MR^2: A Benchmark for Multimodal Retrieval-Augmented Rumor Detection in Social Media

Xuming Hu
Tsinghua University
hxm19@mails.tsinghua.edu.cn

Zhijiang Guo*
University of Cambridge
zg283@cam.ac.uk

Junzhe Chen
Tsinghua University
chenjz20@mails.tsinghua.edu.cn

Lijie Wen*
Tsinghua University
wenlj@tsinghua.edu.cn

Philip S. Yu
University of Illinois at Chicago
psyu@cs.uic.edu

ABSTRACT

As social media platforms are evolving from text-based forums into multi-modal environments, the nature of misinformation in social media is also transforming accordingly. Misinformation spreaders have recently targeted contextual connections between the modalities e.g., text and image. However, existing datasets for rumor detection mainly focus on a single modality i.e., text. To bridge this gap, we construct MR^2, a multimodal multilingual retrieval-augmented dataset for rumor detection. The dataset covers rumors with images and texts , and provides evidence from both modalities that are retrieved from the Internet. Further, we develop established baselines and conduct a detailed analysis of the systems evaluated on the dataset. Extensive experiments show that MR^2 will provide a challenging testbed for developing rumor detection systems designed to retrieve and reason over social media posts. Source code and data are available at: https://github.com/THU-BPM/MR2.

CCS CONCEPTS

• **Computing methodologies** → **natural language processing**; **Knowledge representation and reasoning**.

KEYWORDS

Rumor Detection Benchmark, Social Media, Multimodal Retrieval-Augmented Methods

ACM Reference Format:
Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S. Yu. 2023. MR^2: 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), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3539618.3591896

1 INTRODUCTION

Nowadays, billions of multimodal posts containing texts, images, and videos are shared throughout the Internet, mainly via social media. Misleading rumors circulated in these platforms are not limited to one specific modality, despite the majority of existing efforts for rumor detection focusing on textual data [2, 42, 74]. It is more challenging to detect rumors presented in different modalities, as it necessitates the evaluation of each modality and the credibility of the combination [1, 6, 53]. For instance, consider the anti-vaccination tweet in Figure 1: the text reads "COVID vaccines do this", and an image of a dead person is attached. Although the image and text are not individually misinformative, combined they create misinformation.

Figure 1: Examples of Twitter rumors. The top post shows a *manipulated* photo of an orca in a city, suggesting it emerged in Tampa Bay on Thursday. The bottom *mismatched* post shows doctors with dead bodies. The text implies that the COVID vaccine is lethal. Metadata such as publication dates, retweet counts, and likes can help model the social context.

Datasets for rumor detection often focus on a single modality, such as text [10, 14, 35, 72, 76], thus missing crucial information conveyed by other modalities. There are a few multimodal datasets for rumor detection [8, 23, 42, 71] available, but these datasets are usually small or contain limited evidence such as metadata. Machine-generated texts or manipulated images can be detected based on their contents [63, 69]. However, identifying mismatched image-text pairs requires understanding across the same and different modalities. A rumor detection system that only models the

*Corresponding authors.

SIGIR '23July 23–27, 2023Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591896

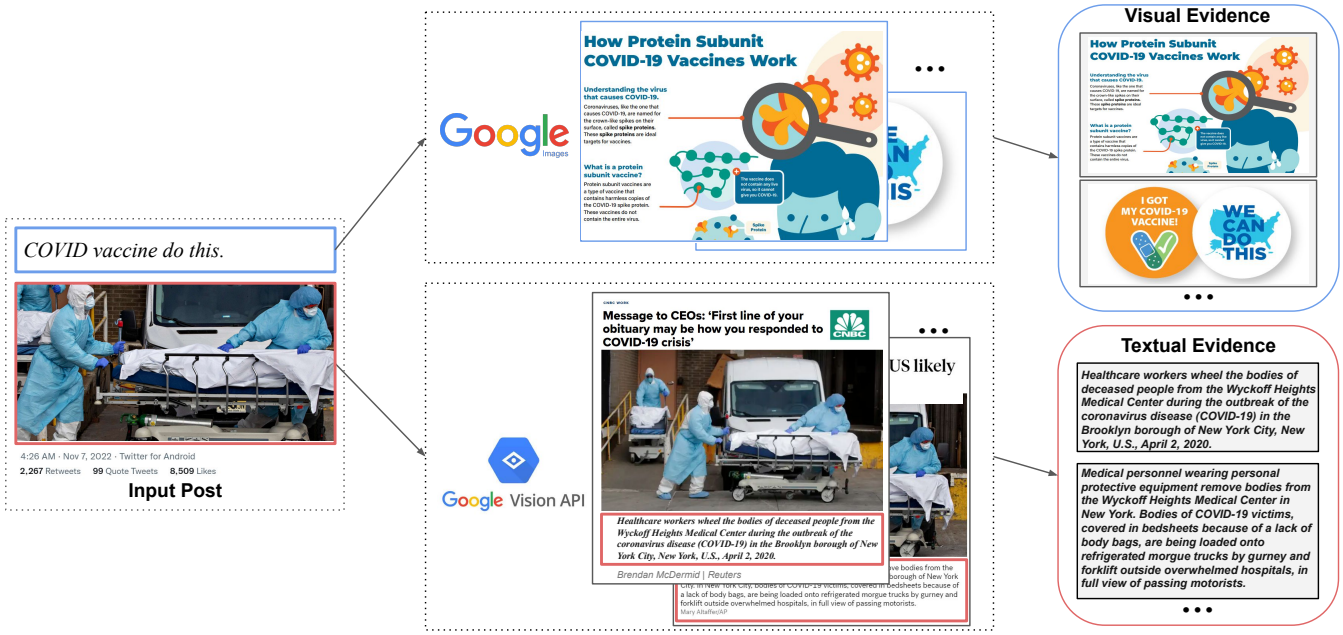


Figure 2: Overview of the proposed formulation. The image in the post is used to retrieve texts by using reverse image search, which returns similar images on the Internet. Descriptions of the top N images are viewed as textual evidence. Similarly, we use the text in the same post to retrieve top N images as visual evidence.

content of the post may not be able to determine its veracity, while incorporating additional metadata (e.g. number of reposts, comments) is helpful but not sufficient to provide grounding evidence for the post [74, 76].

To bridge this gap, we propose a novel approach that incorporates retrieved texts and images as evidence for better misinformation detection. As shown in Figure 2, we first use the image in the post to find its other occurrences via reverse image search. We then retrieve the textual evidence (i.e., descriptions) and compare it against the text in the post. Similarly, we use the text to find other images as visual evidence. Retrieving evidence from the Internet incorporates world knowledge that helps to detect well-presented misinformation. For example, the textual evidence in Figure 2 indicates that the image is mainly about the outbreak in New York, which has no connection to the COVID vaccines. The visual evidence illustrates how the vaccine functions rather than causing death. Our approach can be applied for early detection since retrieved texts and images do not rely on the proliferation process over time as social context (e.g., metadata, comments). To verify the effectiveness of the proposed formulation, we constructed a sizable multimodal dataset MR^2, which consists of 14,700 English and Chinese posts with textual and visual evidence retrieved from the Internet. To characterize the challenge of the dataset presented, we conducted a thorough analysis and demonstrated the utility of the dataset by developing established baselines. Our key contributions are summarized as:

- We incorporate the retrieved multimodal evidence for detection. Compared with prior efforts, ours is better resilient to mismatched rumors and amenable to early detection.
- We construct a sizable multimodal dataset that consists of 14,700 real-world English and Chinese posts with textual and visual evidence retrieved from the Internet.
- We develop established baselines and conduct a detailed analysis of the systems evaluated on the dataset, identifying challenges that need to be addressed in future research.

2 RELATED WORK

2.1 Rumor Detection Datasets

We review the existing rumor detection dataset as summarized in Table 1. As shown in the table, most existing datasets focus on rumors that only contain textual statements. Early efforts in rumor detection do not use any evidence beyond the textual content itself [62]. Content-based approaches that only rely on linguistic (lexical and syntactical) features are applied to capture deceptive cues or writing styles to detect rumors. However, many linguistic features are language-dependent, limiting the generality of these approaches. Recent developments in natural language generation have exacerbated this issue [5, 47], with machine-generated text sometimes being perceived as more trustworthy than human-written text [69]. Recently, some studies empirically show that rumor and non-rumor spread differently on social media, forming propagation patterns that could be harnessed for the rumor detection [36, 40, 67, 76]. Therefore, metadata is incorporated as propagation-based evidence, including post statistics (e.g. publication date, hashtag, URL) [35, 44], user demographics (such as age, gender, location, and education) [36, 39], social network structure (in the form of connections between users such as friendship or follower/followee relations) [31, 76], and user reactions (e.g. number of re-posts or likes) [41, 72]. Such propagation-based statistics can serve as indicators of rumorousness [54, 56, 77]. Apart from metadata, other datasets further incorporate a set of relevant posts (e.g. re-posts and replies of the source tweets) to model the propagation pattern for better detection [10, 14, 30, 50].

Table 1: Summary of rumor detection datasets. Input can be a textual statement or paired with an image. Evidence includes metadata and comments on the post (e.g. repost, reply). Our dataset includes visual and textual evidence from the search engine. Output is a multi-class label. ♣ denotes only a subset of the dataset has images.

Dataset	# Input	Inputs	Evidence	Output	Sources	Language
Suspicious [62]	Text	131,584	None	2/5 Classes	Twitter	En
COVIDLIES [20]	Text	6,761	Misconcepts	3 Classes	Twitter	En
Rumor-has-it [44]	Text	10,417	Metadata	2 Classes	Twitter	En
CredBank [39]	Text	1,049	Metadata	5 Classes	Twitter	En
Realtime [31]	Text	842	Metadata	2 Classes	Twitter	En
DualEmotion [72]	Text	6,362	Metadata	2 Classes	Weibo	Zh
Multidomain [41]	Text	9,128	Metadata	2 Classes	Weibo	Zh
Microblogs [35]	Text	5,656	Metadata/Comment	2 Classes	Weibo	Zh
PHEME [76]	Text	6,068	Metadata/Comment	2 Classes	Twitter	En/De
PropagtaionStruct [36]	Text	1,154	Metadata/Comment	2 Classes	Twitter	En
RumorEval17 [10]	Text	325	Metadata/Comment	3 Classes	Twitter	En
RumorEval19 [14]	Text	446	Metadata/Comment	3 Classes	Twitter/Reddit	En
DAST [30]	Text	220	Metadata/Comment	3 Classes	Reddit	Da
Stanker [50]	Text	5,802	Metadata/Comment	2 Classes	Twitter	Zh
Weibo [23]	Image/Text	9,528	Metadata	2 Classes	Weibo	Zh
FauxBuster [71]	Image/Text	917	Metadata/Comment	2 Classes	Twitter/Reddit	En
ExFaux [27]	Image/Text	263	Metadata/Comment	2/4 Classes	Twitter	En
MuMIN♣ [42]	Image/Text	984	Metadata/Comment	3 Classes	Twitter	En
MR^2	Image/Text	14,700	Image/Text/Metadata/Comment	3 Classes	Twitter/Weibo	En/Zh

Nowadays multimodal misinformation can be easily generated by using many neural network-based editing tools. For example, deepfakes can be used to manipulate or fabricate visual content [38, 58], such as editing the objects, replacing backgrounds, or changing captions. Recent efforts in realistic images and art generation from natural language descriptions have exacerbated this issue [21, 48, 49, 52]. However, existing efforts in multimodal rumor detection are limited, both in size and scope. For example, ExFaux [27] collects 263 image-based tweets, and FauxBuster [71] includes 917 posts from Twitter and Reddit. Though ModalFusion [23] includes 9,528 textual statements paired with images, it only contains metadata as the evidence for detection. Relying on the visual and textual contents of rumors without considering the state of the world may not be able to identify well-presented misinformation, such as mismatch rumors. On the other hand, while metadata offers information complementary to the textual content which is useful when the latter is unavailable, it does not provide sufficient evidence grounding the rumor [74, 76]. Compared to existing datasets, MR^2 further incorporates world knowledge by retrieving evidence from the Internet. Rumor detection systems trained on MR^2 can synthesize additional multimodal contexts for better detection. There exist datasets focused on fact-checking [3, 22], but A claim can be factual regardless of whether it is a rumour [16, 75], as it is based on language subjectivity and growth of readership [44].

2.2 Rumor Detection Models

We group existing rumor detection models into two categories: content-based methods and propagation-based methods. Content-based methods only rely on the content of the post itself, while propagation-based methods further incorporate social contexts (e.g. metadata, comments). Early content-based systems employ supervised classifiers with feature engineering, relying on surface features such as Reddit karma and up-votes [7, 43], Twitter-specific types [4, 17], named entities and verbal forms in political transcripts [78]. However, these studies don't consider the visual features that would be beneficial. Multi-modal data have been exploited by a set of studies to facilitate the detection. Jin et al. [23] employs recurrent neural networks to encode textual and visual features and fused them based on attention mechanism. Wang et al. [64] further incorporates event-invariant features by using the adversarial network to model Twitter events. In order to capture the relationships among multiple modalities, recent efforts explore various approaches to jointly learn the representations across modalities. Khattar et al. [24] leverages a variational auto-encoder to learn a shared representation of visual and textual contents. Qian et al. [45] adopts the hierarchical attention and Wu et al. [66] leverages the co-attention mechanism to fuse the multi-modal representations.

Propagation-based approaches based on sequence or graph modeling have recently become popular, as they allow models to use the context of surrounding social media activity to inform decisions. These approaches often exploit the ways in which information is discussed and shared by users, which are strong indicators of rumorousness [77]. Kockina et al. [26] uses an LSTM [19] to model branches of tweets, processing sequences of posts and outputting a label at each time step. Ma et al. [37] employs Tree-LSTMs [57] to directly encode the structure of threads, and Guo et al. [15] models the hierarchy by using attention networks. Lu and Li [33] learn the representations of user interactions and their correlation with source tweets. Recent works have explored fusing more domain-specific features into neural networks [72]. Graph Neural Networks [25] have also been adopted to model the propagation behavior of a potentially rumorous claim [29, 40, 67]. Although the propagation uncertainty between different nodes has been considered [65], multimodal fusion for graphs has yet to be explored. Recent work has thus proposed to perform multimodal alignment for better fusing the features [73]. While propagation-based approaches can be useful for rumor detection, they are not ideal for

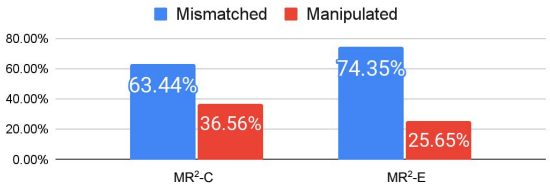


Figure 3: Distribution of mismatched and manipulated post in MR²-E and MR²-C.

early rumor detection. MR² can play a crucial role in early detection, since retrieved texts and images do not rely on propagation-based features. Rumor detection systems trained on MR² can be applied before misinformation spreads widely; this is also known as pre-bunking and has been shown to be more effective than post-hoc debunking [28, 51, 59].

3 DATASET CONSTRUCTION

MR² contains two rumor detection datasets in English and Chinese. The English dataset MR²-E is constructed by using posts from Twitter and the Chinese dataset MR²-C includes posts from Weibo.

3.1 Data Analysis

In order to investigate the effectiveness of the retrieval-based method, we conduct a human evaluation of retrieval-based, content-based, and propagation-based methods. We first analyze the distribution of multimodal rumors circulated on social media platforms by sampling 200 verified rumors from Twitter and Weibo, and categorize them into two types: mismatched and manipulated. Figure 3 show that most rumors spread over social media platforms are mismatched posts, with 63.44% from Weibo and 74.35% from Twitter being mismatched posts. This is likely due to the deceptive nature of mismatched posts, which are true if we only consider one modality (text or image). Then we ask the annotators to verify the posts based on the content, social context, and retrieved evidence. 500 rumors are randomly sampled from Weibo, and each annotator is given 50 rumors with their corresponding social contexts (i.e. comments) and retrieved visual and textual evidence from the Internet. The annotation team has 15 members, and 5 members are only involved in data validation. All annotators are native Chinese speakers. Annotators are required to answer (yes or no) if the content of the post, the social context, or the retrieved evidence provides sufficient information to predict the label of the post. To ensure the annotation quality, they are trained by the authors and go through several pilot annotations. We conduct an additional inter-annotator agreement and manual validation to ensure annotation consistency. For inter-annotator agreement, we randomly select 20% (n = 100) of rumors to be annotated by 5 annotators. We calculate the Fleiss K score [13] to be 0.78, which demonstrates that the annotation results are largely invariant through data validation [1].

We report the average results in Figure 4, only 69% of contents provide sufficient information to verify the claim. Our same analysis suggests that for 86% of the instances, using retrieved evidence provides sufficient details to determine the factuality. In the second phase, each annotator is given a post with different contexts and asked to infer the labels of 25 posts. Human predictions are more accurate when retrieval evidence is given (81% vs. 72%). However,

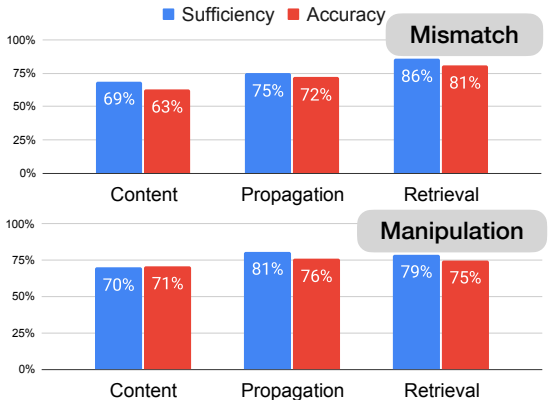


Figure 4: Comparison among information sufficiency and prediction accuracy when annotators are given the post (content), social context (propagation), retrieved evidence (retrieval) for mismatched / manipulated rumors from Weibo.

Table 2: Percent of claims from different fact-checking agencies. Percentage less than 1% are excluded.

Website	Percent	Website	Percent
factcheck.afp.com	11.26%	thequint.com	2.96%
politifact.com	9.17%	factcheck.org	2.89%
leadstories.com	7.89%	thelogicalindian.com	2.49%
boomlive.in	7.62%	usatoday.com	2.49%
checkyourfact.com	6.06%	vishvasnews.com	2.18%
altnews.in	5.60%	polygraph.info	1.71%
factly.in	5.13%	africacheck.org	1.66%
newsmeter.in	4.08%	dubawa.org	1.50%
newsmobile.in	3.82%	verafiles.org	1.48%
indiatoday.in	3.54%	pesacheck.org	1.16%
fullfact.org	3.33%	washingtonpost.com	1.02%
misbar.com	3.16%	healthfeedback.org	1.01%

we notice that the performance gap between propagation and retrieval is smaller in prediction accuracy. One reason is that the retrieved evidence contains more irrelevant information that may affect the prediction accuracy. Overall, our findings suggest that the retrieval-based method is better suited for mismatched posts, and is amenable to early detection when social context is missing. Additionally, if a rumor detection system is able to extract relevant information from retrieved evidence, it will gain benefits from contextual clues.

3.2 Posts Collection

3.2.1 English Dataset. Following the RumorEval shared tasks [10, 14], we adopt a three-way classification scheme: Rumors, Non-Rumors, and Unverified Posts as many posts spreading over social media cannot be verified with existing information [10, 36]. To collect rumors, we utilize the Google Fact Check Tools API[1], a resource that gathers verified claims from fact-checking agencies from around the world. We scrape 48,070 verified claims from active English fact-checking agencies until July 2022. Table 2 displays the distribution of their sources. From these claims, we manually review the corresponding fact-checking articles, and extract 6,287 tweets from them. Fact-checkers usually employ fine-grained labels

[1]https://toolbox.google.com/factcheck

Table 3: Mappings from the veracity labels (from fact-checking agencies) to our three standardized labels.

Standardized	Fact-checker Labels
Rumor	Altered, Altered Image, Altered Photo, Bad Math, Bad Science, Clickbait, Commontion, Death Hoax, Distorts the Facts, Doctored, FAKE, Fake News, Fake Quotes, Fake Quoto, Fake Tweet, False, False and Misleading, False Headline, False!, Flawed-Reasoning, Flip-flop, Full Flop, Hoax, Hoax!, Inaccurate, Incorrect, It's A Joke, Lacks Context, Likely False, Misattributed, Misleading, Misleading and False, Misleading!, Misleading., Misplaced Context, Misrepresented, Missing Context, Mixed, Mixture, Mostly False, Needs Context, Not Legit, Not the full story, Not the Whole Story, Out of Context, Pants on Fire, Partly False, Photoshopped, Selective, Spins the Facts, Staged Skit, Suspicious, This claim is False., This is misleading., This is not true., Three Pinocchios, Totally Fake, Totally False, Trolling, Two Pinocchios, Wrong, Wrong Number, Wrong Numbers
Non-Rumor	Accurate, Close to the mark, Correct, Mostly correct, Mostly True, Mostly-Accurate, No Fraud, Partially True, Partially-Correct, True
Unverified Post	Baseless, Myth, No Evidence, No Proof, Not Supported, This lacks evidence., Unclear, Unproven, Unsubstantiated, Unsupported

Table 4: Statistics of the MR^2 datasets.

Statistics	MR^2-C	MR^2-E
Source	Weibo	Twitter
Number of rumors	1,754	1,418
Number of non-rumors	2,609	2,318
Number of unverified rumors	3,361	3,240
Number of source posts	7,724	6,976
Number of threads	497,233	576,254
Number of users	47,319	60,573
Average length of texts	66.46	67.14
Average length of time	83	76
Average number of comments	64.38	82.61
Max number of comments	393	674

Table 5: Statistics of different labels of MR^2 datasets. "R", "NR" and "UP" denote rumor, non-rumor and unverified post, respectively. "A." denotes average and "M." denotes max. "com" means comments.

Stats	MR^2-C			MR^2-E		
	R	NR	UP	R	NR	UP
#threads	68,406	133,059	295,768	110,604	197,030	268,620
#users	13,342	15,237	18,740	16,681	19,370	24,522
A.text	67.44	76.16	58.42	69.22	56.49	73.85
A.time	76	86	88	73	82	69
A.com	39	51	48	78	85	83
M.com	233	335	393	566	674	579

to represent degrees of truthfulness (true, mostly-true, mixture, etc.). To unify these tweets in the three-way classification scheme, we design a mapping (as shown in Table 3) to standardize the original labels. The majority of fact-checked tweets are labeled as rumors due to journalists usually verifying claims that spread misinformation. The numbers of unverified tweets and non-rumors are limited (167 and 103, respectively). To compensate for this, we sample extra non-rumors and unverified posts that provoke a similar number of reposts [76]. The threshold is based on the median number of reposts. For the non-rumors, we use tweets from authoritative news agencies on Twitter, while for unverified posts, we gather the posts of general threads that are not reported as rumors in the same period. As shown in Table 5, rumors, non-rumors and unverified posts have similar number of threads and users.

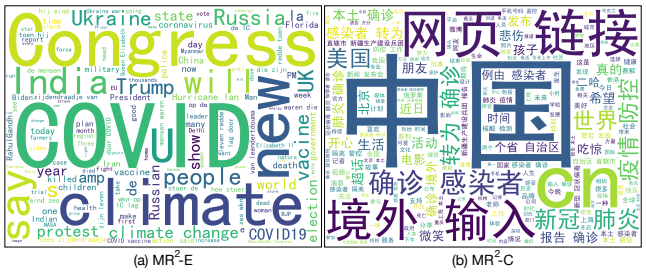
Aiming at building a multimodal dataset, we remove text-only posts. Next, we perform data deduplication to remove posts with similar textual and visual content to prevent data leakage in the training procedure. We remove duplicated images from the raw set with a near-duplicated image detection algorithm based on locality sensitive hashing [55]. Small or long images in terms of the resolution are also removed to maintain good quality. We identify posts concerning the same events based on one-pass text clustering[2] and make sure they are not contained in both training and testing sets. The training, development, and testing sets contain approximately a number of posts with a ratio of 8:1:1. Social contexts of posts are also included in the dataset. We gather the corresponding metadata, such as publication dates, number of reposts, replies and likes, names of users, locations, hashtags, and URLs. All comments to the post including reposts and replies are also collected. While Weibo does not provide an API endpoint to retrieve conversational threads provoked by source posts, it is possible to collect them by scraping posts through the web client interface. We develop a script that enable us to collect and store complete threads for all the source posts. As shown in Table 4, the resulting Twitter dataset consists of 1,418 rumors, 2,318 non-rumors, 3,240 unverified posts, and 576,254 threads in total.

3.2.2 Chinese Dataset. For the Chinese dataset, we obtain a set of verified rumors from March 2017 to July 2022 from the official rumor debunking center of Weibo[3]. Suspicious posts that provoke a high number of reposts are reported to the center, then auditors from Weibo would examine the posts and verify them as rumors or non-rumors. This system serves as an authoritative source to collect rumors in prior efforts [23, 35, 72]. We adopt a similar strategy to collect non-rumors and unverified posts as described in the construction of the English dataset. Table 5 provides more details, the resulting MR^2-C dataset consists of 1,754 rumors, 2,609 non-rumors, 3,361 unverified posts, and 497,233 threads in total.

Detailed statistics of MR^2-E and MR^2-C are presented in Table 4 and 5 respectively. Meanwhile, Figure 5 shows the word clouds of these two datasets. The datasets have a similar proportion to the three classes, with MR^2-C having fewer and shorter comments or threads. This is possible because Chinese social media have more non-political and non-scientific content. The analysis of word frequency revealed that MR^2-E mainly focuses on topics such as politics, public health, social news, and natural science, and MR^2-C is mainly about social news and public health. This indicates that the domains of rumor dissemination on Twitter and Weibo overlap.

[2]https://scikit-learn.org

[3]https://weibo.com/weibopiyao



(a) MR^2-E (b) MR^2-C

Figure 5: Comparison of word clouds on MR^2-E and MR^2-C. The words with the highest frequency in MR^2-C are mainly related to COVID-19, including China, Overseas, Infected, Covid, Web Page, Link.

3.3 Evidence Retrieval

3.3.1 Textual Evidence. We use the image of the input post as the query to retrieve textual evidence by using Google Reverse Image Search[4]. The Search Engine returns a list of images similar to the query image. In addition, the URLs of the web pages that contain these images are returned. We developed a web crawler to crawl the descriptions of the top 20 images based on the URLs. Concretely, the crawler first visits the web page and saves the title, then searches for the tag of the image using its URL and image content matching based on perceptual hashing. After locating the image, we are able to retrieve the description. We scrape the <figcaption> tag, as well as the <img> tag's textual attributes such as alt, image-alt, caption, data-caption, and title. From each page, we collect all the non-redundant text snippets that we found. The search engine returns up to 20 search results. We create a list of websites that spread misinformation based on the list of fake news websites by Wikipedia[5]. We filter out the web URLs that appear on this list. We further discard a page if the detected language of the title is non-English, using the fastText library[6] for language identification. In practice, we keep the descriptions from the top N=5 search results as textual evidence.

3.3.2 Visual Evidence. Next, we use the text of the input post as the query to retrieve visual evidence by using the Google Programmable Search Engine[7]. The Search Engine returns a list of images based on the query text. We saved the top 10 images retrieved by the search engine. We further filter out the images from the list of misinformation websites. In practice, we keep the top N=5 images as visual evidence after removing such images. It is important to note that, unlike the inverse image search, the search results here do not always correspond to the exact match of the textual query. Therefore, the visual evidence might be more loosely related to the image of the post. However, even if it is not exactly related to the input, it potentially provides implicit information to verify the post. Take the doctor with dead bodies in Figure 1 as an example, if the vaccine is causing death, top images returned by the search engine should reflect this event. However, the top results are images introducing how vaccine functions. Comparing the retrieved images against the image of the post is beneficial to reason over the overlap of regions and objects between the images for better prediction.

[4]https://cloud.google.com/vision
[5]https://en.wikipedia.org/wiki/List_of_fake_news_websites
[6]https://fasttext.cc/
[7]https://programmablesearchengine.google.com

4 BASELINE SYSTEMS

4.1 Retrieval-Based Model

In this section, we will present the baseline using retrieved texts and images as evidence for rumor detection. We define an instance $I = \{T, V, E_T^n, E_V^n\}$ as a tuple representing two different modalities of contents: the textual content T and the visual content V of the post, and a set of textual evidence E_T^n and visual evidence E_V^n, where n is the number of evidence. The proposed model consists of three components, including a textual encoder, visual encoder, and classifier.

4.1.1 Textual Encoder. This module aims to encode the texts of the post and retrieved evidence, then select relevant evidence based on the post. Given the textual content of a post and corresponding retrieved textual evidence. We use a context encoder to encode the sentences into contextualized representations. The context encoder can be a Transformer [60] or a BERT [11]. Here we use the BERT as an example. We first feed the text of the post T into BERT and use the representation of the CLS tokens as the textual post representation. Similarly, we feed the textual evidence E_T^n independently to the BERT, and concatenate the CLS of each piece of evidence as the textual evidence representation:

$$\begin{aligned} \boldsymbol{h}_t &= BERT(T), \\ \boldsymbol{e}_t^n &= BERT(E_T^n), \end{aligned} \tag{1}$$

$\boldsymbol{h}_t \in \mathbb{R}^d$ is the representation of the textual post, where d is the embedding size. $\boldsymbol{e}_t^n \in \mathbb{R}^{d \times n}$ is the representation of the textual evidence, where d is the embedding size and n is the number of textual evidence. After obtaining the representations of the textual post and evidence, we use the attention mechanism [11] to select relevant evidence for later prediction. The calculation involves a query and a set of key-value pairs. The output is computed as a weighted sum of the values, where the weight is computed by a function of the query with the corresponding key. Here we view the representation of the post $\boldsymbol{h}_t$ as the query and the representation of the evidence $\boldsymbol{e}_t$ as the key:

$$\boldsymbol{o}_t = softmax(\frac{\boldsymbol{h}_t \mathbf{W}_t \times (\boldsymbol{e}_t \mathbf{W}_e)^T}{\sqrt{d}})\boldsymbol{e}_t, \tag{2}$$

where $\mathbf{W}_t \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_e \in \mathbb{R}^{d \times d}$ are trainable projection matrices. $\boldsymbol{o}_t \in \mathbb{R}^{d \times n}$ is the output of the text encoder.

4.1.2 Visual Encoder. This module aims to encode the visual content of the post and retrieved evidence, then select relevant evidence based on the post. First, we represent the visual context by using an encoder, which can be a ResNet [18] or Vision Transformer [12], pretrained on the ImageNet dataset [9]. Here we use ResNet as an example. We feed the visual content of the post V and visual evidence E_V^n into the ResNet:

$$\begin{aligned} \boldsymbol{h}_v &= ResNet(V), \\ \boldsymbol{e}_v^n &= ResNet(E_V^n), \end{aligned} \tag{3}$$

$\boldsymbol{h}_v \in \mathbb{R}^d$ is the representation of the visual post, where d is the embedding size. $\boldsymbol{e}_v^n \in \mathbb{R}^{d \times n}$ is the representation of the visual

evidence, where n is the number of visual evidence. Similar to Eqn. 2, we compute the output of the enhanced visual representations as $\boldsymbol{o}_v \in \mathbb{R}^{d \times n}$.

4.1.3 Classifier. We concatenate the textual representation and visual representation to get the multimodal representation for classification. Then a feed-forward neural network (FFNN) is applied over the concatenated representations:

$$\boldsymbol{h}_{final} = \text{FFNN}([\boldsymbol{o}_t; \boldsymbol{o}_v]), \tag{4}$$

where h_{final} is then fed into a linear layer followed by a softmax operation to obtain a probability distribution over three labels.

4.2 Content-Based Model

It only relies on the textual and visual contents of the post itself. We adopted three baselines, including models that only use textual content, visual content, and both of them.

Text-Based Model. only considered the textual content of the post. BERT [11] and RoBERTa [32] are used as contextualized encoders to encode the textual content. Similar to the retrieved-based model, the representation of the `CLS` token is used for prediction.

Image-Based Model. only considered the visual content of the post. ResNet [18] and Vision Transformer [12] are used as visual encoders. Both are pretrained on ImageNet [9]. The output representation of the visual encoder is used for classification.

Multi-Modal Model. encodes both the textual and visual contents of the input post. We include two baselines: MVAE [24] and CLIP [46]. MVAE reconstructs both modalities from the shared multi-modal representation based on a variational auto-encoder. CLIP is an image-language pretrained model. We pass the image and text of the post to CLIP and normalize their representations. A joint representation is produced by using a dot product of the output visual and textual representations. The resulting joint representation is then used for prediction.

4.3 Propagation-Based Model

It models the propagation structure of the input post based on its social contexts. We include tree-based, graph-based, and modal-fusion models.

Tree-Based Model. models the information flow from the input post and its comments as a tree structure to capture complex propagation patterns. We included two competitive rumor detection models as baselines, including Tree-RvNN [37] and Tree-Transformer [34]. Tree-RvNN defines a top-down tree, feature vectors of posts are generated based on their propagation paths. A tree-lstm [57] is employed to directly encode the tree, then a max-pooling layer is applied over the leaf posts to form the final representation. Tree-Transformer leverages the attention mechanism [60] to select important reply posts and combine the representation of the top-down and bottom-up trees.

Graph-Based Model. views the propagation structure as the graph and aggregates the information via a message-passing scheme. We included two graph-based models: GLAN [68] and EBGCN [65]. GLAN constructs a heterogeneous graph to capture the interactions among the input post, reposts, and users. Then graph attention networks [61] is used to get the representations for classification. ENGCN formulates the propagation structure as a top-down propagation graph and a bottom-up dispersion graph. The graph embedding is generated for prediction by using the edge-weighted graph convolutional networks [25].

Modal-Fusion Model. combines the information from the content of the post and its social context. We adopted two baselines: Att-RNN and MFAN. Att-RNN [23] employed recurrent neural networks to encode textual and visual features and fused them together with metadata based on the attention mechanism. MFAN [73] jointly uses textual, visual, and social graph features, involving multi-modal alignment for better fusion, and utilizing potential relationships to enhance the graph features [73].

5 EXPERIMENTS AND ANALYSES

5.1 Experimental Setup

Following Jin et al. [23] and Zhang et al. [70], we computed the Accuracy and Macro F1 as the evaluation metric. The hyper-parameters are chosen based on the development set. Results are reported with mean and standard deviation based on 5 runs. For the textual encoder of the retrieval-based model, we use the BERT-Base default tokenizer with a max-length of 256 to preprocess data. For the visual encoder of the retrieval-based model, we use ResNet 152 to encode the visual images. We scale the image proportionally so that the short side is 256, and crop the center to $224 * 224$. For the feed-forward neural network of the classifier, we set the layer dimensions as h_R-1024-verification_labels, where $h_R = 768 * 5 + 2048 * 5$. We use BertAdam [11] with $9e{-}6$ learning rate, warmup with 0.1 to optimize the cross-entropy loss and set the batch size as 16.

5.2 Main Results

Content-based Method. Table 6 shows that multimodal baselines consistently outperform text-based and image-based baselines on both datasets. Results from CLIP demonstrate the highest F1 score of 81.46% on MR^2-E and 81.54% on MR^2-C, indicating that solely relying on text or visual modality cannot accurately identify misinformation. Furthermore, learning a good shared representation between modalities is necessary for identifying multimodal rumors. Generally, text-based baselines perform better than image-based counterparts, due to the higher ratio of misinformation presented in the textual modality. The proposed retrieval-based model achieved an F1 score of 85.34% on MR^2-E and 85.03% on MR^2-C, which is significantly higher than the content-based systems. This implies that retrieving multimodal evidence from the internet provides world knowledge to help detect well-presented misinformation. The retrieval-based model has less performance decline on MR^2-C, which contains more challenging mismatched posts, as additional evidence is better resilient to mismatched cases.

Propagation-based Method. MFAN outperforms tree-based and graph-based methods that mainly focus on modeling the input text and its surrounding social contexts, since MFAN fuses information from different modalities (text, image, and social contexts). Compared to the best content-based model CLIP, MFAN achieves better results on both MR^2-E (81.56% v.s. 81.40%) and MR^2-C (82.33% v.s. 80.54%), which suggests that incorporating social contexts improves

Table 6: Results of baseline systems on Twitter and Sina Weibo. † means we replace the text encoder with BERT and the image encoder with ResNet for a fair comparison.

Baseline Systems			MR²-E				MR²-C			
			Accuracy	Precision	Recall	Macro F1	Accuracy	Precision	Recall	Macro F1
Content-Based	Text Based	BERT	78.34±1.23	78.69±1.11	77.02±1.01	77.57±1.13	82.42±1.07	79.21±0.98	77.45±0.86	78.05±1.18
		RoBERTa	78.22±1.05	78.26±0.92	78.87±1.07	78.56±0.92	83.64±0.88	80.43±0.95	78.77±0.89	78.56±0.94
	Image Based	ResNet	74.83±1.07	76.22±1.23	74.37±1.02	75.12±1.23	72.33±1.10	65.16±1.25	62.13±1.18	63.39±1.13
		Vision Transformer	72.14±1.29	74.33±1.28	72.29±1.36	73.28±1.33	70.75±1.26	65.24±1.27	64.35±1.21	64.89±1.23
	Multi Modal	MVAE†	79.04±1.76	79.12±1.52	78.02±1.61	78.33±1.57	82.14±1.68	79.64±1.47	78.36±1.59	78.95±1.53
		CLIP	81.29±1.22	82.05±1.27	81.06±1.32	81.40±1.22	83.87±1.25	81.77±1.31	81.29±1.23	80.54±1.29
Propagation-Based	Tree Based	Tree-RvNN	70.53±1.63	69.74±1.54	65.06±1.38	67.25±1.45	73.66±1.54	70.13±1.63	66.75±1.51	68.85±1.58
		Tree-Transformer	78.92±1.66	80.13±1.66	76.57±1.59	78.41±1.60	82.88±1.67	79.34±1.62	77.97±1.69	78.90±1.66
	Graph Based	GLAN	72.39±1.68	71.78±1.79	69.47±1.72	70.53±1.77	75.53±1.68	74.24±1.77	70.95±1.58	72.76±1.66
		EBGCN	79.45±1.51	79.03±1.57	78.44±1.52	78.64±1.55	82.53±1.45	80.13±1.77	78.32±1.52	79.31±1.64
	Modal Fusion	Att-RNN	72.46±1.32	71.33±1.31	65.92±1.24	68.93±1.28	75.84±1.28	73.05±1.32	69.93±1.22	71.22±1.27
		MFAN	82.32±1.51	82.22±1.58	81.03±1.55	81.56±1.57	83.98±1.43	83.04±1.52	81.89±1.57	82.33±1.54
Retrieval-Based			**85.62±1.32**	**85.63±1.38**	**84.98±1.33**	**85.34±1.35**	**85.95±1.38**	**85.25±1.41**	**84.77±1.34**	**85.03±1.39**
Retrieval-Based *w/o Visual Evidence*			83.36±1.12	83.67±1.09	82.96±1.27	83.32±1.18	84.13±1.08	82.94±1.02	82.58±1.17	82.79±1.13
Retrieval-Based *w/o Textual Evidence*			82.87±1.28	83.13±1.26	82.58±1.32	82.90±1.29	83.55±1.41	82.53±1.53	82.04±1.19	82.31±1.33
Retrieval-Based *w/o Both Evidence*			78.48±1.22	79.46±1.04	78.88±1.37	79.24±1.19	80.05±1.18	79.64±1.33	79.16±1.15	79.40±1.23

the performance of the multimodal model. It was also observed that propagation-based methods do not suffer the same performance decrease on MR²-C as content-based systems. This might be due to the fact that social contexts provide more information to identify mismatched posts. However, content-based models have higher robustness than other models, likely because irrelevant information is also presented in social contexts. The retrieval-based model achieved higher accuracy and F1 score than MFAN, further proving the effectiveness of retrieving multimodal evidence. Moreover, the standard deviation of the retrieval-based model was lower than propagation-based models and comparable to content-based methods, showing that retrieved evidence maintains a better balance between relevant and irrelevant information.

5.3 Analyses and Discussions

Ablation Study. We conduct an ablation study to show the effectiveness of different modules of the retrieval-based model on the test set. Retrieval-based model *w/o Visual Evidence*, Retrieval-based model *w/o Textual Evidence*, and Retrieval-based model *w/o Both Evidence* mean that textual, visual and all evidences are removed from the retrieved evidence, and only the remaining evidence and the original post are used to detect whether it is a rumor or not. The results from Table 6 demonstrate that the two types of evidence all contributes positively to the performance and can bring 4.13% (Visual Evidence), 4.58% (Textual Evidence), and 5.87% (Both Evidence) Macro F1 improvements, respectively. Among them, text evidence obtained from image retrieval brings better benefits, which may be related to the fact that textual evidence usually describes the objects in the image, and the relationship between objects can often explain whether the original post is a rumor.

Effects of Evidence. In Figure 7, we vary the numbers of retrieved visual and textual evidence from 1 ∼ 10 and report the Macro F1 on the test set of MR²-E and MR²-C. The fluctuation results indicate that both the quantity and quality of retrieved visual and textual

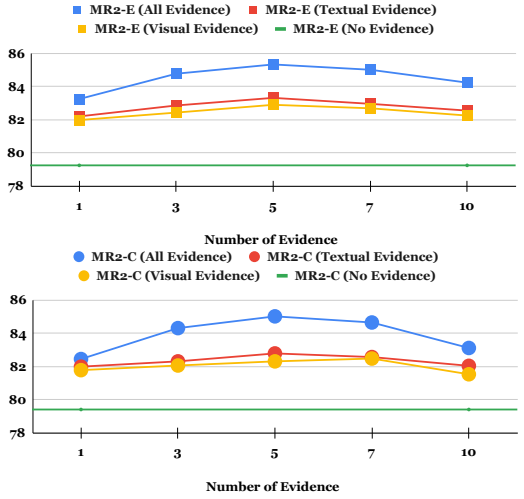


Figure 6: The effectiveness of retrieved visual and textual evidence on MR²-E and MR²-C.

evidence affect the performance. Using insufficient textual or visual evidence will give the original post less explanatory information, thus affecting the effect of the model on rumor detection. Using too much textual or visual evidence will introduce irrelevant or erroneous noise and affect the performance of the model. From Figure 7, we can observe that the damage to model performance caused by introducing more evidence is slightly less than the impact of insufficient evidence. A direction that can be studied on our benchmark in the future is how to screen a large amount of noisy evidence to obtain more informative and effective evidence.

No matter how much evidence is employed, our method consistently outperforms the baseline model: No Evidence, which shows the effectiveness of adding evidence. In our model, we adopt 5 textual and visual evidence for each post to achieve the best performance. Another interesting finding is that adding textual evidence

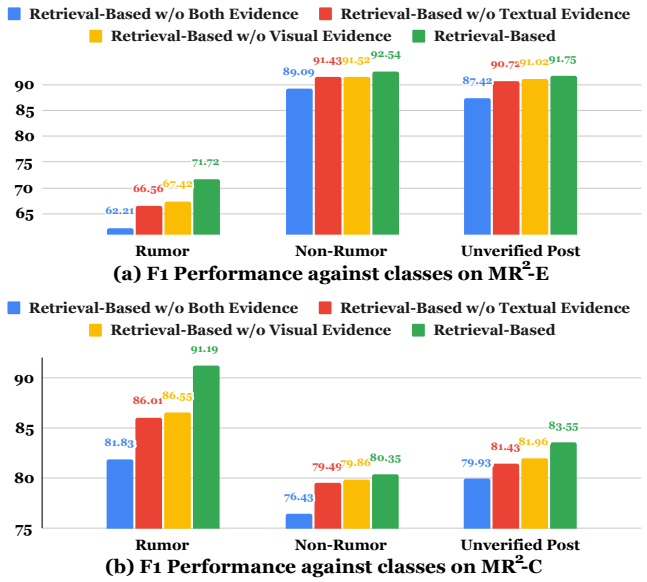


Figure 7: The F1 performance against classes on MR^2-E/ MR^2-C.

works better than adding visual evidence, which may be related to the fact that textual evidence retrieval is easier to obtain high-quality results than visual evidence retrieval. Therefore, how to improve the retrieval ability of visual evidence is also a research direction in the future.

Performance against Classes. We study the F1 performance changes against classes after adding retrieval evidence in Figure 7. A general conclusion is that all classes can obtain an average 5.70% improvement in F1 performance from retrieved evidence, among which textual evidence has a greater improvement than visual evidence (3.57% vs. 3.02%). An interesting finding is that the F1 performance of "Rumor" class that performs poorly in MR^2-E and that performs better in MR^2-C improves more, which fully shows that a gap that affects the performance of the "Rumor" class is the amount of information in the retrieved evidence, and more sufficient auxiliary evidence can lead to more improvements.

Case Study. We give the case study in Figure 8. For the original post 1 from Twitter, since the text of the original post is similar to the semantics expressed in the image, and the language style of the post is similar to the news post, the content-based systems will recognize it as "Non-Rumor". Propagation-based systems will also be identified as "Non-Rumor" based on the original post because no useful information is given in the comments. According to the "False" annotations in the retrieved visual evidence and the "Fabricated" and "No" content in the textual evidence, retrieval-based systems can correctly predict the post as "Rumor". For the original post 2 from Sina Weibo, the style of the original text of this post is close to the news expression, and the comments are also related to the original post. According to the evidence retrieved in the web, it can be mutually confirmed with the original post, so all systems incorrectly predict it as "Non-Rumor". However, all systems ignore the actual casualties in the original post are still an unconfirmed matter, so the original post should be labeled as "Unverified Post".

Figure 8: Case study on MR^2-E and MR^2-C.

6 CONCLUSION AND FUTURE WORK

In this paper, we explore multimodal rumor detection using visual and textual evidence. Our approach is more resistant to image manipulation and allows early detection compared to content-based methods, as it doesn't depend on propagation-based features. We create large multimodal datasets (MR^2) and develop baseline models, expecting MR^2 to be a stimulating challenge for rumor detection.

This proposed approach achieves competitive benchmark results but faces three main challenges. First, using search engines provides related knowledge but introduces noisy evidence, potentially misleading model predictions. Second, not all evidence is trustworthy, and reliable sources may contradict, posing difficulties for machine learning systems, including ours. Obtaining reliable, relevant evidence is a crucial future research direction. Lastly, evidence can be found in other modalities beyond texts and images, such as tables, info lists, knowledge graphs, videos, and audio. Human experts can extract information from these diverse sources, but our system only handles textual and visual evidence. Incorporating multi-modal evidence is another important future research direction.

ACKNOWLEDGMENTS

The work was supported by the National Key Research and Development Program of China (No. 2019YFB1704003), the National Nature Science Foundation of China (No. 62021002), NSF under grants III1909323, Tsinghua BNRist and Beijing Key Laboratory of Industrial Bigdata System and Application.

REFERENCES

[1] Sara Abdali. 2022. Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. *CoRR* abs/2203.13883 (2022). https://doi.org/10.48550/arXiv.2203.13883 arXiv:2203.13883

[2] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A Survey on Multimodal Disinformation Detection. *arXiv preprint arXiv:2103.12541* (2021).

[3] 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*, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/68d30a9594728bc39aa24be94b319d21-Abstract-round1.html

[4] 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*. 743–748.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[6] 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 (2020). arXiv:2003.05096 https://arxiv.org/abs/2003.05096

[7] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar (Eds.). ACM, 675–684. https://doi.org/10.1145/1963405.1963500

[8] 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, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 962–979. https://doi.org/10.18653/v1/2022.findings-naacl.72

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[10] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 69–76. https://doi.org/10.18653/v1/S17-2006

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[12] 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. https://openreview.net/forum?id=YicbFdNTTy

[13] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[14] Genevieve Gorrell, Ahmet Aker, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2019. SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, Jonathan May, Ekaterina Shutova, Aurélie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad (Eds.). Association for Computational Linguistics, 845–854. https://doi.org/10.18653/v1/s19-2147

[15] Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. 2018. Rumor Detection with Hierarchical Social Attention Network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (Eds.). ACM, 943–951. https://doi.org/10.1145/3269206.3271709

[16] Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Trans. Assoc. Comput. Linguistics* 10 (2022), 178–206. https://doi.org/10.1162/tacl_a_00454

[17] 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*, Leslie Carr, Alberto H. F. Laender, Bernadette Farias Lóscio, Irwin King, Marcus Fontoura, Denny Vrandecic, Lora Aroyo, José Palazzo M. de Oliveira, Fernanda Lima, and Erik Wilde (Eds.). International World Wide Web Conferences Steering Committee / ACM, 729–736. https://doi.org/10.1145/2487788.2488033

[18] 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 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[20] Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.11

[21] Xuming Hu, Zhijiang Guo, Yu Fu, Lijie Wen, and Philip S. Yu. 2022. Scene Graph Modification as Incremental Structure Expanding. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*. International Committee on Computational Linguistics, 5707–5720. https://aclanthology.org/2022.coling-1.502

[22] Xuming Hu, Zhijiang Guo, Guanyu Wu, Aiwei Liu, Lijie Wen, and Philip S. Yu. 2022. CHEF: A Pilot Chinese Dataset for Evidence-Based Fact-Checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 3362–3376. https://doi.org/10.18653/v1/2022.naacl-main.246

[23] 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 25th ACM international conference on Multimedia*. 795–816.

[24] 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*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 2915–2921. https://doi.org/10.1145/3308558.3313552

[25] 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. https://openreview.net/forum?id=SJU4ayYgl

[26] Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 475–480. https://doi.org/10.18653/v1/S17-2083

[27] 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*, Xintao Wu, Chris Jermaine, Li Xiong, Xiaohua Hu, Olivera Kotevska, Siyuan Lu, Weija Xu, Srinivas Aluru, Chengxiang Zhai, Eyhab Al-Masri, Zhiyuan Chen, and Jeff Saltz (Eds.). IEEE, 631–636. https://doi.org/10.1109/BigData50022.2020.9378019

[28] Stephan Lewandowsky and Sander van der Linden. 2021. Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology* 0, 0 (2021), 1–38. https://doi.org/10.1080/10463283.2021.1876983 arXiv:https://doi.org/10.1080/10463283.2021.1876983

[29] Jiawen Li, Yudianto Sujana, and Hung-Yu Kao. 2020. Exploiting Microblog Conversation Structures to Detect Rumors. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 5420–5429. https://doi.org/10.18653/v1/2020.coling-main.473

[30] Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint Rumour Stance and Veracity Prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Linköping University Electronic Press, Turku, Finland, 208–221. https://www.aclweb.org/anthology/W19-6122

[31] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM international on conference on information and knowledge management*. 1867–1870.

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[33] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648* (2020).

[34] Jing Ma and Wei Gao. 2020. Debunking Rumors on Twitter with Tree Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 5455–5466. https://doi.org/10.18653/v1/2020.coling-main.476

[35] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting Rumors from Microblogs with Recurrent Neural Networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, Subbarao Kambhampati (Ed.). IJCAI/AAAI Press, 3818–3824. http://www.ijcai.org/Abstract/16/537

[36] Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

[37] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor Detection on Twitter with Tree-structured Recursive Neural Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1980–1989. https://doi.org/10.18653/v1/P18-1184

[38] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)* 54, 1 (2021), 1–41.

[39] Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, Meeyoung Cha, Cecilia Mascolo, and Christian Sandvig (Eds.). AAAI Press, 258–267. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10582

[40] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake News Detection on Social Media using Geometric Deep Learning. *CoRR* abs/1902.06673 (2019). arXiv:1902.06673 http://arxiv.org/abs/1902.06673

[41] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3343–3347.

[42] 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*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 3141–3153. https://doi.org/10.1145/3477495.3531744

[43] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility Assessment of Textual Claims on the Web. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, Snehasis Mukhopadhyay, ChengXiang Zhai, Elisa Bertino, Fabio Crestani, Javed Mostafa, Jie Tang, Luo Si, Xiaofang Zhou, Yi Chang, Yunyao Li, and Parikshit Sondhi (Eds.). ACM, 2173–2178. https://doi.org/10.1145/2983323.2983661

[44] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying Misinformation in Microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., 1589–1599. https://www.aclweb.org/anthology/D11-1147

[45] 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*. 153–162.

[46] 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 (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[48] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR* abs/2204.06125 (2022). https://doi.org/10.48550/arXiv.2204.06125 arXiv:2204.06125

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. http://proceedings.mlr.press/v139/ramesh21a.html

[50] Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. STANKER: Stacking Network based on Level-grained Attention-masked BERT for Rumor Detection on Social Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 3347–3363. https://doi.org/10.18653/v1/2021.emnlp-main.269

[51] Jon Roozenbeek, Sander van der Linden, and Thomas Nygren. 2020. Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *The Harvard Kennedy School Misinformation Review* 1, 2 (2020). arXiv:https://doi.org/10.37016/mr-2020-008

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR* abs/2205.11487 (2022). https://doi.org/10.48550/arXiv.2205.11487 arXiv:2205.11487

[53] Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Y. Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and Understanding Harmful Memes: A Survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, Luc De Raedt (Ed.). ijcai.org, 5597–5606. https://doi.org/10.24963/ijcai.2022/781

[54] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor.* 19, 1 (2017), 22–36. https://doi.org/10.1145/3137597.3137600

[55] Malcolm Slaney and Michael A. Casey. 2008. Locality-Sensitive Hashing for Finding Nearest Neighbors [Lecture Notes]. *IEEE Signal Process. Mag.* 25, 2 (2008), 128–131. https://doi.org/10.1109/MSP.2007.914237

[56] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some Like it Hoax: Automated Fake News Detection in Social Networks. *CoRR* abs/1704.07506 (2017). arXiv:1704.07506 http://arxiv.org/abs/1704.07506

[57] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 1556–1566. https://doi.org/10.3115/v1/P15-1150

[58] Quang-Tien Tran, Thanh-Phuc Tran, Minh-Son Dao, Tuan-Vinh La, Anh-Duy Tran, and Duc Tien Dang Nguyen. 2022. A Textual-Visual-Entailment-based Unsupervised Algorithm for Cheapfake Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7145–7149.

[59] Sander van der Linden, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach. 2017. Inoculating the Public against Misinformation about Climate Change. *Global Challenges* 1, 2 (2017), 1600008. https://doi.org/10.1002/gch2.201600008 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/gch2.201600008

[60] 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*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[61] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=rJXMpikCZ

[62] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 647–653. https://doi.org/10.18653/v1/P17-2102

[63] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 8692–8701. https://doi.org/10.1109/CVPR42600.2020.00872

[64] 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*, Yike Guo and Faisal Farooq (Eds.). ACM, 849–857. https://doi.org/10.1145/3219819.3219903

[65] Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. 2021. Towards Propagation Uncertainty: Edge-enhanced Bayesian Graph Convolutional Networks for Rumor Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3845–3854. https://doi.org/10.18653/v1/2021.acl-long.297

[66] 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*. 2560–2569.

[67] Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2020. Rumor Detection on Social Media with Graph Structured Adversarial Learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, Christian Bessiere (Ed.). ijcai.org, 1417–1423. https://doi.org/10.24963/ijcai.2020/197

[68] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly Embedding the Local and Global Relations of Heterogeneous Graph for Rumor Detection. In *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, Jianyong Wang, Kyuseok Shim, and Xindong Wu (Eds.). IEEE, 796–805. https://doi.org/10.1109/ICDM.2019.00090

[69] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 9051–9062. https://proceedings.neurips.cc/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html

[70] Amy X. Zhang, Aditya Ranganathan, S. Metz, S. Appling, Connie Moon Sehat, Norman Gilmore, Nick B. Adams, E. Vincent, J. Lee, Martin Robbins, Ed Bice, Sandro Hawke, D. Karger, and An Xiao Mina. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. *Companion Proceedings of the The Web Conference 2018* (2018).

[71] 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, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*, Naoki Abe, Huan Liu, Calton Pu, Xiaohua Hu, Nesreen K. Ahmed, Mu Qiao, Yang Song, Donald Kossmann, Bing Liu, Kisung Lee, Jiliang Tang, Jingrui He, and Jeffrey S. Saltz (Eds.). IEEE, 891–900. https://doi.org/10.1109/BigData.2018.8622344

[72] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining Dual Emotion for Fake News Detection. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 3465–3476. https://doi.org/10.1145/3442381.3450004

[73] 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*, Luc De Raedt (Ed.). ijcai.org, 2413–2419. https://doi.org/10.24963/ijcai.2022/335

[74] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Comput. Surv.* 51, 2 (2018), 32:1–32:36. https://doi.org/10.1145/3161603

[75] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Comput. Surv.* 51, 2 (2018), 32:1–32:36. https://doi.org/10.1145/3161603

[76] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *International conference on social informatics*. Springer, 109–123.

[77] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11, 3 (2016), e0150989.

[78] Chaoyuan Zuo, Ayla Karakas, and Ritwik Banerjee. 2018. A Hybrid Recognition System for Check-worthy Claims Using Heuristics and Supervised Learning. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018 (CEUR Workshop Proceedings, Vol. 2125)*, Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2125/paper_143.pdf