Multimodal Automated Fact-Checking: A Survey

Mubashara Akhtar[1,*], **Michael Schlichtkrull**[2], **Zhijiang Guo**[2], **Oana Cocarascu**[1], **Elena Simperl**[1] and **Andreas Vlachos**[2]

[1]Department of Informatics, King's College London
[2]Department of Computer Science and Technology, University of Cambridge

{mubashara.akhtar,oana.cocarascu,elena.simperl}@kcl.ac.uk
{mss84,zg283,av308}@cam.ac.uk

Abstract

Misinformation is often conveyed in multiple modalities, e.g. a miscaptioned image. Multimodal misinformation is perceived as more credible by humans, and spreads faster than its text-only counterparts. While an increasing body of research investigates automated fact-checking (AFC), previous surveys mostly focus on text. In this survey, we conceptualise a framework for AFC including subtasks unique to multimodal misinformation. Furthermore, we discuss related terms used in different communities and map them to our framework. We focus on four modalities prevalent in real-world fact-checking: text, image, audio, and video. We survey benchmarks and models, and discuss limitations and promising directions for future research.

1 Introduction

Motivated by the challenges presented by misinformation in the modern media ecosystem, previous research has commonly modelled automated fact-checking (AFC) as a pipeline consisting of different stages, surveyed in a variety of axes (Thorne and Vlachos, 2018; Kotonya and Toni, 2020a; Zeng et al., 2021; Nakov et al., 2021; Guo et al., 2022). However, these surveys focus on a single modality, text. This is different to real-world misinformation that often occurs via several modalities.

In AFC, the term *multimodal* has been used to refer to cases where the claim and/or evidence are expressed through different or multiple modalities (Hameleers et al., 2020; Alam et al., 2022; Biamby et al., 2022). Examples of multimodal misinformation include: (i) claims about digitally manipulated content (Agarwal et al., 2019; Rössler et al., 2018) such as photos depicting former US president Trump's arrest (Figure 1); (ii) combining

Figure 1: Manipulated image depicting arrest of former US president Donald Trump (source: BBC[1]).

content from different modalities and contexts, e.g. using video footage in a misleading context (Aneja et al., 2021; Biamby et al., 2022; Abdelnabi et al., 2022); (iii) embedding a claim in another modality, e.g. a meme, an image with embedded text (Qu et al., 2022a), with notable real-world examples including a Brexit Vote Leave poster[2] and TikTok videos with COVID misinformation (Shang et al., 2021); (iv) verifying a claim with evidence from a different modality than the input claim, e.g. verifying images against text (Shao et al., 2023), audio against textual metadata (Kopev et al., 2019), and text against images (Akhtar et al., 2023).

Fact-checking multimodal misinformation is important for a number of reasons. First, multimodal content is perceived as more credible compared to text containing a similar claim (Newman et al., 2012). For example, previous research shows that visual content exhibits a "photo truthiness"-effect (Newman and Zhang, 2020), biasing readers to believe a claim is true. Second, multimodal content spreads faster and has a higher engagement than text-only posts (Li and Xie, 2020). Third, with recent advances in generative machine learning models (Rombach et al., 2022), the generation of multimodal misinformation has been simplified.

To validate the importance of multimodal fact-

[*] This work was partially done during Mubashara's research visit at Cambridge.

[1] https://www.bbc.com/news/world-us-canada-65069316

[2] https://www.itv.com/news/2019-01-18/boris-johnson-under-attack-over-turkey-claim/

Claim Modality	Percentage
Image	20.07%
Video	8.06%
Audio	0.55%
Total	28.68%

Table 1: Non-textual modalities present and/or used in addition to text in our manually annotated snapshot of real-world claims from the Google ClaimReview API.

checking, we manually annotated 9,255 claims from the AVeriTeC dataset (Schlichtkrull et al., 2023), which were collected with the Google FactCheck ClaimReview API[3]. For each claim, we identified the modalities present in it and evidence strategies (e.g. identification of geolocation) used for fact-checking. We find that more than $2,600$ (28.68%) claims either contain multimodal data or require multimodal reasoning for verification, with 20.07% involving images, 8.06% videos, and 0.55% audios (see Table 1).[4] These claims can neither be fact-checked by a text-only model, nor by a model with no text capabilities.

In this survey, we introduce a three-stage framework for multimodal automated fact-checking: claim detection and extraction, evidence retrieval, and verdict prediction encompassing veracity, manipulation and out-of-context classification, as well as justification production. The input and output data of each stage can have different or multiple modalities. For each stage, we discuss related terms and definitions developed in different research communities. In contrast to previous surveys on multimodal fact-checking that focus on individual subtasks (Cao et al., 2020; Alam et al., 2022; Abdali, 2022), we consider all subtasks surveying benchmarks and modeling approaches for them.

We focus on four prevalent modalities of real-world fact-checking identified in our annotations: text, image, audio, and video. While tables and knowledge graphs are increasingly used as evidence for benchmarks (Chen et al., 2020; Aly et al., 2021; Akhtar et al., 2022), they have been covered in previous surveys (Thorne and Vlachos, 2018; Zeng et al., 2021; Guo et al., 2022). Finally, we discuss the extent to which current approaches to AFC work for multimodal data, and promising directions for further research (Section 4). We accompany the survey with a repository,[5] which lists the resources mentioned in our survey.

2 Task Formulation

This section introduces a conceptualisation of multimodal AFC as a three-stage process, including claim detection and extraction, evidence retrieval, and production of verdicts and justifications for various types of misinformation (Figure 2). Compared to the text-only pipeline presented in Guo et al. (2022), our framework extends their first stage with a claim *extraction* stage, and generalises their third stage to cover tasks that fall under multimodal AFC, thus accounting for its particular challenges.

Terminology. A number of works (Singhal et al., 2022; Fung et al., 2021) use the term *multimedia*, which is also more common in public discussions instead of *multimodal* (Lauer, 2009). However in in this survey we adopt the latter, following other surveys that use multimodal data (Liang et al., 2022; Guo et al., 2019). Adopting the terminology of previous surveys (Thorne and Vlachos, 2018; Alam et al., 2022) and following advice from institutions such as the UNO (Ireton and Posetti, 2018), we avoid *multimodal fake news* (Meel and Vishwakarma, 2021; Amri et al., 2021; Patwa et al., 2022) due to the term's ambiguous use.

Stage 1: Claim Detection and Extraction. The first pipeline stage aims to find *checkable* (i.e. factually-verifiable) and *check-worthy* (i.e. important factual claims (Hassan et al., 2015b)) claims. Debunking a typical claim and writing the fact-checking article takes approximately one day for a human fact-checker (Hassan et al., 2015a). This stage aims to focus the AFC process on claims which are verifiable and most impactful. Multimodal claims can be diverse and include: (1) a written claim embedded in another modality (Prabhakar et al., 2021; Maros et al., 2021) such as an image or a spoken claim in an audio or video; (2) a claim that a piece of content is authentic, e.g. that a video footage is from a specific geographic location (Zhang et al., 2018; Heller et al., 2018); (3) a claim for which the evidence is manipulated to support it, e.g. through lip-syncing (Rössler et al., 2018). While in some cases the claim is clearly specified (e.g. in form of a headline), in often multiple modalities are required to understand and ex-

[3] https://toolbox.google.com/factcheck/apis

[4] Annotations at http://github.com/MichSchli/AVeriTeC.

[5] https://github.com/Cartus/Automated-Fact-Checking-Resources

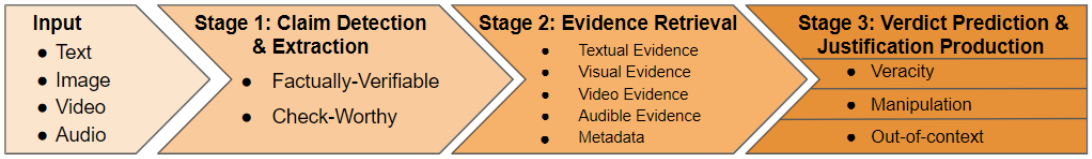

Figure 2: Multimodal fact-checking pipeline.

tract a claim at this stage. Simply *detecting* potentially misleading content is often not enough – it is necessary to *extract* the claim before fact-checking it in the subsequent stages. For example, detecting text in images or videos and understanding it given the context (Qu et al., 2022b) or verifying audios by transcribing and extracting claims (Maros et al., 2021).

Stage 2: Evidence Retrieval. Similarly to fact-checking with text, multimodal fact-checking often relies on evidence to make judgments, similar to the process followed by human fact-checkers (Silverman, 2013; Nakov et al., 2021). Two main approaches have been used in the past: (i) using the claim to-be-checked as evidence itself, e.g. to detect manipulation (Qi et al., 2019; Bonettini et al., 2020); this can be seen as the multimodal version of evidence-free fact-checking of text claims by checking logical fallacies in the text (Jin et al., 2022), and (ii) retrieving additional evidence (Abdelnabi et al., 2022). In multimodal fact-checking, the evidence modality can be different from the claim modality. For example, to retrieve evidence for image or audio fact-checking, previous works have also used text e.g. metadata, social media comments, or captions (Gupta et al., 2013; Huh et al., 2018; Müller-Budack et al., 2020; Kopev et al., 2019).

Stage 3: Verdict Prediction and Justification Production. Following the fact-checking process of professional fact-checkers, the final stage comprises verdict prediction and the production of justification that explains the fact-check to humans (Graves, 2018). Verdict prediction is decomposed into three tasks considering prevalent multimodal misinformation types: manipulation, using content out-of-context, and veracity classification.

Stage 3.1: Manipulation Classification. Manipulation classification commonly addresses (i) misinformative claims with manipulated content; (ii) correct claims accompanied by manipulated content (e.g. to increase credibility). Many methods exist to manipulate text, visual and audio content. While some require more knowledge to use (e.g. speech synthesis), other manipulations can be achieved with simple tools (e.g. changing speed of videos) (Paris and Donovan, 2019). Different terms have been used for manipulated content in recent years. A *deepfake* is commonly defined as "the product of artificial intelligence (AI) applications that [...] create fake videos that appear authentic" (Maras and Alexandrou, 2019), with popular examples including realistic-looking videos where the speaker's voice or face is modified (Paris and Donovan, 2019). On the other hand, *cheap fake* defines manipulated content created through more accessible methods (Paris and Donovan, 2019), e.g. changing captions or speed of videos (La et al., 2022). The term *fauxtography* was first coined in journalism for images manipulated to "convey a questionable (or outright false) sense of the events they seem to depict" (Cooper, 2007; Kalb and Saivetz, 2007). Other terms used in the literature for manipulated content are *fake* (Cheema et al., 2022), *forgery* (Cozzolino et al., 2021), and *splice* (Zampoglou et al., 2015).

Stage 3.2: Out-of-context Classification. Using unchanged content out-of-context is one of the most common and easiest methods to create multimodal misinformation (Luo et al., 2021; Aneja et al., 2021), and involves (possibly misinformative) textual claims paired with content (e.g. a video) taken out of context (Zhang et al., 2018; Abdelnabi et al., 2022; Garimella and Eckles, 2020). Recent work has also studied the applicability of traditional multimodal misinformation detection methods to identify out-of-context content (Zhang et al., 2023). Other terms used for combining multimodal content in a misleading way include *cross-modal (in-) consistency* (Müller-Budack et al., 2020) and *repurposing* (Luo et al., 2021).

Stage 3.3: Veracity Classification. This task is the multimodal counterpart to classifying the veracity of textual claims given retrieved evidence

(Thorne and Vlachos, 2018). Veracity classification of claims embedded in audio is also commonly referred to as *deception detection* (Kopev et al., 2019; Kamboj et al., 2021). While earlier work considered mostly claims recorded in staged setups (Newman et al., 2003) or from court trials (Pérez-Rosas et al., 2015), more recently real-world political debates have become popular. (Kopev et al., 2019; Kamboj et al., 2021).

Stage 3.4: Justification Production. Different to previous research on automated justification production (Kotonya and Toni, 2020a), human fact-checkers also give justifications for fact-checks involving images, audios, or videos (Silverman, 2013). Justifications for multimodal misinformation can be grouped in three categories: (i) identifying which part of the claim input is misleading (e.g. specific areas in a visual claim or words in a textual one) (Kou et al., 2020; Purwanto et al., 2021; Lourenço and Paes, 2022); (ii) providing natural language justifications following human fact-checkers (Yao et al., 2022); (iii) selecting and highlighting evidence parts used for verification (Atanasova et al., 2020; Shang et al., 2022). Justifications serve purposes beyond explaining veracity classification, e.g. human fact-checkers also use them to discuss uncertainties and potential errors – especially needed in fact-checking for rapidly developing events (Silverman, 2013).

3 Datasets and Modeling Approaches

3.1 Stage 1: Claim Detection and Extraction

Input. Typical inputs to claim detection are unimodal, including image (Garimella and Eckles, 2020; Qu et al., 2022a), audio (Maros et al., 2021), and video (Shang et al., 2021; Qi et al., 2022), which are collected from social media platforms such as WhatsApp and TikTok (see Table 2). The written or spoken claim is extracted from the input at this stage before fact-checking it.

Output. Claim detection is typically framed as a classification task. Models predict if a claim is checkable or check-worthy (Prabhakar et al., 2021; Cheema et al., 2022; Barrón-Cedeño et al., 2023). The verdict for factual-verifiability is often binary (Jin et al., 2017; Shang et al., 2021). For check-worthiness, Prabhakar et al. (2021) defines three categories of multimodal claims: statistical/numerical claims, claims about world events/places/noteworthy individuals, and other factual claims. Cheema et al. (2022) extend the binary labels for textual check-worthiness (Hassan et al., 2015b) with images to be considered as well. A tweet is considered check-worthy if it is potentially harmful, breaking news, or up-to-date.

Modeling Approaches. Detecting claims is a challenging task due to the vast number of posts that are published every day. Existing claim detection methods primarily rely on input content since the large volume of potentially check-worthy inputs makes it difficult to retrieve and use evidence. The early multimodal method directly concatenated visual and textual features for detection (Jin et al., 2017; Wang et al., 2018). However, simple modality fusion may not be sufficient to capture the complex relationships among multimodal information. As a result, later efforts focused on jointly learning representations across modalities. For instance, Khattar et al. (2019) leverage a variational autoencoder (Kingma and Welling, 2014) to learn a shared representation of visual and textual content. Various attention mechanisms have also been developed to fuse multimodal representations (Qian et al., 2021; Wu et al., 2021; Liu et al., 2023b; Qi et al., 2023). Another popular approach is to use graph neural networks (Kipf and Welling, 2017) to model the interactions among different modalities (Zheng et al., 2022; Sun et al., 2023).

Multimodal content can implicitly provide claims, as seen in images and videos on social media that often have accompanying text. To extract claims from visual input, OCR systems are commonly used (Garimella and Eckles, 2020; Prabhakar et al., 2021). Qu et al. (2022b) use Google Vision API to identify text in memes. Claim extraction becomes more challenging when dealing with video inputs. Shang et al. (2021) address this challenge by extracting captions and audio chunks after sampling video frames. These captions and audio chunks were then encoded into representations to guide the visual feature extraction process. For audio inputs, Maros et al. (2021) use Google's Speech-to-Text API to produce transcripts.

3.2 Stage 2: Evidence Retrieval

Previous work uses different types of evidence and retrieval methods given the modalities involved. Evidence data and retrieval approaches can be grouped into (i) content-based and (ii) retrieval-based (see column *evidence* in Table 3).

Content-based. Content-based approaches use the

Dataset	Input	Context	Output	#Input	Lang	Source
Weibo (Jin et al., 2017)	Img/Txt	Meta	2	9,528	Zh	Weibo/News
FauxBuster (Zhang et al., 2018)	Img/Txt	Txt/Meta	2	917	En	Twitter/Reddit
Exfaux (Kou et al., 2020)	Img/Txt	Txt	2/4	263	En	Twitter/Reddit
MuMIN (Nielsen and McConville, 2022)	Img/Txt	Meta	3	12,914	Mul	Twitter
MMClaims (Cheema et al., 2022)	Img/Txt	-	4	3,400	En	Twitter
ContrastFaux (Zong et al., 2023)	Img/Txt	-	2	1,841	En	Twitter/Reddit
CLEF2023 (Barrón-Cedeño et al., 2023)	Img/Txt	-	4	6,000	Mul	Twitter
MR2 (Hu et al., 2023)	Img/Txt	Txt/Img/Meta	3	14,700	Mul	Twitter/Weibo
IndiaWApp (Garimella and Eckles, 2020)	Img	Meta	2	2,500	Mul	WhatsApp
DisinfoMeme (Qu et al., 2022a)	Img	-	2	1,170	En	Reddit
WhatsApp (Maros et al., 2021)	Aud	Meta	2	42,689	Pt	WhatsApp
TikTok (Shang et al., 2021)	Vid	Txt/Meta	2	891	En	TikTok
COVID-VTS (Liu et al., 2023a)	Vid	Txt/Aud	2	10,000	En	Twitter
FakeSV (Qi et al., 2022)	Vid	Txt/Meta	2	3,654	Zh	TikTok/Kuai
MisDissem (Resende et al., 2019)	Vid/Aud/Img/Text	Meta	2	121,781	Pt	WhatsApp
CheckMate (Prabhakar et al., 2021)	Vid/Img/Text	Meta	3	2,200	Hi	Sharechat

Table 2: Datasets for claim detection. Img, Txt, Vid, Aud, and Meta denote image, text, video, audio, and metadata, respectively. Output indicates the number classification labels. Mul indicates that the input has multiple languages.

claim and its context (i.e. the same information that is used for claim detection and extraction) as evidence instead of retrieving additional data. This is particularly common for audio and video misinformation (Table 3). Acoustic or visual features extracted from the input are used as evidence for verdict prediction (Wu et al., 2015; Yi et al., 2021; Ismael Al-Sanjary et al., 2016; Jiang et al., 2020). Most approaches use audio (or video) features and accompanying data (e.g. metadata, transcripts if available) as evidence to identify inconsistencies (Kopev et al., 2019; Rössler et al., 2018; Li et al., 2020b). Several datasets with image/text claims (Tan et al., 2020; Luo et al., 2021; Aneja et al., 2021) also do not retrieve additional evidence (Table 3) but rely on the given claim input or use accompanying metadata (Jaiswal et al., 2017; Sabir et al., 2018). Metadata is also often used as evidence for verdict prediction with images as input (Table 3). Jaiswal et al. (2017) and Sabir et al. (2018) use metadata (e.g. image timestamps) to provide additional information. Similarly, Huh et al. (2018) incorporate EXIF metadata (e.g. camera version, focal length, resolution settings) to detect manipulation. Image captions are also used as evidence sometimes (Shao et al., 2023).

Retrieval-based. Retrieved evidence external to the claim is mostly used for fact-checking text claims, text/image and image claims while audio and video fact-checks often don't retrieve additional evidence data (Table 3) but rely on the content of the video/audio input. Fung et al. (2021) leverage a knowledge base for additional background knowledge. They first construct a knowledge graph of the input news article using its text and images. They extract entities/relations from this knowledge graph with an Information Extraction system (Li et al., 2020a; Lin et al., 2020) and map the entities to Freebase (Bollacker et al., 2008) as their background knowledge base. Two recent datasets scrape claims from fact-checking websites, and include text/image/video from those articles as evidence (Singhal et al., 2022; Yao et al., 2022). Akhtar et al. (2023) used chart images as evidence to verify textual claims. To determine if an image is used out-of-context, previous works also use *(reverse) image search* (Müller-Budack et al., 2020; Abdelnabi et al., 2022), to find evidence sources with images similar to or same as the claim image. Müller-Budack et al. (2020) query search engines and the *WikiData* knowledge graph using named entities from the claim text. Abdelnabi et al. (2022) use the claim image caption and the image itself as query.

3.3 Stage 3: Verdict Prediction

As introduced in Section 2, the verdict prediction stage includes manipulation, out-of-context, and veracity classification as sub-tasks.

Input. As shown in Table 3, inputs of **manipulation classification** datasets usually focus on one modality. For dataset creation, manipulated images are often collected from social media platforms such as Twitter, Reddit, and YouTube (Gupta et al., 2013; Heller et al., 2018). For verdict prediction datasets with videos, in addition to social media (Ismael Al-Sanjary et al., 2016), film clips (Guera and Delp, 2018), facial expressions (Rössler et al., 2018), and interviews (Li et al., 2020b) are used. Some works record videos to simulate real-world

Dataset	Input	Evidence	Output	Tasks	#Input	Lang	Source
MAIM (Jaiswal et al., 2017)	Img/Txt	Meta	2	O	239,968	En	Flickr
MEIR (Sabir et al., 2018)	Img/Txt	Meta	2	O	140,096	En	Flickr
TNews (Müller-Budack et al., 2020)	Img/Txt	Img	2	O	72,561	En	News
News400 (Müller-Budack et al., 2020)	Img/Txt	Img	2	O	400	En/De	News
NeuralNews (Tan et al., 2020)	Img/Txt	-	4	O	128,000	En	Grover/GoodNews
COSMOS (Aneja et al., 2021)	Img/Txt	-	2	O	201,700	En	News/Snopes
NewsCLIPings (Luo et al., 2021)	Img/Txt	-	2	O	988,283	En	CLIP/VisualNews
InfoSurgeon (Fung et al., 2021)	Img/Txt	KB/Meta	2	O	30,000	En	VoA
Factify (Suryavardan et al., 2023b)	Img/Txt	Txt	5	O	50,000	En	Twitter
FakingSandy (Gupta et al., 2013)	Img	Txt/Meta	2	M	16,117	-	Twitter
MediaEval (Boididou et al., 2014)	Img	Txt/Meta	2	M	13,924	-	Twitter
In-the-Wild (Huh et al., 2018)	Img	Meta	2	M	201	-	Reddit/Onion
PS-Battles (Heller et al., 2018)	Img	Txt/Meta	2	M	103,028	-	Reddit
DGM (Shao et al., 2023)	Img	Txt	2	M	230,000	-	News
VTD (Ismael Al-Sanjary et al., 2016)	Vid	-	2	M	33	En	YouTube
Faceforensics (Rössler et al., 2018)	Vid	-	2	M	1,004	En	YouTube
DeepfakeDetect (Guera and Delp, 2018)	Vid	-	2	M	600	En	Vid Webs./HOHA
DFDC (Dolhansky et al., 2019)	Vid	-	2	M	128,154	En	Recorded
DeeperForensics-1.0 (Jiang et al., 2020)	Vid	-	2	M	60,000	En	Recorded
Celeb-DF (Li et al., 2020b)	Vid	-	2	M	6,229	En	YouTube
KoDF (Kwon et al., 2021)	Vid	-	2	M	237,942	Ko	Recorded
DF-Platter (Narayan et al., 2023)	Vid	-	2	M	133,260	En	YouTube
ASVspoof (Wu et al., 2015)	Aud	-	2	M	16,375	En	SAS
Phonespoof (Lavrentyeva et al., 2019)	Aud	-	2	M	34,407	En	ASVspoof
FoR (Reimao and Tzerpos, 2019)	Aud	-	2	M	53,868	En	TTS Systems
DeepSonar (Wang et al., 2020)	Aud	-	2	M	18,614	En/Zh	TTS Systems/VCC
HAD (Yi et al., 2021)	Aud	-	3	M	88,035	Zh	AISHELL-3
FakeAVCeleb (Khalid et al., 2021)	Vid/Aud	-	4	M	20,000	En	VoxCeleb2
MedVideo (Hou et al., 2019)	Vid	-	2	VC	250	En	YouTube
CLEF2018 Audio (Kopev et al., 2019)	Aud	Meta	3	VC	286	En	Debates
FactDrill (Singhal et al., 2022)	Txt	Vid/Aud/Img/Txt/Meta	5	VC	22,435	Mul	FC webs.
MMM (Gupta et al., 2022)	Txt	Img/Meta	2	VC	10,473	Mul	FC webs.
ChartFC (Akhtar et al., 2023)	Txt	Img	2	VC	15,886	En	TabFact
Fauxtography (Zlatkova et al., 2019)	Img/Txt	Meta	2	VC	1,233	En	Snopes/Reuters
MOCHEG (Yao et al., 2022)	Img/Txt	Img/Txt	3	VC	21,184	En	FC webs.
r/Fakeddit (Nakamura et al., 2020)	Img/Txt	Meta	2/3/6	O/M/VC	1,063,106	En	Reddit

Table 3: Datasets for manipulation, out-of-context, and veracity classification. O, M and VC denote out-of-context, manipulation and veracity classification, respectively. Mul indicates the input has multiple languages.

scenarios (Dolhansky et al., 2019; Jiang et al., 2020; Kwon et al., 2021). To create datasets of manipulated content, altering methods based on GANs have also been applied in earlier works (Zakharov et al., 2019; Nirkin et al., 2019; Karras et al., 2019). For audio manipulations, most benchmarks (Wu et al., 2015; Kinnunen et al., 2017; Reimao and Tzerpos, 2019; Wang et al., 2020; Yi et al., 2021) use speech synthesis and voice conversion algorithms to collect manipulated audios. To assess real-world audio manipulations, Lavrentyeva et al. (2019) emulate realistic telephone channels.

Most **out-of-context classification** datasets have image-caption pairs as input (Table 3). Jaiswal et al. (2017) replace captions of Flickr images to get mismatched pairs. As replacing the entire caption can be easy to detect, later efforts (Sabir et al., 2018; Müller-Budack et al., 2020) propose to change specific entities in them. Luo et al. (2021) show that such text manipulations introduce linguistic biases and can be solved without the images. They use CLIP (Radford et al., 2021) to filter out pairs that do not require multimodal modeling. Popular sources for out of context datasets with text and image claims include Flickr and news/fact-checking websites (Aneja et al., 2021; Jaiswal et al., 2017; Sabir et al., 2018).

The primary input to multimodal **veracity classification** is the content-to-be-checked itself – typically text, audio or video in past benchmarks. Kopev et al. (2019) include verified speeches from the CLEF-2018 Task 2 (Nakov et al., 2018) while Hou et al. (2019) collect videos about prostate cancer verified by urologists. Zlatkova et al. (2019) and Yao et al. (2022) collect viral images with texts verified by dedicated agencies. Nakamura et al. (2020) collect image-text pairs from Reddit via distant supervision, e.g. labeling a post from the subreddit "fakefacts" as *misleading* and from "photoshopbattles" as *manipulated*. For veracity classification of spoken claims, real-world political debates are popular sources for claims (Kopev et al., 2019; Kamboj et al., 2021). For example, Kopev et al. (2019) and Kamboj et al. (2021) use claims labelled by fact checking organizations, and video recordings as well as transcripts of the respective political debates.

Output. Most manipulation and out-of-context

classification datasets use binary labels: "out-of-context/not out-of-context" (Müller-Budack et al., 2020; Luo et al., 2021), "pristine/falsified" (Boididou et al., 2014; Heller et al., 2018), "manipulation/no manipulation" (Dolhansky et al., 2019; Li et al., 2020b). Following fact-checkers, veracity classification datasets (Singhal et al., 2022; Nakamura et al., 2020) sometimes employ multi-class labels to represent degrees of truthfulness (e.g. true, mostly-true, half-true) (see Table 3). Mishra et al. (2022) adopt labels to denote the entailment between different claim and evidence modalities, e.g. the label *support text* denotes that only the textual part of the evidence supports the claim but not the accompanying image while *support multimodal* includes both modalities.

Modeling Approaches. To detect visual manipulations, early approaches mostly use CNN models, such as VGG16 (Amerini et al., 2019; Dang et al., 2020), ResNet (Amerini et al., 2019; Sabir et al., 2019), and InceptionV3 (Guera and Delp, 2018). Some works extend them to capture temporal aspects of video **manipulation classification**. Amerini et al. (2019) adopt optical flow fields to capture the correlation between consequent video frames and detect dissimilarities caused by manipulation. Guera and Delp (2018) model temporal information with an LSTM model and a sequence of features vectors per video frame to classify manipulated videos. Sabir et al. (2019) similarly extract features for video frames and detect discrepancies between frames using a recurrent convolution network. Some recent models also integrate transformer-based components (Vaswani et al., 2017; Zheng et al., 2021). For example, Wang et al. (2022) combine CNNs and vision transformers (ViTs) (Dosovitskiy et al., 2021) while Wodajo and Atnafu (2021) introduce a multi-scale ViT with variable patch sizes.

Models for **out-of-context** and **veracity classification** typically consist of unimodal encoders, a fusion component to obtain joint, multimodal representations, and a classification component. To obtain text representations, early approaches used combinations of word2vec models (Mikolov et al., 2013), LSTMs (Hochreiter and Schmidhuber, 1997), and TF-IDF scores for n-grams (Jin et al., 2017; Tanwar and Sharma, 2020; Hou et al., 2019). More recent efforts use pretrained language models (Fung et al., 2021; Aneja et al., 2021; La et al., 2022). To encode visual data, many approaches first detect objects in visual content using a Mask R-CNN model (He et al., 2017) before extracting visual features (Aneja et al., 2021; La et al., 2022; Shang et al., 2022). Visual representations for images and videos are commonly obtained using CNN models such as ResNet (He et al., 2016; Garimella and Eckles, 2020; Abdelnabi et al., 2022), VGG (Simonyan and Zisserman, 2015; Jin et al., 2017; Sabir et al., 2018), and Inception (Szegedy et al., 2015; Guera and Delp, 2018; Roy and Ekbal, 2021). To obtain audio features for voice quality, loudness, and tonality, Shang et al. (2021) extract the Mel-frequency cepstral coefficient, Kopev et al. (2019) use the INTERSPEECH 2013 ComParE feature set (Eyben et al., 2013), and Hou et al. (2019) use the openEAR toolkit (Eyben et al., 2009). Various approaches have been used to obtain **multimodal representations**. Early fusion, which joins representations immediately after the encoding step (Baltrusaitis et al., 2019) is more common (Aneja et al., 2021; Tanwar and Sharma, 2020; La et al., 2022) than late fusion (Yao et al., 2022). Moreover, model-agnostic methods (e.g. concatenation and dot product) are more prevalent (Aneja et al., 2021; Kopev et al., 2019; Jin et al., 2017; La et al., 2022) than model-based approaches (e.g. neural networks) (Jaiswal et al., 2017; Shang et al., 2022). Also popular for out-of-context classification are *cross-modality checks* that compare modalities present in a claim to each other, e.g. a video and its caption (Müller-Budack et al., 2020; Fung et al., 2021).

3.4 Stage 3: Justification Production

A small number of datasets is available for multimodal justification production. Previous work can be grouped into two categories: (1) highlighting parts of the input, and (2) generating natural language justifications.

Highlighting Input. The first category highlights input parts as justification which contribute to models' results. A popular approach for this are Graph Neural Networks (Kipf and Welling, 2017). Several papers encode multimodal data as graph elements, combining entities and their relations in and between modalities. Models are trained to detect inconsistencies between different modalities, or to detect relations (i.e., between entities) that may be misinformative. This detection could be based on the local graph structure, or on an external knowledge base (Fung et al., 2021; Shang et al., 2022;

Kou et al., 2020). Highlighted entities and relations serve as explanations for the potential misinformativeness of the entire graph. Conversely, Zhou et al. (2018) and Wu et al. (2019) use a multitask model for manipulation classification and identification of manipulated regions. Rather than labeled data, some papers rely on attention mechanisms to highlight areas as explanations. Bonettini et al. (2020); Dang et al. (2020) use this approach to highlight manipulated image regions; Purwanto et al. (2021) also include captions.

Natural Language Justifications. Yao et al. (2022) recently introduced a multimodal dataset with natural language justifications. They scrape text and visual content from web pages referenced by fact-checking articles. The dataset includes summaries in the fact-checking articles as gold justifications for the verdicts. However, such a setting is not realistic, as fact-checking articles are not available when verifying a new claim.

4 Challenges and Future Directions

Claim extraction from multimodal content. Multimodal claims, e.g. manipulated videos, are often embedded in specific contexts and framed as (part of) larger stories. For example, countering the misinformation in Figure 1 requires not only classifying if the image is manipulated, but understanding that it depicts the arrest of the former president in one of the cases he is being charged in. Only then can relevant evidence data be extracted and used to verify the story of Trump's arrest.To determine what is being claimed is a challenging first step in multimodal automated fact-checking. However, current efforts for multimodal claim extraction are limited to text extraction from visual content or transcribing audios and videos (Qu et al., 2022b; Garimella and Eckles, 2020; Maros et al., 2021). Addressing this challenge will require modeling approaches to effectively align and integrate all modalities present in and around the claim. For example, methods for pixel-based language modeling have recently been introduced to better align visually situated language with image content (Lee et al., 2022). Such approaches considering modalities beyond text and vision for multimodal data alignment can be useful for claim extracting from multimodal input.

Multimodal evidence retrieval. Evidence retrieval for audio and video fact-checking remains a major challenge. Different to other modalities, they cannot be easily searched on the web or social media networks (Silverman, 2013). Fact-checkers often use text accompanying the videos to find evidence (Silverman, 2013). Reverse image search engines, e.g. Google Lens or TinEye, require screenshots from the video as input – and thus require the correct timeframe, which can be challenging to extract. A dedicated adversary can render current tools very difficult to use. Very often evidence for image or audio fact-checking is retrieved using text accompanying them , e.g. metadata, social media comments, or captions (Gupta et al., 2013; Huh et al., 2018; Müller-Budack et al., 2020; Kopev et al., 2019). While incorporating the textual information and the other modality (e.g. audio/image) in retrieval would provide more information, this is missing currently. How to best retrieve evidence data that is non-textual or has a different modality than the claim, also remains a challenge.

Multilinguality and multimodality. While there is increasing work on multilingual fact-checking (Gupta and Srikumar, 2021; Shahi and Nandini, 2020; Hammouchi and Ghogho, 2022), it is mostly limited to text-only benchmarks and models. Surveying benchmarks for different pipeline stages (Figure 2), we found limited multimodal datasets for non-English languages (see Table 3). Previous work on multilingual multimodality shows that training and testing on English data alone introduces biases, as models fail to capture concepts and images prevalent in other languages and cultures (Liu et al., 2021). Moreover, some types of multimodal misinformation exploit cross-lingual sources to mislead, e.g. images or videos from non-English newspapers appearing as out-of-context data for English multimodal misinformation (Silverman, 2013). To prevent false conclusions and biases, it is thus necessary to take approaches that are both multimodal *and* multilingual (Ruder et al., 2022). Construction of large-scale multimodal, multilingual AFC datasets would facilitate futures research in this direction, similar to benchmarks and shared tasks created for automated fact-checking tasks in English (Thorne et al., 2018; Suryavardan et al., 2023a).

Generalizing detection of visual manipulations. The recent popularity of diffusion models (DMs) for visual manipulation have raised questions regarding the generalizability of manipulation detectors developed for earlier models (e.g.

GANs (Goodfellow et al., 2020)). Detection models are biased towards specific manipulation models and struggle to generalize (Wu et al., 2023a; Ricker et al., 2022). A recent study (Ricker et al., 2022) shows that detectors initially developed for GANs, have average performance drops of around 15% for image by DMs. While new detection approaches for DM manipulations are already being developed (Guarnera et al., 2023; Wu et al., 2023b), the question how to generalize and increase robustness of manipulation detectors for potential future manipulation models remains open. Potential solutions can include evidence-based approaches, where the manipulated content is used to retrieve evidence data (e.g. the original video or counterfactual evidence) to prove the manipulation.

Justifications for multimodal fact-checking. While explainable fact-checking has received attention recently (Kotonya and Toni, 2020b; Atanasova et al., 2020), there is limited work on producing justifications for multimodal content. Previous efforts on multimodal justification production have mostly focused on highlighting parts of the input to increase *interpretability* (Kou et al., 2020; Shang et al., 2022). Natural language justifications that explain the fact-check of multimodal claims so that it is accessible to non-technical have not been developed yet. To develop solutions, we first need appropriate benchmarks to measure progress. Moreover, with the recent advances of neural models for visual and audio generation and editing, another so far unexplored direction presents itself: editing input images/videos/audios or generating entirely content to explain fact-checking results. This could include, for example, the generation of infographics or video clips to explanation fact-checks. Such a system, especially if guided by human fact-checkers (Nakov et al., 2021), would be a potent tool. As noted in Lewandowsky et al. (2020), "well-designed graphs, videos, photos, and other semantic aids can be helpful to convey corrections involving complex or statistical information clearly and concisely".

5 Conclusion

We survey research on multimodal automated fact-checking and introduce a framework that combines and organizes tasks introduced in various communities studying misinformation. We discuss common terms and definitions in context of our framework. We further study popular benchmarks and modeling approaches, and discuss promising directions for future research.

Limitations

While we cite many datasets and modeling approaches for multimodal fact-checking, we describe most of them only briefly due to space constraints. Our aim was to provide an overview of multimodal fact-checking and organise previous works in a framework. Moreover, the presented survey focuses primarily on four modalities. While there are other modalities we could have included, we concentrated on those prevalent in real-world fact-checking that have not been discussed as part of a fact-checking framework in previous surveys.

Ethics Statement

As we mention in Section 4, most datasets for multimodal fact-checking tasks are available only in English. Thus, models are evaluated based on their performance on English benchmarks only. This can lead to a distorted view about advancements on multimodal automated fact-checking as it is limited to a single language out of more than 7000 world languages. While we call for future work on a variety of languages, this survey provides an overview on the state-of-the-art of mostly-English research efforts. Finally, we want to point out that multimodal fact-checking works we cite in this survey might include misleading statements or images given as examples.

Acknowledgements

Zhijiang Guo, Michael Schlichtkrull and Andreas Vlachos are supported by the ERC grant AVeriTeC (GA 865958). This paper is produced as part of the MuseIT project which has been co-funded by the EU under the Grant Agreement number 101061441. MuseIT has supported the work of Mubashara Akhtar. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency, REA. Neither the EU nor the granting authority can be held responsible for them.

References

Sara Abdali. 2022. Multi-modal misinformation detection: Approaches, challenges and opportunities. *CoRR*, abs/2203.13883.

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 (CVPR)*, pages 14940–14949.

Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. Protecting world leaders against deep fakes. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 38–45. Computer Vision Foundation / IEEE.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16, Seattle, United States. Association for Computational Linguistics.

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. 2019. Deepfake video detection through optical flow based CNN. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 1205–1207. IEEE.

Sabrine Amri, Dorsaf Sallami, and Esma Aïmeur. 2021. EXMULF: an explainable multimodal content-based fake news detection system. In *Foundations and Practice of Security - 14th International Symposium, FPS 2021, Paris, France, December 7-10, 2021, Revised Selected Papers*, volume 13291 of *Lecture Notes in Computer Science*, pages 177–187. Springer.

Shivangi Aneja, Christoph Bregler, and Matthias Nießner. 2021. Catching out-of-context misinformation with self-supervised learning. *CoRR*, abs/2101.06278.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, Andrea Galassi, Fatima Haouari, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, Gullal S. Cheema, Dilshod Azizov, and Preslav Nakov. 2023. The CLEF-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III*, volume 13982 of *Lecture Notes in Computer Science*, pages 506–517. Springer.

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

Christina Boididou, Symeon Papadopoulos, Yiannis Kompatsiaris, Steve Schifferes, and Nic Newman. 2014. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 743–748.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada, June 10-12, 2008*, pages 1247–1250. ACM.

Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2020. Video face manipulation detection through ensemble of cnns. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 5012–5019. IEEE.

Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *CoRR*, abs/2003.05096.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Stephen Cooper. 2007. A concise history of the fauxtography blogstorm in the 2006 lebanon war. *American Communication Journal*, 9.

Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. 2021. Id-reveal: Identity-aware deepfake video detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15088–15097. IEEE.

Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. 2020. On the detection of digital face manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5780–5789. Computer Vision Foundation / IEEE.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. 2019. The deepfake detection challenge (DFDC) preview dataset. *CoRR*, abs/1910.08854.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Florian Eyben, Felix Weninger, Florian Groß, and Björn W. Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, pages 835–838. ACM.

Florian Eyben, Martin Wöllmer, and Björn W. Schuller. 2009. Openear - introducing the munich open-source emotion and affect recognition toolkit. In *Affective Computing and Intelligent Interaction, Third International Conference and Workshops, ACII 2009, Amsterdam, The Netherlands, September 10-12, 2009, Proceedings*, pages 1–6. IEEE Computer Society.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *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)*, pages 1683–1698, Online. Association for Computational Linguistics.

Kiran Garimella and Dean Eckles. 2020. Images and misinformation in political groups: Evidence from whatsapp in india. *CoRR*, abs/2005.09784.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM*, 63(11):139–144.

D Graves. 2018. Understanding the promise and limits of automated fact-checking.

Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2023. Level up the deepfake detection: a method to effectively discriminate images generated by GAN architectures and diffusion models. *CoRR*, abs/2303.00608.

David Guera and Edward J. Delp. 2018. Deepfake video detection using recurrent neural networks. In *15th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2018, Auckland, New Zealand, November 27-30, 2018*, pages 1–6. IEEE.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394.

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

Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 729–736. International World Wide Web Conferences Steering Committee / ACM.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Vipin Gupta, Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2022. MMM: an emotion

and novelty-aware approach for multilingual multimodal misinformation detection. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Online only, November 20-23, 2022*, pages 464–477. Association for Computational Linguistics.

Michael Hameleers, Thomas E Powell, Toni GLA Van Der Meer, and Lieke Bos. 2020. A picture paints a thousand lies? the effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2):281–301.

Hicham Hammouchi and Mounir Ghogho. 2022. Evidence-aware multilingual fake news detection. *IEEE Access*, 10:116808–116818.

Naeemul Hassan, Bill Adair, James T Hamilton, Chengkai Li, Mark Tremayne, Jun Yang, and Cong Yu. 2015a. The quest to automate fact-checking. In *Proceedings of the 2015 computation+ journalism symposium*. Citeseer.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015b. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1835–1838. ACM.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Silvan Heller, Luca Rossetto, and Heiko Schuldt. 2018. The ps-battles dataset - an image collection for image manipulation detection. *CoRR*, abs/1804.04866.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Rui Hou, Verónica Pérez-Rosas, Stacy L. Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. In *International Conference on Multimodal Interaction, ICMI 2019, Suzhou, China, October 14-18, 2019*, pages 235–243. ACM.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 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 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2901–2912. ACM.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 106–124. Springer.

Cherilyn Ireton and Julie Posetti. 2018. *Journalism, fake news & disinformation: handbook for journalism education and training*. Unesco Publishing.

Omar Ismael Al-Sanjary, Ahmed Abdullah Ahmed, and Ghazali Sulong. 2016. Development of a video tampering dataset for forensic investigation. *Forensic Science International*, 266:565–572.

Ayush Jaiswal, Ekraam Sabir, Wael Abd-Almageed, and Premkumar Natarajan. 2017. Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1465–1471. ACM.

Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2886–2895. Computer Vision Foundation / IEEE.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 795–816. ACM.

Marvin Kalb and Carol Saivetz. 2007. The israeli—hezbollah war of 2006: The media as a weapon in asymmetrical conflict. *Harvard International Journal of Press/Politics*, 12(3):43–66.

Manvi Kamboj, Christian Hessler, Priyanka Asnani, Kais Riani, and Mohamed Abouelenien. 2021. Multimodal political deception detection. *IEEE Multim.*, 28(1):94–102.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. 2021. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2915–2921. ACM.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas W. D. Evans, Junichi Yamagishi, and Kong-Aik Lee. 2017. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2–6. ISCA.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Daniel Kopev, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Detecting deception in political debates using acoustic and textual features. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 652–659. IEEE.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Ziyi Kou, Daniel Yue Zhang, Lanyu Shang, and Dong Wang. 2020. Exfaux: A weakly supervised approach to explainable fauxtography detection. In *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*, pages 631–636. IEEE.

Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. 2021. Kodf: A large-scale korean deepfake detection dataset. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10724–10733. IEEE.

Tuan-Vinh La, Quang-Tien Tran, Thanh-Phuc Tran, Anh-Duy Tran, Duc-Tien Dang-Nguyen, and Minh-Son Dao. 2022. Multimodal cheapfakes detection by utilizing image captioning for global context. In *ICDAR@ICMR 2022: Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, Newark, NJ, USA, June 27 - 30, 2022*, pages 9–16. ACM.

Claire Lauer. 2009. Contending with terms: "multimodal" and "multimedia" in the academic and public spheres. *Computers and Composition*, 26(4):225–239.

Galina Lavrentyeva, Sergey Novoselov, Marina Volkova, Yuri Matveev, and Maria De Marsico. 2019. Phonespoof: A new dataset for spoofing attack detection in telephone channel. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 2572–2576. IEEE.

Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2022. Pix2struct: Screenshot parsing as pretraining for visual language understanding. *CoRR*, abs/2210.03347.

Stephan Lewandowsky, John Cook, and Doug Lombardi. 2020. Debunking Handbook 2020.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020a. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online. Association for Computational Linguistics.

Yiyi Li and Ying Xie. 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.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3204–3213. Computer Vision Foundation / IEEE.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *CoRR*, abs/2209.03430.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. 2023a. COVID-VTS: fact extraction and verification on short video platforms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 178–188. Association for Computational Linguistics.

Hui Liu, Wenya Wang, and Haoliang Li. 2023b. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9781–9796. Association for Computational Linguistics.

Vítor Lourenço and Aline Paes. 2022. A modality-level explainable framework for misinformation checking in social networks. *CoRR*, abs/2212.04272.

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.

Marie-Helen Maras and Alex Alexandrou. 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. *The International Journal of Evidence & Proof*, 23(3):255–262.

Alexandre Maros, Jussara M. Almeida, and Marisa Vasconcelos. 2021. A study of misinformation in audio messages shared in whatsapp groups. In *Disinformation in Open Online Media - Third Multidisciplinary International Symposium, MISDOOM 2021, Virtual Event, September 21-22, 2021, Proceedings*, volume 12887 of *Lecture Notes in Computer Science*, pages 85–100. Springer.

Priyanka Meel and Dinesh Kumar Vishwakarma. 2021. Han, image captioning, and forensics ensemble multimodal fake news detection. *Information Sciences*, 567:23–41.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Shreyash Mishra, Suryavardan S, Amrit Bhaskar, Parul Chopra, Aishwarya N. Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P. Sheth, and Asif Ekbal. 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) co-located with the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2022), Virtual Event, Vancouver, Canada, February 27, 2022*, volume 3199 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*, pages 16–25. ACM.

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 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6149–6157. European Language Resources Association.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 checkthat! lab on automatic identification and verification of political claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, volume 11018 of *Lecture Notes in Computer Science*, pages 372–387. Springer.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4551–4558. ijcai.org.

Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. 2023. Dfplatter: Multi-face heterogeneous deepfake dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9739–9748.

Eryn Newman and Lynn Zhang. 2020. *Truthiness: How Non-Probative Photos Shape Belief*.

Eryn J Newman, Maryanne Garry, Daniel M Bernstein, Justin Kantner, and D Stephen Lindsay. 2012. Nonprobative photographs (or words) inflate truthiness. *Psychonomic Bulletin & Review*, 19(5):969–974.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Dan Saattrup Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3141–3153. ACM.

Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: subject agnostic face swapping and reenactment. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7183–7192. IEEE.

Britt Paris and Joan Donovan. 2019. Deepfakes and cheap fakes. *United States of America: Data & Society*, 1.

Parth Patwa, Shreyash Mishra, Suryavardan S, Amrit Bhaskar, Parul Chopra, Aishwarya Reganti, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, and Chaitanya Ahuja. 2022. Benchmarking multimodal entailment for fact verification.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, pages 59–66. ACM.

Tarunima Prabhakar, Anushree Gupta, Kruttika Nadig, and Denny George. 2021. Check mate: Prioritizing user generated multi-media content for fact-checking. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 1025–1033. AAAI Press.

Christian Nathaniel Purwanto, Joan Santoso, Po-Ruey Lei, Hui-Kuo Yang, and Wen-Chih Peng. 2021. Fakeclip: Multimodal fake caption detection with mixed languages for explainable visualization. In *2021 International Conference on Technologies and Applications of Artificial Intelligence, TAAI 2021, Taichung, Taiwan, November 18-20, 2021*, pages 1–6. IEEE.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2022. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. *CoRR*, abs/2211.10973.

Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 518–527. IEEE.

Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. 2023. Two heads are better than one: Improving fake news video detection by correlating with neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11947–11959. Association for Computational Linguistics.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 153–162.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022a. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *CoRR*, abs/2205.12617.

Jingnong Qu, Liunian Harold Li, Jieyu Zhao, Sunipa Dev, and Kai-Wei Chang. 2022b. Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation. *CoRR*, abs/2205.12617.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ricardo Reimao and Vassilios Tzerpos. 2019. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue, SpeD 2019, Timisoara, Romania, October 10-12, 2019*, pages 1–10. IEEE.

Gustavo Resende, Philipe F. Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara M. Almeida, and Fabrício Benevenuto. 2019. (mis)information dissemination in whatsapp: Gathering, analyzing and countermeasures. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 818–828. ACM.

Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. 2022. Towards the detection of diffusion model deepfakes. *CoRR*, abs/2210.14571.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 10684–10695.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2018. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179.

Arjun Roy and Asif Ekbal. 2021. Mulcob-mulfav: Multimodal content based multilingual fact verification. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics.

Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep multimodal image-repurposing detection. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 1337–1345. ACM.

Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent convolutional strategies for face manipulation detection in videos. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 80–87. Computer Vision Foundation / IEEE.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *CoRR*, abs/2305.13117.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid- A multilingual cross-domain fact check news dataset for COVID-19. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020 Workshops, Atlanta, Georgia, USA [virtual], June 8, 2020*.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for COVID-19 short videos on tiktok. In *2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, December 15-18, 2021*, pages 899–908. IEEE.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2022. A duo-generative approach to explainable multimodal COVID-19 misinformation detection. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 3623–3631. ACM.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Craig Silverman. 2013. Verification handbook.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 1322–1331. AAAI Press.

Tiening Sun, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2023. Graph interactive network with adaptive gradient for multi-modal rumor detection. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, ICMR 2023, Thessaloniki, Greece, June 12-15, 2023*, pages 316–324. ACM.

S. Suryavardan, Shreyash Mishra, Megha Chakraborty, Parth Patwa, Anku Rani, Aman Chadha, Aishwarya Reganti, Amitava Das, Amit P. Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023a. Findings of factify 2: Multimodal fake news detection. *CoRR*, abs/2307.10475.

S Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit P. Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023b. Factify 2: A multimodal fake news and satire news dataset. *CoRR*, abs/2304.03897.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.

Reuben Tan, Bryan A. Plummer, and Kate Saenko. 2020. Detecting cross-modal inconsistency to defend against neural fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2081–2106. Association for Computational Linguistics.

Vidhu Tanwar and Kapil Sharma. 2020. Multi-model fake news detection based on concatenation of visual latent features. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 1344–1348.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. 2022. M2TR: multi-modal multi-scale transformers for deepfake detection. In *ICMR '22: International Conference on Multimedia Retrieval, Newark, NJ, USA, June 27 - 30, 2022*, pages 615–623. ACM.

Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. 2020. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1207–1216. ACM.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Deressa Wodajo and Solomon Atnafu. 2021. Deepfake video detection using convolutional vision transformer. *CoRR*, abs/2102.11126.

Haiwei Wu, Jiantao Zhou, and Shile Zhang. 2023a. Generalizable synthetic image detection via language-guided contrastive learning. *CoRR*, abs/2305.13800.

Xiaoshuai Wu, Xin Liao, and Bo Ou. 2023b. Sepmark: Deep separable watermarking for unified source tracing and deepfake detection. *CoRR*, abs/2305.06321.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569, Online. Association for Computational Linguistics.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2019. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9543–9552. Computer Vision Foundation / IEEE.

Zhizheng Wu, Tomi Kinnunen, Nicholas W. D. Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. 2015. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 2037–2041. ISCA.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2022. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *CoRR*, abs/2205.12487.

Jiangyan Yi, Ye Bai, Jianhua Tao, Haoxin Ma, Zhengkun Tian, Chenglong Wang, Tao Wang, and Ruibo Fu. 2021. Half-truth: A partially fake audio detection dataset. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 1654–1658. ISCA.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. 2019. Few-shot adversarial learning of realistic neural talking head models. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9458–9467. IEEE.

Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2015. Detecting image splicing in the wild (WEB). In *2015 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2015, Turin, Italy, June 29 - July 3, 2015*, pages 1–6. IEEE Computer Society.

Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Lang. Linguistics Compass*, 15(10).

Daniel Yue Zhang, Lanyu Shang, Biao Geng, Shuyue Lai, Ke Li, Hongmin Zhu, Md. Tanvir Al Amin, and Dong Wang. 2018. Fauxbuster: A content-free fauxtography detector using social media comments. In *IEEE International Conference on Big Data (IEEE BigData 2018), Seattle, WA, USA, December 10-13, 2018*, pages 891–900. IEEE.

Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. 2023. Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model. *CoRR*, abs/2304.07633.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. MFAN: multi-modal feature-enhanced attention networks for rumor detection. In *Proceedings of the Thirty-First*

International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, pages 2413–2419. ijcai.org.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. 2021. Exploring temporal coherence for more general video face forgery detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15024–15034. IEEE.

Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

Ruohan Zong, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Contrastfaux: Sparse semi-supervised fauxtography detection on the web using multi-view contrastive learning. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3994–4003. ACM.

A Methodology

We applied the following methodological approach to find and select relevant research papers for the survey.

First, after defining the research scope, we collected pivotal, highly-cited work (e.g. Nakamura et al. (2020)) and related surveys (e.g. (Alam et al., 2022)), resulting in 25 papers, as well as papers citing or cited by these works. We collected further works using the scholarly search engines Google Scholar[6], Semantic Scholar[7], DBLP[8] and ACL Anthology[9], and keyword-based search with Cartesian products of following keyword sets: {"fact checking", "fact verification", "misinformation", "disinformation", "fake news"}, {"multimodal", "text", "image", "audio", "video"}, and {"machine learning", "automated"}. The databases were queried primarily during the time frame July 26, 2022 and August 10, 2022. This step resulted in a collection of 123 papers.

[6] https://scholar.google.com/
[7] https://www.semanticscholar.org/
[8] https://dblp.org/
[9] https://aclanthology.org/

Figure 3: Example from the *FaceForensic* video manipulation dataset (Rössler et al., 2018) showing the manipulation generation approach.

Figure 4: An entry from the *MAIM* dataset (Jaiswal et al., 2017) showing an image/text claim with metadata.

We manually screened and filtered the papers based on abstracts and introduction sections, before creating an overview of papers across the following dimensions: (1) modality; (2) fact-checking task; (3) contribution type (i.e. dataset, modeling approach, demo); (4) paper type (i.e. survey, position paper, solution paper (e.g. introducing a new benchmark or modeling approach), or evaluation paper (e.g. investigating previously proposed approaches)). Papers were mostly excluded because they focused on other tasks than fact-checking (e.g. hate speech detection) or used modalities out of our scope (e.g. tables). Moreover, during the screening process we found and added further related works, and concluded the screening with 84 unique papers.

The taxonomy of tasks (Section 2) was created in an iterative manner starting with the task labels we assigned to works during screening. As a starting point we also used taxonomies of text-only fact-checking surveys (Guo et al., 2022; Thorne and Vlachos, 2018) and adapted them for multimodal fact-checking works.

B Examples: multimodal misinformation

Figure 5: An entry from the *Factify* dataset (Suryavardan et al., 2023b) depicting an image/text claim and supporting image/text evidence document.

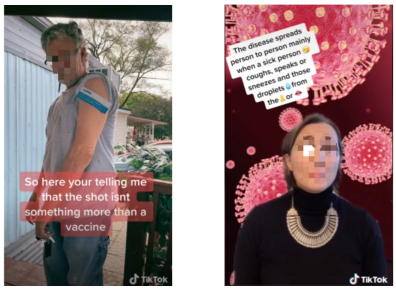

Figure 6: Left a misleading, right a non-misleading video screenshot from the Shang et al. (2021) dataset on COVID-19 TikTok Short Videos.