Fundamental concepts of multimodal deep learning for disinformation detection (hands-on)

Stefanos-Iordanis Papadopoulos
(stefpapad@iti.gr)

AIDA AICET 2025 @ Thessaloniki, 16 July 2025

mever
Media Analysis, Verification and Retrieval Group

Multimodal (Mis)Information?

Claim: *"People in China tearing down a 5G tower in an attempt to stop the spread of COVID-19."*

Fact-check: *“... anti-surveillance protesters tearing down a "smart" lamppost in Hong Kong.”*

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

Claim: *"A news report on Russia's invasion of Ukraine showed a crisis actor sitting up while in a body bag."*

Fact-check: *“A ‘die-in’ climate protest in Austria”*

https://www.snopes.com/fact-check/climate-protest-ukraine-body-bag

Multimodal Misinformation

Misinformation: False or misleading information (regardless of intent to deceive).

Multimodal Misinformation: Misinformation conveyed through multiple modalities (i.e. images and texts) arising through a falsified relationship or mismatch between the modalities.

Visuals can amplify misinformation: Images and videos attract more attention, are shared more widely, and can make false claims more believable.

Li, Y., & Xie, Y. (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. Journal of marketing research, 57(1), 1-19.

Newman, E. J., Garry, M., Bernstein, D. M., Kantner, J., & Lindsay, D. S. (2012). Nonprobative photographs (or words) inflate truthiness. Psychonomic Bulletin & Review, 19(5), 969-974.

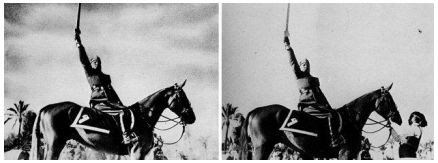
A (very brief) history of multimodal misinformation

Dissemination Technologies can affect speed and reach (e.g., printing press, TV, internet, social media).

Representational Technologies can create new forms of misinformation (photography, film, photoshop, DeepFakes, text-to-image generators).

Witches presenting wax dolls to the devil, featured in The History of Witches and Wizards (1720)

“Composite portraits" by Francis Galton, used to falsely claim visual proof of “criminal types” (1877)



Benito Mussolini on horseback was altered to remove the horse handler, to make it appear more heroic. (1942)

An AI-generated image used to show “dozens of missiles raining down simultaneously on Tel Aviv” (June 2025)

Multimodal Misinformation Detection

The aftermath of environmentalist Greta Thunberg's speech at the Glastonbury Music Festival in June 2022

Input: Image-Text pair under examination

Model

Modality Representation

Modality Fusion

Optimization

Truthful

?

Out-of-context

Output

Claim Source: https://www.snopes.com/fact-check/glastonbury-greta-thunberg

Evidence-based Detection

“An electric-powered BasiGo bus bursting into flames on Karen Road in Kenya.”

Motor1.com

Bus Turns Into Flamethrower When...

See exact matches

Majorca Daily Bulletin

Palma bus fleet being checked after second...

Manchester Evening...

"It was surreal - as we got off we could see th...

*“According to reports, the harrowing incident happened in **Perugia in Italy**. The bus in the video was an **Irisbus Iveco Cityclass** with an internal combustion engine and **compressed natural gas** (CNG) fuel tanks mounted on the roof.”*

Claim

Evidence Retrieval

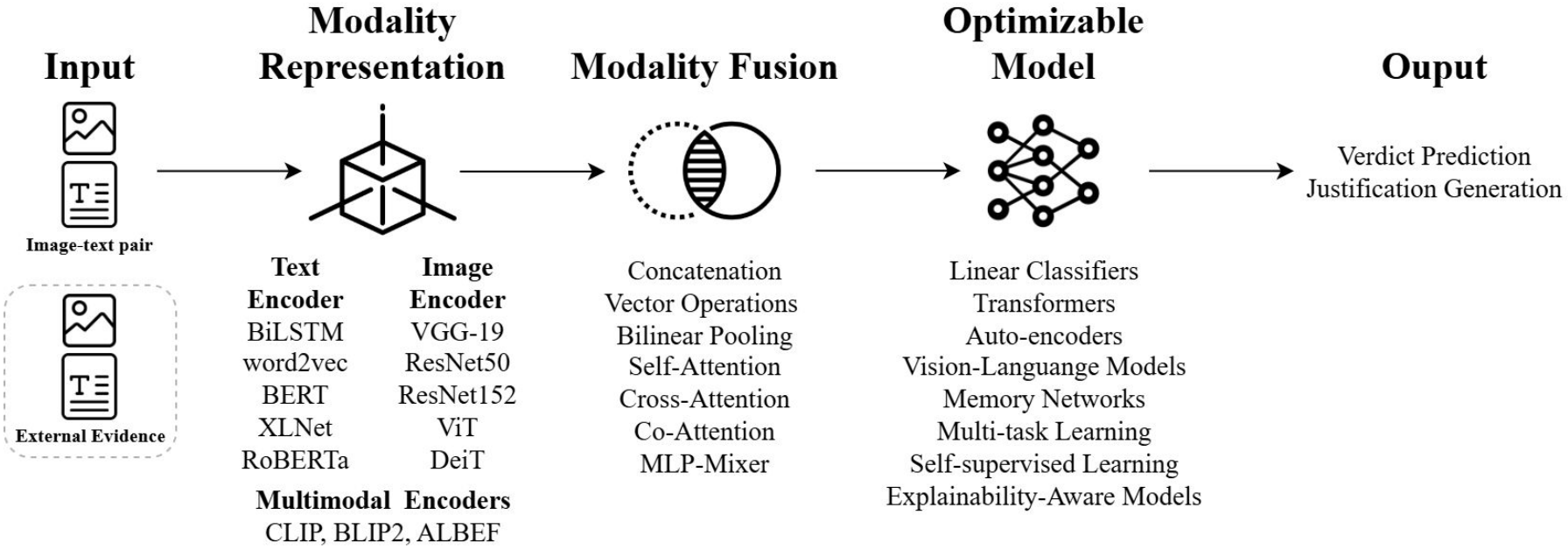
Using Inverse Image Search (Google Images)

Detector

Verdict

Claim Source: https://www.snopes.com/fact-check/electric-bus-fire-kenya

Related Work: Detection methods



Related Work: Datasets

Dataset	# Pairs	Construction	Data Source
Fauxtography [83]	1,233	Annotated	Snopes, Reuters
Weibo [27]	9,528	Weak annotation	Sina Weibo
Twitter (MediaEval 2015) [10]	15,624	Weak annotation	Twitter
Twitter (MediaEval 2016) [11]	17,857	Weak annotation	Twitter
Fakeddit [42]	680,798	Weak annotation	Reddit
MAIM [26]	239,968	Synthetic (OOC)	Flickr30k, MS COCO
COSMOS (Training set) [6]	200,000	Synthetic (OOC)	News Outlets
NewsCLIPpings (Merged) [36]	85,360	Synthetic (OOC)	News Outlets
Twitter-COMMs [9]	2,468,592	Synthetic (OOC)	Twitter
MEIR [58]	140,096	Synthetic (NEM)	Flickr
TamperedNews [41]	72,561	Synthetic (NEM)	News Outlets
CHASMA [47]	291,782	Synthetic (MC)	News Outlets, Reddit
FACTIFY [38]	50,000	Weak annotation	Twitter, News/Fact-check Articles
NewsCLIPpings+ [1]	85,360	Synthetic (OOC)	News Outlets, Search API
COSMOS (Test set) [6]	1,700	Annotated	News/Fact-check Articles
VERITE (Benchmark) [47]	1,000	Annotated	Fact-check Articles, Search API [49]

Annotated datasets: Often limited in scale and diversity.

Weak annotations: Can introduce noise, inaccuracies, and label inconsistencies.

Synthetic data: May lack the complexity and realism of real-world data.

The NewsCLIPpings dataset

(1) Query Caption: Angela Merkel speaks to the German parliament.

Pristine

Semantics / CLIP Text-Image

Semantics / CLIP Text-Text

Person / SBERT-WK Text-Text

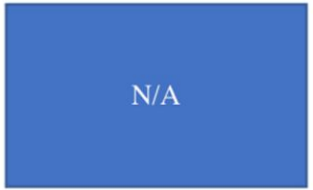
Scene / ResNet Place

(2) Query Caption: Fukushima Daiichi nuclear power plant after Japan s earthquake and tsunami in March.

Pristine

Semantics / CLIP Text-Image

Semantics / CLIP Text-Text



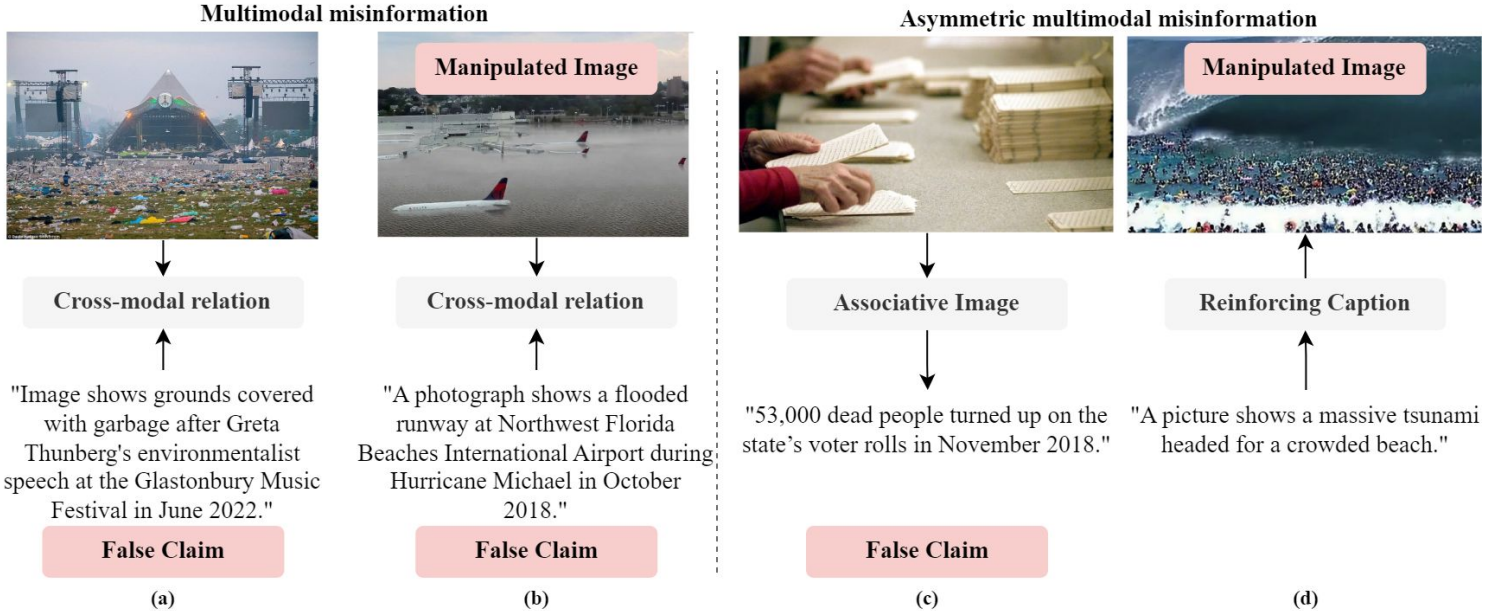
Person / SBERT-WK Text-Text

Scene / ResNet Place

Average human performance (crowdsourced): 65.6% (96.2% truthful, 35.0% out-of-context)

Luo, G., Darrell, T., & Rohrbach, A. (2021, November). NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6801-6817).

VERITE Benchmark (Verification of Image-Text pairs)



338 truthful pairs, 338 “Miscaptioned” pairs, and 324 “Out-of-Context” pairs from Snopes and Reuters

Excluding “Asymmetric” samples and leveraging “Modality Balancing”

Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2024). Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. International Journal of Multimedia Information Retrieval, 13(1), 4.

Pipeline (Today's Task)

- Training Dataset: NewsCLIPpings (Train Set)
- Data Exploration
- Feature Extraction
- Model / Classifier
- Training
- Evaluation: Accuracy on VERITE and NewsCLIPpings (Test Set)

Resources: Data and Google Colab

https://shorturl.at/VENo9

Project Structure

|— hands_on/ (if working with COLAB, “mount” directly to Google Drive)

| |— news_clippings

| | |— train_data.csv

| | |— …

| | |— news_clippings_CLIP_ViT-L14_image_features.npy

| | |— …

| |— VERITE/

| | |— images/

| | | |— false_0.jpg

| | | |— …

| | |— VERITE.csv

|— code.py

1. Data Exploration & Feature Extraction

1. Load Datasets
2. Data Exploration
3. Feature Extraction ***[Only for VERITE!*** *NewsCLIPpings features are provided]*
    a. Load CLIP ViT L/14 (pretrained="openai")
        i. https://github.com/mlfoundations/open_clip
    b. Iterate over the VERITE dataset
    c. Pre-process the images and texts (single items or a batch of items)
    d. Extract the embeddings with CLIP and store them in a list
    e. Also store the IDs of the items
    f. Save them as numpy files:
        i. VERITE_CLIP_ViT-L14_image_features.npy
        ii. VERITE_CLIP_ViT-L14_text_features.npy
        iii. VERITE_CLIP_ViT-L14_ids.npy
4. Load Features
5. Create a Pytorch DataLoader [return image features, text features, label]

2. Training: Modality Fusion

Given:
I = CLIP(img), d_{img} = 768
T = CLIP(txt), d_{txt} = 768
Network input ***d***

Create a function: def combine_features(a, b, fusion_method): that performs simple modality fusion operations:

- Concatenation (“concat”)
  - [***I***;***T***], $d = d_{img} + d_{txt}$ = 1,536
- Addition (“add”)
  - [***I+T***], ***d***=768
- Subtraction (“sub”)
  - [***I-T***], ***d***=768
- Multiplication (“mul”)
  - [***I*T***], ***d***=768
- Combined (“combine_all”)
  - [***I*** ; ***T*** ; ***I+T*** ; ***I-T*** ; ***I*T***], ***d***=3,840 (768 * 5)

Train: a simple neural-network (MLP) with PyTorch with different modality fusion operations
Evaluate: calculate the accuracy on NewsCLIPpings test set and VERITE (true vs out-of-context)

2. Training: Other models

- Self-attention for Modality Fusion:
  - We will use a nn.TransformerEncoder()
  - Input: [CLS, I , T , I+T , I-T , I*T] (d=768, 6 tokens)
    - Where CLS: “classification token” — a trainable nn.Parameter()
  - We use the transformed “CLS” token of the TransformerEncoder to classify the image-text pair
- Similarity-based baseline:
  - Calculate the cosine similarity between (image, text) under verification
  - Use a simple neural network to predict the label only based on the similarity score

3. Discussion

We explored:

- Different Modality Fusion operations
- Self-attention for Modality Fusion

Other approaches?

3. Discussion

We explored:

- Different Modality Fusion operations
- Self-attention for Modality Fusion

Other approaches?

- Experiment with different backbone encoders
- Fine-tune the backbone encoders for the task
- Ensemble learning
- Experiment with data augmentation
- Use of Large Vision-Language Models
- Explore different ways of generating training data
- Integrate external information / evidence
- …

References

VERITE: Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2024). VERITE: a Robust benchmark for multimodal misinformation detection accounting for unimodal bias. International Journal of Multimedia Information Retrieval, 13(1), 4.

NewsCLIPpings: Luo, G., Darrell, T., & Rohrbach, A. (2021, November). NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6801-6817).

Modality Fusion: Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2025). Red-dot: Multimodal fact-checking via relevant evidence detection. IEEE Transactions on Computational Social Systems.

Transformer Encoder: Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. (2023, June). Synthetic misinformers: Generating and combating multimodal misinformation. In Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation (pp. 36-44).

Similarity-based Baseline: Papadopoulos, S. I., Koutlis, C., Papadopoulos, S., & Petrantonakis, P. C. (2025, February). 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) (pp. 5041-5050). IEEE.