summarization, and LEAD-3 baseline.

5.4 Implementation Details

We use 2 Quadro RTX 8000 to run our experiments. The retrieval models cost 15 GB and are trained for about 20 runs with a batch size of 256. The claim verification models cost 3 GB and are trained for about 50 runs with a batch size of 128. The explanation generation model costs 45 GB and is trained for about 30 runs with a batch size of 10. We use grid search to tune the hyperparameters: for evidence retrieval, the learning rage $\in \{10^{-5}, 10^{-6}, 10^{-7}\}$ and batch size $\in \{256, 480, 512\}$; for claim verification, the learning rage $\in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and batch size $\in \{64, 128, 256, 512, 1024, 2048\}$; for explanation generation, the learning rage $\in \{5\times10^{-2}, 5\times10^{-3}, 5\times10^{-4}, 5\times10^{-5}\}$ and batch size $\in \{10, 12, 48, 192\}$.

6 REMAINING CHALLENGES

6.1 Claim Verification

We randomly sample 50 claims with gold evidence that are incorrectly verified from the `Test` set and identify the following remaining challenges for multimodal fact-checking:

Cross-modality Reasoning: Both text evidence and image evidence provide complementary information to verify the truthfulness of the claims. 30% of verification errors are due to deep cross-modality reasoning and evidence fusion. For example, for claim #2 *“'If you just count all the deaths in the red states, we are number two in the world in deaths.”* in Figure 4, since there are two different definitions for the red state, the model needs to refer to the image map to confirm the mentioned states.

Cross Document/Sentence Reasoning: 30% of verification errors are due to the reasoning across multiple pieces of textual evidence or across multiple sentences. For example, given the claim *‘The Biden administration's American Jobs Plan will be 'the biggest non-defense investment in research and development in the history of our country.”*, the model needs to first know the current largest investment is $11 billion by referring to evidence *“The largest increase in research and development came in 1964, and totaled $11 billion”*, and then refer to another piece of evidence *“experts say the plan is likely to far exceed $11 billion in spending on research and development.”* to understand that the Plan will exceed $11 billion.

Deep Visual Understanding: For 6% of wrongly predicted claims, their image evidence is charts, tables, or even maps. The current visual understanding techniques, such as CLIP, cannot deeply understand the content and semantics of such images. For example, given claim #3 *“San Francisco had twice as many drug overdose deaths as COVID deaths last year”* in Figure 4, to determine the truthfulness of this claim, the model needs to obtain the number of drug overdose deaths from the image.

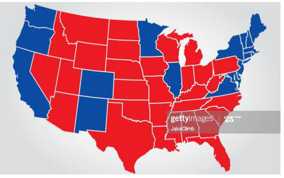
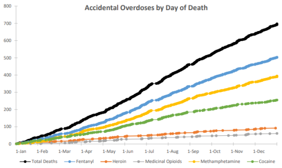
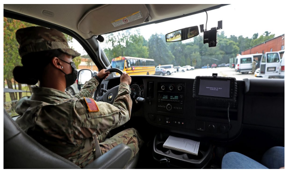
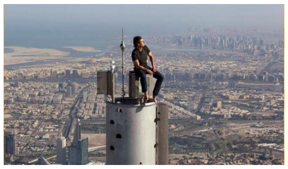
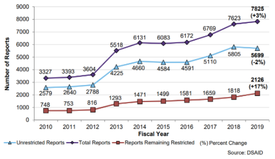
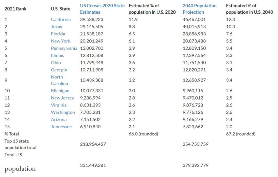
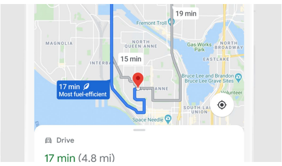
Other Complex Reasoning: Many claims also require various types of complex reasoning, such as mathematical calculation (4% of errors) and commonsense (8% of errors). For instance, the model needs to understand that “*29,000 recipients*” plus “*12,700 recipients*” is “*41,700 recipients*”, “*from 2019 to 1998*” is “*22 years*”, “*there are fifty states in US*”. In addition, the model has difficulty in dealing with claims (12% of errors) that are partially supported or refuted. For example, for the claim *“Since 2010, student debt has increased by 102% and real wages have fallen by over 8%.”*, it's true that *“student debt has increased by 102%”* but the *“real wages have fallen by over 8%”* is not correct.

6.2 Explanation Generation

We also sample 50 system-generated explanations and analyze their error types as follows.

Limited Encoding and Decoding Length: Our approach is based on pre-trained language models, such as BERT and BART, which

[16] https://huggingface.co/facebook/bart-large
[17] https://huggingface.co/bert-base-uncased
[18] https://github.com/pltrdy/extoracle_summarization

Claim	Text Evidence	Image Evidence	Truthfulness
#1: A Boeing B-17E bomber from World War II was found in the jungle	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.		*Supported*
#2: If you just count all the deaths in the red states, we are number two in the world in deaths, just behind Brazil	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.		*Supported*
#3: 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*
#4: 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*
#5: 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*
#6: 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*
#7: By 2040, 70\% of the population is expected to live in just 15 states	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.		*Supported*
#8: A planned update for Google Maps will change the app to no longer show the fastest routes by default.	Soon, Google Maps will default to the route with the lowest carbon footprint when it has approximately the same ETA as the fastest route.		*Supported*
#9: 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	

Figure 4: Examples of Multimodal Fact Checking. The *Truthfulness* column shows gold labels.

Setting	Model	ROUGE 1	ROUGE 2	ROUGE L	BLEU	BERTScore
Gold Evidence ORACLE	-	40.22	23.80	25.97	20.03	86.82
Gold Evidence LEAD-3	-	32.10	16.97	22.17	8.41	86.77
Gold Evidence w/o Generation	-	37.71	21.70	25.62	22.56	87.20
System Evidence w/o Generation	-	28.69	9.93	17.18	7.38	83.95
Gold Evidence + Gold Truthfulness	BART-large	45.51	27.37	35.41	21.84	89.05
Gold Evidence + System Truthfulness	BART-large	43.87	26.37	34.10	20.86	88.87
System Evidence + Gold Truthfulness	BART-large	35.53	17.46	26.05	10.95	87.01
System Evidence + System Truthfulness	BART-large	33.88	16.51	24.83	10.08	86.95

Table 5: Performance of explanation generation. (%)

can only encode or decode a limited length of the sequence. In our dataset, some evidence and ruling statements exceed the maximal length. For those cases, we truncate the sequence and lose part of the information.

Missing Evidence: As we construct the evidence source collection based on the evidence links listed on Snopes and PolitiFact, some evidence used in the ruling statements is not included. For example, given the claim "*By revoking the Keystone pipeline permit, Biden is destroying 11,000 jobs*" the gold explanation contains the information "*A 2014 report found that the company would need only 50 employees to maintain the Keystone XL pipeline*" which is not covered in any of the background documents. In addition, our current explanation generation approach only leverages text evidence while image evidence can also provide complementary information.

Logical Coherence: One critical challenge for explanation generation is to determine the logical connection among the evidence sentences and organize them coherently, a common issue in long-form text generation [26, 27]. For example, given the claim, "*A new, independent study found that at least 55 of our largest corporations used various loopholes to pay zero federal income tax in 2020.*", our explanation generation approach fails to correctly organize the following two evidence: "*many of the relevant provisions are deliberate attempts to set incentives*" and "*Some critics say the financial disclosures used to compile the report are imperfect estimates.*"

7 CONCLUSION

We created Mocheg, an end-to-end multimodal fact-checking and explanation generation benchmark dataset which consists of 15,601 claims annotated with truthfulness labels, together with 33,880 text evidence, 12,112 image evidence as well as explainable statements. We explore the state-of-the-art neural architectures to set up the baseline performance on three sub-tasks (i.e., multimodal evidence retrieval, claim verification, and explanation generation). Our experimental results show that the performance of all three sub-tasks is still far from enough. For future work, an obvious next step is to explore more advanced techniques to improve the three subtasks and deep visual understanding. Furthermore, open-domain fact-checking is another promising direction to detect hallucination errors in large language models like ChatGPT [46]. In the open-domain setting, evaluating the trustworthiness of evidence will play a critical role.

8 ETHICAL STATEMENT

For dataset release, we have obtained permission from both Snopes and Politifact to publish the data for the research purpose. Our dataset is licensed under the CC BY 4.0[19], while the associated codes to Mocheg for data crawler and baseline are licensed under Apache License 2.0[20]. Our dataset contains 2,916 tweets. In accordance with the Twitter developer terms[21], we will only share the Twitter IDs and scripts to crawl tweets based on Twitter API. Our work can be used to predict the truthfulness of various claims in the web 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.

Given the importance of fact-checking in secular societies, we introduce the fact-checking process of Snopes and Politifact to show how our data sources reduce bias. According to Politifact[22] and Snopes[23], they always attempt to contact the person, website, or organization that made the statement they are fact-checking. They will have consultations with a variety of expertise. They seek direct access to government reports, academic studies, and other data. They also have one to two rounds of reviews. Finally, they will accept the error correction from the public and mark the corrected articles. According to Politifact, PolitiFact journalists avoid the public expression of political opinion and public involvement in the political process to set their own opinions aside as they work to uphold principles of independence and fairness. 23 of 36 journalists are women. According to Snopes, members of their editorial staff are precluded from donating to or participating in political campaigns, political party activities, or political advocacy organizations. 6 of 10 journalists are women.

ACKNOWLEDGMENTS

This research is based upon work supported by U.S. DARPA KMASS Program # HR001121S0034. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

[19] https://creativecommons.org/licenses/by/4.0/
[20] https://www.apache.org/licenses/LICENSE-2.0
[21] https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases
[22] https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/
[23] https://www.snopes.com/transparency/

REFERENCES

[1] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14940–14949.

[2] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*. 85–90.

[3] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707* (2021).

[4] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. *arXiv preprint arXiv:2004.05773* (2020).

[5] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242* (2019).

[6] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[7] Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management* 56, 5 (2019), 1849–1864.

[8] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, et al. 2015. Verifying multimedia use at mediaeval 2015. *MediaEval* 3, 3 (2015), 7.

[9] Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating Literal and Implied Subquestions to Fact-check Complex Claims. *arXiv preprint arXiv:2205.06938* (2022).

[10] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164* (2019).

[11] Shih-Chieh Dai, Yi-Li Hsu, Aiping Xiong, and Lun-Wei Ku. 2022. Ask to know more: Generating counterfactual explanations for fake claims. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2800–2810.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Vol. 1. Association for Computational Linguistics (ACL), 4171–4186. arXiv:1810.04805 https://github.com/tensorflow/tensor2tensor

[14] Wei-Wei Du, Hong-Wei Wu, Wei-Yao Wang, and Wen-Chih Peng. 2023. Team Triple-Check at Factify 2: Parameter-Efficient Large Foundation Models with Feature Representations for Multi-Modal Fact Verification. *arXiv preprint arXiv:2302.07740* (2023).

[15] Murray Edelman and Murray Jacob Edelman Edelman. 2001. *The politics of misinformation*. Cambridge University Press.

[16] Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. *arXiv preprint arXiv:2011.05448* (2020).

[17] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[18] Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1683–1698.

[19] Mohamed H Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 87–95.

[20] Peter Godfrey-Smith. 1989. Misinformation. *Canadian Journal of Philosophy* 19, 4 (1989), 533–550.

[21] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint arXiv:2301.04246* (2023).

[22] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206. https://doi.org/10.1162/tacl_a_00454 arXiv:2108.11896

[23] Sai Gurrapu, Lifu Huang, and Feras A Batarseh. 2022. ExClaim: Explainable Neural Claim Verification Using Rationalization. In *2022 IEEE 29th Annual Software Technology Conference (STC)*. IEEE, 19–26.

[24] Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, Laura Freeman, and Feras A Batarseh. 2023. Rationalization for Explainable NLP: A Survey. *arXiv preprint arXiv:2301.08912* (2023).

[25] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214* (2019).

[26] Zhe Hu, Hou Pong Chan, and Lifu Huang. 2022. MOCHA: A Multi-Task Training Approach for Coherent Text Generation from Cognitive Perspective. *arXiv preprint arXiv:2210.14650* (2022).

[27] Zhe Hu, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Hua Wu, and Lifu Huang. 2022. PLANET: Dynamic Content Planning in Autoregressive Transformers for Long-form Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2288–2305.

[28] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. 2020. COVID-19–related infodemic and its impact on public health: A global social media analysis. *The American journal of tropical medicine and hygiene* 103, 4 (2020), 1621.

[29] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[30] Sarthak Jindal, Raghav Sood, Richa Singh, Mayank Vatsa, and Tanmoy Chakraborty. 2020. NewsBag: a multi-modal benchmark dataset for fake news detection. In *CEUR Workshop Proc.*, Vol. 2560. 138–145.

[31] Manvi Kamboj, Christian Hessler, Priyanka Asnani, Kais Riani, and Mohamed Abouelenien. 2020. Multimodal Political Deception Detection. *IEEE MultiMedia* 28, 1 (2020), 94–102.

[32] Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, and Rada Mihalcea. 2021. Extractive and Abstractive Explanations for Fact-Checking and Evaluation of News. *arXiv preprint arXiv:2104.12918* (2021).

[33] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. *The World Wide Web Conference* (2019).

[34] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*. 441–450.

[35] Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926* (2020).

[36] 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 (may 2021), 484–494. https://doi.org/10.48550/arxiv.2105.06947 arXiv:2105.06947

[37] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs.CL]

[38] Lily Li, Or Levi, Pedram Hosseini, and David A Broniatowski. 2020. A multi-modal method for satire detection using textual and visual cues. *arXiv preprint arXiv:2010.06671* (2020).

[39] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[40] Nicholas Micallef, Marcelo Sandoval-Castañeda, Adi Cohen, Mustaque Ahamad, Srijan Kumar, and Nasir Memon. 2022. Cross-Platform Multimodal Misinformation: Taxonomy, Characteristics and Detection for Textual Posts and Videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 651–662.

[41] Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2022. Factify: A multi-modal fact verification dataset. In *Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY)*.

[42] Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. 2023. Factify 2: A multimodal fake news and satire news dataset. In *proceedings of defactify 2: second workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR*.

[43] Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854* (2019).

[44] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. arXiv:1808.08745 [cs.CL]

[45] Dan Saattrup Nielsen and Ryan McConville. 2022. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. *arXiv preprint arXiv:2202.11684* (2022).

[46] OpenAI. 2022. *OpenAI: Introducing ChatGPT.* https://openai.com/blog/chatgpt

[47] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2018. A corpus of debunked and verified user-generated videos. *Online Information Review* 43 (11 2018). https://doi.org/10.1108/OIR-03-2018-0101

[48] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2018. A corpus of debunked and verified user-generated videos. *Online information review* 43, 1 (2018), 72–88.

[49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics.* 311–318.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

[51] Chahat Raj and Priyanka Meel. 2022. ARCNN framework for multimodal infodemic detection. *Neural Networks* 146 (2022), 36–68. https://doi.org/10.1016/j.neunet.2021.11.006

[52] Frederic Rayar, Mathieu Delalandre, and Van-Hao Le. 2022. A large-scale TV video and metadata database for French political content analysis and fact-checking. (2022).

[53] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[54] Nils Reimers and Iryna Gurevych. 2021. The Curse of Dense Low-Dimensional Information Retrieval for Large Index Sizes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers).* Association for Computational Linguistics, Online, 605–611. https://arxiv.org/abs/2012.14210

[55] Julio CS Reis, Philipe Melo, Kiran Garimella, Jussara M Almeida, Dean Eckles, and Fabrício Benevenuto. 2020. A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 903–908.

[56] Arjun Roy and Asif Ekbal. 2021. MulCoB-MulFaV: Multimodal Content Based Multilingual Fact Verification. In *2021 International Joint Conference on Neural Networks (IJCNN).* IEEE, 1–8.

[57] Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19. *arXiv preprint arXiv:2006.11343* (2020).

[58] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* 8, 3 (2020), 171–188.

[59] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13915–13916.

[60] Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing and Management* 58, 1 (2021), 102437. https://doi.org/10.1016/j.ipm.2020.102437

[61] Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management* 58, 1 (2021), 102437. https://doi.org/10.1016/j.ipm.2020.102437

[62] Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)* (2020), 32–43.

[63] Reuben Tan, Bryan A Plummer, and Kate Saenko. 2020. Detecting cross-modal inconsistency to defend against neural fake news. *arXiv preprint arXiv:2009.07698* (2020).

[64] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355* (2018).

[65] Joseph E. Uscinski and Ryden W. Butler. 2013. The Epistemology of Fact Checking. *Critical Review* 25, 2 (June 2013), 162–180. https://doi.org/10.1080/08913811.2013.843872

[66] Joseph E. Uscinski and Ryden W. Butler. 2013. The Epistemology of Fact Checking. *Critical Review* 25, 2 (2013), 162–180. https://doi.org/10.1080/08913811.2013.843872 arXiv:https://doi.org/10.1080/08913811.2013.843872

[67] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv e-prints* (2018), arXiv–1807.

[68] Jingzi Wang, Hongyan Mao, and Hongwei Li. 2022. FMFN: Fine-Grained Multimodal Fusion Networks for Fake News Detection. *Applied Sciences* 12, 3 (2022). https://doi.org/10.3390/app12031093

[69] Shuohang Wang and Jing Jiang. 2016. A compare-aggregate model for matching text sequences. *arXiv preprint arXiv:1611.01747* (2016).

[70] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

[71] Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. 2021. Multimodal Emergent Fake News Detection via Meta Neural Process Networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (aug 2021), 3708–3716. https://doi.org/10.1145/3447548.3467153 arXiv:2106.13711

[72] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[73] Jing Yang, Didier Vega-Oliveros, Taís Seibt, and Anderson Rocha. 2022. Explainable Fact-Checking Through Question Answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 8952–8956.

[74] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[75] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and Predict, and then Predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining.* 418–426.

[76] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. [... formula...]: Similarity-Aware Multi-modal Fake News Detection. *Advances in Knowledge Discovery and Data Mining* 12085 (2020), 354.

[77] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).

[78] Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. *arXiv preprint arXiv:1908.11722* (2019).