RED-DOT: Multimodal Fact-checking via Relevant Evidence Detection

Stefanos-Iordanis Papadopoulos*[1,2], Christos Koutlis[1], Symeon Papadopoulos[1], and Panagiotis C. Petrantonakis[2]

[1]Information Technology Institute, Centre for Research & Technology, Hellas.
[2]Department of Electrical & Computer Engineering, Aristotle University of Thessaloniki.
{stefpapad,ckoutlis,papadop}@iti.gr, ppetrant@ece.auth.gr

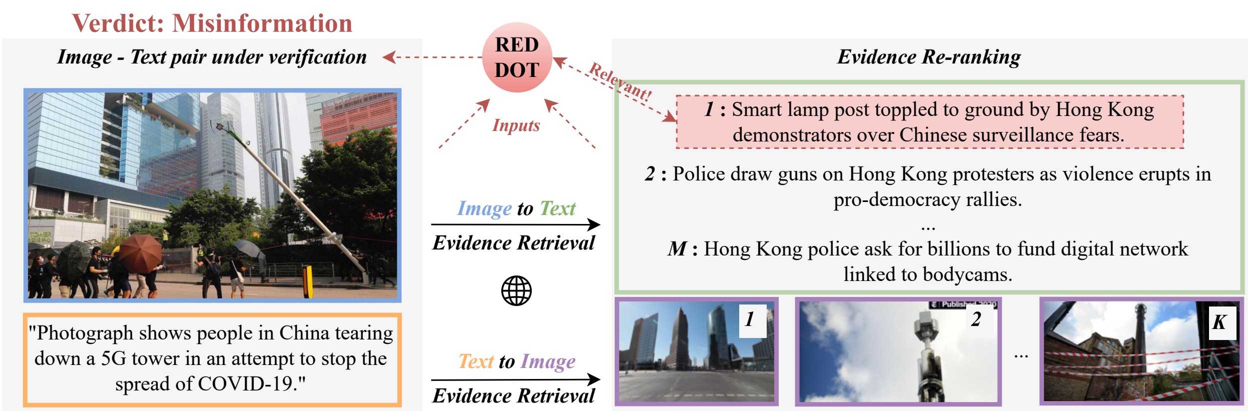


Figure 1: **Image**-**text** pair under verification with external evidence, both **images** and **texts**, collected from the web. The proposed framework retrieves and re-ranks the evidence while *RED-DOT* determines which pieces of information are most relevant to (support or refute) the image-text pair and then uses those to determine the pair's veracity.

Abstract

Online misinformation is often multimodal in nature, i.e., it is caused by misleading associations between texts and accompanying images. To support the fact-checking process, researchers have been recently developing automatic multimodal methods that gather and analyze external information, *evidence*, related to the image-text pairs under examination. However, prior works assumed all collected evidence to be relevant. In this study, we introduce a "Relevant Evidence Detection" (RED) module to discern whether each piece of evidence is relevant, to support or refute the claim. Specifically, we develop the "Relevant Evidence Detection Directed Transformer" (*RED-DOT*) and explore multiple architectural variants (e.g., single or dual-stage) and mechanisms (e.g., "guided attention"). Extensive ablation and comparative experiments demonstrate that *RED-DOT* achieves significant improvements over the state-of-the-art on the VERITE benchmark by up to 28.5%. Furthermore, our evidence re-ranking and element-wise modality fusion led to *RED-DOT* achieving competitive and even improved performance on NewsCLIPings+, without the need for numerous evidence or multiple backbone encoders. Finally, our qualitative analysis demonstrates that the proposed "guided attention" module has the potential to enhance the architecture's interpretability. We release our code at: `https://github.com/stevejpapad/relevant-evidence-detection`.

Keywords Multimodal Learning · Deep Learning · Misinformation Detection · Multimodal Fact-checking

*Corresponding author

1 Introduction

The dissemination of misinformation has greatly intensified in the digital age with the advent of the internet and social media platforms [1] and entails numerous adverse outcomes [2, 3, 4, 5, 6, 7]. Generative AI advances, such as Large Language Models and Stable Diffusion, can now be exploited to create convincing and hyper-realistic misinformation [8, 9]. Nevertheless, decontextualization or image re-purposing, where images are taken out of their original context in order to support false narratives, remain a prevalent challenge in multimodal misinformation detection [10, 11]. For instance, Fig.1 provides a real-world example, where a legitimate image is used as evidence to promote a false narrative, linking 5G technology to the transmission of COVID-19 when in reality, the image captures an anti-surveillance protest in Hong Kong, China[2].

Recent years have witnessed a surge in research attention towards automated multimodal misinformation detection [12] with efforts focusing on the creation of multimodal misinformation datasets [10, 11, 13, 14, 15, 16, 17, 18] and detection methods [19, 20, 21, 22, 23, 23]. Nevertheless, solely relying on the analysis of an image-text pair is not always sufficient for detecting misinformation. More often that not, the incorporation and cross-examination of external information, i.e., evidence, is necessary [24]. To this end, researchers have explored methods for automated and evidence-based fact-checking, initially focusing on text-based approaches [25, 26, 27, 28] and, more recently, venturing into the realm of Multimodal Fact-Checking (MFC) [29, 30, 31]. Certain works on MFC, collect external evidence from fact-checked articles [29, 30], which can be considered to be “leaked evidence” [32]. Conversely, the NewsCLIPings+ dataset, comprises external information collected from the web [31], does not suffer from “leaked evidence” and has recently been used in numerous studies [31, 33, 34]. Nevertheless, prior works on MFC operated under the assumption that all collected evidence were relevant and only addressed the task of verdict prediction.

Motivation. We aim to simulate a more realistic scenario, akin to what fact-checkers encounter daily, where multiple pieces of information are amassed and examined, but not all of them are necessarily relevant for supporting or refuting a claim. Thus, the challenge lies in effectively distinguishing between relevant and irrelevant evidence in order to assist the overall verdict prediction process.

Contributions. We incorporate the step of Relevant Evidence Detection (RED), as part of the MFC process. Specifically, we propose the Relevant Evidence Detection Directed Transformer (*RED-DOT*) which comprises: “Evidence re-ranking”, “Modality Fusion” and “RED” modules. “Evidence re-ranking” leverages CLIP-based [35] intra-modal similarity to re-rank the evidence based on relevance and to create hard negative “irrelevant” samples. “Modality Fusion” leverages element-wise operations between the two modalities in order to cross-check their relation. Finally, “RED” examines all provided pieces of information, determines their relevance to the image-text pair and leverages the relevant ones to assess the pair’s veracity.

We explore numerous variants of *RED-DOT*, with different architectures (e.g., single or dual stage) and mechanisms (e.g., guided attention). *RED-DOT* is optimized via multi-task learning on the NewsCLIPings+ dataset [31] while its performance is also evaluated on the VERITE benchmark [36] comprising real-world multimodal misinformation. Finally, we propose the Out-Of-Distribution Cross-Validation (OOD-CV) evaluation protocol in order to capture models with superior generalizability across algorithmically created training data (NewsCLIPings+) and real-world data (VERITE).

Findings. We conduct extensive ablation and comparative experiments that demonstrate several key findings: (1) OOD-CV captures models with improved performance compared to in-distribution validation (ID-V). (2) By utilizing evidence re-ranking, a single piece of evidence per modality suffices while additional information can introduce noise; degrading overall performance. (3) The combination of element-wise operations can significantly boost detection accuracy, especially when contrasted with simple late-fusion concatenation. (4) *RED-DOT* surpasses prior methods that do not incorporate external evidence by a substantial margin of up to 28.5% and the *RED-DOT* Baseline, which does not utilize the RED module, by 4.5%. Moreover, it competes and even surpasses the current state-of-the-art on NewsCLIPings+ without requiring multiple backbone encoders, numerous evidence or additional features. (5) Finally, we demonstrate that leveraging “guided attention” -where the loss function is directly applied on the attention weights- has the potential to improve interpretability.

2 Related Work

In recent years, automated misinformation detection has gained significant research attention, with ongoing exploration of methods for identifying false information across textual [37], visual [38], and multimodal formats [39]. While several datasets have been developed for text-based fact-checking [25, 26, 27, 28], Glockner et al. [32] showed that

[2] https://www.snopes.com/fact-check/5g-tower-torn-down-china-covid

most employ “leaked evidence” from fact-checked articles. This leads to an unrealistic setting that is not applicable for early detection and emerging misinformation detection. In the context of multimodal misinformation, the majority of research only considers the image-text pairs [10, 11, 14, 15, 16, 19, 20, 21, 22, 36, 40, 41, 42] and do not incorporate external information or evidence; with few notable exceptions [29, 30, 31].

MOCHEG [29] was developed as an end-to-end dataset for MFC. However, it is important to note that it only accommodates textual claims -not multimodal inputs- while examining multimodal evidence. Moreover, the provided evidence can be considered “leaked” since they are collected solely from fact-check articles. Likewise, FACTIFY2 [30] contains “leaked evidence” in the “Refute” class; resulting in high scores of 99-100% even by simple baseline models [43]. To the best of our knowledge, FACTIFY2 is the only dataset addressing multimodal evidence entailment. The authors distinguish between “multimodal” (both modalities contribute to the verdict) or “text” entailment (only the textual modality aligns with or exhibits similarity to the claim). Nevertheless, this framework does not account of the “refute” class or situations where both modalities are irrelevant or only the images are entailed.

On the other hand, Abdelnabi et al. [31] leverage the NewsCLIPings dataset [11], augmenting it with external information; which we refer to as NewsCLIPings+ for simplicity. NewsCLIPings is a algorithmically generated dataset, where image-text pairs are decontextualized using the multimodal encoder CLIP [35], as well as Scene and Person matching computer vision models. As the “falsified” pairs in NewsCLIPings are algorithmically generated, it secures the absence of “leaked” evidence. Recent studies have utilized NewsCLIPings+ for evidence-based multimodal fact-checking [31, 33, 34]; see Section 4.2 for more information. However, this algorithmic generation introduces uncertainty regarding the ability of models trained on NewsCLIPings to generalize effectively to real-world multimodal misinformation.

Recently, researchers have been experimenting with training multimodal misinformation detection models on algorithmically generated datasets like NewsCLIPings [11] and MEIR [40] and evaluating their generalizability to real-world misinformation [16]. Due to unimodal biases in widely adopted benchmarks, such as VMU Twitter [13] and COSMOS [10], Papadopoulos et al. [36] developed VERITE, an evaluation benchmark for real-world multimodal misinformation detection that accounts for unimodal biases.

3 Methodology

3.1 Problem Formulation

We define the tasks of Multimodal Fact-Checking (MFC) and Relevant Evidence Detection (RED) as follows: given a dataset $(I_i^v, T_i^v, \mathcal{I}_i^e, \mathcal{T}_i^e, y_i^v, \boldsymbol{y}_i^e)_{i=1}^N$, where (I_i^v, T_i^v) represents the image-text pair under verification (pair, from now on), $\mathcal{I}_i^e = [I_{i1}^e, I_{i2}^e, \ldots, I_{i2K}^e]$ and $\mathcal{T}_i^e = [T_{i1}^e, T_{i2}^e, \ldots, T_{i2M}^e]$ represent an array of image evidence with $2 \cdot K$ elements (K relevant + K irrelevant) and text evidence with $2 \cdot M$ elements (M relevant + M irrelevant), respectively, $y_i^v \in 0, 1$ denotes the overall verdict label, truthful (0) or misinformation (1) and $\boldsymbol{y}_i^e = (y_{i1}^e, y_{i2}^e, \ldots, y_{i2\cdot(M+K)}^e)$ is an array of binary labels of $2 \cdot (M + K)$ size denoting whether each piece of evidence is relevant (1) or irrelevant (0) to the pair; with the primary objective of training a classifier $f : (\mathcal{I}^v, \mathcal{T}^v, \mathcal{I}^e, \mathcal{T}^e) \rightarrow (\hat{y}^v, \hat{\boldsymbol{y}}^e)$.

3.2 Evidence Retrieval and Re-ranking

We leverage the NewsCLIPings+ dataset [31] comprising 85,360 image-text pairs, balanced in terms of truthful and misinformation pairs. The authors use the text T_i^v to retrieve visual evidence $\mathcal{I}_i^e$ and T_i^v to retrieve textual evidence $\mathcal{T}_i^e$. Each pair is associated with up to 19 textual evidence and up to 10 visual evidence, with a total of 146,032 text evidence and 736,731 image evidence; collected via Google API. While we expect some level of relevance between the collected evidence and the pair, we undertake the task of re-ranking all gathered evidence to heighten the likelihood of their relevance.

Following [31], we use the pre-trained CLIP ViT B-32 [35] as the encoder in order to extract visual $F_I \in R^{512\times1}$ and textual features $F_T \in R^{512\times1}$ for $\mathcal{I}^v, \mathcal{I}^e$ and $\mathcal{T}^v, \mathcal{T}^e$. We signify the ranked or “relevant” evidence with the superscript “+”, and rank them based on the intra-modal cosine similarity sim as follows:

$$
\begin{aligned}
\mathcal{I}_i^+ &= \operatorname*{argsort}_{I_j^e \in \mathcal{I}_i^e} sim(F_{I_i^v}, F_{I_j^e}) \\
\mathcal{T}_i^+ &= \operatorname*{argsort}_{T_j^e \in \mathcal{T}_i^e} sim(F_{T_i^v}, F_{T_j^e})
\end{aligned}
\tag{1}
$$

For the “irrelevant” evidence class, denoted by the superscript “-”, we employ hard negative sampling, instead of depending on random sampling. More specifically, we calculate the most similar item based on text-text similarity, fetch

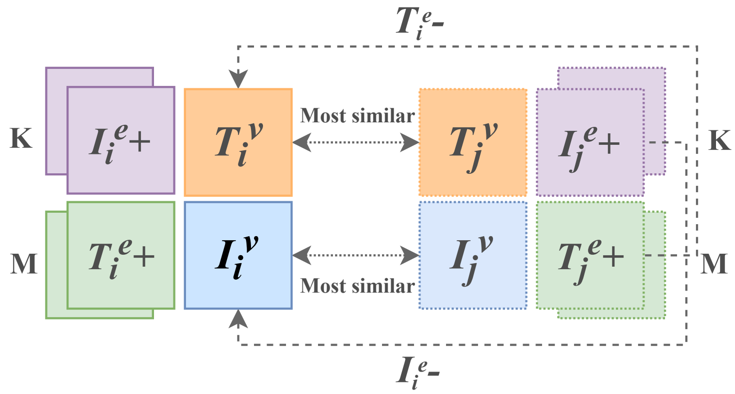


Figure 2: Visualization of Eq.2. Hard negative sampling for retrieving "irrelevant" evidence.

its ranked evidence and employ them as $\mathcal{I}_i^-$, and use image-image similarity for $\mathcal{T}_i^-$. We incorporate the 'opposite' modality in similarity calculations in order to mimic the evidence retrieval process of NewsCLIPings+. The process of hard negative sampling "irrelevant" evidence is illustrated in Fig.2 and can be expressed as follows:

$$\begin{aligned} \mathcal{I}_i^- &= \mathcal{I}_j^+, \text{ where } j = \underset{T_j^v \in \mathcal{T}^v, j \neq i}{\operatorname{argmax}} \; sim(F_{T_i^v}, F_{T_j^v}) \\ \mathcal{T}_i^- &= \mathcal{T}_j^+, \text{ where } j = \underset{I_j^v \in \mathcal{I}^v, j \neq i}{\operatorname{argmax}} \; sim(F_{I_i^v}, F_{I_j^v}) \end{aligned} \tag{2}$$

After retrieving relevant evidence $\mathcal{E}^+ = (\mathcal{T}^+, \mathcal{I}^+)$ and irrelevant $\mathcal{E}^- = (\mathcal{T}^-, \mathcal{I}^-)$, we also define $\boldsymbol{y}^e + = [1, 1, \ldots, 1]$ and $\boldsymbol{y}^e - = [0, 0, \ldots, 0]$, each having size M+K. In order to avoid overfitting based on positional patterns during training, we combine all evidence $\mathcal{E}' = \mathcal{E}^+ \cup \mathcal{E}^-$ and labels $\boldsymbol{y'}^e = [1, 1, ...1, 0, 0, ..., 0]$, shuffle their positions with P representing the permutation positions of the elements and the final evidence and labels become: $\mathcal{E} = \mathcal{E}'[P]$ and $\boldsymbol{y}^e = \boldsymbol{y'}^e[P]$. Evidence $\mathcal{E}$ should consist of $2 \cdot (M + K)$ items. For instance, setting $M = 1$ signifies our expectation of having 1 relevant and 1 irrelevant textual evidence. If this requirement is not met, we randomly sample additional items from the dataset.

3.3 Modality Fusion module

To enhance the cross-examination of image-text pairs and accentuate specific consistencies or inconsistencies, we utilize a range of multimodal fusion operations. In addition to simple concatenation of the two modalities' features, denoted as $[F_{I^v}; F_{T^v}]$, we also employ "addition", "subtraction" and "multiplication". Our rationale is that "addition" emphasizes complementarity, "subtraction" accentuates differences and "multiplication" underscores shared aspects. We express the modality fusion module as follows:

$$F^v = [F_{I^v}; F_{T^v}; F_{I^v} + F_{T^v}; F_{I^v} - F_{T^v}; F_{I^v} * F_{T^v}] \tag{3}$$

with $F^v \in R^{512 \times 5}$. Prior research in multimodal fusion has explored different element-wise operations, including multiplication [44] and outer product [45]. However, to the best of our knowledge, the combination of multiple fusion operations remains unexplored, especially for MFC.

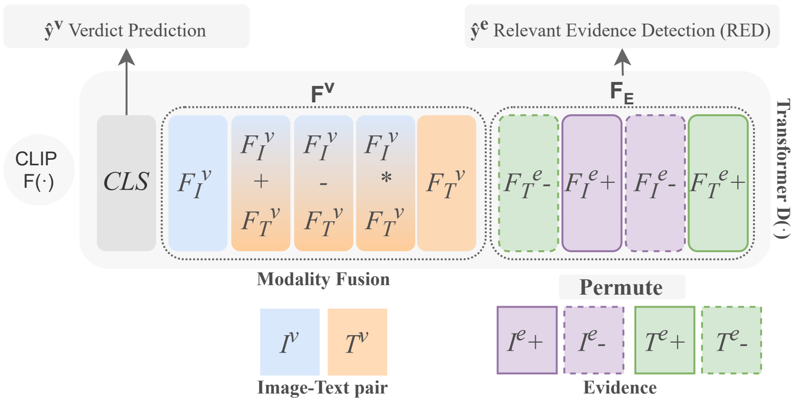
3.4 Verdict Prediction module

We employ a Transformer encoder $D(\cdot)$ to obtain the predicted verdict $\hat{y}^v$. Our *RED-DOT*-Baseline experiments that only leverage relevant evidence are expressed by Eq.4 while the final verdict is predicted by Eq.5.

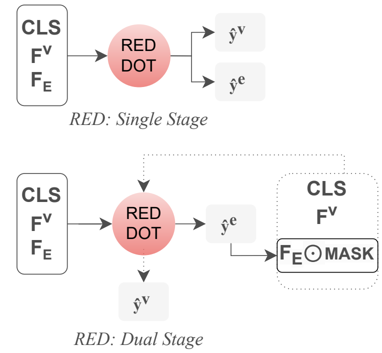
$$[\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F^v}}, \mathbf{d}_{\mathbf{F}_{\mathcal{E}+}}] = D([\text{CLS}; F^v; F_{\mathcal{E}+}]) \tag{4}$$

$$\hat{y}^v = \mathbf{W}_1 \cdot \text{GELU}(\mathbf{W}_0 \cdot \text{LN}(\mathbf{d}_{\text{CLS}})) \tag{5}$$

where LN stands for Layer Normalization, CLS stands for classification token, $\mathbf{W}_0 \in \mathbb{R}^{1 \times 512}$ is a GELU activated fully connected layer and $\mathbf{W}_1 \in \mathbb{R}^{256 \times 1}$ is the final verdict classification layer. $D(\cdot)$ is optimized for binary classification based on the binary cross entropy (BCE) loss function L^v, after applying the sigmoid activation function on $\hat{y}^v$.



(a) Overview of the proposed Transformer $D(\cdot)$ in *RED-DOT*, employing "Modality Fusion".



(b) High-level overview of the single and dual stage *RED-DOT* variants. Dotted lines represent the second stage in DSL.

Figure 3

3.5 Relevant Evidence Detection

The Relevant Evidence Detection (RED) module is responsible for discerning which evidence in $\mathcal{E}$ are relevant or irrelevant to the claim under verification. All methods utilize Transformer $D(\cdot)$ as shown in Eq.6 with $\mathbf{d}_{F_\mathcal{E}}$ as shown in Eq.7 and Eq.5 to predict $\hat{y_v}$. Moreover, all methods leverage multi-task learning to predict both binary $\hat{y}^v$ and multi-label $\hat{\boldsymbol{y}}^e$ and are optimized based on $L = L^v + L^e$ with L^e being the average BCE loss for multi-label classification, after applying the sigmoid activation function on $\hat{\boldsymbol{y}}^e$. An overview of the *RED-DOT* architecture can be seen in Fig.3.

$$[\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F^v}}, \mathbf{d}_{\mathbf{F}_\mathcal{E}}] = D([CLS; F^v; F_\mathcal{E}]) \tag{6}$$

$$\mathbf{d}_{F_\mathcal{E}} = [\mathbf{d}_{\mathbf{F}_{\mathcal{E}_1}}, \ldots, \mathbf{d}_{\mathbf{F}_{\mathcal{E}_{(\mathbf{M}+\mathbf{K})\cdot\mathbf{2}}}}] \tag{7}$$

(1) Single-stage Learning (SSL): employs Eq.6 and Eq.8 to predict $\hat{\boldsymbol{y}}^e$ in a single stage.

$$\hat{y}^e = [\mathbf{W}_3 \cdot \text{GELU}(\mathbf{W}2 \cdot \text{LN}(\mathbf{d}_{\mathbf{F}_{\mathcal{E}_\mathbf{i}}}))], \quad \text{for } i = 1 \text{ to } 2 \cdot (M + K) \tag{8}$$

with $\mathbf{W}_2 \in \mathbb{R}^{1\times512}$ is a GELU activated fully connected layer and $\mathbf{W}_3 \in \mathbb{R}^{256\times1}$.

(2) Single-stage Learning with Guided Attention (SSL+GA): Similar to SSL but instead of Eq.8, we apply a type of "guided attention" where the L_v is directly applied onto the attention weights. Consider the vector $\mathbf{d} = [\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F^v}}, \mathbf{d}_{\mathbf{F}_\mathcal{E}}]$, to calculate the attention scores, we use Eq.9 and Eq.10 to predict $\hat{\boldsymbol{y}}^e$

$$\mathbf{a} = \frac{\mathbf{d}\cdot\mathbf{d}^\mathbf{T}}{512} \tag{9}$$

$$\hat{y}^e = \mathbf{a}[< CLS >][-2 \cdot (M + K) :] \tag{10}$$

where $< CLS >$ represents the position of the classification token, providing a global representation of attention and $-2 \cdot (M + K)$ denotes the last $2 \cdot (M + K)$ items in $\mathbf{d}$, corresponding to the evidence.

(3) Dual-stage Learning (DSL): first employs Eq.6 and Eq. 8. Afterwards, evidence $\mathcal{E}$ are masked by MASK $\in \{0, 1\}^{512\times2\cdot(M+K)}$ where 0s denote predicted irrelevant and 1s predicted relevant evidence. During training, we employ teacher enforcing. At the second stage, we apply MASK onto the evidence and re-process with Transformer $D(\cdot)$ with Eq.11 and then use Eq.5 to predict $\hat{y}^v$.

$$[\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F^v}}, \mathbf{d}_{\mathbf{F}_\mathcal{E}}] = D([\text{CLS}; F^v; F_\mathcal{E} \odot \text{MASK}]) \tag{11}$$

(4) Dual-stage Learning with Guided Attention (DSL+GA): similar to DSL, but in the first stage it utilizes guided attention (as in Eq.9 and Eq.10) instead of Eq.8 to predict $\hat{\boldsymbol{y}}^e$. The second stage remains identical to DSL.

(5) Dual-stage Learning with two Transformers (DSL+D2): similar to DSL, but in the second stage we employ a second identical Transformer encoder $D2(\cdot)$ with $[\mathbf{d}_{\text{CLS}}, \mathbf{d}_{\mathbf{F^v}}, \mathbf{d}_{\mathbf{F}_\mathcal{E}}] = D2([\text{CLS}; F^v; F_\mathcal{E} \odot \text{MASK}])$. The first stage remains identical to DSL.

4 Experimental Setup

4.1 Evaluation Protocol

Our goal is to examine the generalizability of *RED-DOT* onto real-world misinformation. For this reason, we train *RED-DOT* on the NewsCLIPings+ dataset and evaluate its performance on the Out-Of-Context (OOC) pairs from the VERITE benchmark. To gather "external evidence" via Google API for VERITE, we followed the process outlined in [31]; which we make publicly available in order to facilitate future research and ensure fair comparability across studies. Our evaluation protocol departs from "In Distribution Validation" (ID-V) which involves validation and check-pointing on NewsCLIPings and final testing on VERITE [36]. Instead, we leverage "Out-Of-Distribution Cross Validation" (OOD-CV) with k-fold cross-validation and check-pointing directly on a VERITE fold (k-fold=3), while training *RED-DOT* on the NewsCLIPings+ training set. Partially inspired by [46], we hypothesize that due to NewsCLIPings+ and VERITE following different distributions -the first comprising algorithmically created samples and the latter real-world misinformation- OOD-CV can capture a version of the model with improved generalizability onto the "out-of-distribution" VERITE. Finally, we report "Accuracy" on NewsCLIPings+ based on ID-V and "True vs. OOC Accuracy" and the standard deviation on VERITE based on OOD-CV.

4.2 Competing Methods

On NewsCLIPings+, we compare *RED-DOT* against the Consistency Checking Network (*CCN*) [31], the Explainable and Context-Enhanced Network (ECENet) [33] and the Stance Extraction Network [34]. ***CCN*** leverages two attention-based memory networks for visual and textual reasoning, with ResNet152 and BERT, respectively, as well as a fine-tuned CLIP (ViT B/32) component for additional feature extraction. ***SEN*** leverages the overall architecture and encoders of CCN but also extracts the "stance" of external evidence towards the image-text pair and calculates the "support-refutation score" based on the co-occurrence of named entities between the claim and the textual evidence. ***ECENet*** employs ResNet50, BERT and CLIP ViT-B/32 for feature extraction on images, texts, evidence and named entities within a "Coarse- and Fine-grained Attention" network for intra-modal and cross-modal feature reasoning. Lastly, we also include a few methods that do not incorporate external evidence, namely: CLIP [47], DT (Detector Transformer) [16], MNSL (Multimodal Neural Symbolic Learning) [42] and SSDL (Self-Supervised Distilled Learning) [41]. On VERITE, we compare *RED-DOT* against the ***Detector Transformer (DT)*** [36] employing features from CLIP ViT-L/14 after being trained on "R-NESt + CHASMA-D + NC-t2t"; a combination of three different algorithmically created datasets with 437,673 samples in total. To ensure comparability, we faithfully replicate the *DT* and train it on NewsCLIPings+ without external evidence. Within our framework, the *DT* can be expressed as $D([\text{CLS}; F_{I^v}; F_{T^v}])$ with features from CLIP ViT-L/14 and is trained under the ID-V protocol.

4.3 Implementation Details

We train *RED-DOT* for a maximum of 100 epochs with early stopping at 10 epochs, using the Adam optimizer and a learning rate of $lr = 1e-4$. When tuning hyperparameters, we consider the following configurations: $t \in \{4, 6\}$ transformer layers, $z \in \{128, 2048\}$ for the dimension of the feed-forward network, and $h \in \{2, 8\}$ for the number of attention heads. The dropout rate is set at 0.1 and the batch size to 512. For faithful reproduction of *DT*, we follow the implementation details of [36] which uses $lr = 5e-5$ and $t \in \{1, 4\}$. Finally, to ensure experiment reproducibility, we use a constant random seed (0) for PyTorch, Python random, and NumPy.

5 Experimental Results

5.1 Out-of-distribution Cross-Validation

Table 1 showcases the performance of *RED-DOT* variants, with $M = 1$ texts and $K = 1$ images as evidence, when trained and validated under two different protocols: ID-V or the proposed OOD-CV. We observe that OOD-CV consistently gives better results than ID-V, with 72.3% and 70.4% mean accuracy, respectively. These results suggest that OOD-CV is more effective at capturing a model with better generalizability from algorithmically generated samples (NewsCLIPings+) to real-world out-of-context misinformation (VERITE). Based on these findings, we recommend that future research endeavors involving VERITE adopt the OOD-CV protocol to ensure fairer and more representative model performance on the dataset. We employ OOD-CV in all following experiments.

Table 1: Variants of *RED-DOT* trained under the OOD-CV protocol with M texts and K images as evidence. We report accuracy on VERITE and the standard deviation for OOD-CV (k-fold=3). For $M, K = 1$ we also report detection accuracy under ID-V. Underline denotes the best performance per method.

Method	**M,K=1 (ID-V)**	**M,K=1**	**M,K=2**	**M,K=4**
Baseline	69.7	70.7 (1.5)	65.9 (2.3)	66.4 (2.3)
SSL	70.0	71.8 (0.1)	72.1 (3.7)	67.2 (1.8)
SSL + GA	71.2	72.6 (1.7)	71.9 (3.0)	65.8 (2.7)
DSL + D2	69.1	71.3 (0.5)	66.5 (2.6)	65.3 (0.7)
DSL + GA	71.3	73.6 (1.7)	69.9 (2.5)	62.7 (1.7)
DSL	70.9	73.9 (0.5)	70.0 (2.4)	64.7 (1.6)
Mean	70.4	**72.3** (1.0)	69.4 (2.8)	65.3 (1.8)

Table 2: Ablation with different Modality Fusion combinations. F^v followed by '-' signifies removal of one fusion operation.

	$F_{I^v}; F_{T^v}$	F^v	F^v - "-"	F^v - "+"	F^v - "*"
Baseline	66.7 (0.9)	70.7 (1.5)	68.3 (1.2)	70.3 (1.7)	71.6 (4.1)
SSL	71.6 (1.2)	71.8 (0.1)	73.2 (1.2)	73.9 (1.1)	72.9 (0.3)
SSL + GA	70.6 (0.8)	72.6 (1.7)	72.0 (1.5)	71.6 (2.4)	70.5 (2.1)
DSL + D2	68.8 (1.6)	71.3 (0.5)	70.0 (0.8)	69.8 (1.7)	69.3 (1.7)
DSL + GA	69.8 (1.1)	73.6 (1.7)	71.2 (2.0)	73.6 (1.7)	72.8 (1.1)
DSL	71.0 (0.4)	73.9 (0.5)	71.5 (2.2)	72.9 (1.1)	74.2 (1.5)
Mean	69.8 (1.0)	**72.3** (1.0)	71.0 (1.5)	72.0 (1.6)	71.8 (1.8)

5.2 Impact of multiple ranked evidence

Table 1 also illustrates the impact of leveraging $M, K \in \{1, 2, 4\}$ ranked texts and images as evidence. We observe that the best performance is yielded with $M, K = 1$, with an average accuracy of 72.3%, while the performance tends to decline with the inclusion of additional evidence. *RED-DOT* is explicitly trained to detect evidence relevance. Expectedly, the inclusion of less relevant information in $\mathcal{E}+$ with their labels defined as $\boldsymbol{y}^e+$, introduces noise and tends to perplex the network. Nevertheless, we observe that even *RED-DOT*-Baseline that does not leverage RED and, consequently, irrelevant evidence, still demonstrates a reduction in accuracy with the addition of more information. We can infer that the process of re-ranking evidence, as described in Section 3.2, can be advantageous compared to leveraging all collected evidence both in terms of detection accuracy and decreased computational complexity; since the network has fewer items to process. We employ $M, K = 1$ in all following experiments.

5.3 Ablation on Modality Fusion

Table 2 presents the ablation results for different Modality Fusion configurations within *RED-DOT* variants. Our findings indicate that the simple concatenation ($F_{I^v}; F_{T^v}$) results in the lowest average performance (69.8%) across all *RED-DOT* variants while employing all fusion operations (F^v) produces the highest (72.3%). However, there are individual instances where alternative fusion operations demonstrate superior performance. For instance, *RED-DOT*-DSL attains the highest overall accuracy while utilizing $F^v -$ "*" (removing the multiplication operation), yielding 74.2% and surpassing the 73.9% accuracy achieved with F^v. Notably, *RED-DOT*-SSL yields 73.9% without "addition" ($F^v -$ "+") and 71.8% with F^v; a +2.9% improvement. Based on these outcomes, we can conclude that leveraging F^v can lead to adequate results, but if optimal performance is required, it is advisable to experiment with multiple multimodal fusion configurations.

5.4 Comparative Study

Table 3 presents the comparison between *RED-DOT* and the current state-of-the-art on NewsCLIPings+ and VERITE. When *Detector Transformer (DT)* is trained on NewsCLIPings+ it yields 57.5% accuracy on VERITE. Integrating external evidence in *RED-DOT*-Baseline can significantly improve performance, reaching 66.7% (+16.0%) with simple concatenation ($F_{I^v}; F_{T^v}$) and 70.7% (+22.3%) with modality fusion F^v (Eq.3). By leveraging proposed Relevant Evidence Detection module, *RED-DOT* (specifically the DSL variant) can further improve performance up to 73.9%, a

Table 3: Comparison with the state-of-the-art. All methods are trained on NewsCLIPings+ ("News" for short) with the exception of † denoting "R-NESt + CHASMA-D + NC-t2t" [36]. Here, $M, K+$ and $M, K-$ represent the number of relevant and irrelevant evidence, respectively, provided as inputs to the models.

Method	**M,K+**	**M,K-**	**News**	**VERITE**
CLIP [11]	0	0	60.2	-
DT [16]	0	0	65.7	-
MNSL [42]	0	0	68.2	-
SSDL [41]	0	0	71.2	-
CCN [31]	MAX	0	84.7	-
SEN [34]	MAX	0	87.1	-
ECENet [33]	MAX	0	87.7	-
DT [36]	0	0	-	57.5
DT [36] †	0	0	-	72.7
RED-DOT-Baseline ($F_{I^v}; F_{T^v}$)	1	0	80.8	66.7 (0.9)
RED-DOT-Baseline	1	0	**87.8**	70.7 (1.5)
RED-DOT-DSL	1	1	84.5	**73.9** (0.5)

notable +28.5% improvement over *DT* and +4.5% over *RED-DOT*-Baseline. If we consider the ablation of modality fusion (Table 2), the best performances of *RED-DOT*-DSL and *RED-DOT*-Baseline were yielded by F^v-"*", with 74.2% and 71.6% respectively; which translates to a 3.5% improvement. These findings validate our hypothesis that leveraging multitask learning with both "verdict prediction' and "relevant evidence detection" can improve the overall detection accuracy on real-world misinformation. When the *DT* is trained on a larger dataset ("R-NESt + CHASMA-D + NC-t2t") we observe a reduction in the performance gap, with *RED-DOT* outperforming it by 1.7%. Based on these results, we deem highly likely that collecting evidence and training *RED-DOT* on a larger dataset has the potential to further improve performance. However, we do not explore this here, since NewsCLIPings+ is the currently largest publicly available dataset that provides external information as evidence.

When considering all modality fusion ablations, Table 2 highlights the top three performances, with DSL leading at 74.2%, followed by SSL at 73.9%, and DSL+GA at 73.6%. Our interpretation suggests that the dual-stage architecture holds an advantage over SSL by masking irrelevant evidence and employing teacher enforcing. Furthermore, the Verdict Prediction network (Eq.5) in both DSL and SSL provides additional pattern recognition capabilities that contribute to enhanced performance when compared to the "guided attention".

With respect to NewsCLIPings+, we observe that all methods leveraging external information significantly outperform those that do not [11, 16, 42, 41]. Moreover, *RED-DOT*-Baseline yields 87.8% on NewsCLIPings+, outperforming CCN (84.7%), SEN (87.1%) and ECENet (87.7%); while only leveraging a single piece of evidence per modality and without requiring fine-tuning of the CLIP ViT-B/32 encoder, nor multiple encoders (e.g., BERT and ResNet on top of CLIP) nor additional features (e.g., named entities). These results highlight the advantage of the proposed "Evidence re-ranking" and "Modality Fusion" modules. *RED-DOT*-DSL (84.5%) competes with CCN (84.7%) but is outperformed by its "Baseline" counterpart. This can be explained by the fact that *RED-DOT*-Baseline is exclusively trained on Verdict Prediction while *RED-DOT*-DSL is trained for two tasks at the same time, Verdict Prediction and Relevant Evidence Detection. As a result, the model's "focus" is split into the two tasks. Nevertheless, we observe that better performance on the NewsCLIPings+ test set - consisting of algorithmically created negative pairs- does not necessarily translate to improved performance on the annotated misinformation of VERITE. We deem the latter to be the more important metric, since it gives a better indication of the model's ability to detect real world misinformation.

5.5 Qualitative Analysis

Figure 4 depicts the inference process on NewsCLIPings+ test set samples when using two *RED-DOT* variants, DSL and DSL+GA. We observe that DSL+GA tends to assign higher attention scores the "relevant" items. For instance, in the "Truthful" pair (left), the evidence retrieval and re-ranking module was able to retrieve the same image I^e+ as I^v+ and DSL+GA correctly exhibits higher attention scores on I^e+ (0.182) followed by T^e+ (0.109) and lower on irrelevant evidence (0.100). On the contrary, the DSL variant shows lower attention scores on I^e+ (0.081) and I^e+ (0.078) and displays higher attention scores on the irrelevant evidence (0.111). In the "Misinformation" pair, again, DSL+GA assigns higher attention scores to the relevant evidence over irrelevant ones, while DSL only assigns a higher score on T^e+ and the same score (0.08) on all other items, including I^e+. If the DSL model were to be deployed and used by general users or fact-checkers, it might be challenging to trust the model, even if its predictions were accurate,

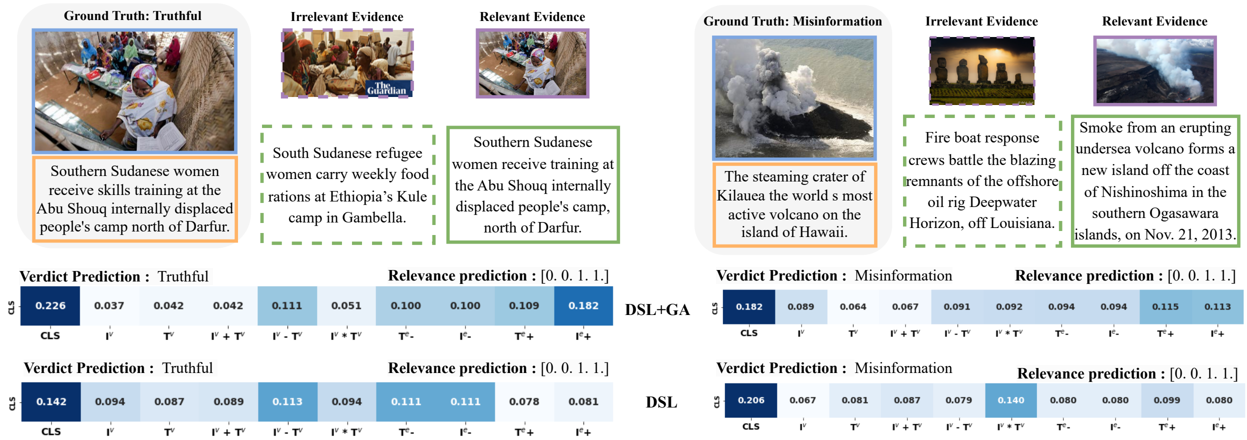


Figure 4: Inference by *RED-DOT* variants: DSL and DSL+GA on samples from NewsCLIPings+. We report the Attention scores assigned by each method. "Relevance Ground truth" is set to [0, 0, 1, 1] for simplicity, regarding [T^e-, I^e-, T^e+, I^e+], respectively.

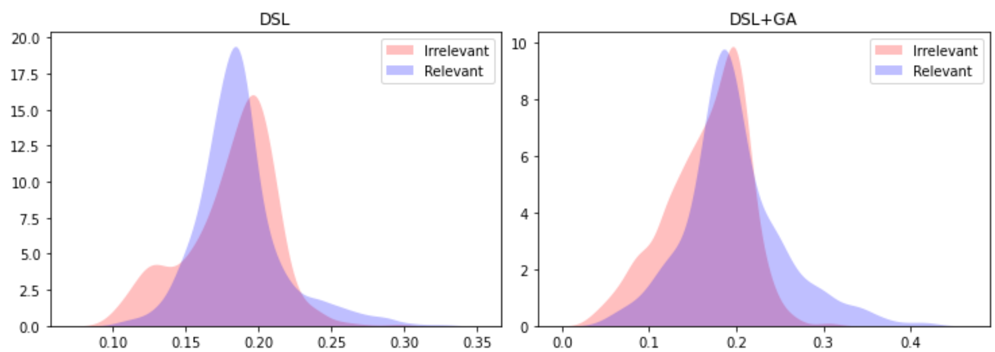


Figure 5: Distributions of attention scores assigned by DSL and DSL+GA per "relevant" and "irrelevant" evidence.

as it consistently assigns higher attention to irrelevant evidence. In contrast, DSL+GA assigns higher attention scores to relevant texts T^e+ and even higher to the relevant images I^e+, thereby enhancing interpretability. In a more systematic way, Fig.5 illustrates the distributions of attention scores for both relevant and irrelevant evidence, as assessed by DSL and DSL+GA, based on a sample size of 1000. Notably, we observe that the median attention scores assigned by DSL to relevant evidence are lower than those for irrelevant evidence, which is not ideal for interpretability. On the other hand, DSL+GA exhibits two distributions with a similar median; however, irrelevant evidence follow a negatively skewed distribution, while relevant evidence follow a positively skewed distribution. Drawing from these observations, we can deduce that the inclusion of guided attention, while not offering optimal interpretability, it holds the potential to enhance it.

6 Discussion and Limitations

Despite recent and notable improvements in the field of automated Multimodal Fact-Checking, it is essential to discuss certain limitations of our research and the field more broadly. Following previous studies [31, 34, 33], we conducted our experiments on NewsCLIPings+, which, albeit being the sole publicly available dataset with external evidence and no "leaked evidence" concerns, comes with certain constraints. Specifically, the dataset's size, consisting of 85,000 samples, may not encompass the full spectrum of diversity concerning historical events, prominent figures, and other crucial contextual elements pertinent to misinformation. Moreover, NewsCLIPings+ consists of algorithmically generated "out-of-context" instances lacking the complexity and nuances present in real-world misinformation scenarios. As evidenced in our experiments, high accuracy on NewsCLIPings+ does not necessarily translate to high accuracy on real-world misinformation (VERITE). For this reason, we suggest that future studies also evaluate the performance of their models on real-world misinformation. Additionally, NewsCLIPings+ exclusively focuses on "re-purposed images" or "out-of-context misinformation", omitting categories like "miscaptioned images" [36]. Subsequent research endeavors should aim at the collection of more extensive and more diverse datasets, encompassing various forms of multimodal misinformation.

While our "evidence re-ranking" module yielded satisfactory results, it's essential to acknowledge that the external information used as evidence is retrieved from search engines, which might introduce noise and irrelevant data. These limitations may impact the realization of the full potential of *RED-DOT* and similar methods. Hence, future research should explore improved methods for gathering, filtering, and assessing the relevance of external information employed as evidence. Our work represents a significant first step towards this direction by providing a novel methodological framework for assessing the relevance of external evidence.

On a more technical level, future research could explore the use of larger multimodal foundation models, including ALBEF [48], BLIP2 [49], LLaVA [50] or the Meta-Transformer [51]. Moreover, our "Modality Fusion" module employs element-wise operations on the image-text pair, but these operations are not extended to external evidence. Inquiry into various modality fusion methods [52, 53, 21] and application to both external evidence and image-text pairs could further improve performance.

7 Conclusions

In this study we address the challenge of evidence-based Multimodal Fact-Checking (MFC) and incorporate Relevant Evidence Detection (RED) as part of the process, where the model, first has to determine which pieces of evidence are relevant to, support or refute, the claim under verification and then proceed to assess its veracity. We conduct extensive ablation and comparative experiments to show that the proposed Relevant Evidence Detection Directed Transformer (*RED-DOT*) is capable of outperforming its counterparts that are not optimized for RED on VERITE. Moreover, it competes and even outperforms the state-of-the-art on NewsCLIPings+ without requiring numerous evidence, multiple backbone encoders or additional features. Finally, leveraging "guided attention" -where the loss function is directly applied onto the Transformer's attention weights- has the potential to improve the model's interpretability.

8 Acknowledgements

This work is partially funded by the project "vera.ai: VERification Assisted by Artificial Intelligence" under grant agreement no. 101070093.

References

[1] Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. Fake news on social media: the impact on society. *Information Systems Frontiers*, pages 1–16, 2022.

[2] W Lance Bennett and Steven Livingston. The disinformation order: Disruptive communication and the decline of democratic institutions. *European journal of communication*, 33(2):122–139, 2018.

[3] Michele Cantarella, Nicolò Fraccaroli, and Roberto Volpe. Does fake news affect voting behaviour? *Research Policy*, 52(1):104628, 2023.

[4] Jon Roozenbeek, Claudia R Schneider, Sarah Dryhurst, John Kerr, Alexandra LJ Freeman, Gabriel Recchia, Anne Marthe Van Der Bles, and Sander Van Der Linden. Susceptibility to misinformation about covid-19 around the world. *Royal Society open science*, 7(10):201199, 2020.

[5] Israel Junior Borges Do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization*, 100(9):544, 2022.

[6] José Gamir-Ríos, Raquel Tarullo, Miguel Ibáñez-Cuquerella, et al. Multimodal disinformation about otherness on the internet. the spread of racist, xenophobic and islamophobic fake news in 2020. *Anàlisi*, pages 49–64, 2021.

[7] Hilda Ruokolainen and Gunilla Widén. Conceptualising misinformation in the context of asylum seekers. *Information Processing & Management*, 57(3):102127, 2020.

[8] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2023.

[9] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. Combating misinformation in the era of generative ai models. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 9291–9298, 2023.

[10] Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: Catching out-of-context image misuse using self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14084–14092, 2023.

[11] Grace Luo, Trevor Darrell, and Anna Rohrbach. Newsclippings: Automatic generation of out-of-context multi-modal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, 2021.

[12] Sakshini Hangloo and Bhavna Arora. Combating multimodal fake news on social media: methods, datasets, and future perspective. *Multimedia systems*, 28(6):2391–2422, 2022.

[13] Christina Boididou, Stuart E Middleton, Zhiwei Jin, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, and Yiannis Kompatsiaris. Verifying information with multimedia content on twitter: a comparative study of automated approaches. *Multimedia tools and applications*, 77:15545–15571, 2018.

[14] Kai Nakamura, Sharon Levy, and William Yang Wang. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, 2020.

[15] Ayush Jaiswal, Ekraam Sabir, Wael AbdAlmageed, and Premkumar Natarajan. Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1465–1471, 2017.

[16] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, pages 36–44, 2023.

[17] Giscard Biamby, Grace Luo, Trevor Darrell, and Anna Rohrbach. Twitter-comms: Detecting climate, covid, and military multimodal misinformation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1530–1549, 2022.

[18] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 16–25, 2020.

[19] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.

[20] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 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, 2018.

[21] Chuanming Yu, Yinxue Ma, Lu An, and Gang Li. Bcmf: A bidirectional cross-modal fusion model for fake news detection. *Information Processing & Management*, 59(5):103063, 2022.

[22] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE, 2019.

[23] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference 2022*, pages 726–734, 2022.

[24] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[25] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, 2017.

[26] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, 2019.

[27] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, 2018.

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

[29] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 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*, pages 2733–2743, 2023.

[30] S Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. Factify 2: A multimodal fake news and satire news dataset. *arXiv preprint arXiv:2304.03897*, 2023.

[31] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 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*, pages 14940–14949, 2022.

[32] Max Glockner, Yufang Hou, and Iryna Gurevych. Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, 2022.

[33] Fanrui Zhang, Jiawei Liu, Qiang Zhang, Esther Sun, Jingyi Xie, and Zheng-Jun Zha. Ecenet: Explainable and context-enhanced network for muti-modal fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1231–1240, 2023.

[34] Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. Support or refute: Analyzing the stance of evidence to detect out-of-context mis-and disinformation. *arXiv preprint arXiv:2311.01766*, 2023.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[36] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Verite: A robust benchmark for multimodal misinformation detection accounting for unimodal bias. *arXiv preprint arXiv:2304.14133*, 2023.

[37] Muhammad F Mridha, Ashfia Jannat Keya, Md Abdul Hamid, Muhammad Mostafa Monowar, and Md Saifur Rahman. A comprehensive review on fake news detection with deep learning. *IEEE Access*, 9:156151–156170, 2021.

[38] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE Access*, 2022.

[39] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, Preslav Nakov, et al. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643. International Committee on Computational Linguistics, 2022.

[40] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1337–1345, 2018.

[41] Michael Mu, Sreyasee Das Bhattacharjee, and Junsong Yuan. Self-supervised distilled learning for multi-modal misinformation identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2819–2828, 2023.

[42] Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model. *arXiv preprint arXiv:2304.07633*, 2023.

[43] S Suryavardan, Shreyash Mishra, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha, Aishwarya Reganti, Amitava Das, Amit Sheth, Manoj Chinnakotla, et al. Findings of factify 2: multimodal fake news detection. *arXiv preprint arXiv:2307.10475*, 2023.

[44] Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. Memefier: Dual-stage modality fusion for image meme classification. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 586–591, 2023.

[45] Gokul Karthik Kumar and Karthik Nandakumar. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 171–183, 2022.

[46] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[47] Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 97–106, 2022.

[48] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[49] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[51] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.

[52] Jing Jing, Hongchen Wu, Jie Sun, Xiaochang Fang, and Huaxiang Zhang. Multimodal fake news detection via progressive fusion networks. *Information processing & management*, 60(1):103120, 2023.

[53] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2560–2569, 2021.