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

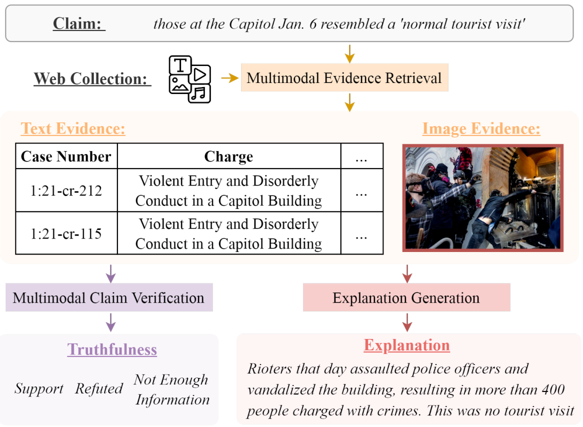
Lichao Sun
Lehigh University
lis221@lehigh.edu

Jin-Hee Cho
Virginia Tech
jicho@vt.edu

Lifu Huang
Virginia Tech
lifuh@vt.edu

Abstract

We propose the 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 (i.e., *support*, *refute* and *not enough information*), and generate a rationalization statement to explain the reasoning and ruling process. To support this research, we construct MOCHEG, a large-scale dataset that consists of 21,184 claims where each claim is assigned with a truthfulness label and ruling statement, with 58,523 evidence in the form of text and images. 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 the current state-of-the-art performance of end-to-end multimodal fact-checking is still far from satisfying. 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 justification.



those at the Capitol Jan. 6 resembled a 'normal tourist visit'

Case Number	Charge	...
1:21-cr-212	Violent Entry and Disorderly Conduct in a Capitol Building	...
1:21-cr-115	Violent Entry and Disorderly Conduct in a Capitol Building	...

Rioters that day 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 serious negative impacts on daily human life, especially making it difficult to find reliable information online. For example, as Islam et al. (2020) shows, the misinformation about COVID-19 has widely spread and led people to distrust medical treatment and even refuse to get vaccinated. 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 some reliable and relevant documents. However, it is time-consuming and hard to generalize to more broad claims.

In recent years, researchers from natural language processing and computer vision have started to investigate automatic misinformation detection and fact-checking by developing various benchmark datasets (Thorne et al., 2018; Wang, 2017; Shu et al., 2020; Nakamura et al., 2019; Papadopoulou et al., 2018a) as well as start-of-the-art neural network architectures (Tan et al., 2020; Song et al., 2021a; Li et al., 2020; Zhou et al., 2020). However, there are at least three limitations with the current fact-checking studies: (1) most fact-checking studies only consider text while ignoring the multi-media nature of online articles. Multimedia information, such as images, videos, and audio, is essential and beneficial for predicting the truthfulness of claims. (2) While current studies simply predict a *support* or *refute* label, it's also necessary to provide a textual explanation to rationalize the judgment. These explanations are vital to justify how the conclusion is reached step by step, and the public can analyze the reasoning pro-

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

cess and share it with others. (3) Last but not least, some prior studies assume that a short piece of evidence text is already identified, based on which the models can directly predict the truthfulness of the claims, which is not realistic in the practice of end-to-end fact-checking.

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, videos and tweets, and the goal is to automatically retrieve information sources that are 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 is shown in Fig. 1. To support research in this direction, we introduce MOCHEG, a new benchmark dataset with 21,184 claims annotated with truthfulness labels, together with a large collection of web sources, including 61,475 articles, 108,673 images, 903 videos, and 4,661 tweets. 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 huge room for future 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 end-to-end multi-modal fact-checking and explanation generation task.
- We also create the first benchmark dataset for end-to-end multi-media fact checking and explanation generation. The baseline performance of the state-of-the-art language models demonstrate that the task is still challenging and there is a huge space to improve.

2 Dataset Construction

2.1 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 and sources. Considering this, we use these two websites as the data sources. Specifically, we develop scripts based on (Hanselowski et al., 2019) to collect claims which may consist of text and/or images, truthfulness labels, evidence references that are relevant to the claims and help determine their truthfulness labels, and ruling articles that explain and justify the truthfulness of the claims and can be viewed as a short summary of the various evidence sources. Generally, the claims are from online speeches, public statements, news articles, and social media platforms, such as Facebook, Twitter, Instagram, TikTok, or some blogs. The truthfulness labels, evidence references, and ruling articles are provided by humans.

Based on the evidence references, we also develop scripts to collect the evidence sources, which consist of text, images, and videos. Since the evidence sources are from thousands of websites with distinct HTML templates, we use boilerplate removal tools to efficiently crawl their contents. In detail, we utilize (Kohlschütter et al., 2010) to extract text and *newspaper* (Ou-Yang, 2013) to get all image links contained in the webpages of the relevant evidence and download the images based on urllib[3]. In addition, some evidence sources are from social media, such as Twitter or Facebook. To collect them, we first extract the Tweet IDs from the evidence references and then apply Twitter API[4] to collect the text, images, and videos from the corresponding Tweets.

2.2 Data Preprocessing

The initial data contains more than 75 truthfulness labels, making it hard for machine learning models to predict them. Given that, we refer to (Hanselowski et al., 2019) and manually map 68 of the labels to three, 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 with one of the three target labels.

The initial dataset contains a lot of advertisement images. And some instances do not contain all information we need. To clean the dataset, we design several rules, including: (1) remove an image if its name contains any of the keywords, including "-ad-", "logo", ".svg", ".gif", ".ico", "lazyload", ".cgi", "Logo"," .php", "icon", "Bubble", "svg", "rating-false", "rating-true", "banner", "-line" or its size is smaller than 400*400; (2) remove a claim if we can not crawl any evidence sources or the ruling

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

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

article. 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-validation conclusion, thus we use this paragraph as the explanation.

As a result, we collect 21,184 claims from Snopes and Politifact with 43,148 textual evidence, 15,375 image evidence, 61,475 textual articles, and 108,673 images from relevant documents.

2.3 Task Definition

We name the dataset as MOCHEG and propose the End-to-End Multimodal Fact-Checking and Explanation Generation, which consists of three subtasks:

Task 1. Evidence Retrieval: Given a claim and a collection of web sources which is created by mixing the evidence of all claims and in the form of text, images, and videos, the Evidence Retrieval task is to determine which text/image/video is related to the claim, and then further extract the top-5 pieces of text and the top-5 images as the evidence which can be used to further determine the truthfulness of the claim.

Task 2. Multimodal Claim Verification: As we have retrieved the top-5 relevant text passages and top-5 relevant images, the Multimodal Claim Verification task is to predict the truthfulness label (*Supported*, *Refuted* or *NEI*) of the claim. As both the input claim and the retrieved evidence contain both text and images, this task requires cross-modality understanding and reasoning.

Task 3. Explanation Generation: Given an input claim, the evidence retrieved from Task 1, as well as the truthfulness label predicted from Task 2, the goal of Explanation Generation is to generate a short paragraph to explain the ruling process and justify the truthfulness label.

2.4 Train / Dev / Test Split

We split the whole dataset into training, development, and test sets. For some claims, their relevant evidence sources or ruling outline may not be fully collected due to the diverse HTML templates they use. Thus we put all these claims into the training set to ensure the high quality of development and test sets. In addition, we also keep the truthfulness labels of development and test sets to be balanced. Table 1 shows the detailed statistics for each split.

Data	Train	Dev	Test
# Claims	18,583	600	2,001
Ave. # Tokens in Claim	20	20	21
Max. # Tokens in Claim	84	58	89
# Textual Evidences	36,358	1,562	5,228
# Textual Relevant Document	56,553	2,500	6,774
# Images Evidence	13,206	519	1,650
# Images from Relevant Document	88,464	4,046	16,163
# Supported Labels	6,936	200	667
# Refuted Labels	7,137	200	667
# NEI Labels	4,510	200	667
Ave. # Tokens in Explanation	306	114	167
Max. # Tokens in Explanation	6,340	2,471	5,235

Table 1: Dataset Statistics of MOCHEG

3 Approach

Figure 2 illustrates the framework for End-to-End Multimodal Fact-checking and Explanation Generation, which consists of three components, each corresponding to a particular sub-task. Next, we will describe the details of each of the components.

3.1 Evidence Retrieval

As the first step, *Evidence Retrieval* aims to retrieve the relevant evidence, including textual passages and images, from a large collection of web sources to support the verification of each input claim. To solve this task, we apply two baseline models from (Reimers and Gurevych, 2019) for retrieving text and image evidence separately.

3.1.1 Text Evidence Retrieval

The top left in Fig. 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 (Reimers and Gurevych, 2019) to take in the input claim and a sentence from the document corpus and output a similar score. Based on these similarity scores, we rank all the sentences and select the top-25 as the candidate evidence. We further design a re-ranking model based on BERT (Devlin et al., 2018), which encodes each pair of the input claim and a piece of candidate evidence, 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-5 as the text evidence. During training, we fix the parameters of pre-trained SBERT to select candidate evidence, and only fine-tune the BERT-based re-ranking model and the linear classification layer based on our own training set.

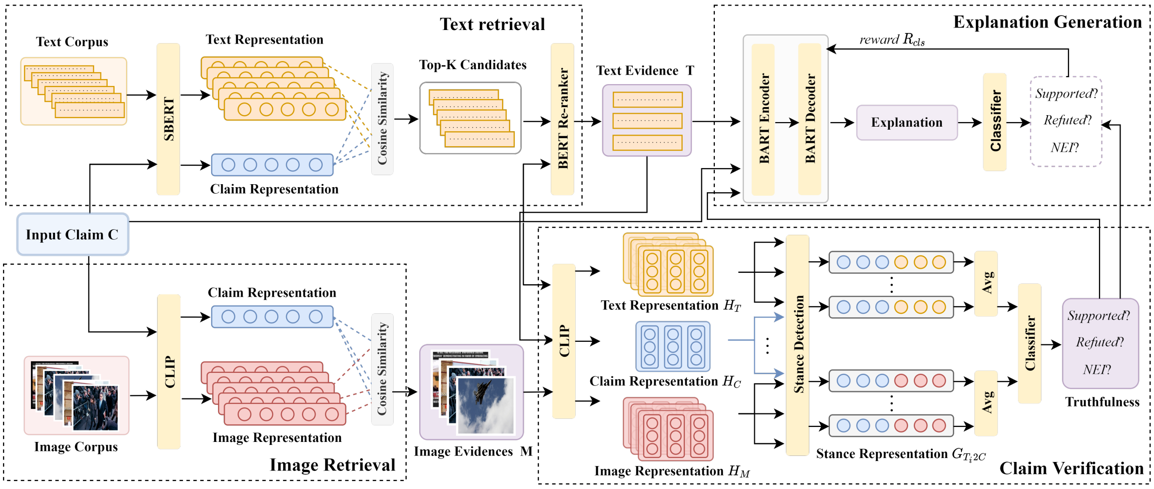


Figure 2: Overview of framework. It consists of a text retrieval module (top left), a image retrieval module(bottom left), a claim verification module(bottom right), and an explanation generation module(top right)

3.1.2 Image Evidence Retrieval

As shown in the bottom left of Figure 2, given an input claim and the image corpus, we use CLIP (Radford et al., 2021) as 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-5 as the candidate image evidence.

3.2 Claim Verification

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

3.2.1 Encoding with CLIP

We formulate an input claim as $C = \{c_0, c_1, ..., c_n\}$, a text evidence as $T_i = \{t_{i0}, t_{i1}, ..., t_{is}\}$, an image evidence as $M_j = \{m_{j0}, m_{j1}, ..., m_{jq}\}$, where c_k denotes the k-th word of the claim, t_{ik} is the k-th word 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 an input claim C and its text evidence $\{T_0, T_1, ...\}$ and image evidence $\{M_0, M_1, ...\}$, we append claim to the text evidence list to form a text list, then feed this text list and image evidence list into CLIP (Radford et al., 2021) to get their contextual representations: $\boldsymbol{H}_C = \{\boldsymbol{h}_{c_0}, \boldsymbol{h}_{c_1}, ..., \boldsymbol{h}_{c_n}\}$, $\boldsymbol{H}_{T_i} = \{\boldsymbol{h}_{t_{i0}}, \boldsymbol{h}_{t_{i1}}, ..., \boldsymbol{h}_{t_{is}}\}$, and $\boldsymbol{H}_{M_j} = \{\boldsymbol{h}_{m_{j0}}, \boldsymbol{h}_{m_{j1}}, ..., \boldsymbol{h}_{m_{jq}}\}$.

3.2.2 Stance detection

We then pair each evidence with the input claim and detect stance of the evidence towards the claim. As Fig. 3 shows, taking text evidence as 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, $\boldsymbol{H}_{T_i} = \{\boldsymbol{h}_{t_{i0}}, \boldsymbol{h}_{t_{i1}}, ..., \boldsymbol{h}_{t_{is}}\}$ as key and value to compute cross attention and obtain an updated claim representation $\boldsymbol{H}_{T_i2C} = \{\boldsymbol{h}_{\tilde{c}_0}, \boldsymbol{h}_{\tilde{c}_1}, ..., \boldsymbol{h}_{\tilde{c}_n}\}$.

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

We then fuse the updated claim representation $\boldsymbol{H}_{T_i2C}$ with its original representation H_C by two arithmetic operations, and obtain the stance representation of evidence T_i towards the claim C based on max pooling.

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

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

As we have multiple text and image evidences, we further compute the average of the stance representations of all text evidences and image evidences, respectively, and concatenate the overall stance representation from both modalities to predict the truthfulness label with cross-entropy objective.

$$\begin{aligned}\boldsymbol{G}_{T2C} &= \text{Mean\_Pooling}(\boldsymbol{G}_{T_i2C}) \\ \boldsymbol{G}_{M2C} &= \text{Mean\_Pooling}(\boldsymbol{G}_{M_j2C}) \\ \hat{\boldsymbol{y}}_{cls} &= \boldsymbol{W}_h^\top \cdot [\boldsymbol{G}_{T2C} : \boldsymbol{G}_{M2C}] + \boldsymbol{b}_h \\ \mathcal{L}(y_i|C) &= -\log(\frac{\exp(\hat{\boldsymbol{y}}_{cls,i})}{\sum_{j=0}^{2}\exp(\hat{\boldsymbol{y}}_{cls,j})})\end{aligned}$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{cls}}$ denotes the probabilities over all possible classes, y_i is the corresponding truthfulness label of claim C.

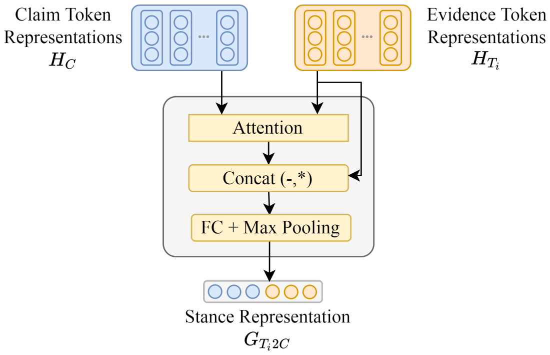


Figure 3: Overview of Stance Detection (Taking text evidence as the example)

3.3 Explanation Generation

To explain the prediction of the truthfulness of the input claim, we use BART (Lewis et al., 2019) to generate a ruling statement by considering the input claim, the predicted truthfulness label as well as the text evidence. To ensure the generated explanation is consistent with the truthfulness label, we incorporate a truthfulness reward (Lai et al., 2021) based on a classification layer and optimize the generation model with reinforcement learning. The top right of Fig. 2 illustrates the overall architecture for explanation generation.

Specifically, given an input claim C, its truthfulness label y_C, and text evidences $\{T_1, T_2, ..., T_5\}$, we concatenate them into an overall sequence X with a separator `</s>`. Then we feed this sequence as input to BART (Lewis et al., 2019), which is a state-of-the-art pre-trained sequence-to-sequence model, and optimize BART for generating $S = \{s_1, s_2, ..., s_q\}$ close to the ground truth ruling statement $\tilde{S} = \{\tilde{s}_1, \tilde{s}_2, ..., \tilde{s}_q\}$. During training, we use the gold truthfulness label of the claim as input, while during evaluation, we use the truthfulness label predicted by the claim verification model. The training objective is to minimize the following negative log-likelihood:

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

To ensure the generated ruling statement is consistent with the truthfulness label of the claim, we design a truthfulness reward. Specifically, we pretrain a truthfulness classification model based on BERT (Devlin et al., 2019), which takes the ruling statement as input and outputs a confidence score for each candidate's truthfulness label. We use the difference between the confidence score of the correct truthfulness label and the confidence score of the wrong truthfulness labels as the reward R_{cls} and apply it for policy learning.

$$\begin{aligned}\boldsymbol{p}(\tilde{y}|S) &= \text{Softmax}_i(\text{classifier}_\theta(S)) \\ R_{cls} &= \boldsymbol{p}(\tilde{y}_C|S) - \sum_{\tilde{y}_j != \tilde{y}_C, \tilde{y}_j \in \{0,1,2\}} \boldsymbol{p}(\tilde{y}_j|S) \\ \nabla_\phi \mathcal{J}(\phi) &= \mathbb{E}[\lambda \cdot R_{cls} \cdot \nabla_\phi \sum_i \log(\boldsymbol{p}(s_i|s_{1:i-1}, X; \phi))]\end{aligned}$$

where $\tilde{y}_C$ is the gold truthfulness label of C, S is the generated explanation, λ is a coefficient weight for the reward, X is the concatenated sequence of claim, truthfulness label and text evidences, and ϕ are the model parameters.

4 Experiments

4.1 Evidence Retrieval

We build the text and image corpus by combining the relevant articles and images of all claims in respectively. For each claim, we retrieve the top-5 text and image evidence from the corresponding text and image corpus. To evaluate the retrieval performance, we refer to (Thorne et al., 2018; Hanselowski et al., 2019; Nie et al., 2019) to measure the Precision, Recall, and F-score of on five highest-ranked sentences or images. These scores are computed in a BERTScore-like (Zhang et al., 2019) manner. In detail, the precision of each retrieved evidence is based on the highest similarity between the retrieved evidence and all the gold evidence, while the similarity is measured by SBERT (Reimers and Gurevych, 2019) and cosine similarity. The overall precision is computed by the average precision of all the retrieved evidence. Similarly, the recall of each gold evidence is based on the highest similarity between the gold evidence and all the retrieved evidence, and we use the average recall of all gold evidence as the overall recall.

Dataset	Media	re-ranking?	Precision	Recall	F-score
Train	Image	-	58.97	66.14	62.34
Dev	Image	-	60.39	68.97	64.40
Test	Image	-	56.37	64.46	60.14
Train	Text	w/	52.84	37.93	44.16
Dev	Text	w/	52.98	39.61	45.33
Test	Text	w/	53.15	41.22	46.43
Train	Text	w/o	52.46	37.60	43.80
Dev	Text	w/o	52.50	39.39	45.01
Test	Text	w/o	53.12	41.11	46.35

Table 2: Performance of Text and Image Evidence Retrieval on Training, Development, and Test Sets. (%)

Setting	F-score
w/o Evidence	33.98
w/ Text Evidence (Gold)	45.18
w/ Image Evidence (Gold)	40.93
w/ Text and Image evidence (Gold)	49.43
w/ Text Evidence (System)	41.03
w/ Image Evidence (System)	38.68
w/ Text and Image evidence (System)	46.78

Table 3: Performance of Claim Verification based on Gold and System-retrieved Evidence. (%)

We show the performance of text and image evidence retrieval on training, development, and test sets in Table 2. We can see that the performance of both image and text evidence retrieval is very low, indicating the difficulty of both tasks. Taking text evidence retrieval as an example, the model needs to retrieve 2.6 text evidence on average for each claim from a collection of 2,792,639 sentences. The performance of image evidence retrieval is higher than text evidence retrieval, especially for recall, which is understandable as the number of text evidence is usually higher than that of image evidence. Finally, we have also done the ablation study to explore the effect of the re-ranking module mentioned in the text retrieval model (Section. 3.1.1). Although we do observe some improvements after adding the re-ranking module, the improvements are tiny.

4.2 Claim Verification

For claim verification, the model needs to detect the stance from the text and image evidence regarding a particular claim, and predict a truthfulness label for the claim, i.e., *refuted*, *supported*, and *NEI (not enough information)*. 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 also compare the performance of our based on the system-retrieved evidence and the gold evidence to show the impact of evidence retrieval.

Table 3 shows the results. Without considering any evidence, the model can still achieve a decent F-score on claim verification due to the fact that some refuted 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, which demonstrates the usefulness of the evidence. The text evidence provides more significant gain than images evidence due to two reasons: (1) for about 30% of the claims (623 out of 2,001) in the evaluation set, they only have text evidence without any associated image evidence, while our approach always returns the top-5 most relevant texts and images as evidence, thus it may introduce noise; (2) it's also intuitive that texts usually carry more information than images. However, we also observe many examples that the image evidence complements the text evidence. For example, for the claim #1 *"San Francisco had twice as many drug overdose deaths as COVID deaths last year"* in Figure 4, its image evidence plays a crucial role since we can only obtain the number of drug overdoes deaths from the image. According to the results, there is still huge room for further improvements. We organize important error types and discuss them in detail in the Section 5.2.

4.3 Explanation Generation

We fine-tune BART based on a pre-trained `bart-large`[5] checkpoint (Wolf et al., 2019) to generate the ruling statement, and use ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and Bertscore (Zhang et al., 2019) as the evaluation metrics. The BERT-based[6] classifier is pre-trained on the gold explanation to evaluate the truthfulness label and reach an F-score of 0.865 after 49 epochs. During training the generation model, we fix the classifier. We set the reward coefficient weight λ as 1. To evaluate the impact of the evidence retrieval and claim verification on explanation generation,

[5] https://huggingface.co/facebook/bart-large

[6] https://huggingface.co/bert-base-uncased

Setting	Model	Rouge1	Rouge2	RougeL	BLEU	BERTScore
Gold Evidence w/o Generation	-	36.47	19.04	23.78	16.25	86.60
System Evidence w/o Generation	-	26.36	7.15	15.35	5.11	83.32
Gold Evidence + Gold Truthfulness	BART-large	46.21	26.52	35.59	16.73	86.67
Gold Evidence + System Truthfulness	BART-large	39.93	22.43	27.58	16.70	86.67
System Evidence + Gold Truthfulness	BART-large	28.75	10.73	17.33	7.03	83.31
System Evidence + System Truthfulness	BART-large	28.74	10.72	17.29	7.00	83.31

Table 4: Performance of Explanation Generation. (%)

we compare the performance of our approach based on gold evidence and/or gold truthfulness with the system-based evidence and truthfulness. 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.

The results are shown in Table 4, from which we have several observations: (1) without generation, the explanation is directly from the concatenation of all the 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, which is understandable as the evidence carries most of the content in the explanation and the truthfulness is usually implicitly implied when comparing the evidence and the input claim.

5 Discussion

5.1 How Text and Image Evidence Complement Each Other?

We explore how each modality contributes to the overall claim verification.

Impact of Text Evidence Using only image evidence, the model gives 1,182 false predictions for claims. By further adding text evidence, 536 of these 1,182 claims can be correctly predicted. In Figure. 4, for the claim #2 *"Massachusetts Gov. Charlie Baker directed National Guard troops to help transport K-12 students to school"*, the image evidence can only show "one guard is driving", and the text evidence can further confirm that it is to help with school transportation.

Impact of Image Evidence In the text evidence only setting, 1,097 claims are incorrectly predicted. By further considering image evidence, 252 of them are correctly predicted. For example, for the claim #3 *"A photograph shows actor Tom Cruise sitting on top of the Burj Khalifa skyscraper without a harness"* in the Figure. 4, the text evidence only describes the height of the building without explicitly mention the actor Tom Cruise, while the image evidence can show us Tom Cruise was sitting on top of a building.

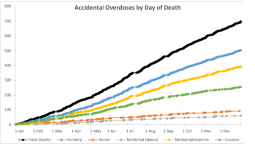
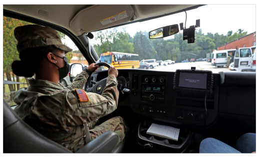
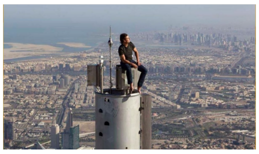
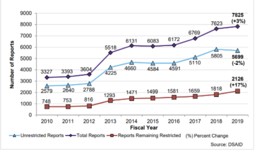
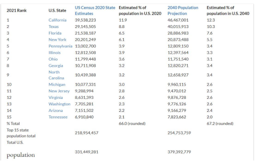
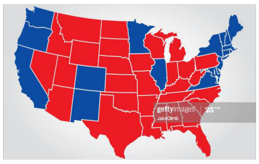
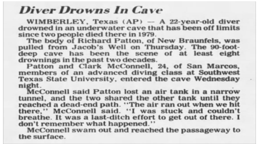
5.2 Remaining Challenges

5.2.1 Claim Verification

We randomly sampled 300 claims that are not correctly verified. By analyzing their input claims, retrieved text, image evidence and the predicted truthfulness, we identify the following remaining challenges for the task of multimodal fact-checking:

Text Evidence Retrieval: The evidence retrieval we proposed is based on similarity matching. However, in many cases, it's more important to find evidence that is relevant to the claim but indicates different opinions or is against the claim. This is especially important to retrieve the evidence for the false claims. For example, given an input claim *"If you look at some of these places that (reduced police funding), they've already seen crime go up"*, the retrieval model missed a piece of important evidence *"Murder and gun violence was already up nationwide in 2020 before cities reduced police funding. Cities that did not cut police budgets also saw murder go up in 2020"*, which is against the claim and has low similarity to the claim but is important to the prediction of the truthfulness. In addition, for many claims, their evidence comes from the comprehension of long paragraphs instead of several sentences. Though our approach successfully retrieves several relevant sentences, they are not enough to cover all the background and indicate the truthfulness of the claims.

Deep Visual Understanding: For some claims, their image evidence is charts, tables or even maps. The current visual understanding techniques, such as CLIP, cannot deeply understand the content and

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 4: Examples of Multimodal Fact Checking

semantics of such images. For example, given the claim #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*" in the Figure 4, in order to determine the truthfulness of this claim, the model needs to analyze the trend of the three curves presented in the image evidence, which is very challenging for the current visual encoders. In addition, many image evidence also contains the text. Without performing Optical Character Recognition (OCR), the current visual encoders cannot fully understand the content of the images.

Cross-modality Reasoning: Both text evidence and image evidence can provide complementary information to verify the truthfulness of the input claims. This requires deep cross-modality reasoning and evidence fusion. For example, in the claim #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"* in Figure. 4, we need to know *"The man next to Mike Pompeo"* is *"Abdul Ghani Baradar"* by referring to the image evidence, and then confirm the claim by referring to the text evidence which is about *"Abdul Ghani Baradar"*.

Other Complex Reasoning: Many claims also require various types of complex reasoning, such as mathematical calculation, commonsense, etc. For example, the model needs to understand that "*14.2 billion*" is "*approximately 15 billion*", "*5,000*" is larger than "*1,500*", "*2018, 2019 and 2017*" is "*3 years*", "*4th of July*" is the "*Independence Day*". In addition, the model has difficulty in dealing with the claims that are partially supported and refuted. For example, in the claim *"Amy Klobuchar vows to resettle 500 percent more refugees"*, the correct part is *"Amy Klobuchar vows to resettle more refugees"*, and the wrong part is *"500 percent"* because she has not specified how many refugees.

5.2.2 Explanation Generation

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

Limited Encoding and Decoding Length: Our approach is based on the pre-trained language models, such as BERT, CLIP, BART, which can only encode or decode a limited length of the sequence. While in our dataset, some evidence and ruling statements exceed the maximal length. In this case, we have to truncate the sequence and lose part of information.

Missing Evidence: As we construct the background document and image corpus based on the source links listed in the Snopes and PolitiFact websites, it's possible that some evidence used in the ruling statement is not included in the background document or image corpus. For example, given the claim "*Opening the schools is a local determination, but it is not a state determination*", the gold explanation contains the information "*If districts do not work with the state, and seek its approval for their reopening plans, they could lose state funding*" 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. For example, given the claim "*Nike donated three times more to Republicans than Democrats during the 2018 federal election cycle, up to August 2018*" and its truthfulness label "NEI", our explanation generation approach failed to correctly organize the following two evidence: "*This is consistent with our history as a non-partisan company*", "*Nike and its employees have spent more than three times as much supporting Republicans*".

6 Related work

Multimodal Fake News Detection and Fact Checking Most previous benchmark datasets (Wang, 2017; Thorne et al., 2018; Hanselowski et al., 2019; Kotonya and Toni, 2020; Augenstein et al., 2019; Alhindi et al., 2018; Vlachos and Riedel, 2014; Nørregaard et al., 2019) for fake news detection or fact-checking are mainly based on text. As only information is naturally in multi-modality, recent studies start to take images (Boididou et al., 2015; Zlatkova et al., 2019; Shu et al., 2020; Nakamura et al., 2019; Jindal et al., 2020; Reis et al., 2020; Fung et al., 2021; Mishra et al., 2022) and videos (Papadopoulou et al., 2018b) into consideration. Most of the methods for multimodal fake news detection or fact-checking are based on cross-modality consistency checking (Tan et al., 2020; Zhou et al., 2020; Song et al., 2021a; Wang et al., 2021; Abdelnabi et al., 2021) or computing a fused representation of multimodal (textual + visual) information for final classification (Khattar et al., 2019; Jin et al., 2017; Song et al., 2021b; Wang et al., 2018, 2022). Compared with these studies, our work considers the end-to-end and explainable fact-checking which requires the system to automatically select relevant sources from a large collection of source document and images, and provide a detailed ruling statement to explain the truthfulness prediction of the input claim.

Explainable Fact-Checking Providing explanations to the model predictions is beneficial for humans to understand the truthfulness of the claims. Current explainable fact-checking studies can be divided into three categories. The first is based on the evidence (Thorne et al., 2018; Alhindi et al., 2018; Hanselowski et al., 2019; Fan et al., 2020) that is used for claim verification. However, the evidence usually consists of several individual sentences extracted from a large collection of documents, which may still be hard for humans to interpret. The second is to incorporate external knowledge graphs to compute a set of semantic traces that start from the claim (Gad-Elrab et al., 2019). The semantic traces will serve as explanations to justify the truthfulness of the claims. The third is to apply natural language generation to generate a paragraph to describe the reasoning process (Zhang et al., 2021; Atanasova et al., 2020; Kotonya and Toni, 2020), which is the most interpretable to humans. Our work is similar to this line of research, however, we consider a more realistic setting where the system needs to sequentially or jointly perform all three sub-tasks including evidence retrieval, multimodal fact verification and explanation generation.

7 Conclusion

We introduce MOCHEG, an end-to-end multimodal fact-checking and explanation generation benchmark dataset which consists of 21,184 claims annotated with truthfulness labels, together with a large collection of web sources including 61,475 articles, 108,673 images, 903 videos and 4661 tweets. We explore the state-of-the-art neural architectures to set up the baseline performance on three sub-tasks, including multimodal evidence retrieval, claim validation and explanation generation. Experimental results show that the performance of all three subtasks is still far from enough. For future work, we will explore more advanced techniques to understand the visual information from image evidence and incorporate it into the explanation generation.

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2021. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. *arXiv preprint arXiv:2112.00061*.

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)*, pages 85–90.

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

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*.

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):7.

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*.

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*, volume 1, pages 4171–4186. Association for Computational Linguistics (ACL).

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*.

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)*, pages 1683–1698.

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*, pages 87–95.

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*.

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):1621.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 795–816, New York, NY, USA. Association for Computing Machinery.

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.*, volume 2560, pages 138–145.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. *The World Wide Web Conference*.

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*, pages 441–450.

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

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.

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.

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*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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)*.

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*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Jeppe Nørregaard, Benjamin D Horne, and Sibel Adalı. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 630–638.

Lucas Ou-Yang. 2013. Newspaper. `https://newspaper.readthedocs.io/en/latest/`.

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2018a. A corpus of debunked and verified user-generated videos. *Online Information Review*, 43.

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

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*, pages 311–318.

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.

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.

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*, volume 14, pages 903–908.

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):171–188.

Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021a. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing and Management*, 58(1):102437.

Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021b. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1):102437.

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

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*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22.

Jingzi Wang, Hongyan Mao, and Hongwei Li. 2022. Fmfn: Fine-grained multimodal fusion networks for fake news detection. *Applied Sciences*, 12(3).

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

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pages 849–857.

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*, pages 3708–3716.

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*.

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

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*, pages 418–426.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. [... formula...]: Similarity-aware multi-modal fake news detection. *Advances in Knowledge Discovery and Data Mining*, 12085:354.

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