SIMILARITY OVER FACTUALITY: ARE WE MAKING PROGRESS ON MULTIMODAL OUT-OF-CONTEXT MISINFORMATION 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

ABSTRACT

Out-of-context (OOC) misinformation poses a significant challenge in multimodal fact-checking, where images are paired with texts that misrepresent their original context to support false narratives. Recent research in evidence-based OOC detection has seen a trend towards increasingly complex architectures, incorporating Transformers, foundation models, and large language models. In this study, we introduce a simple yet robust baseline, which assesses MUltimodal SimilaritiEs (MUSE), specifically the similarity between image-text pairs and external image and text evidence. Our results demonstrate that MUSE, when used with conventional classifiers like Decision Tree, Random Forest, and Multilayer Perceptron, can compete with and even surpass the state-of-the-art on the NewsCLIPpings and VERITE datasets. Furthermore, integrating MUSE in our proposed "Attentive Intermediate Transformer Representations" (AITR) significantly improved performance, by 3.3% and 7.5% on NewsCLIPpings and VERITE, respectively. Nevertheless, the success of MUSE, relying on surface-level patterns and shortcuts, without examining factuality and logical inconsistencies, raises critical questions about how we define the task, construct datasets, collect external evidence and overall, how we assess progress in the field. We release our code at: `https://github.com/stevejpapad/outcontext-misinfo-progress`.

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

1 Introduction

In recent decades, we have witnessed the proliferation of new types of misinformation, beyond fake news [1] and manipulated images [2], including AI-generated "DeepFakes" [3] and misinformation that spans multiple modalities such as images and texts[4, 5]. In an effort to assist the work of human fact-checkers, researchers have been leveraging the power of deep learning to automate certain aspects of the fact-checking process [6], such as claim and stance detection, evidence and fact-check retrieval, and verdict prediction, among others[7, 8, 9, 10, 11].

In this study, we focus on multimodal fact-checking, specifically targeting evidence-based out-of-context (OOC) detection, a topic that has recently gained significant attention from researchers. OOC misinformation involves the presentation of images with captions that distort or misrepresent their original context[12]. Due to the lack of large-scale, annotated datasets for OOC detection, researchers have turned to algorithmic generation of OOC datasets [13, 14, 15] which have been used to train numerous methods for OOC detection [16, 17, 18, 19, 20], some of which leverage external information or evidence to further enhance detection accuracy [21, 22, 23, 24]. Overall, there is a trend towards increasingly complex architectures for OOC detection, including the integration of Transformers and memory networks [21], fine-tuning foundational vision-language models [18], incorporating modules for detecting relevant evidence [24] and leveraging instruction tuning and large language models [23], which generally translate to marginal improvements in performance.
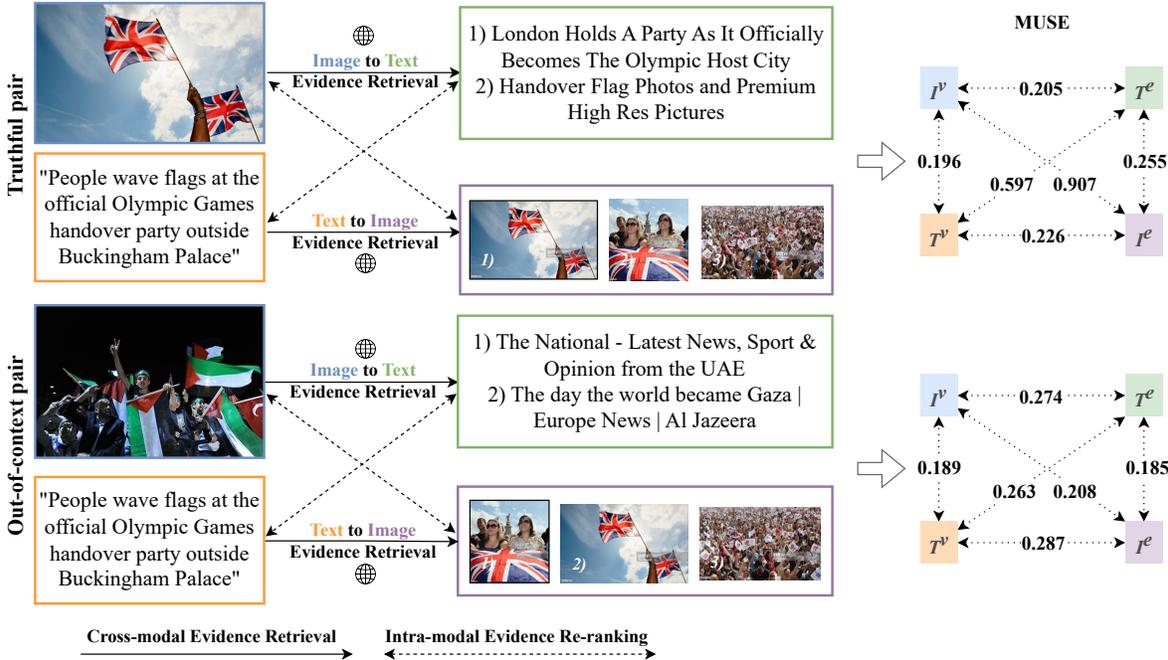
---

[*]Corresponding author

Figure 1: Samples from the NewsCLIPpings dataset, along with their retrieved and re-ranked evidence and their multimodal similarities.

We develop a simple yet robust baseline that leverages MUultimodal SimilaritiEs (MUSE), specifically CLIP-based [25] similarities between image-text pairs under verification and across external image and text evidence. Our findings show that training machine learning classifiers, such as Decision Tree, Random Forest and Multi-layer Perceptron with MUSE can compete and even outperform much more complex architectures on NewsCLIPpings [14, 21] and VERITE [24, 26] by up to 4.8%. Furthermore, integrating MUSE within complex architectures, such as our proposed "Attentive Intermediate Transformer Representations" (AITR) can further improve performance, by 3.3% on NewsCLIPpings and 7.5% on VERITE, over the state-of-the-art (SotA).

Nevertheless, our analysis reveals that the models primarily rely on shortcuts and heuristics based on surface-level patterns rather than identifying logical or factual inconsistencies. For instance, as illustrated in Fig.1, given a Truthful image-text pair, we use the text to retrieve image evidence and the image to retrieve text evidence from the web. Due to the popularity of the NewsCLIPpings' sources (USA Today, The Washington Post, BBC, and The Guardian) search engines often retrieve the exact same or highly related images and texts as those under verification, which, after re-ranking, are selected as the likely evidence. This results in a high 'image-to-evidence image' (0.907) 'text-to-evidence text' (0.597) similarities. In contrast, given the OOC image, we retrieve unrelated text evidence, leading to significantly lower similarity scores. Consequently, a model can learn to rely on simple heuristics such as, if the image-text pair exhibits significant similarity both internally and with the retrieved (and re-ranked) evidence, then the pair is likely truthful; otherwise, it is OOC. Furthermore, we show that these models yield high performance only within a limited definition of OOC misinformation, where legitimate images are paired with otherwise truthful texts from different contexts. In contrast, their performance deteriorates when dealing with 'miscaptioned images,' where images are de-contextualised by introducing falsehoods in their captions i.e. by altering named entities such as people, dates, or locations.

These findings raise critical questions about how realistic and robust the current frameworks are, how we define the task, create datasets, collect external evidence, and, more broadly, how we assess progress in OOC detection and multimodal fact-checking. In summary, we recommend future research to: 1) avoid training and evaluating methods solely on algorithmically created OOC datasets; 2) incorporate annotated evaluation benchmarks; 3) broaden the definition of OOC to include miscaptioned images, named entity manipulations, and other types of de-contextualization, and 4) to expand training datasets and collect external evidence accordingly.
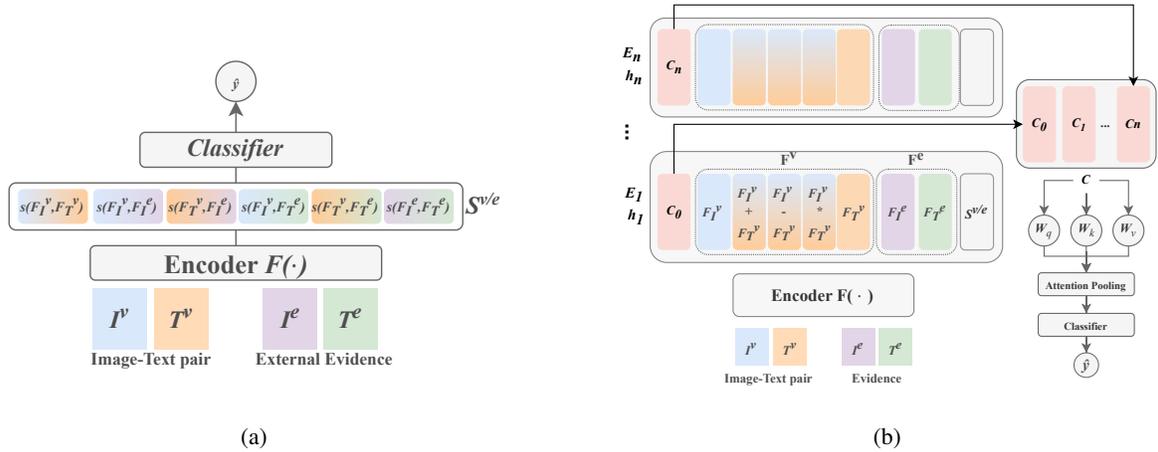
Figure 2: Overview of the (a) MUSE and (b) AITR architectures.

2 Related Work

Out-Of-Context (OOC) detection, also known as image re-purposing, multimodal mismatching, or "CheapFakes", involves pairing legitimate, non-manipulated, images with texts that misrepresent their context. Due to the lack of manually annotated and large-scale datasets, initial attempts to model out-of-context misinformation relied on randomly re-sampling image-text pairs [13, 27], while more sophisticated methods now rely on hard negative sampling, creating out-of-context pairs that maintain semantic similarity [14, 15, 28]. In turn, multiple methods have been proposed for OOC detection that cross-examine and attempt to identify inconsistencies within the image-text pair without leveraging external evidence [16, 17, 18, 19, 20]. Another strand of research has focused on constructing multimodal misinformation datasets through weak annotations [29] or named entity manipulations [15, 30, 31] but, to the best of our knowledge, such datasets have not yet been enhanced with external evidence and used for multimodal fact-checking. Nevertheless, professional fact-checkers[2] rarely rely solely on internal inconsistencies between modalities and instead collect relevant external information, or evidence, that support or refute the claim under verification [6, 32]. Furthermore, prior studies on evidence-based OOC detection demonstrate significant performance improvements when leveraging external information [21, 23, 24].

Specifically, Abdelnabi et al. [21] enhanced the NewsCLIPpings dataset [14] by collecting external evidence (See Section 4.1) and developed the Consistency Checking Network (**CCN**) which examines image-to-image and text-to-text consistency using attention-based memory networks, that employ ResNet152 for images and BERT for texts, as well as a fine-tuned CLIP (ViT B/32) for additional multimodal features. The Stance Extraction Network (**SEN**) employs the same encoders as CCN but enhances performance by semantically clustering external evidence to determine their stance toward the claim. It also integrates the co-occurrence of named entities between the text and textual evidence [33]. The Explainable and Context-Enhanced Network (**ECENet**) combines a coarse- and fine-grained attention network leveraging ResNet50, BERT and CLIP ViT-B/32 for multimodal feature extraction along with textual and visual entities [22]. **SNIFFER** examines the "internal consistency" of image-text pairs and their "external consistency" with evidence with the use of a large language model, InstructBLIP, that is first fine-tuned for news captioning and then for OOC detection, utilizing GPT-4 to generate instructions that primarily focus on named entities while the Google Entity Detection API is used for extracting visual entities [23]. Finally, the Relevant Evidence Detection Directed Transformer (**RED-DOT**) utilizes evidence re-ranking, element-wise modality fusion, guided attention and a Transformer encoder optimized with multi-task learning to predict the weakly annotated relevance of retrieved evidence [24].

On the whole, there is a noticeable trend toward increasing architectural complexity which typically translates into limited improvements in performance. In this study, we show how simple machine learning approaches can compete and even surpass complex SotA methods by simply leveraging multimodal similarities, which raises critical questions on how we define the task, collect data, external evidence and how we access progress in the field.

---

[2]`https://www.factcheck.org/our-process`

3 Methodology

3.1 Problem Formulation

We define the task of evidence-based out-of-context detection as follows: given dataset $(I_i^v, T_i^v, I_i^e, T_i^e, y_i)_{i=1}^N$ where I_i^v, T_i^v represent the image-text pair under verification, I_i^e, T_i^e image and textual external information, or evidence, retrieved for the pair and $y_i \in \{0, 1\}$ is the pair's ground-truth label, being either *truthful* (0) or *out-of-context* (1), the objective is to train classifier $f : (\mathcal{I}^v, \mathcal{T}^v, \mathcal{I}^e, \mathcal{T}^e) \to \hat{y}^v$.

3.2 Multimodal Similarities

As shown in Fig.2a, given feature extractor $F(\cdot)$ and extracted features $F_{I^v}, F_{T^v}, F_{I^e}, F_{T^e}$, we use cosine similarity s to calculate the Multimodal Similarities (MUSE) vector $S^{v/e}$ between $s(F_{I^v}, F_{T^v})$ image text pairs, $s(F_{I^v}, F_{I^e})$ image to image evidence, $s(F_{T^v}, F_{I^e})$ text to image evidence, $s(F_{I^v}, F_{T^e})$ image to text evidence, $s(F_{T^v}, F_{T^e})$ text to text evidence and $s(F_{I^e}, F_{T^e})$ image evidence to text evidence.

Afterwards, the $S^{v/e}$ vectors are used to train a machine learning classification such as Decision Tree (DT), Random Forest (RF) and Multi-layer Perceptron (MLP), denoted as MUSE-DT/RF/MLP, respectively, or are integrated within the "Attentive Intermediate Transformer Representations" (AITR) network.

3.3 Attentive Intermediate Transformer Representations

Attentive Intermediate Transformer Representations (AITR) attempts to model how human fact-checkers may iterate multiple times over the claim and collected evidence during verification, drawing various inferences and interpretations at each pass, exploring both general and fine-grained aspects and finally reassessing the entire process while assigning different weights to different aspects at each stage of analysis.

As shown in Fig.2b, AITR utilizes a stack of n Transformer encoder layers $E(\cdot) = [E_1, E_2, \cdots, E_n]$ with $h = [h_1, h_2, \cdots, h_n]$ number of multi-head attention enabling both stable attention (e.g., $h = [8, 8, 8, 8]$) and granular attention, ranging from general to fine-grained (e.g., $h = [1, 2, 4, 8]$) or from fine-grained to general (e.g., $h = [8, 4, 2, 1]$). Given initial input:

$$x_0 = [C_0; F^v; F^e; S^{v/e}] \tag{1}$$

where C_0 is a learnable classification token, F^v represents element-wise modality fusion [24] defined as $F^v = [F_{I^v}; F_{T^v}; F_{I^v} + F_{T^v}; F_{I^v} - F_{T^v}; F_{I^v} * F_{T^v}]$, $F_e = [F_{I^e}; F_{T^e}]$ and ";" denoting concatenation, intermediate Transformer outputs are given by:

$$x_i = E_i(x_{i-1}) \quad \text{for} \quad i \in \{1, 2, \ldots, n\} \tag{2}$$

From each intermediate output, we extract the processed classification tokens $\mathcal{C} = [C_1, C_2, \ldots, C_n]$ and apply the scaled-dot product self-attention mechanism:

$$\mathcal{C}_a = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V \tag{3}$$

with fully connected layers $Q = \mathbf{W}_q \cdot \mathcal{C}$, $K = \mathbf{W}_k \cdot \mathcal{C}$, $V = \mathbf{W}_v \cdot \mathcal{C}$ and $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d \times d}$. Afterwards, we use average pooling to calculate $\mathcal{C}_p = \frac{1}{n} \sum_{i=1}^n \mathcal{C}_a[:, i, :]$ and a final classification layer to predict $\hat{y}^v = \mathbf{W}_1 \cdot \text{GELU}(\mathbf{W}_0 \cdot \mathcal{C}_p)$ with $\mathbf{W}_0 \in \mathbb{R}^{1 \times d}$ and $\mathbf{W}_1 \in \mathbb{R}^{d \times 1}$.

4 Experimental Setup

4.1 Datasets

We utilize the NewsCLIPpings Merged/Balanced dataset, comprising 85,360 samples in total [14], 42,680 "Pristine" or truthful $\mathcal{I}^v, \mathcal{T}^v$ pairs sourced from credible news sources -as provided by the VisualNews dataset [34]- and 42,680 algorithmically created OOC pairs. Specifically, OOC pairs are generated by mismatching the initial image or text with another, utilizing semantic similarities, either CLIP text-to-image or text-to-text similarities, SBERT-WK for text-to-text person mismatching, and ResNet Place for scene mismatching. Furthermore, we utilize the VERITE evaluation benchmark [26] comprising 1,000 annotated samples, 338 truthful pairs, 338 miscaptioned images and 324 out-of-context pairs.

4.2 External Evidence

For NewsCLIPpings, we use the external evidence $\mathcal{I}^e, \mathcal{T}^e$ as provided by [21] comprising up to 19 text evidence and up to 10 image evidence for each I_i^v, T_i^v pair, collected via Google API; totaling to 146,032 and 736,731 textual and image evidence, respectively. Specifically, the authors employ cross-modal retrieval, namely the text T_i^v is used to retrieve potentially relevant image evidence I_i^e and image I_i^v to retrieve potentially relevant textual evidence T_i^e. We use the same Training, Validation and Testing sets as prior works to ensure comparability. For VERITE, we employ the external evidence as provided by [24].

Instead of utilizing all provided evidence as in [21], we follow [24], in re-ranking the external evidence based on CLIP [25] intra-modal similarities (image-to-image evidence, text-to-text evidence). We only select the top-1 items, as leveraging additional items was shown to degrade performance by introducing less relevant and noisy information into the detection model.

4.3 Backbone Encoder

Following [21, 24] we use the pre-trained CLIP ViT B/32 and ViT L/14 [25] as the backbone encoders in order to extract visual $F_{I^v}, F_{I^e} \in R^{d \times 1}$ and textual features $F_{T^v}, F_{T^e} \in R^{d \times 1}$ with dimensionality $d = 512$ or $d = 768$ for CLIP ViT B/32 and L/14, respectively. Unless stated otherwise, we employ L/14 while using B/32 only for comparability purposes with some older works. We use the "*openai*" version of the models as provided by OpenCLIP[3].

4.4 Evaluation Protocol

We train each model on the NewsCLIPpings train set, tune the models' hyper-parameters on the validation set and report the best version's accuracy on the NewsCLIPpings test set and unless stated otherwise, as in Table 5, we report the "True vs OOC" accuracy for VERITE.

To ensure comparability with [24] on VERITE, we report the mean "out-of-distribution cross-validation" (OOD-CV) accuracy for VERITE in Table 2. Specifically, we validate and checkpoint a model on a single VERITE fold (k=3) while evaluating its performance on the other folds. We then retrieve the model version (hyper-parameter combination) that achieved the highest mean validation score and report its mean performance of the testing folds.

4.5 Implementation Details

We train AITR for a maximum of 50 epochs, with early stopping and check-pointing set at 10 epochs to prevent overfitting. The AdamW optimizer is utilized with $\epsilon = 1e - 8$ and weight decay $= 0.01$. We employ a batch size of 512 and a transformer dropout rate of 0.1 During hyperparameter tuning, we explore learning rates $lr \in \{1e - 4, 5e - 5\}$, transformer feed-forward layer dimension $z \in \{256, 1024, 2048\}$ and for h we try the following values $[4, 4, 4, 4], [8, 8, 8, 8], [1, 2, 4, 8], [8, 4, 2, 1]$. In the ablation experiments that do not leverage intermediate transformer representations, we exclude the $h = [1, 2, 4, 8]$ and $[8, 4, 2, 1]$ configurations. To ensure reproducibility of our experiments, we use a constant random seed of 0 for PyTorch, Python Random, and NumPy.

5 Experimental Results

5.1 Ablation and Comparative Studies

Table 1 presents the ablation study results for AITR which consistently achieves the highest performance among all ablation configurations, underscoring the importance of each component. Specifically, substituting the attention mechanism with max pooling or weighted pooling leads to a notable reduction in performance across both datasets. Similarly, using the default transformer encoder (Pooling = None) without leveraging intermediate representations lowers performance. Notably, the most critical component of AITR appears to be MUSE, as removing it significantly deteriorates the model's performance, especially on VERITE, in both AITR and the default Transformer encoder. The best AITR performance was achieved with $h = [1, 2, 4, 8]$, $z = 2048$ and learning rate $5e - 5$.

In comparison with the current SotA, as shown in Table 2, MUSE-MLP competes and even outperforms much more complex architectures on NewsCLIPpings. Specifically, MUSE-MLP (90%) performs similar to RED-DOT (90.3%) while surpassing SNIFFER (88.4%), ECENet (87.7%), SEN (87.1%) and CCN (84.7%). Notably, MUSE-MLP also significantly outperforms RED-DOT on VERITE, with +4.8% relative improvement. Furthermore, integrating MUSE within AITR, significantly outperforms the SotA on NewsCLIPpings by +3.3% and VERITE by +7.5%.

---

[3]`https://github.com/mlfoundations/open_clip`

Pooling	MUSE	NewsCLIPpings	VERITE
None	No	88.28	69.98
None	Yes	93.19	78.13
Attention	No	89.39	71.04
Max	Yes	92.94	78.88
Weighted	Yes	93.03	78.73
Attention	Yes	93.31	81.00

Table 1: Ablation experiments of AITR using CLIP L/14 features.

Method	NewsCLIPpings	VERITE
CCN [21]	84.7	-
SEN [33]	87.1	-
ECENet [22]	87.7	-
SNIFFER [23]	88.4	-
RED-DOT (B/32) [24]	87.8	73.9 (0.5)
RED-DOT (L/14) [24]	90.3	76.9 (5.4)
MUSE-MLP (B/32)	85.4	74.8 (3.8)
MUSE-MLP (L/14)	90.0	80.6 (4.1)
AITR (B/32)	89.8	76.5 (2.4)
AITR (L/14)	**93.3**	**82.7** (6.1)

Table 2: Comparative analysis of evidence-based approaches for OOC detection. We report the "True vs OOC" accuracy and standard deviation (in parentheses) on VERITE under the OOD-CV protocol.

While this study primarily focuses on evidence-based approaches, we may also note that MUSE-MLP with $s(F_{I^v}, F_{T^v})$ and no external evidence, achieves 80.7% on NewsCLIPpings, as seen in Table 4, and thus outperforms complex and resource-intensive architectures such as the Self-Supervised Distilled Learning [18] (71%) that uses a fully fine-tuned CLIP ResNet50 backbone on NewsCLIPpings, the Detector Transformer [15] (77.1%) and even competes against RED-DOT without evidence (81.7%) [24].

5.2 Similarity Importance

Furthermore, we examine the contribution of each similarity measure within $S^{v/e}$. Table 3 illustrates the performance and feature importance by the Decision Tree and Random Forest classifiers. We observe that both classifier put the highest emphasis on the image-text pair similarity $s(F_{T^v}, F_{I^e})$ followed by image to image evidence $s(F_{I^v}, F_{I^e})$ and text to text evidence $s(F_{T^v}, F_{T^e})$.

Table 4 demonstrates an ablation of MUSE-MLP classifier on NewsCLIPpings (N) and VERITE (V), while excluding certain similarities. We observe that employing $S^{v/e}$ with all 6 similarity measures consistently achieves the highest overall accuracy (N=89.86, V=80.54) on both datasets. Therefore, each similarity measure contributes to some extend to the overall performance. Nevertheless, among single similarity experiments, we observe that $s(F_{T^v}, F_{I^e})$ and $s(F_{I^v}, F_{T^e})$ yield near-random performance while image-text pair similarities $s(F_{I^v}, F_{T^v})$ yield the highest performance (N=80.69, V=70.89), followed by image to image evidence $s(F_{I^v}, F_{I^e})$ (N=79.86, V=68.02) and then text to text evidence $s(F_{T^v}, F_{T^e})$ (N=71.83, V=52.19), where performance, especially on VERITE, drops significantly. Similarly, removing the image-text pair $s(F_{I^v}, F_{T^v})$ results in a notable drop performance with N=85.64% and V=69.68%. Again, similar to the Random Forest and Decision Tree classifiers, it is $s(F_{I^v}, F_{T^v})$ that has the highest contribution and is followed by $s(F_{I^v}, F_{I^e})$ and $s(F_{T^v}, F_{T^e})$.

5.3 Performance with Limited Data

As shown in Fig. 3, MUSE-MLP maintains high performance on both dataset with only using 25% of the NewsCLIPpings training set. Notably, MUSE-RF maintains high performance even when trained with 1% of the training set, which translates to only 710 samples. Surprisingly, even with 0.1% and 0.05% of the dataset, or 71 and 36 samples, respectively, the performance of MUSE-RF does not completely deteriorate. This means that the patterns that MUSE-RF relies on are simple enough that can be learned from even from a few tens or hundreds of samples.

	Random Forest	Decision Tree
$s(F_{I^v}, F_{T^v})$	0.3424	0.5918
$s(F_{I^v}, F_{I^e})$	0.2426	0.2149
$s(F_{T^v}, F_{I^e})$	0.0871	0.0178
$s(F_{I^v}, F_{T^e})$	0.0700	0.0547
$s(F_{T^v}, F_{T^e})$	0.1807	0.1154
$s(F_{I^e}, F_{T^e})$	0.0773	0.0054
NewsCLIPpings	90.16	88.44
VERITE	79.93	77.37

Table 3: Feature importance and performance by MUSE-RF and MUSE-DT.

$s(F_{I^v}, F_{T^v})$	$s(F_{I^v}, F_{I^e})$	$s(F_{T^v}, F_{I^e})$	$s(F_{I^v}, F_{T^e})$	$s(F_{T^v}, F_{T^e})$	$s(F_{I^e}, F_{T^e})$	NewsCLIPpings	VERITE
✓	-	-	-	-	-	80.69	70.89
-	✓	-	-	-	-	79.86	68.02
-	-	✓	-	-	-	55.52	55.81
-	-	-	✓	-	-	54.01	50.68
-	-	-	-	✓	-	71.83	52.19
-	-	-	-	-	✓	68.72	56.11
-	✓	-	-	✓	-	84.33	69.38
✓	-	-	-	✓	-	84.99	74.06
✓	✓	-	-	-	-	85.39	77.07
✓	✓	-	-	✓	-	88.12	77.22
✓	✓	-	-	✓	✓	88.04	78.43
✓	✓	✓	✓	-	✓	88.31	77.98
✓	-	✓	✓	✓	✓	88.46	76.62
-	✓	✓	✓	✓	✓	85.64	69.68
✓	✓	✓	✓	✓	✓	**89.96**	**80.54**

Table 4: Ablation of MUSE-MLP.

5.4 Pattern Analysis

By examining Fig.4, illustrating the distributions of the 6 similarity measures in NewsCLIPpings, we observe clear differences between Truthful and OOC distributions, primarily on $s(F_{I^v}, F_{T^v})$, $s(F_{I^v}, F_{I^e})$, $s(F_{T^v}, F_{T^e})$ and $s(F_{I^e}, F_{T^e})$. Indicatively, the median values of $s(F_{I^v}, F_{T^v})$, $s(F_{I^v}, F_{I^e})$, $s(F_{T^v}, F_{T^e})$ are 0.27, 0.91, 0.63 for Truthful pairs and 0.19, 0.69, 0.32 for OOC pairs. In contrast, $s(F_{T^v}, F_{I^e})$ and $s(F_{I^v}, F_{T^e})$ demonstrate mostly overlapping distributions between True and OOC classes which explains why they result in near-random performance in single-similarity experiments of Table 4.

In Fig. 5 we observe that the similarity distributions of VERITE exhibits relatively similar "True vs OOC" distributions with NewsCLIPpings in terms of $s(F_{I^v}, F_{T^v})$ and $s(F_{I^v}, F_{I^e})$ but not $s(F_{T^v}, F_{T^e})$ that have mostly overlapping distributions. Indicatively, the median values of $s(F_{I^v}, F_{T^v})$, $s(F_{I^v}, F_{I^e})$, $s(F_{T^v}, F_{T^e})$ are 0.31, 0.83, 0.32 for Truthful pairs and 0.24, 0.69, 0.28 for OOC pairs.

Importantly, we also observe that "True vs Miscaptioned" distributions are overlapping on $s(F_{I^v}, F_{I^e})$ and that $s(F_{T^v}, F_{T^e})$ similarities of the "Miscaptioned" class is skewed towards higher similarity, with median values of 0.29, 0.82 and 0.46 for $s(F_{I^v}, F_{T^v})$, $s(F_{I^v}, F_{I^e})$ and $s(F_{T^v}, F_{T^e})$, respectively, thus inverting the pattern found in NewsCLIPpings. As a result, as seen in Table 5, while MUSE and AITR exhibit high performance on VERITE in terms of "True vs OOC", their performance completely degrades on the "True vs Miscaptioned" evaluation.

6 Discussion

Overall, the experimental results indicate that while our methods surpass the SotA, they primarily rely on shortcuts and simple heuristics rather than detecting logical and factual inconsistencies. This raises critical questions about the realism and robustness of the current OOC detection framework, as well as how we define the task, collect data and external information.
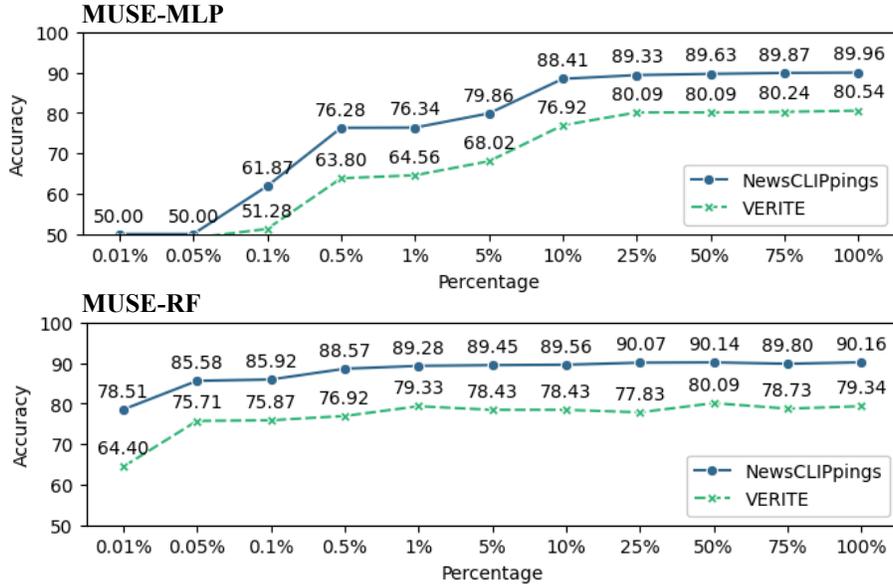
Figure 3: Performance of MUSE-MLP and MUSE-RF with limited training data.

	MUSE-MLP	AITR
NewsCLIPpings	89.96	93.31
True	92.21	93.34
OOC	87.72	93.28
VERITE "True vs OOC"	80.54	81.00
VERITE "True vs Miscaptioned"	51.18	51.78
True	89.94	92.31
OOC	70.77	69.23
Miscaptioned	12.43	13.02

Table 5: Per-class performance of MUSE-MLP and AITR on NewsCLIPpings and VERITE.

As discussed in Section 5.1, our proposed methods, MUSE and AITR, reach high accuracy scores on NewsCLIPpings, surpassing the SotA. It is important to note that OOC samples in NewsCLIPpings are generated by misaligning the original, truthful image-text pairs with other semantically similar images or texts, based on similarities by CLIP, ResNet and S-BERT features. Consequently, the truthful pairs tend to exhibit relatively higher cross-modal similarity, while OOC pairs demonstrate lower similarity, as seen in Fig.4. By relying on this simple relation, MUSE-MLP achieved a high accuracy of 81%, without incorporating any external information.

Integrating multimodal similarities with external evidence increased the detection accuracy to 90-93%. To understand this result, it is essential to consider the role of the evidence retrieval process. Following [21], external evidence is gathered through cross-modal retrieval, where the image I^v is used to retrieve text evidence T^e and the text T^v is used to retrieve image evidence I^e. Afterwards, we re-rank the retrieved items based on intra-modal similarity, meaning image-to-image and text-to-text comparisons.

Considering that the original truthful image-text pairs in NewsCLIPpings are sourced from VisualNews, which in turn collected pairs from four meainstream sources —USA Today, The Washington Post, BBC, and The Guardian— during the evidence collection process, it is highly likely that the same source article, or a highly related one, is retrieved. These conditions contribute significantly to the high accuracy observed. For instance, as illustrated in Fig.1 and discussed in Section 1, the Truthful pair exhibits very high $s(F_{I^v}, F_{I^e})$ (0.907) and a relatively high $s(F_{T^v}, F_{T^e})$ (0.597) similarity scores with the retrieved evidence while the OOC sample exhibits significantly lower scores. Although relying on such heuristics leads to high performance on the NewsCLIPpings dataset, the performance on the annotated OOC samples of VERITE is more limited, particularly for the OOC class (70%), which is the primary focus of this task.
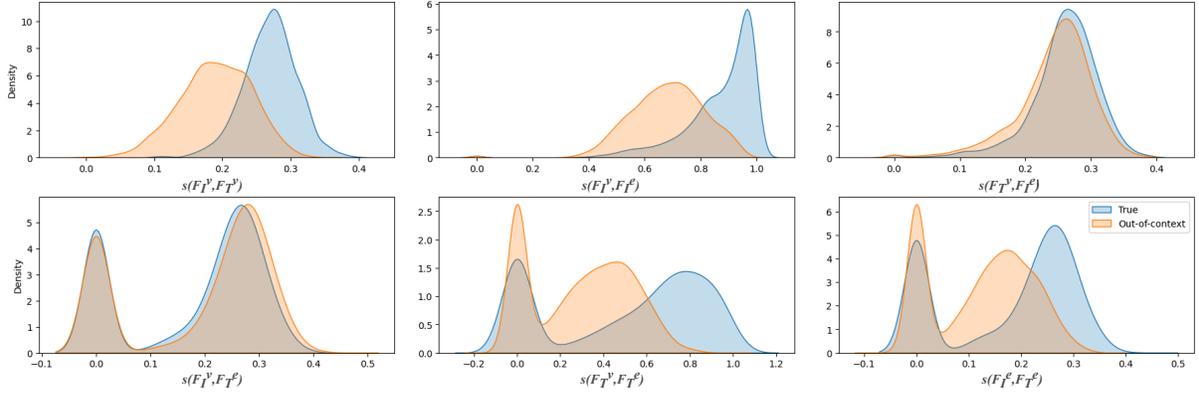
8

Figure 4: Distributions of similarity measures on NewsCLIPpings True and OOC classes.
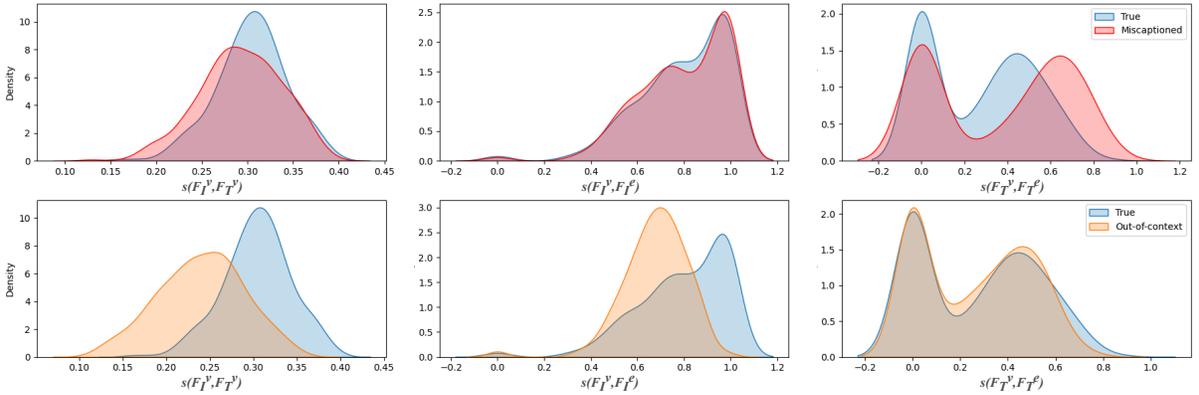

Figure 5: Distributions of similarity measures on VERITE, True vs OOC and True vs Miscaptioned classes.

In terms of the "True vs OOC" evaluation on VERITE, our methods consistently outperform the SotA, though they display lower accuracy compared to NewsCLIPpings, achieving scores around 80-82%. Additionally, there is a notable imbalance with higher accuracy for the Truthful class (90-92%) compared to the OOC class (70%). More importantly, as discussed in relation to Fig.5, MUSE and AITR can not generalize to 'Miscaptioned' samples of VERITE. This is because miscaptioned images, as defined by Snopes and Reuters, typically involve images and texts that are highly related, but with some key aspect being misrepresented in the text, such as a person, date, or event. CLIP features and similarities do not capture the subtle linguistic differences necessary to detect such cases.

Nevertheless, there is certainly room for further improving OOC detection. Firstly, we recommend that future research in this field not only utilize algorithmically generated misinformation (e.g., NewsCLIPpings) but also incorporate annotated evaluation benchmarks such as VERITE. Additionally, it is crucial to implement evaluation tests and analyses that demonstrate the models' reliance on factuality and its ability to detect logical inconsistencies, rather than merely exploiting shortcuts and simple heuristics. Furthermore, we find the current working definition of OOC to be rather limiting, as it focuses solely on truthful texts combined with mismatched (out-of-context) images. This definition may not fully capture the complexity of real-world OOC misinformation, where the texts themselves often contain falsehoods[4]. We recommend that future research in the field of automated fact-checking and evidence-based OOC detection expand their methods and training datasets to also include 'miscaptioned images,'[26] which encompass cases where an image is decontextualized but key aspects of the image, such as the person, date, or event, are misrepresented within the text. To this end, weakly annotated datasets such as Fakeddit [29] and algorithmically created datasets based on named-entity manipulations, such as MEIR, TamperedNews and CHASMA [15, 26, 30, 31], can prove useful if they are augmented with external evidence and combined with existing OOC datasets such as NewsCLIPpings. Finally, we

---

[4]"Miscaptioned: photographs and videos that are "real" (i.e., not the product, partially or wholly, of digital manipulation) but are nonetheless misleading because they are accompanied by explanatory material **that falsely describes their origin, context, and/or meaning**." https://www.snopes.com/fact-check/rating/miscaptioned

recommend future researchers to consider the problem of "leaked evidence" while collecting external information from the web [35, 36].

7 Conclusions

In this study, we adress the challenge of out-of-context (OOC) detection by leveraging multimodal similarities (MUSE) between image-text pairs and external image and text evidence. Our results indicate that MUSE, even when used with conventional machine learning classifiers, can compete against complex architectures and even outperform the SotA on the NewsCLIPpings and VERITE datasets. Furthermore, integrating MUSE within our proposed "Attentive Intermediate Transformer Representations" (AITR) yielded further improvements in performance. However, we discovered that these models predominantly rely on shortcuts and simple heuristics for OOC detection rather than assessing factuality. Additionally, we found that these models excel only under a narrow definition of OOC misinformation, but their performance deteriorates under other types of de-contextualization. These findings raise critical questions about the current direction of the field, including the definition of OOC misinformation, dataset construction, and evidence collection and we discuss potential future directions to address these challenges.

Acknowledgments

References

[1] Vaishali Vaibhav Hirlekar and Arun Kumar. Natural language processing based online fake news detection challenges–a detailed review. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 748–754. IEEE, 2020.

[2] Rahul Thakur and Rajesh Rohilla. Recent advances in digital image manipulation detection techniques: A brief review. *Forensic science international*, 312:110311, 2020.

[3] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4):3974–4026, 2023.

[4] Anna Wilson, Seb Wilkes, Yayoi Teramoto, and Scott Hale. Multimodal analysis of disinformation and misinformation. *Royal Society Open Science*, 10(12):230964, 2023.

[5] Carmela Comito, Luciano Caroprese, and Ester Zumpano. Multimodal fake news detection on social media: a survey of deep learning techniques. *Social Network Analysis and Mining*, 13(1):101, 2023.

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

[7] Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37, 2020.

[8] 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.

[9] Nguyen Vo and Kyumin Lee. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, 2020.

[10] Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *European Conference on Information Retrieval*, pages 499–507. Springer, 2020.

[11] Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Mária Bieliková. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500, 2023.

[12] Sijia Qian, Cuihua Shen, and Jingwen Zhang. Fighting cheapfakes: using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. *Journal of Computer-Mediated Communication*, 28(1):zmac024, 2023.

[13] 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.

[14] 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.

[15] 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.

[16] Shivangi Aneja, Cise Midoglu, Duc-Tien Dang-Nguyen, Michael Alexander Riegler, Paal Halvorsen, Matthias Nießner, Balu Adsumilli, and Chris Bregler. Mmsys' 21 grand challenge on detecting cheapfakes. *arXiv preprint arXiv:2107.05297*, 2021.

[17] 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.

[18] 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.

[19] Duc-Tien Dang-Nguyen, Sohail Ahmed Khan, Michael Riegler, Pål Halvorsen, Anh-Duy Tran, Minh-Son Dao, and Minh-Triet Tran. Overview of the grand challenge on detecting cheapfakes at acm icmr 2024. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1275–1281, 2024.

[20] Yimeng Gu, Mengqi Zhang, Ignacio Castro, Shu Wu, and Gareth Tyson. Learning domain-invariant features for out-of-context news detection. *arXiv preprint arXiv:2406.07430*, 2024.

[21] 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.

[22] 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.

[23] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13052–13062, 2024.

[24] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Reddot: Multimodal fact-checking via relevant evidence detection. *arXiv preprint arXiv:2311.09939*, 2023.

[25] 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.

[26] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4, 2024.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 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, 2018.

[33] 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4268–4280, 2023.

[34] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, 2021.

[35] 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.

[36] Zacharias Chrysidis, Stefanos-Iordanis Papadopoulos, Symeon Papadopoulos, and Panagiotis Petrantonakis. Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, pages 73–81, 2024.