Multimedia Tools and Applications (2026) 85:588
https://doi.org/10.1007/s11042-026-21713-9

An In-Depth Survey on Multimodal Automatic Fact-Checking Datasets

Ian Marco Gallegos Carvajal[1] · **Beatrice Portelli**[1,2] · **Leonardo Zini**[3] · **Lorenzo Baraldi**[3] · **Giuseppe Serra**[1]

Received: 24 September 2025 / Revised: 13 March 2026 / Accepted: 25 May 2026

Abstract
The rapid spread of misinformation poses a significant challenge in the digital age, with false claims appearing across multiple modalities, particularly combinations of textual and visual content such as images and videos. While automatic fact-checking plays a crucial role in countering misinformation, traditional approaches predominantly rely on textual data, often neglecting the multimodal nature of modern misinformation. In this survey, we provide a comprehensive evaluation of multimodal datasets designed for automatic fact-checking that combine textual and visual information, systematically analyzing their sources, annotation methodologies, and key statistical properties, such as class distribution, topic diversity, and label availability. Additionally, we assess the usability of these datasets in real-world scenarios, discussing their limitations, biases, and potential risks, such as information leakage. Motivated by the practical difficulties we encountered when attempting to integrate existing datasets into our own multimodal fact-checking pipeline, our work also offers concrete guidance to help researchers choose the most suitable resources. By identifying gaps and challenges in existing datasets, our survey aims to support the development of more reliable and scalable multimodal fact-checking systems.

Keywords Multilingual corpus · Fact checking · Survey · Reproducibility · Data analysis

1 Introduction

The increasing complexity of online misinformation has made automated fact-checking (AFC) a crucial task, particularly in the context of multimodal content. While traditional AFC approaches heavily rely on textual data, leveraging additional modalities such as images and videos can significantly enhance the detection and verification process. However, while developing our own multimodal AFC system, we repeatedly encountered severe practical obstacles when attempting to integrate existing datasets into our AFC pipeline. Issues such as missing or inaccessible multimedia files, evidence texts that inadvertently reveal the label and make the task unrealistic, heterogeneous label schemas across datasets, and incomplete dataset releases often made seemingly suitable datasets difficult to use in

Extended author information available on the last page of the article

Published online: 01 July 2026

practice. These firsthand challenges motivated us to compile this survey, so that others do not face the same struggles.

In this paper, we focus exclusively on datasets, not on model architectures or verification methods, and provide a comprehensive analysis of datasets that include both textual and visual content (e.g., images or video frames), as these currently represent the most widely available resources for multimodal AFC. While other modalities such as audio signals or social network structures may also contribute to misinformation detection, datasets integrating these modalities remain relatively limited and heterogeneous, and are therefore outside the scope of this survey. Although multimodal AFC has attracted growing interest over the past decade, most survey efforts still concentrate on unimodal text-based AFC datasets [1], making it difficult for practitioners to locate and compare resources suitable for real-world pipelines. Moreover, sharing multimodal data introduces unique challenges and privacy concerns, often resulting in limited availability or inconsistent usability of datasets.

Even when datasets appear accessible, they may suffer from class imbalance, annotation bias, or unintended information leakage. Models trained on such imperfect resources can inadvertently exploit dataset-specific artifacts, such as URL counts or user metadata, instead of learning to verify content, as evidenced by the De-Factify 2 workshop results, where top systems achieved near-perfect accuracy by incorporating superficial textual features [2].

To fill the gap in the literature, in this work we present a detailed survey of the most recent datasets for multimodal AFC, focusing on their *data availability*, *origin*, and providing *data analysis* for selected resources. We unify information from various papers by introducing a standardized description of tasks and data sources to facilitate dataset comparisons and provide valuable insights for future researchers. In addition, we examine the limitations of the available datasets, discuss important considerations for their use and for the creation of future multimedia resources for AFC, and offer practical guidelines for the design of robust multimodal fact-checking (FC) datasets. All supplementary materials, including analysis scripts and dataset download files, are available at the following GitHub link: https://github.com/beatrice-portelli/multimodal-afc-survey.

2 Definitions

Based on the literature, we define a list of key terms which will be used in the subsequent analyses:

- **Claim** piece of information (textual and/or visual) to be verified.
- **Evidence** piece of external information (textual and/or visual) which can be used as contextual information for the claim. May consist of a set of resources.
- **Sample** basic input unit for a task, it may consist of a claim only, or a combination of claim and evidence.
- **Information leak** situation in which the evidence provided in the dataset contains/relies on human-generated FC articles (concept introduced in [3]). These articles have specific structures, writing styles, and usually state explicitly if the claim is true or false. This makes them unrealistic sources of evidence to train a completely automatic FC model (e.g., to use in early fake news detection), leading to models that rely on the presence of FC articles to function correctly.

- **Factuality assessment task (FAC)** given a claim (and possibly an evidence), determine whether the claim is true (or factual, reliable, etc.).
- **Stance detection task (STA)** given a claim and an evidence, determine if the evidence supports or refutes the claim. This does not address the truthfulness or reliability of the claim.
- **Crossmodal inconsistency task (INC)** given a claim with at least two parts / modalities, determine if the way the parts are presented distorts the original message.
- **Explanation generation task (EXP)** given a claim, an evidence, and a verdict, generate a text summarizing the reasons leading to the verdict.
- **Evidence retrieval task (RET)** given a claim and a set of documents, retrieve relevant evidences.

3 Related work

AFC has been extensively studied in the context of textual datasets, with surveys such as [1, 4] and [5] providing comprehensive overviews of AFC datasets, categorizing them based on task formulation, label types, and sources. However, these works often focus on unimodal text-based verification, leaving the area of multimodal datasets under-explored. Other surveys tend to address only a single task [6, 7] or are not focused specifically on datasets, including sections about models and methods too [8, 9]. Among recent surveys, only [10] investigates datasets across different tasks, but without focusing on important aspects such as the label origin, leading to broad yet superficial analyses.

Table 1 compares our survey with previous ones about AFC, highlighting the key aspects discussed in each publication. None of the previous surveys provides a detailed statistical analysis of AFC datasets ("dataset stats."), nor does it systematically investigate the labels origin, a critical factor when deploying models in real-world scenarios. Furthermore, a fundamental aspect of AFC is the presence of evidence, which serves as the foundation for fact verification. Despite its importance, no prior multimodal survey fully examines the data availability and quality of the evidences ("evidence analysis"), leaving a significant gap in

Table 1 Comparison with previous surveys. "≈" means the analysis is done partially

Name	pub. year most recent dataset	multimodal	claim origin	label origin	num. of samples	evidence analysis	datasets stats.	datasets avail.	dataset topic	multiple tasks
Kotonya and Toni [5]	2020	-	✓	-	✓	≈	-	-	-	-
Zeng et al. [4]	2021	-	-	-	✓	≈	-	-	✓	-
Guo et al. [1]	2021	-	✓	-	✓	≈	-	-	-	-
Hangloo and Arora [9]	2019	✓	✓	-	✓	-	-	-	✓	-
Tufchi et al. [8]	2022	✓	≈	-	-	-	-	≈	≈	-
Akhtar et al. [10]	2023	✓	✓	-	✓	-	-	-	-	✓
Ours	2025	✓	✓	✓	✓	✓	✓	✓	✓	✓

understanding how datasets support AFC decisions. Our work fills this gap by covering the latest datasets and systematically analyzing their structure, label distributions, and thematic coverage, offering a more in-depth perspective on multimodal AFC resources.

4 Method

We identified multimodal AFC datasets using an iterative citation-based approach, illustrated in Fig. 1. Starting from all the survey articles listed in Table 1, we extracted the datasets they cited and treated those references as *seed* articles. For each seed article, we first verified that it involved multimedia information (specifically visual content). We then examined its background, related work, and experimental sections to identify any additional cited papers introducing datasets or proposing AFC methods, adding these to the growing set of seed articles. This process was repeated iteratively until no new articles were added to the set.

To address the bias of this method toward older, well-known datasets contemporary to the surveys, we complemented our search with targeted queries on Google Scholar. We searched for papers published after 2023 with the following terms: "multimodal fact checking dataset", "multimodal fake news dataset", and "multimodal fact verification dataset". We selected all matching dataset papers, added them to the seed set, and iterated on their citation trees accordingly. This resulted in the inclusion of several newer datasets.

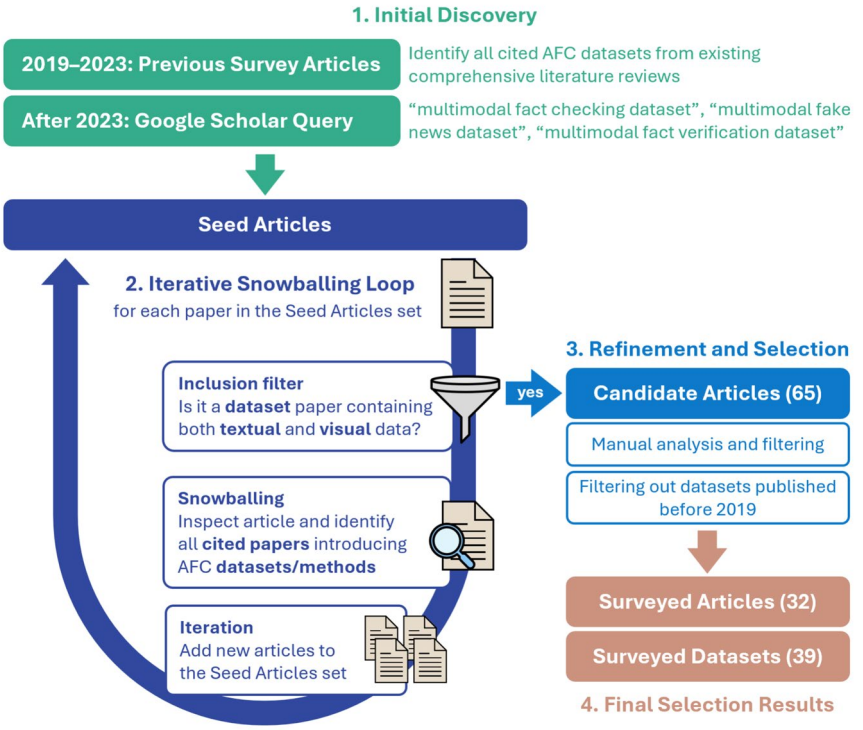

Fig. 1 Flowchart illustrating the dataset identification and selection process used in this survey

In total, we retrieved approximately 65 candidate articles. Each article was analyzed to identify the introduced dataset(s), assess whether they were actually multimodal, and document their characteristics. Datasets published before 2019 were removed because of their small sizes, outdated information and negligible impact on the overall landscape.

This final filtering led to the selection of **32 papers** introducing **39 multimodal datasets** regarding AFC, fake news detection, and related tasks.

For each dataset, we recorded a comprehensive set of characteristics that include: the sample structure (that is, whether samples consist solely of **claims** or also include **claim-evidence** pairs), the presence of **multimodal** information and the specific media they contain, the intended **task** along with the employed **labeling scheme**, and the overall **scale** measured by the number of claims, evidences, and samples. We also documented the **source** and **topic** of the data (to assess reliability and potential biases), whether the evidences may lead to **"information leak"**, the **languages** included in the dataset, and the dataset **availability**.

Finally, for a subset of fully available datasets, we performed an analysis examining various structural and formal characteristics. Specifically, we evaluated the distribution of **text length** by label, the distribution of **topics** by label, and the overall distribution of **samples** across labels. Additionally, we verified the validity of the provided data access links using a set of random samples to ensure that the datasets are currently accessible.

5 Results overview

5.1 Sample format

Out of the 39 analyzed datasets, 13 contain claim-evidence pairs (Table 2, top) while 26 consist of claim-only samples (Table 2, bottom). Fauxtography [11] is the oldest considered dataset and the only one published in 2019, while the most recent one is [12], the only one published in 2025. 2020 was the most prolific year, with 11 new datasets, while only 3 datasets were published in 2021. In 2022–2024, the community introduced 7-9 new datasets per year, reflecting an active interest in the creation and use of multimodal AFC datasets.

5.2 Size and creation methods

Dataset sizes vary widely (see Table 2, # Samples), ranging from 263 samples ExFaux [41] to 2.8 million samples CLIP-NESt [27]. However, the largest datasets are often automatically generated and may contain synthetic data, as it is extremely difficult and time-consuming to annotate thousands or millions of data points. All datasets in the range of 1 million samples (e.g., CLIP-NESt, CHASMA [25], and TamperedNews [37]) contain programmatically generated ones. For example, some of them consist of purely synthetic data, where fake news are generated by automatically altering real facts, while other datasets use embeddings and similarity metrics to automatically pair real images and texts, creating mismatched samples. The largest dataset which does not rely on this data augmentation is r/Fakeddit (1.0 million samples), which instead uses the reputation of the data sources (subreddits) to label a large number of scraped samples, relying on subreddit moderators. While the number of samples may seem really high for some datasets, they might consist of combinations of a smaller set

Table 2 List of the multimodal datasets with (top) and without (bottom) evidence, ordered by publication year

Dataset	Ref./ Year	# Samples	MM type Claim	MM type Evid.
FakeClaim	2024	755	v	
FineFake	2024	16,909	i	
WarClaim	2024	2,773	i v	
RD-E	2024	32,892	i *	
MR^2	2023	14,700	i	i
Factify2	2023	50,000	i	i
MOCHEG	2023	15,601		i
OoCMMFC	2022	85,360	i	i
STVD-FC	2022	1,200		v a
Factify	2022	50,000	i	i
MuMiN	2022	12,914	i *	i
PolitifactSnopes	2020	13,239	i	i
Fauxtography	2019	1,305	i	
ChileCP	2025	300	i *	
VERITE	2024	1,000	i	
CHASMA	2024	2,015,488	i	
CHASMA-D	2024	291,782	i	
MFD-Task1	2023	1,795	i	
MFD-Task2	2023	1,460	i	
CLIP-NESt	2023	2,838,082	i	
COSMOS	2023	200,000 + 1,700	i	
Twitter-COMMs	2022	2,468,592	i	
Evons	2022	92,969	i *	
CovID I	2022	2,369	i	
CovID II	2022	2,474	i	
COVID5G	2022	6,000	v	
NewsCLIPpings	2021	988,283	i	
VOA-KG2txt	2021	30,000	i	
Weibo C	2021	10,130	i	
NeuralNews	2020	128,000	i	
TamperedNews	2020	1,079,523	i	
News400	2020	6,360	i	
ReCOVery	2020	2,029 + 140,820	i	
r/Fakeddit	2020	1,063,106	i *	
FakeNewsNet	2020	23,196	i	
ExFaux	2020	263	i	
NewsBag	2020	215,000	i	
NewsBag++	2020	589,000	i	
NewsBag Test	2020	29,000	i	

Claim-only datasets, no evidence →

i image v video a audio

* only part of the samples is multimodal

of claims and evidences (or texts and images). For more information about unique number of claims, evidences, texts, and multimedia pieces of information, see Appendix A.

5.3 Multimodality

All the 26 claim-only datasets contain multimodal claims, while others present different combinations of multimodal inputs (Table 2, MM type). Regarding claim-evidence datasets, 6 of them provide multimodal information for both claims and evidences (MR ^{2}, Factify2, OoCMMFC, Factify, MuMiN, and PolitifactSnopes), two for the evidence only (MOCHEG and STVD-FC), and the remaining 5 have only multimodal claims. Images i, as expected, are the most common type of multimodal content, as they are the easiest data type to annotate, process, and share. Four datasets introduce videos v in their samples. FakeClaim [13], WarClaim [15], and COVID5G [32] all deal with YouTube videos, while STVD-FC incorporates both video v and audio a, as it comprises French TV program recordings (news and politics).

5.4 Tasks

Each dataset provides annotations for up to three different tasks.

Factuality assessment (FAC) The most common task is **FAC**, supported by 64% of the datasets (see Table 3, Tasks). In particular, 85% of the datasets containing evidence support the FAC task, with the only two exceptions being OoCMMFC (INC) and PolitifactSnopes (STA and RET). Although most datasets do not provide evidences to prove or disprove the truthfulness of the claims, 54% of them supply FAC labels. This leaves researchers with the choice of either relying solely on the textual and visual features of the claim or incorporating external sources of information to fact-check the claims.

Crossmodal inconsistency (INC) The second most frequent task is **INC**, which focuses on detecting inconsistencies between different modalities. Overall, 41% of the datasets and 50% of the claim-only datasets support this task. **INC** is closely related to the phenomenon of fauxtography, where multimedia content is misleading due to mismatches between textual and visual information.

Stance detection (STA) The third most frequent task is **STA**, accounting for 38% of claim-evidence datasets. It can be regarded as a simplified version of the FAC task, as it aims to classify the relationship (e.g., support or refute) between a claim and the evidence without producing a final verdict on the claim (e.g., real or fake). Interestingly, two claim-only datasets can also be categorized under STA, as they involve classifying the relationship between a claim's textual content and its accompanying multimedia elements. For example COVID5G contains social media posts referring to YouTube videos about COVID-5G conspiracy theories. One of the dataset labeling schemes specifies the relation between the post and the video (e.g., Supports, Related but does not support, Contradiction) making it an example of STA.

Explanation generation and evidence retrieval (EXP and RET) The remaining tasks appear less frequently. **EXP** is supported by MOCHEG and VOA-KG2txt [34], while **RET** is addressed by PolitifactSnopes. Although evidence retrieval and explanation generation are

Table 3 Task distribution and number of labels

Dataset	Tasks FAC	INC	STA	EXP	RET	# Labels
FakeClaim	✓		✓			2
FineFake	✓	✓				2 or 6
WarClaim	✓					1
RD-E	✓					6
MR^2	✓					3
Factify2	✓		✓			5
MOCHEG	✓		✓	✓		3
OoCMMFC		✓				2
STVD-FC	✓					3
Factify	✓		✓			5
MuMiN	✓					2
PolitifactSnopes			✓		✓	1 or 2
Fauxtography	✓	✓				2
ChileCP	✓					3
VERITE		✓				2 or 3
CHASMA		✓				2
CHASMA-D		✓				2
MFD-Task1		✓				3

Table 3 (continued)

Dataset	Tasks FAC	INC	STA	EXP	RET	# Labels
MFD-Task2	✓					4
CLIP-NESt		✓				3
COSMOS		✓	✓			2
Twitter-COMMs		✓				2
Evons	✓					2
CovID I	✓					2
CovID II	✓					2
COVID5G	✓		✓			3 or 5 or 6
NewsCLIPpings		✓				2
VOA-KG2txt				✓		2
Weibo C	✓					2
NeuralNews	✓	✓				2 or 4
TamperedNews		✓				2
News400		✓				2
ReCOVery	✓					2
r/Fakeddit	✓	✓				2 or 3 or 6
FakeNewsNet	✓					2
ExFaux		✓				2 or 5
NewsBag	✓					2
NewsBag++	✓					2
NewsBag Test	✓					2

commonly pursued in unimodal (textual) AFC, only a few of the analyzed multimodal datasets were specifically designed to support these tasks. This highlights a potential research direction for future dataset development.

5.5 Labeling scheme

Most of the considered tasks are annotated using categorical labels, ranging from 2 to 6 classes (Table 3, # Labels). Some of the datasets (e.g., Fakeddit [39] and FineFake [14]) provide different levels of granularity for the same task, allowing the researchers to choose between a coarse-grained annotation (true/fake) or a fine-grained one. Most of the datasets (67%) provide binary annotations for at least one of their tasks, and this annotation scheme is the most prevalent among claim-only datasets (77%). Among datasets that include evidences, 23% adopt a 3-way labeling scheme and, in these cases, the third class is typically neutral to capture uncertainty (e.g., the failure of an eventual evidence retrieval component or the early stages of a fake news spreading). The third class is usually labeled as "not enough information", "neutral", "unverified", or similar terms. Two of the datasets are

marked as having one label, as they only contain fake claims (WarClaim and PolitifactSnopes). Table 8 in Appendix B provides in-depth details of all labeling schemes.

5.6 Languages, data sources, leaking, and topics

87% of the datasets are monolingual (Table 4, Language), with the only exceptions being FakeClaim, WarClaim, MR^2 , MuMiN, and ReCOVery [38]. 85% of the datasets (33 out of 39) include the English language (EN). The other languages for which monolingual datasets are available are: French (FR, STVD-FC), Spanish (ES, ChileCP), Italian (IT, MFD-Task1 and MFD-Task2[26]), Chinese (ZH, Weibo C [35]), and German (DE, News400 [37]).

As regards the origin of the samples, in 67% of the cases claims are usually sourced from social media (Table 4, Claim origin), with the most frequent ones being (in order): Twitter (X), Reddit, Facebook, YouTube, TikTok, Instagram, Weibo, Telegram, and WhatsApp. Frequently, interesting social media posts are identified using FC websites, as they contain links to the unreliable posts which discuss fake and unreliable news. Evidences are sourced directly from FC websites in most of the cases (62%), but several datasets use google search to automatically retrieve web articles as evidences (Evidence origin), filtering out FC websites to avoid information leaks and/or unreliable websites to avoid creating unreliable evidences. The Label origin of most of the claim-evidence datasets comes directly from FC websites or human annotations, while most of the claim-only datasets rely on automatics methods (By construction). When evidences and labels are sourced from FC websites, this often leads to verified ❗ or possible ❓ information leaks, which researchers should be aware of when using the datasets.[1]

Most of the datasets (59%) were created without specific topic filtering, and therefore include various ones. The most frequent topic is politics (28%), followed by COVID19 (15%), and general health misinformation (13%) (details in Appendix D).

5.7 Data availability

The availability is summarized in Table 4, while links are provided in Appendix C, Table 9. Most studies on multimodal FC make their datasets available through repositories or websites that include download instructions for texts and images. Texts can often be provided directly in compliance with privacy regulations, while visual data may be shared in different ways, due to its larger size.

✅ Datasets such as r/Fakeddit, Evons [30], FineFake, MOCHEG, MR ^{2}, and PolitifactSnopes are among the most accessible, as both texts and images can be downloaded via compressed folders hosted on file-sharing services such as Dropbox and Google Drive, or directly from the authors' websites. Other datasets are still accessible, but only texts are provided directly while images are supplied as individual web links (e.g., Fauxtography, VERITE [25], Factify2, ReCOVery, CovID I, CovID II [31], TamperedNews, News400, and Weibo C). Other, like VOA-KG2txt, may include data from different sources, one which is completely available and one which provides image links only. Further limitations are evident in datasets such as Neural-

[1] Information leaks are marked as verified ❗ when datasets exhibit leaks by construction in virtually all samples, due to the way the data was collected. In contrast, datasets marked as having potential leaks ❓ fall into one of two categories: either the methodology is unclear, leaving room for potential leaks or leaks have been confirmed through dataset inspection but only affect a subset of the data.

Table 4 Origin of dataset samples, along with information leakage, dataset availability and language(s)

Dataset	Social media	News websites	Image search	Text manual editing	Text auto-tampering	Text auto-generation	Txt-img auto-pairing
FakeClaim	✓						
FineFake	✓	✓					
WarClaim	✓						
RD-E	✓						
MR^2	✓						
Factify2	✓	✓[1]	✓				
MOCHEG	✓						
OoCMMFC		✓					
STVD-FC	✓						
Factify	✓		✓				
MuMiN	✓						
PolitifactSnop	✓						
Fauxtography	✓						
ChileCP		✓					
VERITE	✓			✓			
CHASMA	✓	✓					✓
CHASMA-D	✓	✓					✓
MFD-Task1	✓	✓					
MFD-Task2	✓	✓					
CLIP-NESt		✓			✓		✓

Table 4 (continued)

Dataset	Social media	News websites	Image search	Text manual editing	Text auto-tampering	Text auto-generation	Txt-img auto-pairing
COSMOS	✓	✓					
Twitter-COMMs	✓						
Evons		✓[2]					
CovID I	✓	✓					
CovID II	✓						
COVID5G	✓						
NewsCLIPpings		✓					✓
VOA-KG2txt		✓			✓		
Weibo C	✓	✓					
NeuralNews		✓				✓	✓
TamperedNews		✓			✓		✓
News400		✓			✓		✓
ReCOVery		✓[2]					
r/Fakeddit	✓						
FakeNewsNet	✓	✓					
ExFaux	✓						
NewsBag		✓[1]					
NewsBag++		✓[1]				✓	✓
NewsBag Test		✓[1]					

Table 4 (continued)

	Evid. origin FC websites	Articles	Social media	Other
FakeClaim	✓			
FineFake	✓			
WarClaim	✓			
RD-E	✓			✓
MR[2]		✓[3]		
Factify2	✓	✓[4]	✓[4]	
MOCHEG		✓[4]		
OoCMMFC		✓[4]		
STVD-FC				✓
Factify	✓	✓[4]	✓[4]	
MuMiN	✓			
PolitifactSnopes	✓			
Fauxtography		✓[5]		
ChileCP				
VERITE				
CHASMA				
CHASMA-D				
MFD-Task1				
MFD-Task2				
CLIP-NESt				

Table 4 (continued)

	Evid. origin FC websites	Articles	Social media	Other
COSMOS				
Twitter-COMMs				
Evons				
CovID I				
CovID II				
COVID5G				
NewsCLIPpings				
VOA-KG2txt				
Weibo C				
NeuralNews				
TamperedNews				
News400				
ReCOVery				
r/Fakeddit				
FakeNewsNet				
ExFaux				
NewsBag				
NewsBag++				
NewsBag Test				

Table 4 (continued)

	Label origin							
	FC web-sites	Human annota-tion	Crowd-sourcing	Website reputation	By con-struction	Information leak?	Data availal-ibity	Languages
FakeClaim	✓					!	A ≈	30
FineFake	✓	✓					✓	EN
WarClaim	✓					!	A ≈	40
RD-E	✓					?	R	EN
MR^2	✓	✓				?	✓	EN ZH
Factify2	✓				✓	!	✓ R	EN
MOCHEG	✓						✓	EN
OoCMMFC					✓	?	R	EN
STVD-FC	✓						R	FR
Factify	✓				✓	!	✓ R	EN
MuMiN	✓					!	A	41
PolitifactSnopes	✓	✓				!	✓	EN
Fauxtography	✓					?	✓	EN
ChileCP						!	✓	ES
VERITE					✓		✓	EN
CHASMA					✓		≈	EN
CHASMA-D					✓		≈	EN
MFD-Task1			✓				A	IT
MFD-Task2			✓				A	IT
CLIP-NESt					✓		≈	EN

Table 4 (continued)

	Label origin							
	FC web-sites	Human annota-tion	Crowd-sourcing	Website reputation	By con-struction	Information leak?	Data availal-ibity	Languages
COSMOS		✓					R	EN
Twitter-COMMs					✓		A	EN
Evons				✓			✓	EN
CovID I		✓			✓		✓	EN
CovID II		✓			✓		✓	EN
COVID5G		✓					✗	EN
NewsCLIPpings					✓		✓	EN
VOA-KG2txt					✓		✓	EN
Weibo C	✓						✓	ZH
NeuralNews					✓		✓ ≈	EN
TamperedNews					✓		A	EN
News400					✓		A	DE
ReCOVery				✓			✓ A	40
r/Fakeddit				✓			✓	EN
FakeNewsNet	✓						A	EN
ExFaux		✓					✗	EN
NewsBag				✓			✗	EN
NewsBag++				✓	✓		✗	EN
NewsBag Test				✓			✗	EN

[1] Includes articles from satire websites

[2] Includes articles from unreliable/fake news websites

[3] Excluding unreliable sources

[4] Reliable sources only

[5] Excluding FC websites

News [36], which offers only the image caption and a link to the original article, leaving to the user the task to retrieve the visual data independently by scraping the webpage.

R For some datasets, information is partially available, but the remaining part must be requested to the authors (e.g., STVD-FC and RD-E [16]). While this makes the datasets less readily available, it also ensures its integrity and the presence of all multimedia pieces of information.

≈ Other datasets, in particular the ones comprising of synthetic samples, may only release some seed data and require the researcher to run code to generate the full dataset (e.g., CHASMA and CLIP-NESt). This adds a layer of complexity and a possible point of failure, as the code needs to be released and fully reproducible as well to obtain the same dataset.

A Several datasets rely on external APIs to access both visual and textual data. For example, FakeClaim and WarClaim provide video IDs for YouTube videos, while MFD-Task1, MFD-Task2, FakeNewsNet [40], MuMiN, and Twitter-COMMs [29] require the use

of the Twitter (X) API, a factor that significantly limits their usage due to the high associated costs, even for academic research.

✗ Finally, some datasets are currently not accessible through public repositories, direct download, and not explicitly made available to the research community (e.g., ExFaux and COVID5G). In other cases (e.g., NewsBag [42]), although the datasets were supposed to be available after publication, we were not able to recover them even after contacting the authors.

6 Dataset analysis

In this section we provide an analysis of some of the publicly available datasets identified in our survey, focusing on the ones tackling the binary FAC task: FakeClaim, FineFake, WarClaim, MOCHEG, Fauxtography, Evons, CovID I, CovID II, ReCOVery, and r/Fakeddit. Similar considerations can be drawn for the publicly available multiclass datasets, and can be found in Appendix E.

6.1 Text format and label distribution

As previously mentioned, some surface-level characteristics, such as differences in the average text length across labels can unintentionally leak information about the target labels. An example of these issues is Factify2, a 5-way classification dataset part of the De-Factify 2 workshop, for which participants achieved up to 100% accuracy in predicting the Refute class [43]. Refute claims and evidences are shorter on average compared to other classes, while also exhibiting a higher variability (Table 5). In contrast, Support claims and evidences are typically longer, while the Insufficient category lies in-between but remains distinguishable (see Appendix F for data distributions). The Refute category is the only one in that contains almost no URLs in their claims, while Support has a much higher than average number of URLs and user mentions (@). In addition, as noted by Chrysidis et al. [3], Factify2 suffers from information leak, as evidences for the Refute class are FC articles explicitly stating that the claim is false.

Table 6 reports a similar analysis for the other binary datasets (for MOCHEG we exclude the "not enough information" label, for FakeClaim and WarClaim we only consider claims from YouTube videos). We observe a marked difference in the average text length between classes, which could lead to imbalances in the representation of textual content. For instance, in FineFake, the average length of the negative and positive classes are 487.84 and 2,446.14 characters, making it a strong predictor. Similar discrepancies are observed in CovID II, CovID I, and Fauxtography, where the average text length varies considerably between classes. Another issue is the presence of very short texts (min < 20) in datasets such as

Table 5 Factify2 text length and distributions (in characters) for each class

	Claim Length			Evidence Length		
Label	Avg	Min	Max	Avg	Min	Max
Support MM	193.48	43	340	12,347.65	8	126,774
Support Text	186.75	22	333	10,462.37	17	273,682
Insuff. MM	191.38	38	336	5,943.25	8	271,151
Insuff. Text	191.06	27	336	5,281.44	14	1,641,275
Refute	109.45	5	712	2,348.85	109	27,753

Table 6 Statistics of various FC datasets, showing the distribution of samples belonging to the positive and negative classes, the average text length (in characters) of the samples for each class and the range of text lengths (minimum and maximum). Data which present significant issues or discrepancies between the two classes are highlighted

Dataset	Distribution		Avg Length		Min Length		Max Length	
	Neg[1]	Pos[2]	Neg[1]	Pos[2]	Neg[1]	Pos[2]	Neg[1]	Pos[2]
FakeClaim	39.87%	60.13%	75.52	76.78	9	12	149	140
WarClaim	100.00%	0.00%	71.26	–	9	–	366	–
FineFake	44.37%	55.63%	487.84	2,446.14	3	11	100,049	100,096
MOCHEG	53.23%	46.77%	3,677.37	4,046.85	125	175	34,179	23,949
r/Fakeddit	60.61%	39.39%	34.37	53.53	1	1	2,785	297
Fauxtography	54.99%	45.01%	89.06	159.76	31	38	211	962
CovID I	55.30%	44.07%	2,612.91	1,807.62	26	25	18,682	31,863
CovID II	52.67%	47.33%	2,625.92	102.52	26	26	18,682	1,154
Evons	46.64%	53.36%	186.28	125.94	1	4	648	505
ReCOVery	32.77%	67.23%	5,014.27	5,210.35	131	151	91,828	79,503

[1]Negative labels: fake (FakeClaim, WarClaim, FineFake, r/Fakeddit, Evons), false (Fauxtography, CovID I, CovID II), refuted (MOCHEG), unreliable (ReCOVery)

[2] Positive labels: real (FakeClaim, FineFake), true (r/Fakeddit, Fauxtography, CovID I, CovID II, Evons), supported (MOCHEG), reliable (ReCOVery)

FineFake, r/Fakeddit, and Evons. This may result from data collection errors or the inclusion of non-informative texts. Additionally, an excessive range between the minimum and maximum text lengths, observed in datasets like FineFake and CovID I, suggests the presence of noisy data. Some datasets, including FakeClaim, ReCOVery, and r/Fakeddit, have highly unbalanced class distributions, which may impact model performance. In addition, r/Fakeddit data includes bot-generated comments, while some texts automatically scraped from articles still contained HTML tags that needed further cleaning.

6.2 Topics and label distribution

While most of the datasets cover various topics, only few provide them as an explicit feature. Among the ones we analyzed, only FineFake does, while the r/Fakeddit subreddit feature could be used to vaguely infer the theme of the claim (e.g., US news, or digitally altered images). We performed an analysis of the topic distribution among all samples and labels for the two datasets, observing a strong topic unbalance in both of them, favoring Politics, Society, and Entertainment for FineFake, and psbattle_artwork for r/Fakeddit (subreddit on manipulated images and photo editing software). In addition, most of the topics showed unique label distributions (e.g., psbattle_artwork samples are almost exclusively false in r/Fakeddit, while mildlyinteresting samples are true), creating a possible bias between topics and labels (see Appendix F for additional details).

6.3 Missing data

Datasets that rely on links and APIs often suffer from missing data, as seen in WarClaim and FakeClaim, where many entries were lost due to deleted or privatized videos. Notably,

the downloaded versions contained more instances than reported: FakeClaim had 782 false instances (vs. 389 in the paper), and WarClaim had 693 total instances (vs. 501), including 415 real and 278 false samples. Additionally, the two datasets share 271 false instances.

7 Cross-modal semantic alignment

To quantify semantic alignment between textual and visual modalities, we compute the CLIPScore [44] metric for each image-text pair at the sample level. CLIPScore is based on the cosine similarity between image and text embeddings produced by the CLIP model, providing an estimate of their semantic alignment in a shared embedding space. We compute CLIPScore using the pretrained CLIP ViT-B/32 model and the textual claim associated with each sample. We analyze the distribution of scores across classes to identify modality-driven biases that could affect multimodal learning.

Figure 2 shows the box plots for the CLIPScore calculated on six of the analyzed datasets: FineFake, CovID I, CovID II, r/Fakeddit, ReCOVery, and Factify2. The remaining datasets were excluded from the analysis because images could not be reliably retrieved for a large portion of samples (FakeClaim, WarClaim, and Evons) or because the dataset does not provide unique text-image pairs due to the task setup (MOCHEG and Fauxtography). The distributions reveal notable differences across datasets. CovID I and CovID II show a strong class mismatch, with the Fake class achieving higher CLIPScores compared to True

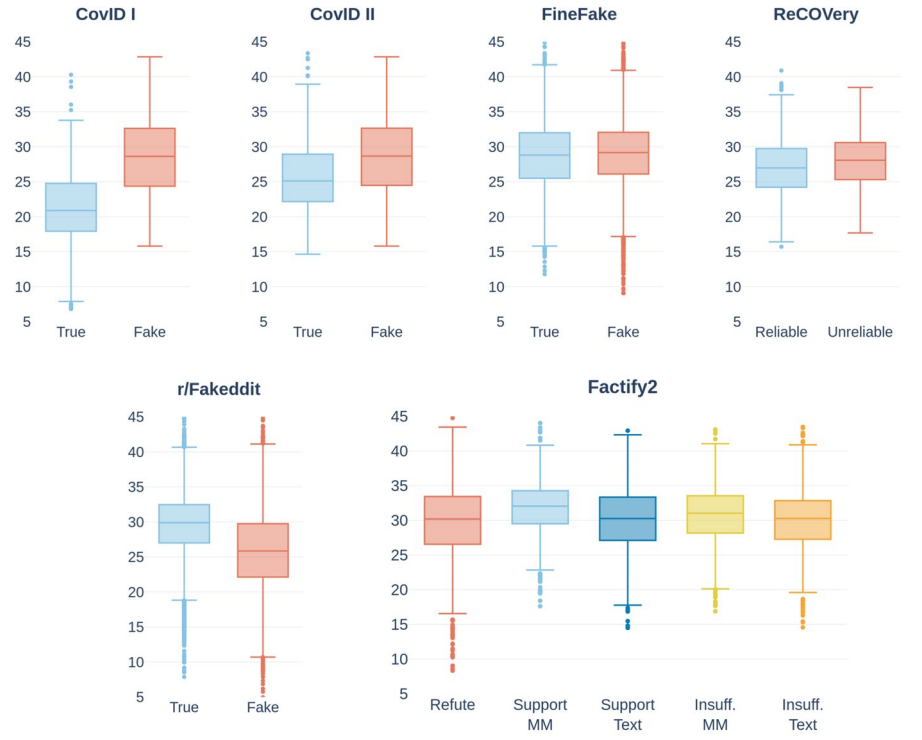

Fig. 2 CLIPScore distributions grouped by class for the analyzed datasets.

samples. This suggests that, contrary to expectations, Fake samples exhibit higher semantic alignment between text and image on average, compared with True samples.

The binary version of the r/Fakeddit dataset also shows clear differences in the semantic alignment of the two classes, but with the opposite trend: the True samples, on average, receive CLIPScore values approximately 5 points higher than Fake samples. Similarly, the analysis of Factify2 reveals that samples belonging to the Support MM class exhibit the highest median CLIPScore, indicating stronger semantic alignment between textual claims and visual content. In contrast, the Support Text class shows lower alignment, which is consistent with the dataset design where the verification decision relies primarily on textual evidence. The Refute class exhibits the lowest median CLIPScore and higher variance, suggesting that refutation examples may involve more diverse or loosely related visual content. The Insufficient classes (MM and Text) display intermediate alignment levels, reflecting the weaker semantic grounding expected when evidence does not clearly support or contradict the claim.

Finally, FineFake and ReCOVery exhibit comparable CLIPScore distributions across classes

Beyond these quantitative differences, the results highlight recurring alignment challenges. Low similarity scores often reflect weak content-claim grounding, where images are only loosely related to the textual statement. Conversely, class-dependent alignment gaps (e.g., in CovID I, CovID II, and r/Fakeddit) suggest that semantic coherence itself may become predictive of the label, potentially enabling multimodal shortcuts during model training, where models rely on alignment patterns rather than genuine claim verification. These findings indicate that multimodal dataset quality depends not only on similarity statistics, but also on structural alignment properties and class composition.

8 Discussion and future directions

Over the last five years, multimodal AFC has developed rapidly, and the number of available datasets has grown accordingly. Most of the resources we reviewed were designed for FAC, STA, and INC, while EXP and RET tasks are still scarcely represented, making them an interesting topic for future research. At the same time, the picture that emerges from this survey is far from uniform. Differences in dataset size, annotation strategy, accessibility, and evidential structure are substantial, and these differences often have direct consequences for the type of models that can be trained and for the conclusions that can be drawn from benchmark results.

Dataset size is highly dependent on the collection and annotation methods: manual labeling, although more accurate, entails high costs, leading many researchers to employ automatic methods or rely the source's reputation as a proxy. Larger data collections are obtained through automatic procedures or weak labeling strategies, which are easier to scale but more exposed to noise, artifacts, and unrealistic sample construction.

A similar limitation appears in the **linguistic** and geographic concentration of the available resources. Most datasets consist of monolingual English data, derived from FC websites and US social media. As a result, current benchmarks offer limited support for the development of AFC methods which generalize to linguistically and culturally diverse contexts.

Another critical aspect concerns **data availability**: suboptimal sharing practices can render a dataset unusable, or only accessible upon completing specific procedures, creating bottle-necks and compromising the reproducibility of studies. Relying solely on image links can create issues in the long term as linked data can be deleted or relocated thus invalidating the link. This problem is alleviated in some cases (e.g., Factify2) where the original images may be obtained by contacting the authors of the dataset.

A recurrent problem in the resources we examined is that the final label may be recoverable from signals that are only marginally related to the actual verification process. The clearest example is the presence of **leaked evidence** which makes it trivial for any NLP model to assign the correct label. In addition, during the dataset analysis, we highlighted other class-level and topic-level **biases**, which models may use as shortcuts, preventing them from generalizing in the wild. These issues underscore the necessity for careful data collection and preprocessing protocols, such as the removal of overly short instances, balancing class distributions, and checking for significant discrepancies in text length.

For these reasons, the construction of multimodal fact-checking datasets requires careful methodological design. The process should begin with the selection of realistic claims drawn from genuine misinformation contexts while filtering out trivial, duplicated, or poorly contextualized instances. Evidence collection must balance informativeness with neutrality, ensuring that the supporting material enables verification without explicitly revealing the final verdict. Similarly, label assignment should rely on transparent and defensible procedures, ideally grounded in professional fact-checking sources, expert annotation, or clearly documented verification protocols. Beyond individual annotations, datasets should also be examined as structured objects. Researchers should verify whether class distributions, topic concentration, textual characteristics, or multimodal alignment introduce unintended biases or shortcuts that models may exploit. Finally, the way a dataset is released plays an important role in its long-term usability: resources that depend on unstable infrastructure, private requests, or external APIs may become progressively unusable and hinder reproducibility.

8.1 Recommendations for future dataset design and use

In summary, it is not sufficient to collect a large number of examples. The dataset must be designed so that it genuinely evaluates verification ability rather than the detection of spurious shortcuts. Based on the issues identified in this survey, the construction of future datasets should follow a set of methodological steps.

Practical checklist for dataset design

- Carefully select claims, collecting them from realistic misinformation contexts and filtering out trivial, duplicated, poorly contextualized, or noisy examples that increase dataset size without improving benchmark quality.
- Design the evidence collection process by retrieving sources that allow a plausible verification process while avoiding materials that directly reveal the final verdict, such as fact-checking articles containing explicit labels.
- Define a transparent labeling protocol, assigning labels through clearly documented procedures and preferably relying on professional fact-checking sources, expert annotation, or verifiable chains of evidence.

- Inspect the dataset for structural artifacts by analyzing class distributions, topic concentration, and textual characteristics in order to detect biases or accidental regularities that may allow models to solve the task through shortcuts.
- Verify multimodal consistency by ensuring that images and textual content are genuinely aligned and that visual material is not missing, duplicated, or only weakly related to the corresponding claim, unless this misalignment is intentional and documented.
- Control for label leakage by checking that labels cannot be inferred from superficial cues such as text length, topic-specific patterns, stylistic markers, or artifacts introduced during dataset construction.
- Ensure reproducible dataset release by providing direct access to the dataset and stable storage whenever legally possible, or by documenting the collection and reconstruction procedure in sufficient detail to enable reuse.

Following such practices can help future datasets better reflect the evidential uncertainty and multimodal complexity of real-world fact-checking, thereby improving their usefulness as reliable evaluation benchmarks.

In addition to guidelines for dataset construction, it is equally important for researchers to critically assess datasets before using them for model training or benchmarking. Based on the issues identified in this survey, we propose a short checklist that may help practitioners detect potential risks when selecting multimodal fact-checking datasets.

Practical checklist for dataset users

- Verify data availability and completeness, ensuring that all multimedia content (images, videos, metadata) can be reliably retrieved and does not depend on unstable external links.
- Check for possible information leakage, such as evidence texts containing explicit verdict markers, URLs pointing to fact-checking pages, or metadata features strongly correlated with the label.
- Inspect class balance and topic distribution to identify whether certain topics or textual characteristics are disproportionately associated with a specific label.
- Evaluate multimodal alignment, verifying that images are semantically related to the associated claim rather than loosely connected or reused contextual material.
- Examine dataset documentation and annotation procedures to ensure that labels are derived from transparent and reliable verification processes.

9 Conclusions

This survey examined 39 multimodal FC datasets, analyzing their sources, annotation methods, statistical properties, and limitations. Despite the growing availability of resources, significant challenges persist. Our findings provide key insights and guidelines to support the development of more reliable and scalable misinformation detection systems.

10 Limitations

One limitation of this work is the high cost associated with accessing social network APIs, which restricts the scope to datasets that directly share the source of information rather than relying on API links. The use of API links presents additional challenges, as some valuable samples may be lost or inaccessible. Another limitation is the temporal span of the datasets considered, with the earliest being published in 2019. As methods and technologies evolve rapidly, reliance on older data may hinder the effectiveness of the findings, as advancements in both the data and the underlying technologies may not be adequately captured. While this survey focuses on the characteristics and limitations of existing datasets, evaluating how these properties influence model performance and generalization would require a dedicated benchmarking study with controlled experimental settings. Such analyses are highly model-dependent and remain an important direction for future research.

A Dataset size

Table 7 provides a more in-depth overview of the scale of all the analyzed datasets, including the number of samples, claims and evidences. One of the most evident aspects is the wide variability in dataset sizes. Some datasets contain millions of samples, such as CLIP-NESt, NewsCLIPpings, and Twitter-COMMs, while others, like ExFaux and ChileCP, are significantly smaller, with only a few hundred samples. This variation in dataset scale can have important implications for model training, as larger datasets may offer better generalization, whereas smaller datasets might be more focused on specific domains.

Another key observation is the availability of supporting evidence. Some datasets, such as Factify, MOCHEG, and FineFake, provide one piece of evidence per claim, ensuring a direct claim-evidence alignment. Others, like MR^2 , offer multiple evidences per claim (e.g., a 5:1 ratio), whereas Fauxtography has a highly variable structure, with up to 50 pieces of evidence per claim.

Finally, a general trend can be observed regarding dataset characteristics. Larger datasets, such as CLIP-NESt and Twitter-COMMs, tend to be more general-purpose, containing a large number of claims without associated evidence. In contrast, smaller datasets often focus on specific domains and include multimodal elements, as seen in ChileCP and Fauxtography. Additionally, certain datasets derive their data from specific sources, such as STVD-FC, which consists of hours of television programs, and recovery, which aggregates both tweets and news articles. These variations in dataset composition provide insights into the diversity of available resources and the different FC scenarios they can support.

Some datasets may appear large when considering the number of samples, such as, VERITE (1k samples), CHASMA-D (over 200k samples), and CHASMA (over 2 million samples). However, samples comprise of combinations of a smaller set of texts and images. For examples, the 1k samples of VERITE are a combination of 388 unique texts and 324 images, while CHASMA contains 145k unique texts and 1.2 million images for 2 million samples. The use of repeated texts and images in different samples may be useful to force models to combine information coming from different modalities, but should be considered carefully when choosing datasets.

Table 7 Detailed information about the number of samples, claims and evidences for each dataset

Dataset	# Samples	# Claim	# Evidence
FakeClaim	755	755	1,370 (1 per claim)
FineFake	16,909	16,909	16,909
WarClaim	2,773	2,773 → 2,251 (AVAIL)	1,362 (1 per claim)
RD-E	32,892	32,892 → 19,162 (M)	32,892
MR^2	14,700	14,700	73,500 (5 per claim)
Factify2	50,000	50,000	50,000
MOCHEG	15,601	15,601	91,822 (T) + 122,246 (I)
OoCMMFC	85,360	85,360	85,360
STVD-FC	1,200	1,200	6,730 programs (6,540 hours)
Factify	50,000	50,000	50,000
MuMiN	12,914	12,914 + 6,573 (I)	10,920
PolitifactSnopes	13,239	13,091	2,170
Fauxtography	1,305	1,305	59,037 (max 50 per claim)
ChileCP	300	300 → 168 (M)	
VERITE	1,000	388 (T) + 324 (I)	
CHASMA	2,015,488	145,891 (T) + 1,259,732 (I)	
CHASMA-D	291,782	145,891 (T) + 145,891 (I)	
MFD-Task1	1,795	1,795	
MFD-Task2	1,460	1,460	
CLIP-NESt	2,838,082	2,838,082	
COSMOS – train	200,000	450,000 (T) + 200,000 (I)	
COSMOS – test	1,700	3,400 (T) + 1,700 (I)	
Twitter-COMMs	2,468,592	884,331	
Evons	92,969	92,969 → 92,657 (M)	
CovID I	2,369	2,369	
CovID II	2,474	2,474	
COVID5G	6,000	6,000	
NewsCLIPpings	988,283	473,663	
VOA-KG2txt	30,000	30,000	
Weibo C	10,130	10,130	
NeuralNews	128,000	128,000	
TamperedNews	1,079,523	72,561 real + 1,006,962 tampered	
News400	6,360	400 real + 5,960 tampered	
ReCOVery	2,029 articles + 140,820 tweets	2,029 → 2,017 (M)	
r/Fakeddit	1,063,106	1,063,106 → 682,996 (M)	
FakeNewsNet	23,196	23,196	
ExFaux	263	263	
NewsBag	215,000	215,000	
NewsBag++	589,000	589,000	
NewsBag Test	29,000	29,000	

(AVAIL) number of samples in downloaded dataset (if different from the one declared in the paper)

(M) number of multimodal samples (if not all)

(T) number of text-only pieces of information

(I) number of image-only pieces of information

B Dataset labels

Table 8 provides a complete list of all annotations schemes and labels used in the analyzed datasets.

Table 8 List of the labeling schemes and label names of all analyzed datasets. Abbreviations used in the labels: OOC – out of context, MC – miscaptioned, NEI – not enough information, SUP – support, REF – refute, INS – insufficient

Dataset	Labels		
FakeClaim	Real, Fake		
FineFake	real, text-image inconsistency, content-knowledge inconsistency, text-based fake, image-based fake, others	Real, Fake	
WarClaim	False		
RD-E	true, mostly true, half true, mostly false, false, pants on fire		
MR^2	Rumor, Non-Rumor, Unverified		
Factify2	SUP_text, SUP_multimodal, INS_text, INS_multimodal, REF		
MOCHEG	SUP, REF, NEI		
OoCMMFC	falsified, pristine		
STVD-FC	False, Imprecise, True		
Factify	SUP_text, SUP_multimodal, INS_text, INS_multimodal, REF		
MuMiN	Misinformation, Factual		
PolitifactSnopes	Related, not related	False	
Fauxtography	True, False		
ChileCP	True, False, Non verified/Others		
VERITE	True, OOC, MC	True, Misinformation	
CHASMA	True, MC		
CHASMA-D	True, MC		
MFD-Task1	Misleading, Not Misleading, Unrelated		
MFD-Task2	Certainly Fake, Probably Fake, Probably Real, Certainly Real		
CLIP-NESt	True, OOC, NEI		
COSMOS	OOC, NOOC		
Twitter-COMMs	Pristine, Falsified		
Evons	Real, Fake		
CovID I	True, False		
CovID II	True, False		
COVID5G	explicit, implicit, neutral, ambivalent, others related, others unrelated	SUP, related but not SUP, contradiction, unrelated, SUP but OOC	misinformation, countering, other
NewsCLIPpings	Pristine, Falsified		
VOA-KG2txt	True, False		
Weibo C	Real, Fake		

Table 8 (continued)

Dataset	Labels		
NeuralNews	Real_Real, Real_Fake, Fake_Real, Fake_Fake	Real, Fake	
TamperedNews	positive, negative		
News400	positive, negative		
ReCOVery	Reliable, Unreliable		
r/Fakeddit	True, Satire/Parody, Misleading Content, Imposter Content, False Connection, Manipulated Content	Real, Fake, Inbetween	Real, Fake
FakeNewsNet	Real, Fake		
ExFaux	True, Fake_img, Fake_text, Fake_img_and_text, Fake_True_img_and_text	True, Fake	
NewsBag	Real, Fake		
NewsBag++	Real, Fake		
NewsBag Test	Real, Fake		

C Dataset availability

For ease of reading, in Table 9 we report the data availability of all datasets and an external URLs to access the data when available. For datasets which require authorization, we provide the URL of the page where such authorization can be requested.

Table 9 Data availability and external link (when available) for each dataset

Dataset	Link	Avail.
FakeClaim	https://github.com/Gautamshahi/FakeClaim	A ≈
FineFake	https://drive.google.com/file/d/16D9ix7ZOisa4VVBznBTBcv1N7TA-jodH	✓
WarClaim	https://github.com/Gautamshahi/WarClaim/	A ≈
RD-E	https://github.com/zhengyang5/RDE	R
MR^2	https://github.com/THU-BPM/MR2	✓
Factify2	https://aiisc.ai/defactify2/factify.html	✓ R
MOCHEG	https://github.com/VT-NLP/Mocheg	✓
OoCMMFC	https://s-abdelnabi.github.io/OoC-multi-modal-fc/	R
STVD-FC	http://mathieu.delalandre.free.fr/projects/stvd/	R
Factify	https://competitions.codalab.org/competitions/35153	✓ R
MuMiN	https://mumin-dataset.github.io/	A
PolitifactSnopes	https://github.com/nguyenvo09/EMNLP2020	✓
Fauxtography	https://gitlab.com/didizlatkova/fake-image-detection	✓
ChileCP	https://github.com/MolodyGs/Multimodal-News-Data-Collection	✓
VERITE	https://github.com/stevejpapad/image-text-verification	✓
CHASMA	https://github.com/stevejpapad/image-text-verification	≈

Table 9 (continued)

Dataset	Link	Avail.
CHASMA-D	https://github.com/stevejpapad/image-text-verification	≈
MFD-Task1	https://sites.google.com/unipi.it/multi-fake-detective	A
MFD-Task2	https://sites.google.com/unipi.it/multi-fake-detective	A
CLIP-NESt	https://github.com/stevejpapad/image-text-verification	≈
COSMOS	https://github.com/shivangi-aneja/COSMOS/tree/main	R
Twitter-COMMs	https://github.com/GiscardBiamby/Twitter-COMMs	A
Evons	https://github.com/krstovski/evons	✓
CovID I	https://drive.google.com/file/d/1bjMrvPIgwAXt_nvtmP0vFqEqEtYq_YmS	✓
CovID II	https://drive.google.com/file/d/1ivBi9T0GoY3vkQiabWEQg6CnPSvkpAh7	✓
COVID5G		✗
NewsCLIPpings	https://huggingface.co/g-luo/news-clippings/tree/main/data	✓
VOA-KG2txt	https://github.com/yrf1/InfoSurgeon	✓
Weibo C	https://github.com/lumen2018/dataset	✓
NeuralNews	https://cs-people.bu.edu/rxtan/projects/didan/	✓ ≈
TamperedNews	https://data.uni-hannover.de/dataset/tamperednews	A
News400	https://data.uni-hannover.de/dataset/news400	A
ReCOVery	https://github.com/apurvamulay/ReCOVery	✓ A
r/Fakeddit	https://github.com/entitize/Fakeddit	✓
FakeNewsNet	https://github.com/KaiDMML/FakeNewsNet	A
ExFaux		✗
NewsBag		✗
NewsBag++		✗
NewsBag Test		✗

✓ Most of the dataset is available and easily downloadable

R All or part of the dataset is available only on request

≈ Requires running code to obtain the full dataset (e.g., synthetic data generation)

A Relies on external APIs to access data

✗ Dataset not available

D Dataset topics

Table 10 provides more details about the language and topic(s) addressed by all datasets. English is by far the most widely used language across the datasets, which can be attributed to its global prevalence and its position as the primary language of the scientific community. However, the widespread use of English introduces a limitation, as the availability of datasets in other languages remains sparse and AFC heavily relies on language. This imbalance restricts the potential applicability of the research to non-English-speaking populations.

Table 10 Languages and topics for each dataset.

Dataset	Lang.	Topics
FakeClaim	30	2023 Israel-Hamas War
FineFake	EN	Various (incl. Politics, Health, Conflicts)
WarClaim	40	2023 Israel-Hamas War
RD-E	EN	Various (incl. Politics, Health, COVID19)
MR^2	EN ZH	Various (incl. Politics, Health)
Factify2	EN	Various (USA and Indian Politics)
MOCHEG	EN	Various
OoCMMFC	EN	Various
STVD-FC	FR	2022 French Presidential Election
Factify	EN	Various (USA and Indian Politics, Health)
MuMiN	41	Various
PolitifactSnopes	EN	Politics
Fauxtography	EN	Various
ChileCP	ES	Chile's constitutional process
VERITE	EN	Various
CHASMA	EN	Various
CHASMA-D	EN	Various
MFD-Task1	IT	2022 Ukrainian-Russian war
MFD-Task2	IT	2022 Ukrainian-Russian war
CLIP-NESt	EN	Various (incl. Politics, Environment, Law)
COSMOS	EN	Various (incl. Politics, Health, Environment)
Twitter-COMMs	EN	COVID19, Climate, Military Vehicles
Evons	EN	2016 USA Presidential Election
CovID I	EN	COVID19
CovID II	EN	COVID19
COVID5G	EN	COVID19 5G Conspiracy Theories
NewsCLIPpings	EN	Various
VOA-KG2txt	EN	Various
Weibo C	ZH	Various
NeuralNews	EN	Various
TamperedNews	EN	Various
News400	DE	Various (incl. Politics, Economy, Sports)
ReCOVery	40	COVID19
r/Fakeddit	EN	Various
FakeNewsNet	EN	Politics, Entertainment
ExFaux	EN	Various
NewsBag	EN	Various
NewsBag++	EN	Various
NewsBag Test	EN	Various

Multilingual datasets, while present, are limited both in terms of the number of available datasets and the quantity of samples they offer. The few multilingual datasets available often focus on very specific topics and their coverage remains relatively narrow compared to their English counterparts. This highlights the ongoing need for more diverse and inclusive datasets that represent a broader linguistic and cultural spectrum, ensuring that research can benefit from a wider range of sources and viewpoints.

E Multiclass dataset analysis

We extend the two-class overview from Table 6 to multiclass label spaces. First, we consider all datasets that naturally split into three categories (e.g., Support/Neutral/Refute or True/Neutral/False) in Table 11. Next, we examine datasets with six fine-grained labels in Table 12. The tables highlights the same common potential pitfalls identified in the binary analysis, such as rare classes which might benefit from merging and excessive length variance.

Table 11 Statistics on r/Fakeddit and MOCHEG with three classes

		Class	Text Length		
Dataset	Label	Distrib.	Avg	Min	Max
r/Fakeddit	C1(Label 0)	39.39%	53.53	1	297
	C2(Label 2)	58.16%	32.97	1	2,785
	C3(Label 1)	2.45%	67.63	1	290
MOCHEG	C1(refuted)	37.53%	3,677.37	125	34,179
	C2(supported)	32.97%	4,046.85	175	23,949
	C3(NEI)	29.50%	5,256.67	242	37,485

Table 12 r/Fakeddit and FineFake dataset statistics for six labels (0–5)

		Distribution		Text Length		
Dataset	Label	Count	Perc.	Avg	Min	Max
r/Fakeddit	0	268,908	39.39%	53.53	1	297
	1	40,516	5.93%	45.31	1	288
	2	129,795	19.01%	40.59	1	292
	3	14,246	2.09%	77.38	2	285
	4	203,139	29.75%	21.11	1	2,785
	5	26,057	3.82%	66.16	1	290
FineFake	0	7,502	44.37%	2,446.15	11	100,096
	1	1,477	8.73%	1,006.05	3	100,047
	2	1,612	9.53%	439.62	12	36,371
	3	3,444	20.37%	271.27	20	62,861
	4	2,574	15.22%	172.52	9	100,049
	5	300	1.77%	3,387.25	13	97,166

F Binary dataset analysis – additional details

Text format and label distribution Figure 3 illustrates the distribution of claim and document lengths across different classes in the dataset. The top row presents histograms for claim lengths, while the bottom row displays document lengths. The vertical dashed lines indicate the mean length for each class. Claims exhibit relatively consistent lengths across most classes, except for the "Refute" class, which contains shorter claims on average. In contrast, document lengths vary significantly, with "Support_Multimodal" and "Support_Text" containing the longest documents, whereas "Refute" has the shortest. This variation suggests potential biases in textual evidence availability across classes, which may influence model performance.

Topics and label distribution Figure 4 presents the distribution of topics across labels in the FineFake [14] dataset. The heatmap on the left shows the distribution of samples considering the binary labels (0: fake, 1: real) across different topics, while the heatmap on the right provides a finer-grained breakdown of label distributions. The intensity of the color indicates the number of instances, with darker shades representing higher frequencies. Notably, most of the samples belong to the "Politics" and "Society" topics, for which the fake class is also the most frequent. Conversely, categories such as "Health" and "Uncategorized" are under-represented, and "Business" and "Conflict" present an even distribution of fake and real samples. The variation in topic distribution across labels highlights potential biases in the dataset, which may influence model learning and generalization.

Figure 5 presents the distribution of subreddit sources across different labeling schemes in the r/Fakeddit [39] dataset. The three heatmaps correspond to the 2-way, 3-way, and 6-way labeling schemes. Notably, the "psbattle_artwork" and "mildlyinteresting" subreddits contain the highest number of instances, and most of their samples belong to a single label (e.g., 0 and 1 respectively for the 2-way labeling scheme). Other subreddits have more balanced distributions across labels. The variation in label distribution across subreddits suggests potential dataset biases, as certain subreddits contribute disproportionately to specific labels. This imbalance may impact model generalization and requires careful consideration during training and evaluation.

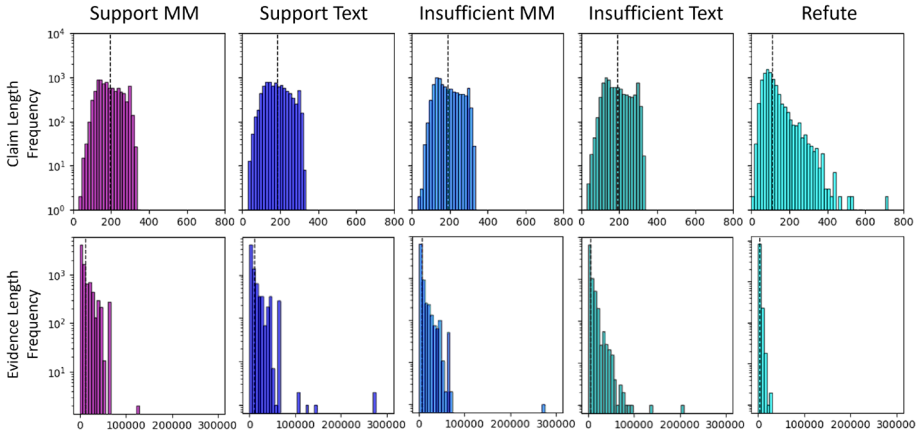

Fig. 3 Factify2 text length and distribution (in characters) for each of the five labels. Left column: claim, right column: evidence

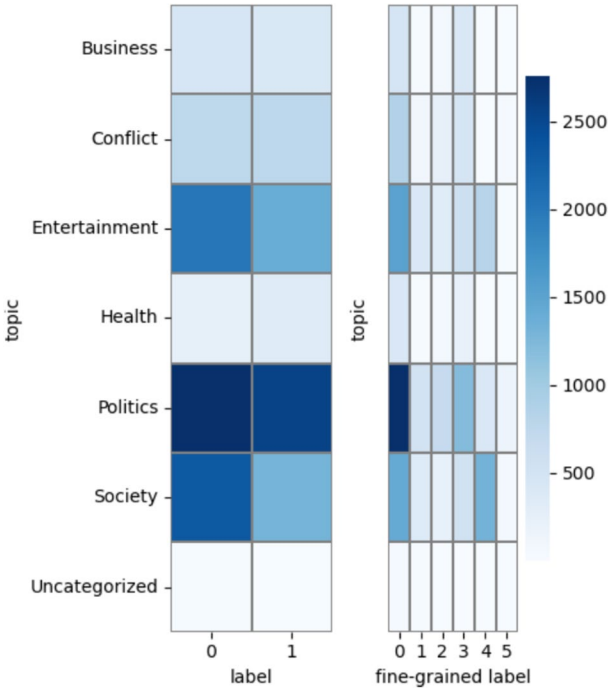

Fig. 4 FineFake label distribution for each topic

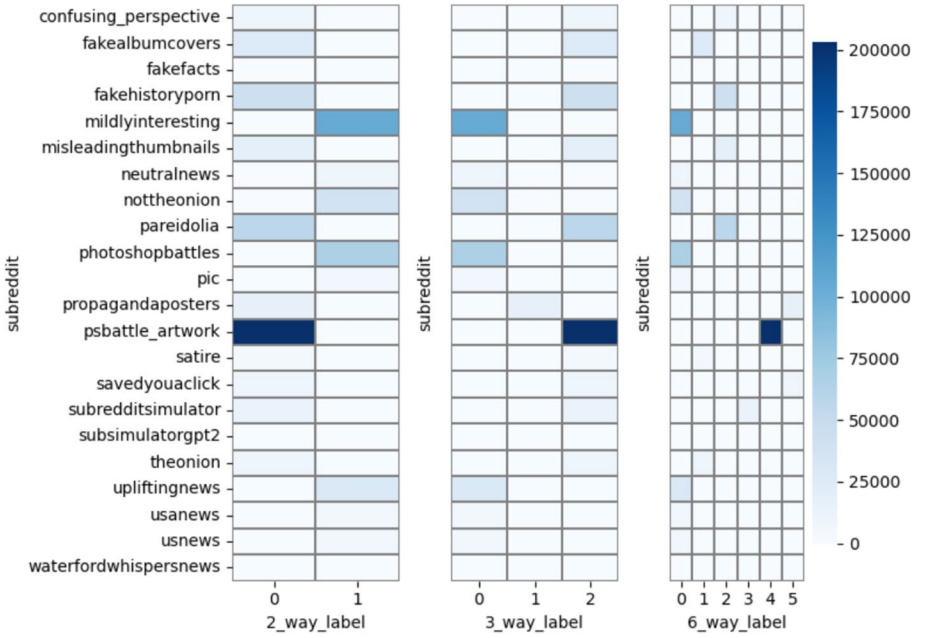

Fig. 5 r/Fakeddit label distribution for each subreddit

Author Contributions **Ian Marco Gallegos Carvajal:** Investigation, Software, Validation, Visualization, Formal analysis, Data curation, Writing - Original draft preparation. **Beatrice Portelli:** Investigation, Visualization, Data curation, Writing - Original draft preparation. **Leonardo Zini:** Investigation, Validation, Formal analysis, Data curation, Writing - Original draft preparation. **Lorenzo Baraldi:** Conceptualization, Methodology, Supervision, Writing - Reviewing and Editing, Funding acquisition. **Giuseppe Serra:** Conceptualization, Methodology, Supervision, Writing - Reviewing and Editing, Funding acquisition.

Funding Open access funding provided by Università degli Studi di Udine within the CRUI-CARE Agreement. This work was supported by the PRIN 2022 "MUSMA" - CUP G53D23002930006 - "Funded by EU - Next-Generation EU – M4 C2 I1.1", and by the Department Strategic Plan (PSD) of the University of Udine – Interdepartmental Project on Artificial Intelligence (2020-25).

Data Availability All materials and instructions on how to access the datasets are publicly available online at https://github.com/beatrice-portelli/multimodal-afc-survey

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Guo Z, Schlichtkrull M, Vlachos A (2022) A survey on automated fact-checking. Trans Assoc Computat Linguist 10:178–206. https://doi.org/10.1162/tacl_a_00454. https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00454
2. Suryavardan S, Mishra S, Chakraborty M, Patwa P, Rani A, Chadha A, Reganti A, Das A, Sheth A, Chinnakotla M et al (2023) Findings of factify 2: Multimodal fake news detection. In: Proceedings of de-factify 2: 2nd workshop on multimodal fact checking and hate speech detection
3. Chrysidis Z, Papadopoulos S-I, Papadopoulos S, Petrantonakis P (2024) Credible, unreliable or leaked?: Evidence verification for enhanced automated fact-checking. In: Proceedings of the 3rd ACM international workshop on multimedia AI against disinformation. MAD '24, pp 73–81. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3643491.3660278
4. Zeng X, Abumansour AS, Zubiaga A (2021) Automated fact-checking: A survey. Lang Linguist Compass 15(10):12438
5. Kotonya N, Toni F (2020) Explainable automated fact-checking: A survey. In: Scott D, Bel N, Zong C (eds) Proceedings of the 28th international conference on computational linguistics, pp 5430–5443. International committee on computational linguistics, Barcelona, Spain (Online). https://doi.org/10.18653/v1/2020.coling-main.474
6. Hardalov M, Arora A, Nakov P, Augenstein I (2022) A survey on stance detection for mis- and disinformation identification. In: Carpuat M, Marneffe M-C, Meza Ruiz IV (eds) Findings of the association for computational linguistics: NAACL 2022, pp 1259–1277. Association for computational linguistics, Seattle, United States. https://doi.org/10.18653/v1/2022.findings-naacl.94
7. Küçük D, Can F (2020) Stance detection: A survey. ACM Comput Surv 53(1). https://doi.org/10.1145/3369026
8. Tufchi S, Yadav A, Ahmed T (2023) A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. Intern J Multimed Inf Retriev 12(2):28

9. Hangloo S, Arora B (2022) Combating multimodal fake news on social media: methods, datasets, and future perspective. Multimedia Syst 28(6):2391–2422
10. Akhtar M, Schlichtkrull MS, Guo Z, Cocarascu O, Simperl E, Vlachos A (2023) Multimodal automated fact-checking: A survey. In: The 2023 conference on empirical methods in natural language processing. https://openreview.net/forum?id=ggTNeg2fem
11. Zlatkova D, Nakov P, Koychev I (2019) Fact-checking meets fauxtography: Verifying claims about images. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp 2099–2108. Association for computational linguistics, Hong Kong, China. https://doi.org/10.18653/v1/D19-1216
12. Molina I, Keith B, Matus M (2025) A multimodal dataset of fact-checked news from chile's constitutional processes: Collection, processing, and analysis. Data 10(2):13
13. Shahi GK, Jaiswal AK, Mandl T (2024) Fakeclaim: A multiple platform-driven dataset for identification of fake news on 2023 israel-hamas war. In: Goharian N, Tonellotto N, He Y, Lipani A, McDonald G, Macdonald C, Ounis I (eds) Adv Inf Retriev. Springer, Cham, pp 66–74
14. Zhou Z, Zhang X, Zhang L, Liu J, Wang S, Liu Z, Zhang X, Li C, Yu PS (2024) FineFake: A Knowledge-enriched dataset for fine-grained multi-domain fake news detection. arxiv:2404.01336
15. Shahi GK (2024) Warclaim: A dataset for fake news on 2023 israel–hamas war. In: Companion publication of the 16th ACM web science conference. Websci Companion '24, pp 19–21. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3630744.3658410
16. Yang Z, Lin J, Guo Z, Li Y, Li X, Li Q, Liu W (2024) Towards rumor detection with multi-granularity evidences: A dataset and benchmark. IEEE Trans Knowl Data Eng 36(11):7188–7200. https://doi.org/10.1109/TKDE.2024.3401700
17. Hu X, Guo Z, Chen J, Wen L, Yu PS (2023) Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval. SIGIR '23, pp 2901–2912. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3539618.3591896
18. Suryavardan S, Mishra S, Patwa P, Chakraborty M, Rani A, Reganti A, Chadha A, Das A, Sheth A, Chinnakotla M et al (2023) Factify 2: A multimodal fake news and satire news dataset. In: Proceedings of de-factify 2: 2nd workshop on multimodal fact checking and hate speech detection
19. Yao BM, Shah A, Sun L, Cho J-H, Huang L (2023) 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. SIGIR '23, pp 2733–2743. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3539618.3591879
20. Abdelnabi S, Hasan R, Fritz M (2022) Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In: Conference on Computer Vision and Pattern Recognition (CVPR)
21. Rayar F, Delalandre M, Le V-H (2022) A large-scale tv video and metadata database for french political content analysis and fact-checking. In: Proceedings of the 19th international conference on content-based multimedia indexing. CBMI '22, pp 181–185. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3549555.3549557
22. Mishra S, Suryavardan S, Bhaskar A, Chopra P, Reganti AN, Patwa P, Das A, Chakraborty T, Sheth AP, Ekbal A (2022) Factify: A multi-modal fact verification dataset. In: Proceedings of the workshop on multi-modal fake news and hate-speech detection (DE-FACTIFY 2022)
23. Nielsen DS, McConville R (2022) Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval. SIGIR '22, pp 3141–3153. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3477495.3531744
24. Vo N, Lee K (2020) Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing (EMNLP), pp 7717–7731. Association for computational linguistics, online. https://doi.org/10.18653/v1/2020.emnlp-main.621
25. Papadopoulos S-I, Koutlis C, Papadopoulos S, Petrantonakis PC (2024) Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. Intern J Multimed Inf Retriev 13(1):4
26. Bondielli A, Dell'Oglio P, Lenci A, Marcelloni F, Passaro LC, Sabbatini M (2023) Multi-fake-detective at evalita 2023: Overview of the multimodal fake news detection and verification task. In: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)

27. Papadopoulos S-I, Koutlis C, Papadopoulos S, Petrantonakis P (2023) Synthetic misinformers: Generating and combating multimodal misinformation. In: Proceedings of the 2nd ACM international workshop on multimedia AI against disinformation. MAD '23, pp 36–44. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3592572.3592842
28. Aneja S, Bregler C, Niessner M (2023) Cosmos: Catching out-of-context image misuse using self-supervised learning. Proceed AAAI Conf Artif Intell 37(12):14084–14092. https://doi.org/10.1609/aaai.v37i12.26648
29. Biamby G, Luo G, Darrell T, Rohrbach A (2022) Twitter-COMMs: Detecting climate, COVID, and military multimodal misinformation. In: Carpuat M, Marneffe M-C, Meza Ruiz IV (eds) Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp 1530–1549. Association for computational linguistics, seattle, United States. https://doi.org/10.18653/v1/2022.naacl-main.110. https://aclanthology.org/2022.naacl-main.110/
30. Krstovski K, Ryu AS, Kogut B (2022) Evons: A dataset for fake and real news virality analysis and prediction. In: Calzolari N, Huang C-R, Kim H, Pustejovsky J, Wanner L, Choi K-S, Ryu P-M, Chen H-H, Donatelli L, Ji H, Kurohashi S, Paggio P, Xue N, Kim S, Hahm Y, He Z, Lee TK, Santus E, Bond F, Na S-H (eds) Proceedings of the 29th international conference on computational linguistics, pp 3589–3596. International committee on computational linguistics, Gyeongju, Republic of Korea. https://aclanthology.org/2022.coling-1.317/
31. Raj C, Meel P (2022) Arcnn framework for multimodal infodemic detection. Neural Netw 146:36–68. https://doi.org/10.1016/j.neunet.2021.11.006
32. Micallef N, Sandoval-Castañeda M, Cohen A, Ahamad M, Kumar S, Memon N (2022) Cross-platform multimodal misinformation: Taxonomy, characteristics and detection for textual posts and videos. Proceed Intern AAAI Conf Web Soc Med 16:651–662. https://doi.org/10.1609/icwsm.v16i1.19323
33. Luo G, Darrell T, Rohrbach A (2021) NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In: Moens M-F, Huang X, Specia L, Yih SW-t (eds) Proceedings of the 2021 conference on empirical methods in natural language processing, pp 6801–6817. Association for computational linguistics, online and Punta Cana, Dominican Republic. https://doi.org/10.18653/v1/2021.emnlp-main.545
34. Fung Y, Thomas C, Gangi Reddy R, Polisetty S, Ji H, Chang S-F, McKeown K, Bansal M, Sil A (2021) InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In: Zong C, Xia F, Li W, Navigli R (eds) Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers), pp 1683–1698. Association for computational linguistics, online. https://doi.org/10.18653/v1/2021.acl-long.133
35. Song C, Ning N, Zhang Y, Wu B (2021) A multimodal fake news detection model based on cross-modal attention residual and multichannel convolutional neural networks. Inform Process Manage 58(1):102437. https://doi.org/10.1016/j.ipm.2020.102437
36. Tan R, Plummer B, Saenko K (2020) Detecting cross-modal inconsistency to defend against neural fake news. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing (EMNLP), pp 2081–2106. Association for computational linguistics, online. https://doi.org/10.18653/v1/2020.emnlp-main.163
37. Müller-Budack E, Theiner J, Diering S, Idahl M, Ewerth R (2020) Multimodal analytics for real-world news using measures of cross-modal entity consistency. In: Proceedings of the 2020 international conference on multimedia retrieval. ICMR '20, pp 16–25. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3372278.3390670
38. Zhou X, Mulay A, Ferrara E, Zafarani R (2020) Recovery: A multimodal repository for covid-19 news credibility research. In: Proceedings of the 29th ACM international conference on information & knowledge management. CIKM '20, pp 3205–3212. Association for computing machinery, New York, NY, USA. https://doi.org/10.1145/3340531.3412880
39. Nakamura K, Levy S, Wang WY (2020) Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the twelfth language resources and evaluation conference, pp 6149–6157. European language resources association, Marseille, France. https://aclanthology.org/2020.lrec-1.755/
40. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2020) Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data 8(3):171–188. https://doi.org/10.1089/big.2020.0062
41. Kou Z, Yue Zhang D, Shang L, Wang D (2020) Exfaux: A weakly supervised approach to explainable fauxtography detection. In: 2020 IEEE international conference on big data (big data), pp 631–636. https://doi.org/10.1109/BigData50022.2020.9378019

42. Jindal S, Sood R, Singh R, Vatsa M, Chakraborty T (2020) Newsbag: A benchmark dataset for fake news detection. In: Proceedings of the workshop on artificial intelligence safety (SafeAI 2020). https://api.semanticscholar.org/CorpusID:246273451
43. Du W-W, Wu H-W, Wang W-Y, Peng W-C (2023) Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. In: Proceedings of de-factify 2: 2nd workshop on multimodal fact checking and hate speech detection
44. Hessel J, Holtzman A, Forbes M, Le Bras R, Choi Y (2021) CLIPScore: A reference-free evaluation metric for image captioning

Authors and Affiliations

Ian Marco Gallegos Carvajal[1] · Beatrice Portelli[1,2] · Leonardo Zini[3] · Lorenzo Baraldi[3] · Giuseppe Serra[1]

✉ Beatrice Portelli
portelli.beatrice@spes.uniud.it

Ian Marco Gallegos Carvajal
gallegoscarvajal.ianmarco@spes.uniud.it

Leonardo Zini
leonardo.zini@unimore.it

Lorenzo Baraldi
lorenzo.baraldi@unimore.it

Giuseppe Serra
giuseppe.serra@uniud.it

[1] Department of Mathematics, Computer Science and Physics, University of Udine, Via delle Scienze, 206, 33100 Udine, Italy

[2] Department of Agricultural, Food, Environmental and Animal Sciences, University of Udine, Via delle Scienze, 206, 33100 Udine, Italy

[3] Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Via P. Vivarelli, 10, 41125 Modena, Italy