

Evidence-Grounded Multimodal Misinformation Detection with Attention-Based GNNs

Sharad Duwal^{1*}, Mir Nafis Sharear Shopnil¹, Abhishek Tyagi², Adiba Mahbub Pruma²

¹Fatima Fellowship ²University of Rochester

Abstract

Multimodal out-of-context (OOC) misinformation is misinformation that repurposes real images with unrelated or misleading captions. Detecting such misinformation is challenging because it requires resolving the context of the claim before checking for misinformation. Many current methods, including LLMs and LVLMs, do not perform this contextualization step. LLMs hallucinate in absence of context or parametric knowledge. In this work, we propose a graph-based method that evaluates the consistency between the image and the caption by constructing two graph representations: an evidence graph, derived from online textual evidence, and a claim graph, from the claim in the caption. Using graph neural networks (GNNs) to encode and compare these representations, our framework then evaluates the truthfulness of image-caption pairs. We create datasets for our graph-based method, evaluate and compare our baseline model against popular LLMs on the misinformation detection task. Our method scores 93.05% detection accuracy on the evaluation set and outperforms the second-best performing method (an LLM) by 2.82%, making a case for smaller and task-specific methods.

1 Introduction

Misinformation has emerged as a major issue with social media (Denniss and Lindberg, 2025). Bad actors disseminate fake information to spread hate, political division, conspiracy theories, health misinformation, and rumors to the disadvantage of targeted groups (Fisher et al., 2016; Islam et al., 2020; Denniss and Lindberg, 2025). Visual content is a more viral vector of misinformation than text: fact-checks collected by Dufour et al. (2024) found 80% of claims contained visual media. Videos are becoming more common in misinformation as of 2022, as are AI-generated media (Dufour et al., 2024).

*Correspondence to: sharad.duwal@gmail.com

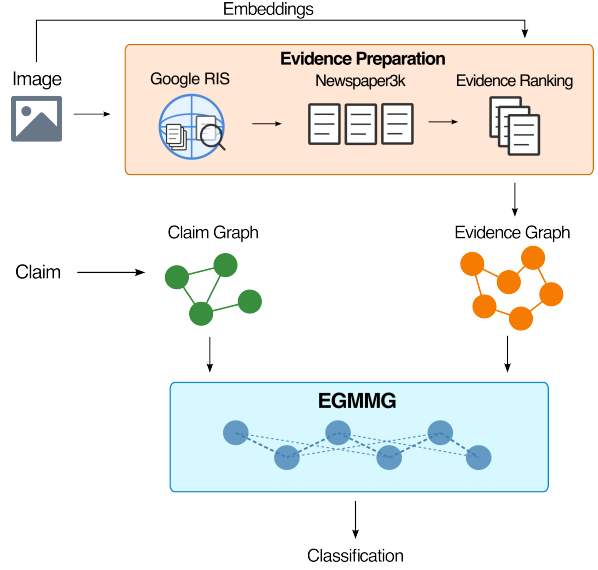


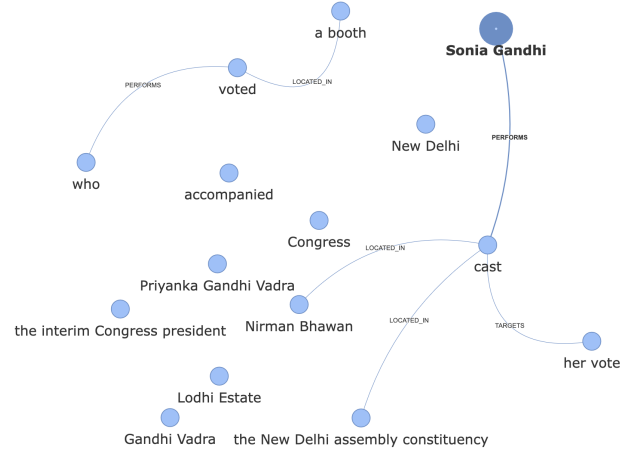
Figure 1: **The EGMMG pipeline.** For an image-claim sample, the pipeline prepares two graphs, evidence graph and claim graph, using online evidence retrieval followed by a rule-based analysis of subject-object relations in the evidence documents. Once we have the two graphs, we use a graph attention-based classifier to detect misinformation.

Images repurposed with different captions to make false claims are categorized as out-of-context (OOC) misinformation (Fazio, 2020). OOC misinformation is particularly egregious because the image is generally authentic and the misinformation stems from context manipulation (Qi et al., 2024). This makes them more believable, and people, used to photojournalism, tend more readily to accept the claims at face value (Fazio, 2020).

Existing methods to detect OOC misinformation have explored traditional classification algorithms, large language models (LLMs), and large vision language models (LVLMs). While classical methods like feature-based classification were a good starting point, LLMs have proved to be particularly good at detecting (and also explaining) OOC misinformation (Qi et al., 2024; Xuan et al., 2024;



(a) Image



(b) Claim graph generated by EGMMG

Figure 2: *Data sample*: Image and claim graph for claim «**Sonia Gandhi, the interim Congress president, cast her vote at Nirman Bhawan in the New Delhi assembly constituency, accompanied by Priyanka Gandhi Vadra, who also voted at a booth in Lodhi Estate.**» A section of the evidence graph is provided in the Appendix (Figure 4).

Tahmasebi et al., 2024). This improvement can be attributed to the world knowledge obtained via pre-training that allows LLMs to be multitask learners (Radford et al., 2019; Brown et al., 2020).

An offshoot in OOC multimodal misinformation detection focuses on contextualizing an image claim before reasoning about its truth. Grounding the claim and supporting visual elements using circumstantial evidence assists veracity detection. Tonglet et al. (2024) adopt the “5 pillars of verification” as described by Urbani (2020), which grounds images on five properties: provenance, source, date, location, and motivation to contextualize the images. Tonglet et al. (2025) use these “pillars” to establish veracity.

However, a major issue with LLMs is that they “hallucinate” when relevant information is not present as parametric information while generating (Shuster et al., 2021; Maynez et al., 2020). Hallucinations are particularly problematic because LLMs can generate explanations that appear credible even when untrue. This is not a desired property in a tool for misinformation detection.

Retrieval-augmented generation (RAG) (Gao et al., 2024), in-context prompting (Brown et al., 2020) and knowledge graphs (KGs) have been introduced to ensure factuality of language models. KGs are effective in adding structured external information; several KG augmented LLMs for misinformation have also been introduced (Lu and Li, 2020; Opsahl, 2024; Tan et al., 2024; Wang and Shu, 2023). Graph-based methods are widely used

in evaluating factuality because they exploit meaningful node relationships (Kim et al., 2023).

In this work, we take a graph-based approach to detect multimodal OOC misinformation. We focus on the image contextualization task discussed earlier. We first create an online evidence retrieval pipeline that *hydrates* image-text datasets by collecting textual evidence for the samples using reverse image search (RIS)¹. The textual evidence found online is used to construct an evidence graph, while the caption in the claim is used to generate a claim graph. We also introduce a baseline graph attention method to learn misinformation detection over the generated graph data.

Our contributions are threefold:

1. We introduce a text-grounding approach for the image contextualization task using evidence graphs, which capture the entities in the image and the relations between them,
2. We introduce a baseline graph attention method to tackle the multimodal OOC misinformation task using the grounding approach,
3. We pass several publicly available misinformation datasets through the pipeline and establish the model’s performance.

2 Related work

2.1 Attention-based GNNs

Veličković et al. (2018) introduced graph attention networks (GAT) that used attention, popular in

¹We use Google Vision API.

sequence-based natural language tasks, to tackle irregular graph structures, which GCN (Kipf and Welling, 2017) before them did not handle effectively.

Other than the standard GAT and GCN, there have been approaches like Graph Transformer (Shi et al., 2021), which uses node features and labels jointly, and GATv2 (Brody et al., 2022), which calculates dynamic attention as opposed to GAT’s static.

GNN (attention) methods are now being specialized for particular tasks like drug discovery, material property prediction, misinformation detection, etc (Zhang et al., 2024; Lu and Li, 2020). More work in refining and making them more adoptable via scalability and interpretability is underway (Kazi et al., 2021).

2.2 OOC Misinformation Detection

Early methods of detecting OOC misinformation focused on image-text similarity and object alignment (Aneja et al., 2023) and researchers designed various LLM-based architectures for it (Qi et al., 2024; Aneja et al., 2023; Xuan et al., 2024; Tahmasebi et al., 2024). However, these models were limited in the information available to them. Architectures were proposed to mitigate this limitation and to incorporate external information from the internet (Abdelnabi et al., 2022). For example, Abdelnabi et al. (2022) suggested gathering external knowledge regarding both the image and the caption of the (image, caption) pair to detect OOC misinformation. Papadopoulos et al. (2025) showed that providing more context by adding external sources improved performance, even with relatively simple models. Our work builds on this concept as we also focus on providing external context to image-caption pairs.

Qi et al. (2024) proposed the SNIFFER model, which not only detects OOC misinformation, but also provides an explanation for the model’s choice, thus improving the interpretability of the model. Tonglet et al. (2024) suggested that providing context to images by asking various questions through an LLM pipeline could establish the factuality of a sample. However, LLM-based models are resource-intensive to train and have the potential to hallucinate. Graph-based approaches make the systems more accurate and explainable using causal methods (Opsahl, 2024; Wang and Shu, 2023; Tan et al., 2024; Lu and Li, 2020).

2.3 Datasets

Misinformation detection on text is well-studied, with a great amount of work. Research in multi-modal misinformation detection is also picking up, due to growing necessity and interest. As a result, there are several datasets, distant-supervised and manually annotated. There are also fine-grained divisions along which these datasets are categorized: textual distortion, visual distortion, edited image, repurposed image, etc. FEVER (Thorne et al., 2018) and Politifact (Shu et al., 2019) focus on textual distortion, especially rumors. Fakeddit (Nakamura et al., 2020) was collected from over 1 million samples and included various categories of fake news, distantly supervised. Factify is another multimodal fact verification dataset, collected from tweets of US and Indian news sources (Mishra et al., 2022). NewsCLIPPings (Luo et al., 2021) and COSMOS (Aneja et al., 2023) focused particularly on OOC misinformation too. More recently, especially to tackle the issues related to distortion using AI (textual, visually altered, generated), LLMFake (Chen and Shu, 2024) and MM-FakeBench (Liu et al., 2024) have been introduced.

3 Method

Problem Formalization Given an image I and a textual claim (usually a caption) C , the task is to determine a veracity score $s \in [0, 1]$ indicating how well the claim supports the image. An image-caption pair will have a high s if the image and the caption are in context (i.e. are related via the subject, object or event).

Misinformation detection that depends solely on images and captions have some issues: i) images might not provide explicit context and ii) captions are often short, single-source, and might also not provide detailed context. These shortcomings present a challenge in establishing veracity.

To tackle this, we focus on the image contextualization task before processing for veracity. We perform reverse image search to obtain resources related to the image with the assumption that these resources (news articles, blog posts, etc.) provide context to the event depicted in the image and claim.

We create an online evidence retrieval pipeline and run it on OOC misinformation datasets to construct contextualized knowledge graph data from the image-caption samples in the datasets.

We finally introduce a baseline graph attention

method to perform misinformation detection on this data.

For our experiments, we focus on positive and negative classes only (for example, Refute and Support_Multimodal for the Factify dataset). We describe below our evidence retrieval pipeline. Where relevant, we use the Factify dataset (Mishra et al., 2022) as a placeholder, but the pipeline can be easily adapted to other image-caption datasets mentioned in §2.3.

3.1 Data

We start by extracting the claim image, claim text, and the misinformation label from the multimodal misinformation dataset.

For the Factify dataset, there are claim images, support “document” images, lemmatized claims, lemmatized related document, and a classification label. For our task, we are only interested in the claim image, claim text, and the label.

To generate the graphs required for our task, we first gather *evidence* for the image (related textual documents on the web) and rank them based on their similarity with the image. Then we use the textual evidence to generate knowledge graphs with entities (subjects and objects) in the text as nodes and relations between them as edges. For an image-caption pair, we follow the steps below to generate the data.

1. **Evidence documents:** We use the Google Vision API to get web pages that use the claim image (full or partial matches). We assume that news articles, essays, and blogs host these images to provide reporting and commentary, which could be useful context. For an image, we try to get at most 30 web pages containing the image. We discard images for which the Vision API does not return at least one web page.

For the web content extraction, we use Newspaper3k. We extract the metadata including text, language, author name, publication date, and time. While the metadata could also be leveraged for detection (similar to (Tonglet et al., 2024)), because our focus is on the main text and the entity-relationship, we extract the text content (“evidence” documents) $E = e_1, e_2, \dots, e_m$ from web pages containing image I .

As quality check before inclusion as evidence, we rank and filter the web pages. We ac-

Dataset	Orig	Final
Factify (Mishra et al., 2022)	14000	4945
Factify Val	3000	1145
COSMOS (Aneja et al., 2023)	1700	813
MMFB Val (Liu et al., 2024)	1000	391
MMFB Test ²	6750	3829

Table 1: Dataset statistics showing sample counts before and after evidence retrieval and processing. We do not focus on large train sets due to limited resources.

complish this by computing similarity between the embeddings of the web page texts and the image. We use the clip-ViT-L-14 model offered by SentenceTransformers that can embed both images and texts (Reimers and Gurevych, 2019). We get the embeddings for the web page documents (and page titles) and the image separately and use cosine similarity to get the top-k evidences per image. For evidence document e_i , we calculate the similarity score

$$\text{sim}(e_i, I) = \text{cos_sim}(\phi_I(I), \phi_T(e_i))$$

where ϕ_I is the image embedding function and ϕ_T is the text embedding function.

After we have the similarity scores for all the evidence documents, we select the top 7 and concatenate them into the final *evidence* for our task.

2. **Graph Construction:** We construct evidence and claim graphs using the final evidence and the claim texts. Nodes for both the graphs are entities, events, and locations as identified by the en_core_web_lg spaCy model (Honnibal et al., 2020). Relations between the entities are identified based on how they are related: The possible relations (edges) are: PERFORMS, EXPERIENCES, TARGETS, LOCATED_IN, HAS_STATE, and SAME_AS. To annotate these edge types, we use the token POS (verb, etc.) and dependencies (nsubjpass, prep, etc.). More details on graph construction are in Appendix B. We construct the claim graphs similarly using the captions.

²The original set has 10000 samples, but we remove samples with AI-generated images.

3.2 Model

We introduce a GNN-based method as baseline for the graph image contextualization task. The method leverages the topological and relational information present in the evidence and claim graphs using cross-graph attention.

The first step is to extract meaningful representations from the created graphs. This can be accomplished by extracting node and edge features.

3.2.1 Node features

We use the node label embeddings and node neighborhood information to obtain the node features. We get the label embedding using a language model (BERT (Devlin et al., 2019), for example). For the neighborhood structure, we utilize properties like in-degrees, out-degrees, total degrees, pagerank, and reverse pagerank.

Since the label’s text embeddings are higher-dimensional (depending on the LM of choice; 768 for BERT-base) than the neighborhood structure information (5, for the five structural properties above), we project these two representations to a common dimension during training so that both contribute meaningfully to the node features. For this, we implement a node features projector inside the graph encoder (described in §3.2.3.) In addition to the linear projections for the label and the structure information, we implement learnable multiplicative coefficients that determine the contribution of each feature to the final node representation.

Thus, the embedding of node v at initialization is:

$$h_v^{(0)} = \alpha \cdot \text{LM}(v) + \beta \cdot \text{NS}(v)$$

where LM and NS are functions that project language model embedding and neighborhood information respectively into a common space and α and β are weight coefficients.

3.2.2 Edge features

Since edges inform node relationships, we obtain the edge features as well. The properties we focus on are: edge centrality, common predecessors, common successors, in-jaccard, out-jaccard, forward path length and backward path length.

However, we find that edge features are not beneficial in their current formulation to the architecture on the misinformation detection task (see ablation in Table 4). They also add computational overhead while data preparation. For a detailed discussion, please refer to Appendix B.

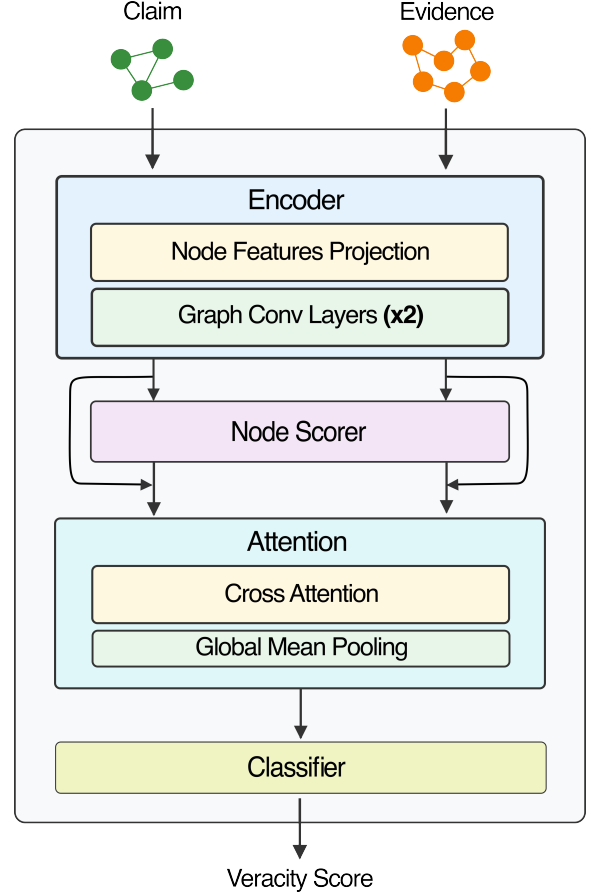


Figure 3: The EGMMG classifier.

3.2.3 Architecture

After initializing the node and edge features for both the evidence and claim graphs as described above, we perform message-passing between the nodes using graph convolutions (GATConv (Veličković et al., 2018), TransformerConv (Shi et al., 2021), GATv2Conv (Brody et al., 2022)) to update the node representations. This allows subgraph neighborhoods to inform each node’s representation. The node representation at layer $\ell + 1$ is given by:

$$h_v^{(\ell+1)} = \text{GraphConv} \left(h_v^{(\ell)}, \{h_u^{(\ell)} : u \in \mathcal{N}(v)\} \right)$$

where GraphConv is the convolution function, $h_v^{(\ell)}$ and $h_u^{(\ell)}$ are node representations at layer ℓ and $\mathcal{N}(v)$ is the neighborhood of the node v .

We experiment with all three convolution methods discussed above and go with TransformerConv for our final model. For experimental results with each convolution type, refer to Table 6 in Appendix.

Since all nodes might not be equally important for the detection task, we assign node importance

with the help of a trainable node scorer. We multiply the node embeddings generated by the convolution layers with the node scorer, which is a score $s_v \in \mathbb{R}$:

$$\hat{h}_v = h_v \cdot s_v$$

where $s_v = \text{NodeScorer}(h_v)$ and \hat{h}_v is the importance-weighted node embedding.

We then perform cross-attention computation between the evidence and claim graphs. Similar to encoder-decoder cross-attention mechanism in transformers, this gives us the attention pattern between the evidence nodes and claim nodes. We use the claim nodes for the query part of the attention calculation and the evidence nodes for the key and value parts:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q is the query projection of the claim graph and K and V are key and value projections of the evidence graph.

After cross-attention, we apply global mean pooling to get the average node feature to represent the graphs in the batch. For each batch, we calculate:

$$\mathbf{g}_X = \frac{1}{|\mathcal{V}_X|} \sum_{v \in \mathcal{V}_X} h_X(v)$$

where $X \in \{\text{evidence, claim, attended}\}$, and $h_X(v)$ is the node feature for v . We calculate the global mean for all three types: evidence graph, claim graphs, and cross-attention. We then concatenate them to get a combined representation of each sample:

$$\mathbf{f} = \mathbf{g}_{\text{evidence}} \parallel \mathbf{g}_{\text{claim}} \parallel \mathbf{g}_{\text{parallel}}$$

The combined features are fed into a classifier layer that makes the decision, outputting a score between 0 and 1 that represents how well the evidence supports the claim:

$$s = \sigma(W \cdot \mathbf{f} + b)$$

where σ is the sigmoid activation function, W is the weight matrix, b is the bias term and $s \in [0, 1]$ represents the claim veracity score.

4 Experimental Setup

Our work focuses on the image contextualization task and generates graph data based on image-caption datasets using reverse image search. Since the graph data generated for our task is effectively new, with different number of samples (depending on availability of online evidence) than the original (Table 1), we compare the performance of our baseline method with frontier LLMs available at the time of writing: Claude Sonnet 3.7, Claude Haiku 3.5, GPT 4o, GPT 4o-mini. We focus our experiments and evaluation on the Factify dataset first, then discuss about robustness to other datasets and generalizability in subsection §4.2.

4.1 Evaluation Sets

For the Factify evaluation set, we use the validation set introduced in the Factify paper. The Factify validation set consists of 7000 samples (1500 for each label type). Because we focus only on the positive and negative labels, we extract 3000 samples (1500 each for Refute and Support_Multimodal classes). We apply the evidence retrieval pipeline described in §3.1. At the end we have 1145 samples in the evaluation set.

We prompt the LLMs with the evidence document and the claim as input, and the LLMs are asked whether the evidence-claim text pairs are misinformation or not. (Prompts can be found in Appendix C and Figure 5.) The models abstain on some samples, and so we construct three evaluation sets based on abstentions.

1. EVAL_ALL ($n = 1145$): We prompt the models to answer regardless if they consider the evidence insufficient. (All models answer for all 1145 samples under this setting, except Sonnet which abstained on 64 samples even when prompted to answer strictly between “True” or “False”.)
2. EVAL_SUFFICIENT: We prompt the models allowing them to abstain on samples that they do not consider answerable with the provided evidence. This has the obvious issue of the models choosing to only answer for samples that are “easy” for them to decide, abstaining from difficult ones. Our method does not abstain, but we include these results for completeness.
3. EVAL_COMMON ($n = 461$): One limitation with EVAL_SUFFICIENT is that each of the mod-

Model	EVAL_ALL ($n = 1145$)		EVAL_SUFFICIENT		EVAL_COMMON ($n = 461$)	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Sonnet 3.7	0.6182	0.4969	0.8515	0.8325	0.8676	0.8571
Haiku 3.5	0.6692	0.5830	0.8695	0.8695	0.8915	0.8826
GPT 4o	0.6872	0.6209	0.8828	0.8841	0.9023	0.8936
GPT 4o-mini	0.6419	0.5438	0.8112	0.7891	0.8741	0.8619
EGMMG (ours)	0.8305	0.8455	-	-	0.9305	0.9219

Table 2: Performance metrics comparison across different test sets. EVAL_SUFFICIENT counts are different for each model.

els may not have the exact same samples, thus making comparison difficult. So we also take an overlap set that all the models consider answerable and have answered. The EVAL_COMMON set is given by

$$\mathbf{ES}_{\text{Sonnet}} \cap \mathbf{ES}_{\text{GPT4oMini}} \cap \mathbf{ES}_{\text{Haiku}}$$

where \mathbf{ES} is EVAL_SUFFICIENT.

4.2 Generalization and Robustness

In addition to evaluating the model on the Factify benchmark, we evaluate the generalizability of our pipeline on other datasets. For these tests, we first obtain online evidence and prepare graph data. We then perform train-test performance analysis on individual datasets.

We focus on:

- evaluating the model on different datasets using a standard 85:15 train-test split, training and evaluating the accuracy of the model on the test set,
- architecture ablation on a particular test set to understand the model’s robustness.

5 Results

Table 2 presents a performance comparison between multiple LLMs and our proposed approach. The evaluation was conducted across the three test sets described in section 4.1. For each evaluation category, we report accuracy and F1 scores.

Our method outperforms the LLMs on both EVAL_ALL and EVAL_COMMON sets. As discussed earlier, our model does not abstain, so we do not have an EVAL_SUFFICIENT set. On the EVAL_ALL test set, EGMMG achieves an accuracy of 0.8305 and F1 score of 0.8455, substantially higher than GPT-4o (0.6872/0.6209), Haiku 3.5 (0.6692/0.5830), GPT4o-mini (0.6419/0.5438) and Sonnet 3.7 (0.6182/0.4969).

On the EVAL_COMMON set, EGMMG has 0.9305 accuracy and 0.9219 F1 scores, compared to GPT-4o’s 0.9023/0.8936. GPT-4o performs marginally better on EVAL_SUFFICIENT (0.8828/0.8841) compared to Haiku 3.5 (0.8695/0.8695).

Table 3 presents our method’s performance on the different datasets we have discussed. We report these numbers to discuss the generalization abilities of our baseline method. The best performance is achieved on the Factify dataset (85:15 train-test split) (0.8248). It performs relatively well on the COSMOS Test dataset (0.7750) as well. But the method struggles with the MMFB Val (0.7100) and MMFB Test (0.6823) sets. For a baseline approach, with no special modifications for individual datasets, the model maintains reasonable generalization capabilities.

Dataset	Acc
Factify	0.8248
MMFB Val	0.7100
MMFB Test	0.6823
COSMOS Test	0.7750

Table 3: Performance of our model on different datasets on 85:15 train-test split. For each dataset, we train on the train split and report performance on the unseen test split.

6 Discussion

Earlier work in image contextualization has focused on using metadata (Tonglet et al., 2024, 2025), image entities extraction (Aneja et al., 2023; Ma et al., 2024) and LLM knowledge (Qi et al., 2024; Tahmasebi et al., 2024), among others. This work focuses on using related online text content only. Below we briefly discuss the performance of the model, its robustness to ablation and datasets

and efficiency.

6.1 Model Performance and Robustness

Table 2 shows that our classifier performs better than frontier LLMs on the Factify evaluation sets. All methods have access to the same amount of data (in text or graph format).

To investigate the model’s robustness and the contribution of individual components, we conduct an ablation study (Table 4). We report the performances on the EVAL_COMMON set.

The full model achieves the best performance with 0.9305 accuracy and 0.9219 F1 score. Adding edge features (§3.2.2) causes a performance drop to 0.9132 accuracy and 0.9 F1 score, probably indicating that edge information adds noise and might need to be processed differently. A more substantial degradation occurs when using unweighted node embeddings (i.e. without a node feature projector and weight coefficients) (0.8741/0.8473) or reduced-dimension 384-dim node embeddings (instead of BERT’s 768-dim) (0.8872/0.8725).

Weighing the contributions of node label embeddings and node neighborhood information seems particularly important, as evidenced by the differences in accuracy and F1 score (especially F1 score). This is due to the 5 dimensional neighborhood structure information in the unweighted setup.

While further improvements seem possible by increasing the node dimensions and dedicated processing of edge features, we can see (from Tables 3 and 4) that our method is robust.

6.2 Efficiency

The results are also encouraging from an efficiency point of view. Our model is significantly smaller compared to the LLMs we compare it against. The 768-dim node embeddings variant of our model has 10M parameters (10, 724, 391) and takes up around 41 megabytes of disk space.

Due to the size and intermediate dimensions, the computational costs are also sizably small. We run our training and tests on a single NVIDIA T4. The inference is similarly cheap.

7 Conclusion

In this work, we introduced a graph-based method to tackle the image contextualization task for multimodal out-of-context misinformation detection. We developed an online evidence retrieval pipeline

Model Configuration	Accuracy	F1 Score
Full Model	0.9305	0.9219
+ edge features	0.9132	0.9
unweighted node embeddings	0.8741	0.8473
384-dim node embeddings	0.8872	0.8725

Table 4: Ablation study results. Each row represents the performance when a specific component is removed or modified.

and graph data generation method to ground images with textual evidence available online. Then we introduced a GNN-based method to learn misinformation classification over the generated graph data. We experimented with several publicly available datasets using our method. Our results show that using relevant text information, in the form of entity-relation graphs, is greatly effective in misinformation detection, evidenced by the performance of our proposed method over frontier LLMs when provided the same information. The effectiveness of the method also highlights possible improvements.

Limitations

Some limitations of this work are highlighted below.

First, our methods do not utilize the images directly. We use reverse search on images to get web pages with matches, but we do not process the image itself for information. Extracting actors and events from the image could potentially improve the model.

Second, the evidence retrieval method can be made more robust. Currently, images that do not have web page matches are discarded. Better (or multiple) image search methods could help improve web page retrieval. We currently also do not have a method to establish the relevance of evidences collected, thus depending completely on the image match and the cosine similarity between the evidence texts and the image embedding.

Similarly, there are aspects of the text-to-graph pipeline that have potential room for improvement. With more rules for entity and relations extraction, the method could extract more information relevant to the veracity detection task.

Our approach currently does not use existing

large knowledge graphs (for example, ConceptNet) to help incorporate real-world logic. It did not fit the current research scope, but might assist the task with common-sense knowledge.

Earlier work has been about using metadata only and this work focuses on using related text content only. Combining these methods for evidence and adding LLMs in the workflow would be a worth-exploring direction.

References

- Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14940–14949.
- Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2023. 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.
- Shaked Brody, Uri Alon, and Eran Yahav. 2022. [How attentive are graph attention networks?](#)
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Canyu Chen and Kai Shu. 2024. [Can LLM-generated misinformation be detected?](#) In *The Twelfth International Conference on Learning Representations*.
- Emily Denniss and Rebecca Lindberg. 2025. [Social media and the spread of misinformation: infectious and a threat to public health](#). *Health Promotion International*, 40(2):daaf023.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Duffield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, and Christoph Bregler. 2024. [Ammeba: A large-scale survey and dataset of media-based misinformation in-the-wild](#).
- Lisa Fazio. 2020. Out-of-context photos are a powerful low-tech form of misinformation. *The Conversation*, 14(1).
- Marc Fisher, John Woodrow Cox, and Peter Hermann. 2016. [Pizzagate: From rumor, to hashtag, to gunfire in d.c.](#)
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. [Retrieval-augmented generation for large language models: A survey](#).
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- S M Hasibul Islam, S M Hasibul Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu Hena Mostafa Kamal, S. M. Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad A. Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, Abrar Ahmad Chughtai, and Holly Seale. 2020. [Covid-19–related infodemic and its impact on public health: A global social media analysis](#). *The American Journal of Tropical Medicine and Hygiene*, 103:1621 – 1629.
- Anees Kazi, Soroush Farghadani, and Nassir Navab. 2021. [Ia-gcn: Interpretable attention based graph convolutional network for disease prediction](#).
- Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. [Factkg: Fact verification via reasoning on knowledge graphs](#).
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#).
- Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. 2024. [Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms](#). *arXiv preprint arXiv:2406.08772*.
- Yi-Ju Lu and Cheng-Te Li. 2020. [GCAN: Graph-aware co-attention networks for explainable fake news detection on social media](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.
- Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. [NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

- Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. [Event-radar: Event-driven multi-view learning for multimodal fake news detection](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821, Bangkok, Thailand. Association for Computational Linguistics.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, A. Sheth, and Asif Ekbal. 2022. [Factify: A multi-modal fact verification dataset](#). In *DE-FACTIFY@AAAI*.
- Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. [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, Marseille, France. European Language Resources Association.
- Tobias A. Opsahl. 2024. [Fact or fiction? improving fact verification with knowledge graphs through simplified subgraph retrievals](#).
- Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2025. [Similarity Over Factuality: Are we Making Progress on Multimodal Out-of-Context Misinformation Detection?](#). In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5041–5050, Los Alamitos, CA, USA. IEEE Computer Society.
- Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. [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 (CVPR)*, pages 13052–13062.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. 2021. [Masked label prediction: Unified message passing model for semi-supervised classification](#).
- Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. [Fakenewsnet: A data repository with news content, social context and spatio-temporal information for studying fake news on social media](#).
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval augmentation reduces hallucination in conversation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. [Multimodal misinformation detection using large vision-language models](#).
- Fiona Anting Tan, Jay Desai, and Srinivasan H. Sengamedu. 2024. [Enhancing fact verification with causal knowledge graphs and transformer-based retrieval for deductive reasoning](#). In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 151–169, Miami, Florida, USA. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [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, New Orleans, Louisiana. Association for Computational Linguistics.
- Jonathan Tonglet, Marie-Francine Moens, and Iryna Gurevych. 2024. ["image, tell me your story!" predicting the original meta-context of visual misinformation](#).
- Jonathan Tonglet, Gabriel Thiem, and Iryna Gurevych. 2025. [Cove: Context and veracity prediction for out-of-context images](#).
- Shaydanay Urbani. 2020. [Verifying online information](#).
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. [Graph attention networks](#).
- Haoran Wang and Kai Shu. 2023. [Explainable claim verification via knowledge-grounded reasoning with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.
- Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R. Fung, and Heng Ji. 2024. [Lemma: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation](#).
- Yang Zhang, Caiqi Liu, Mujiexin Liu, Tianyuan Liu, Hao Lin, Cheng-Bing Huang, and Lin Ning. 2024. [Attention is all you need: utilizing attention in ai-enabled drug discovery](#). *Briefings in Bioinformatics*, 25(1):bbad467.

A Model Details

Our baseline model has 2 convolution layers for message passing between the nodes. For node label embeddings, we experiment with two language models: a) BERT-Base (768 embedding size) and b) SentenceTransformer’s all-MiniLM-L6-v2 (384 embedding size). The node embeddings for our model includes the label embedding and the neighborhood structure information. To accomplish this we use a feature projector to map text embeddings (384- or 768-dim) and structural features (5-dim) to a common (389-dim or 773-dim) space.

We use TransformerConv as our convolution layer after a set of experiments (see Table 6. We also use multiheaded attention. Among the two conv layers, the first layer uses 4 attention heads and the second layer uses 2 attention heads.

The hidden dimension of the model is 1024.

All processing and experiments (graph data generation, training, inference) were run on an NVIDIA T4.

Hyperparameter	
Node label embedding	768
Hidden dimensions	1024
Conv. layers	2
Learning rate	3e-4
Batch size	64
No of parameters	10,724,391

Table 5: Model details.

B Knowledge Graph Construction

Here we describe the rules for graph construction for our system, including node and edge type taxonomies, their extraction and relationship formation.

B.1 Entities as nodes

First let’s define what object types we consider a node: ENTITY, EVENT, STATE, LOCATION, TIME and ATTRIBUTE.

Entity Type Assignment Entities are classified based on their NER labels (when present) into one of the aforementioned subtypes. When the label is not present or doesn’t match any predefined subtype, we default to the ENTITY type.

Node Identification and Deduplication

We map entity label to node IDs using the original- and lowercase variants. For entities not beginning with "the", we additionally map "the [entity]" to the same node ID to handle different references to the same entity to ensure consistency.

B.2 Relations as edges

As mentioned earlier in the document, the edge types we focus on are: PERFORMS, EXPERIENCES, TARGETS, LOCATED_IN, HAS_STATE and SAME_AS.

Verbs as events For tokens with VERB part-of-speech, we create EVENT nodes. We establish different relationship types based on following rules:

- `nsubj` (subject) creates a PERFORMS edge from subject to verb,
- `nsubjpass` (passive subject) creates an EXPERIENCES edge from verb to subject,
- `dobj` or `pobj` (direct/prepositional objects) creates TARGETS edges from verb to object.

Prepositions for locations When a verb has a child with prep dependency and text “in”, “at”, or “on”, we create a LOCATED_IN edge from the verb to location. Similarly, for tokens with prep dependency and text “in” where both head and child exist in the node map, we create a LOCATED_IN edge from head to child.

Attribute and SAME_AS edges If a compound and its head are in the node map and we find the compound phrase exists, we create HAS_STATE edge from the compound to the head entity. And, the SAME_AS edge is used for co-references. This is most likely handled by the node deduplication step.

C Prompts

Since we do not have direct baselines to compare our method against, we use LLMs on the eval sets we prepare (Table 2).

The prompt we use for the LLMs is provided in Figure 5.

For the EVAL_SUFFICIENT set, we edit the prompt to allow the model to choose among three answers: “true”, “false” and “not enough information”.

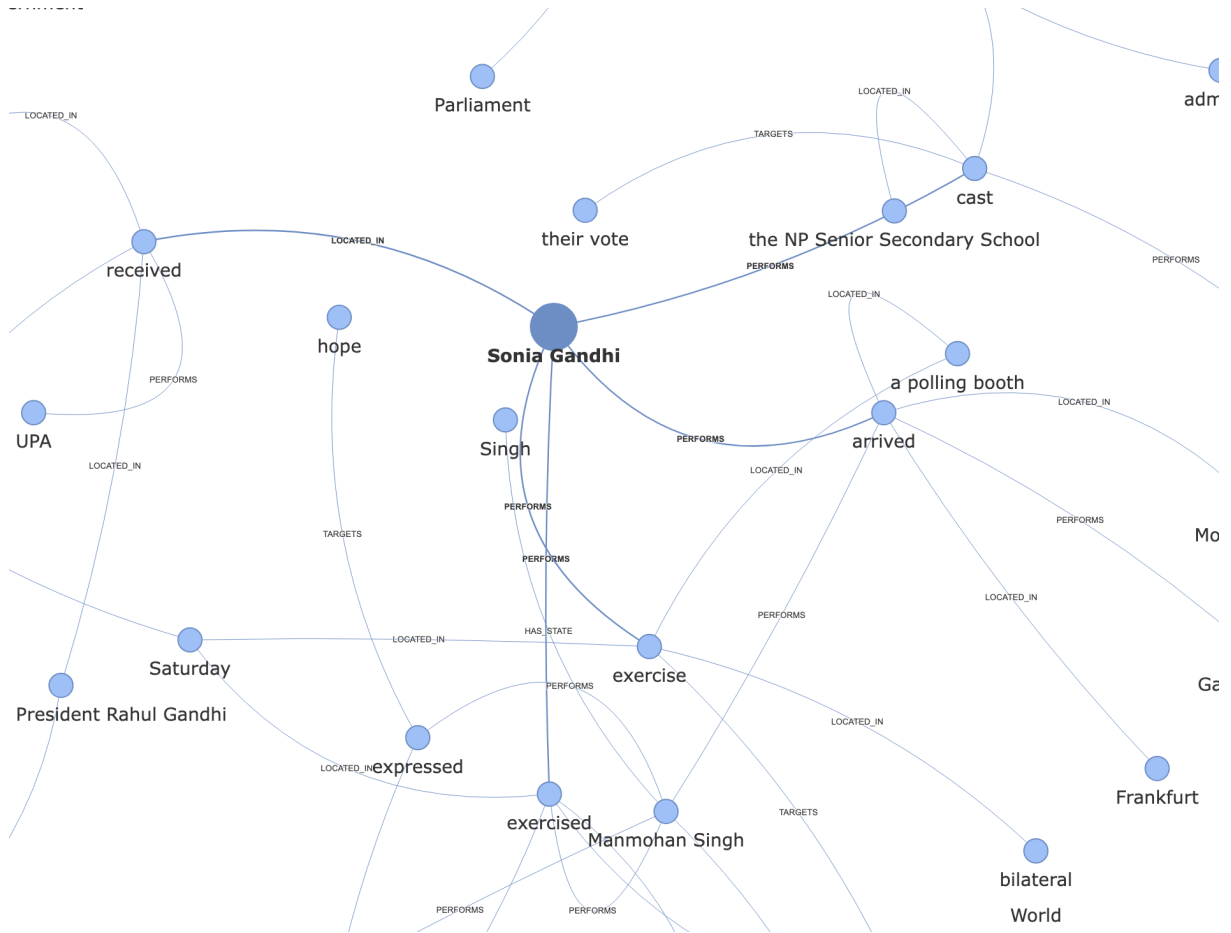


Figure 4: Evidence graph generated by EGMMG for the example in Figure 2

D Data Licenses

- **Factify:** CC BY 4.0
- **MMFakeBench:** CC BY-NC-SA 4.0
- **COSMOS:** Academic research only

All data provided under this work is licensed under CC BY-NC-SA 4.0 because the data we originally use are released under this license.

You are a fact-checking assistant tasked with evaluating the accuracy of a claim based on evidence provided. Your goal is to determine whether the claim is true or false based solely on the evidence. Do not consider external knowledge or information not included in the evidence.

Instructions:

1. Carefully read the evidence document, which consists of excerpts from multiple news articles.
2. Analyze the claim provided and compare it to the evidence.
3. Respond with "true" or "false" based on your analysis. Do not provide explanations or additional commentary.

EVIDENCE: {evidence}

CLAIM: {claim}

Your response should be exactly one of: TRUE, FALSE.

YOUR RESPONSE:

Figure 5: Prompt used to evaluate misinformation detection performance of LLMs (Sonnet, Haiku, GPT). For the EVAL_SUFFICIENT set, we allow one more option: "not enough information".

Conv Type	384-dim embeddings				768-dim embeddings		
	Run 1	Run 2	Run 3	Run 4	Run 1	Run 2	Run 3
GatConv	0.8059	0.8140	0.8086	0.8181	0.8288	0.8235	0.8221
Gatv2Conv	0.8099	0.8207	0.8194	0.8248	0.8221	0.8221	0.8194
TransformerConv	0.8180	0.8005	0.8315	0.8221	0.8221	0.8181	0.8248

Table 6: Performance comparison of graph convolution methods across multiple runs for 384-dim and 768-dim node label embeddings on the Factify 85:15 split. Best performance per embedding dimension is in **bold**.