arXiv:2203.13883v3 [cs.LG] 26 Jul 2022

Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities

SARA ABDALI*, Georgia Institute of Technology

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. Taking advantage of the fact that visual modalities such as images and videos are more favorable and attractive to the users and textual contents are sometimes skimmed carelessly, misinformation spreaders have recently targeted contextual connections between the modalities e.g., text and image. Thus, many research efforts have been put into development of automatic techniques for detecting possible cross-modal discordance in web-based content. In this work, we aim to analyze, categorize and identify existing approaches in addition to challenges and shortcomings they face in order to unearth new opportunities in furthering the research in the field of multi-modal misinformation detection.

Additional Key Words and Phrases: Misinformation Detection, Multi-modal Learning, Fake News Detection, Survey, Multi-modal Datasets

ACM Reference format:
Sara Abdali. 2016. Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities. 1, 1, Article 1 (January 2016), 36 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Nowadays billions of multi-modal posts containing texts, images, videos, sound tracks etc. are shared throughout the web, mainly via social media platforms such as Facebook, Twitter, Snapchat,

*This material is based upon work supported by the National Science Foundation under Grant #: 2127309 to the Computing Research Associate for the CIFellows Project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of funding agencies.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2016 ACM. Manuscript submitted to ACM

Reddit, Instagram, YouTube and so on and so forth. While the combination of modalities allows for more expressive, detailed and user friendly content, it brings about new challenges, as it is harder to accommodate uni-modal solutions to multi-modal environments.

However, in recent years, due to the sheer use of multi-modal platforms, many automated techniques for multi-modal tasks such as Visual Question Answering (VQA) [4, 26, 28, 33, 73], image captioning [15, 29, 44, 91] and more recently for fake news detection including hate speech in multi-modal memes [25, 43, 55, 75] have been introduced by machine learning researchers.

Similar to other multi-modal tasks, it is harder and more challenging to detect fake news on multi-modal platforms, as it requires not only the evaluation of each modality, but also cross-modal connections and credibility of the combination as well. This becomes even more challenging when each modality e.g., text or image is credible but the combination creates misinformative content. For instance, a COVID-19 anti-vaccination misinformation[1] post can have text that reads ”vaccines do this” and then attaches a graphic image of a dead person. In this case, although image and text are not individually misinformative, taken together create misinformation.

Over the past decade, several detection models [13, 35, 67, 69] have been developed to detect misinformation. However, the majority of them leverage only a single modality for misinformation detection, e.g., text [27, 32, 65, 88] or image [1, 17, 34, 56], which miss the important information conveyed by other modalities. There are existing works [2, 3, 30, 41, 64] that leverage ensemble methods which create multiple models for each modality and then combine them to produce improved results. However, in many multi-modal misinformative content, individual modalities combined loosely together are insufficient to identify fake news and as a result, the joint model also fails.

Nevertheless, in recent years, machine learning scientists have come up with different techniques for cross-modal fake news detection which combine information from multiple modalities leveraging cross-modal information, such as the consistency and meaningful relationships between different modalities. Study and analyze of these techniques and identifying existing challenges will give a clearer picture of the state of knowledge on multi-modal misinformation detection and opens the door to new opportunities in this filed.

Even though there are a number of valuable surveys on fake news detection [20, 45, 67], very few of them focus on multi-modal techniques [5, 6]. Since the number of proposed techniques on multi-modal fake news detection has been increasing immensely, the necessity of a comprehensive survey on existing techniques, datasets and emerging challenges is felt more than ever. With that

[1]Misinformation” is false information that spreads unintentionally whereas the term ”Disinformation” refers to false information that malicious users share intentionally and often strategically to affect other audiences’ behaviours toward social, political, and economic events. In this work, regardless of spreaders’ intention, we refer to all sorts of false news i.e., misinformation and disinformation as ”Misinformation” or ”Fake News” interchangeably.

said, in this work, we aim to conduct a comprehensive study on fake news detection in multi-modal environments.

To this end, we classify multi-modal misinformation detection study into three different directions as follows:

- **Multi-modal Data Study:** In this category, the goal is to collect multi-modal fake news data e.g., image, text, social context etc. from different sources of information and use fact checking resources to evaluate veracity of collected data and annotate them accordingly. Comparison and analysis of existing datastes as well as bench-marking are other tasks that are under the umbrella of this category.
- **Multi-modal Feature Study:** In this direction, the main objective is to identify meaningful connections between different data modalities which are often manipulated by misinformation spreaders to falsify, imposter or exaggerate original information. These meaningful connections may be used as clues for detecting misinformation in multi-modal environments such as social media posts. Another goal of this direction is to study and develop strategies for fusing feature of different modalities and create information-rich multi-modal feature vectors.
- **Multi-modal Model Study:** This direction mainly focuses on development of efficient multi-modal machine learning solutions to detect misinformation leveraging multi-modal features and clues. Proposing new techniques and approaches, in addition to improving performance, scalability, interpretability and explicablity of machine learning models are some of the common tasks in this direction.

These three directions create a pipeline in multi-modal misinformation study i.e., output of each study provides an input for the next one. A summary of the aforementioned directions, is illustrated in Fig. 1. In this work, we aim to dig deeper into each direction in order to identify challenges and shortcomings of each study and propose unexplored avenues for addressing them. It is worth mentioning that, for readability purposes, we do not necessarily follow the flow of the pipeline shown in Fig.1.

The rest of this survey is organized as follows: To start with, we discuss multi-modal feature study by introducing some widely spread categories of misinformation in multi-modal settings and commonly used cross-modal clues for detecting them. In addition, we discuss different fusion mechanisms to merge modalities that are involved in such clues. Furthermore, we talk about multi-modal model study by introducing solutions and categorizing them based on the machine learning techniques they utilize. Moreover, we go through multi-modal data study by introducing, analysing and comparing existing datbases for multi-modal fake news detection. Finally, we discuss existing challenges and shortcomings that each direction is facing, and propose new avenues for addressing those shortcomings and furthering the multi-modal misinfrmation detection research.

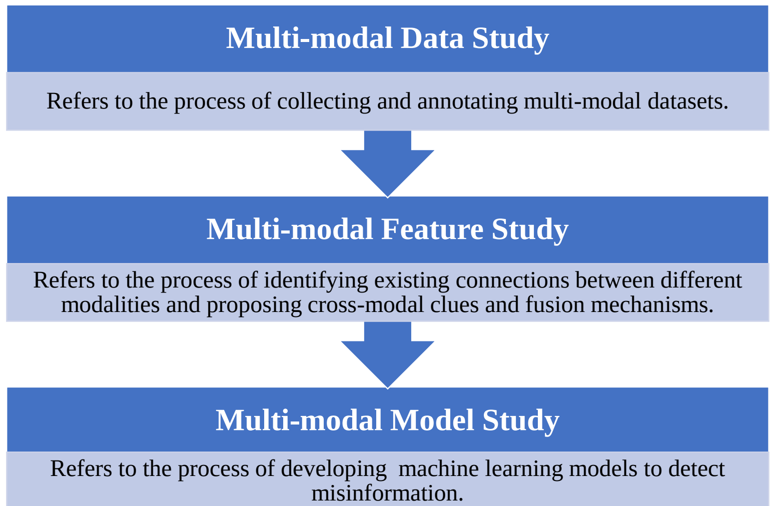


Fig. 1. An overview of multi-modal misinformation detection pipeline.

2 MULTI-MODAL FEATURE STUDY

In this section, we discuss feature-based direction of multi-modal misinformation study. To better understand the rational behind multi-modal features and clues, we start with a brief introduction of some of the common categories of misinformation that spread in multi-modal environments. Furthermore, we discuss some of the commonly used multi-modal features and clues and then we talk about existing fusion mechanisms for combining data modality features. Finally, we debate pros and cons of each fusion mechanism.

2.1 Common categories of misinformation in multi-modal environments

Multi-modal misinformation refers to a package of misleading information which includes multiple modalities such as image, text, video and so on and so forth. In multi-modal misinformation, not all modalities are necessarily false, but sometimes the connections between the modalities are manipulated to deceive the perception of audiences. In what follows, we briefly discuss some of the common categories of misinformation that are widely spread in multi-modal settings. It is worth mentioning that these categories of misinformation are common types of misinformation in both multi-modal and uni-modal environments. However, we provide examples of each category in multi-modal platforms as well.

- **Satire or parody** This category refers to content that convey true information with a satirical tone or added information that makes it false. One of the well-known publishers of this category is TheOnion website[2] which is a digital media organization that publishes satirical articles on a variety of international, national, and local news. A multi-modal example of this category is an image within a satirical news article that contains absurd or ridiculous content or is manipulated to create humorous critique [21, 48]. In this case, the textual content is not necessary false but putting it together with image creates misleading content.
- **Fabricated content** is a category of information which is completely false and is generated to deceive the audiences. The intention to publish fabricated content is usually to mislead people for political, social or economic benefits. A multi-modal instance of this category is a news report that uses auxiliary images or videos that are either completely fake or belong to irrelevant events.
- **Imposter Content** is a category of misinformation that takes advantage of established news agencies by publishing misleading content under their branding. Since audiences trust established agencies, they less likely doubt the validity of content and consequently pay less attention to subtle clues. Imposter content may damage the reputation of agencies and undermines audiences' trust. An example of imposter content is a website that mimics domain features of global news outlets, such as CNN and BBC. To detect this category of misinformation, it is very crucial to identify and pay attention to subtle features of web publishers [1, 2].
- **Manipulated content** is a category of misinformation that is generated by editing a valid information usually in form of images and videos to deceive the audiences. Deepfake videos are well-known examples of this category. Manipulated videos and images have been widely generated to support fabricated contents [61, 79].
- **False connection** This is one of the most common type of misinformation in multi-modal environments. In this category some of the modalities such as captions or titles do not support other modalities e.g., text or video. False connection is designed to catch audiences attention with clickbait headlines or provocative images [49, 53].

[2]https://www.theonion.com/

The above categories are used to spread a variety of fake news content such as ”Junk Science”[3], ”Propaganda”[4], ”Conspiracy Theories”[5],”Hate Speech”, ”Rumors”, ”Bias” etc. In next section, we introduce some of cross-modal clues for detecting them in multi-modal settings.

2.2 Multi-modal features and clues

As mentioned earlier, combinations of features e.g., text and image have been recently used for detecting misinformation in multi-modal environments. In this section, first, we present a non-exhaustive list of commonly used clues that have been exploited by machine learning researchers for detecting misinformation. Furthermore, we discuss categories of fusion mechanisms for merging modalities that are involved in such clues in order to generate multi-modal feature representations.

Image and text mismatch. Combination of textual content and article image is one of the widely used set of features for multi-modal fake news detection. The intuition behind this cue is: some fake news spreaders use tempting images e.g., exaggerated, dramatic or sarcastic graphics which are far from the textual content to attract users' attention. Since it is difficult to find both pertinent and pristine images to match these fictions, fake news generators sometimes use manipulated images to support non-factual scenarios. Researchers refer to this cue as the similarity relationship between text and image [25, 90, 95] which could be captured with a variety of similarity measuring techniques such as cosine similarity between the title and image tags embeddings [25, 95] or similarity measure architectures [90].

Mismatch between video and descriptive writing style. In video-based platforms such as YouTube, TikTok etc. the video content is served with descriptive textual information such as video description, title, users' comments and replies. Different users and video producers use different writing styles in such textual content. These writing styles, could be learned and distinguished from unrecognized patterns by machine learning models. Meanwhile, meaningful relationship between the visual content and the descriptive information e.g., video title is another important clue that could be used for detecting online misbehavior [16]. However, this is a very challenging task, as it is difficult to detect frames that are relevant to the text and discard irrelevant ones e.g., advertisements, opening or ending frames. Moreover, encoding all video frames, is very inefficient in terms of speed and memory.

[3]The term ”Junk Science” refers to inaccurate information about scientific facts that is used to skew opinions or push a hidden agenda

[4]Refers to biased information that is often generated to promote a political point of view. Propaganda ranges form completely false information to subtle manipulation.

[5]Refers to rejecting a widely accepted explanation for an event and offering a secret plot instead.

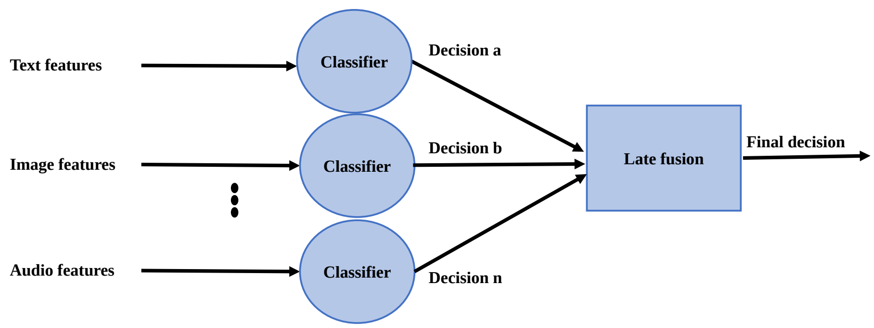
Textual content and propagation network. The majority of the online fact checkers such as BS Detector[6] or News Guard[7] provide labels that pertain to domains rather than articles. Despite this disparity, there are several works [31, 97] that show the weakly-supervised task of using labels pertaining to domains, and subsequently testing on labels pertaining to articles, yields negligible accuracy loss due to the strong correlation between the two [31, 97]. Thus, by recognizing the domain features and behaviours, we might be able to classify articles published by them with admissible accuracy. Some of these feature patterns are the propagation network and word usage patterns of the domains which could be considered [66, 71, 72, 96] as a discriminating signature for different domains. It is empirically shown that not only news articles from different domains have significantly different word usage, but also they follow different propagation patterns [72].

Textual content and overall look of serving domain. Another domain level feature that researchers have recently come up with for detecting misinformation is the overall look of serving webpage [1, 2]. It is shown that, in contrast to credible domains, unreliable web-based news outlets tend to be visually busy and full of events such as advertisements, popups etc [1]. Trustworthy webpages often look professional and ordered, as they often request users to agree to sign up or subscribe, have some featured articles, a headline picture, standard writing styles and so on. On the other hand, unreliable domains tend to have unprofessional blog-post style, negative space and sometimes hard-to-read fonts errors. Considering this discriminating clue, researchers have recently proposed to consider overall look of the webpages in addition to textual content and social context in order to create a multi-modal model for detecting misinformation [2, 3].

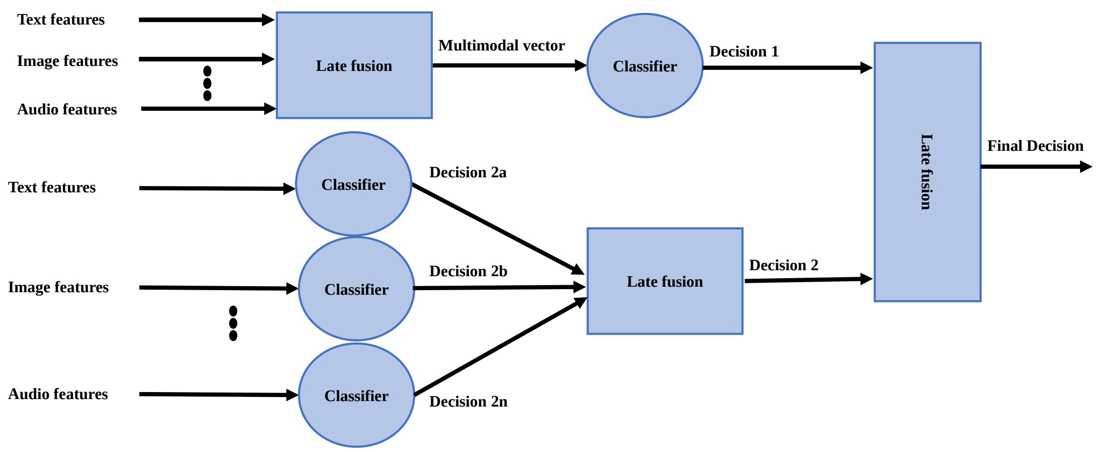
Video and audio mismatch. Due to the ubiquity of camera devices and video-editing applications, video-based frameworks are extremely vulnerable to manipulation e.g., virtual background, anime filter etc. Such kind of visual manipulation introduce a non-trivial noise to the video frames which may lead to mis-classification of irrelevant information from videos [63]. Moreover, manipulated videos often incorporate content in different modalities such as audio and text which sometimes none of them is misinformative while considered individually. However, they mislead the audiences while considered jointly with the video content. In order to detect misleading content that is jointly expressed in video, audio, and text content, researchers have proposed to leverage frame based information along with audio and text content on video-based platforms like Tik Tok [63].

[6]https://github.com/selfagency/bs-detector
[7]https://www.newsguardtech.com/

(a) Early fusion mechanism.



(b) Late fusion mechanism.



(c) A hybrid of early and late fusion mechanisms.

Fig. 2. Different fusion mechanisms in multi-modal learning.

2.3 Fusion mechanisms

Data fusion is the process of combining information from multiple modalities in order to take advantage of all different aspects of the data and extract as much information as possible to improve performance of machine learning models as opposed to using a single data aspect or modality. Different fusion mechanisms have been used to combine features from different modalities, including those we mentioned in the previous section. Fusion mechanisms are often categorized into one of the following groups:

Early fusion. also know as feature-level fusion refers to combining features from different data modalities at an early stage using an operation which is often concatenation. This type of fusion is often performed ahead of classification. If the fusion process is done after feature extraction, it is sometimes referred to as intermediate fusion [12, 47, 51].

Late fusion. also known as decision-level or kernel level fusion, is usually done in the classification stage. Thus, depends on the results obtained by each data modality individually. In other words, the modality-wise classification results are combined using techniques such as sum, max, average, and weighted average. Most of the late fusion solutions use handcrafted rules, prone to human bias, and far from real-world peculiarities [12, 47, 51]

Comparison of fusion mechanisms. In most cases, early fusion is a complex operation, whereas late fusion is easier to perform [8] because unlike the early fusion where the features from different modalities e.g., image and text may have different representation, the decisions at the semantic level usually have the same representation. Therefore, the fusion of decisions is easier than fusion of features. However, the late fusion strategy does not utilize the feature level correlation among modalities which may improve the classification performance. In fact, it is shown that in many cases, the early fusion of different modalities, outperforms multi-modal late fusion while applying deep learning or classic machine learning classifiers [23, 24]. For instance, early fusion of images and texts while using BERT and CNN, on UPMC Food-101 dataset[8] [82] outperforms late fusion of these modalities.

Another advantage of early fusion is that it requires less computation time because training is performed only once but the late fusion needs multiple classifiers for local decisions [8]. However, to have best of both worlds, there are hybrid approaches as well which take advantage of both early and late fusion strategies [8]. A simplified scheme of different fusion mechanisms for multi-modal learning is demonstrated in Fig. 2.

[8]http://visiir.lip6.fr/explore

3 MULTI-MODAL MODEL STUDY

As far as model based study of multi-modal misinformation detection is concerned, machine learning scientists have come up with a variety of solutions. Based on the machine learning techniques that are exploited by these solutions, we may categorize them into two main categories: 1) classic machine learning, and 2) deep leaning based solutions. In this section we discuss each category in detail.

3.1 Classic machine learning solutions

As we discussed earlier, a vast majority of misinformation detection leverage a single modality a.k.a. aspect of news articles e.g., text [27, 32, 65, 88], image [1, 17, 34, 56], user features [68, 70, 89], and temporal properties [45, 66, 77]. However, recently, there have been very few works that incorporate various aspects of a news article using classic machine learning techniques in order to create a multi-modal article representations.

For instance, a work by Shu et al. [41], proposes individual embedded representations for text, user-user interactions, user-article interactions and publisher-article interactions, and define a joint optimization problem leveraging these individual representations. Finally, Alternative Least Square (ALS) algorithm is utilized to solve the proposed optimization problem.

In another work, Abdali et al. propose HiJoD [2], which encodes three different aspects of an article i.e., article text, context of social sharing behaviors and host website/domain features into individual embeddings and extracts shared structures of these embeddings by canceling out the unshared structures using a principled tensor-based framework. The extracted shared structures are utilized for article classification. In this work, the classification performance of the algebraic joint model i.e., HiJoD is compared with the naive concatenation of embedding representations and it is shown that tensor-based representation is more effective in capturing the nuance patterns of the joint structure.

More recently, Meel et al. [30], have proposed an ensemble framework which leverages text embedding, a score calculated by cosine similarity between image caption and news body, and noisy images. Despite the fact that some of the modules of this model e.g., text embedding generator leverage deep attention based architecture, the classification process is done via a classic ensemble technique i.e. max voting.

Summarily, due to the success of deep learning based techniques in feature extraction and classification tasks, classic machine learning based techniques are not commonly used these days. However, considering the fact that deep learning techniques are data hungry and require a lot of effort for training and fine tuning the models, depending on the applications, they are still being used solely or in conjunction with deep learning techniques.

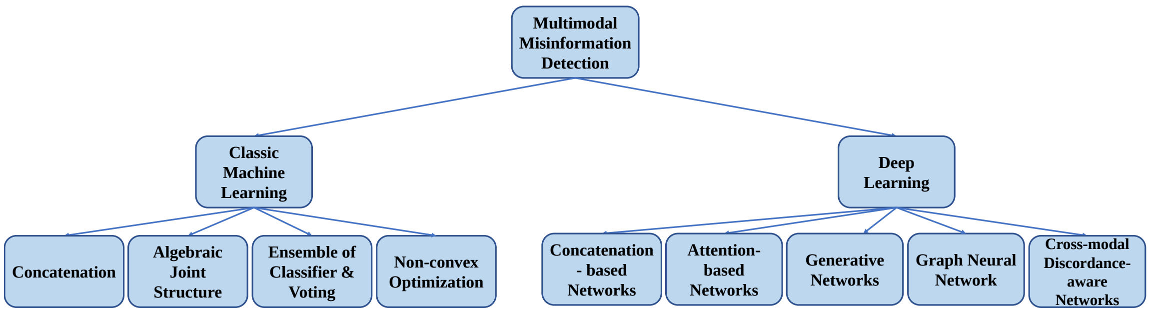


Fig. 3. An overview of the multi-modal solutions.

3.2 Deep learning solutions

Due to the impressive success of deep neural networks in feature extraction and classification of text, image and many other modalities, over the past few years, they have been widely exploited by research scientists for a variety of multi-modal tasks including misinformation detection. We may categorize deep learning based multi-modal misinformation detection into five categories: concatenation-based, attention-based, generative-based, graph neural network-based and cross-modality discordance-aware architectures. In what follows, we summarize and categorize the existing works into the aforementioned categories.

3.2.1 Concatenation-based architectures. The majority of the existing work on multi-modal misinformation detection, embed each modality e.g., text or image into a vector representation and then concatenate them to generate a multi-modal representation which can be utilized for classification tasks. For instance, Singhal et al. propose to use a pretrained XLnet and Vgg-19 models to embed text and image respectively and then classifying the concatenation of resulted feature vectors to detect misinformation [75].

In another work [7], Bartolome et al. exploit a Convolutional Neural Network (CNN) that takes as inputs both text and image corresponding to an article and the outputs are concatenated into a single vector. Qi et al. extract text, Optical Character Recognition (OCR) content, news-related high level semantics of images e.g., celebrities and landmarks, and visual CNN features of the image and then in the stage of multi-modal features fusion, text-image correlations, mutual enhancement, and entity inconsistency are merged by concatenation operation [55].

In another work [60], Rezayi et al. leverage network, textual and relaying features such as hashtags and URLs and classify articles using the concatenation of the feature embeddings. [59, 64] are other examples of this category of deep learning based solutions.

3.2.2 Attention-based architectures. As mentioned above, many architectures simply concatenate vector representations, thereby fail to build effective multi-modal embeddings. Such models are not efficient in many cases. For instance, the entire text of an article does not necessarily need to be false for the corresponding image and vice versa to consider the article as a misinformative content. Thus, some recent works attempt to use attention mechanism to attend to relevant parts of image, text etc. Attention mechanism is a more effective approach for utilizing embeddings, as it produces richer multi-modal representations.

For instance, a work by Sachan et al. [62], proposes Shared Cross Attention Transformer Encoders (SCADE) which exploits CNNs and transformer-based methods to encode image and text information and utilizes cross-modal attention and shared layers for the two modalities. SCADE pays attention to the relevant parts of each modality with reference to the other one.

Another example is a work by Kumari et.al. [46], where a framework is developed to maximize the correlation between textual and visual information. This framework has four different submodules. Attention Based Stacked Bidirectional Long Short Term Memory (ABS-BiLSTM) for textual feature representation, Attention Based Multilevel Convolutional Neural Network–Recurrent Neural Network (ABM-CNN–RNN) for visual feature extraction, multi-modal Factorized Bilinear Pooling (MFB) for feature fusion and finally Multi-Layer Perceptron (MLP) for the classification.

In another work, Qian et al. [57], proposes a Hierarchical Multi-modal Contextual Attention Network (HMCAN) architecture which utilizes a pretrained Bidirectional Encoder Representations from Transformers (BERT) and convolutional ResNet50 to generate word and image embeddings and a multi-modal contextual attention network to explore the multi-modal context information. In this work, HMCAN leverages different multi-modal contextual attention networks to constitute a hierarchical encoding network to explore and capture the rich hierarchical semantics of multi-modal data.

Another example is [38], where Jin et al. fuse features from three modalities i.e., textual, visual and social context using an RNN that utilizes attention mechanism (att-RNN) for feature alignment. Jing et al. propose TRANSFAKE [40] to connect features of text and images into a series, and feed them into vision-language transformer model to learn the joint representation of multi-modal features. TRANSFAKE adopts a preprocessing method similar to BERT for concatenated text, comment and image.

In another work [81], Wang et al. apply scaled dot-product attention on top of image and text features as a fine-grained fusion and use the fused feature to classify articles.

Wang et al. propose a deep learning network for Biomedical informatics that leverages visual and textual information and a semantic- and task-level attention mechanism to focus on the essential contents of a post that signal anti-vaccine messages [87].

Another example is a work by Lu et al. where representations of user interaction, word representations and propagation features are concatenated after applying a dual co-attention mechanism. The rationale is to capture the correlations between users' interaction/propagation and tweet's text [50].

Finally, Song et al. [76], propose a multi-modal fake news detection architecture based on Cross-modal Attention Residual (CARN) and Multichannel convolutional neural Networks (CARMN). Cross-modal Attention Residual or CARN selectively extracts the information related to a target modality from a source modality while maintaining the unique information of the target.

3.2.3 Generative architectures. In this category of deep learning solutions, the goal is to either apply Generative Networks or uses auxiliary networks to learn individual or multi-modal representations, spaces or parameters in order to improve the classification performance of the fake news detector.

As an example, Jaiswal et al. propose a BERT based multi-modal variational Autoencoder (VAE) [37] that consists of an encoder, decoder and a fake news detector. The encoder encodes the shared representations of both the image and text into a multidimensional latent vector. The decoder decodes the multidimensional latent vector into the original image and text and the fake news detector is a binary classifier that takes the shared representation as an input and classifies it as either fake or real.

Similarly, Kattar et.al. propose a deep multi-modal variational autoencoder (MVAE) [42] which learns a unified representation of both the modalities of a tweet's content. Similar to the previous work, MVAE has three main components: encoder, decoder and a fake News detector that utilizes the learned shared representation to predict if a news is fake or not.

Like previous work, a work by Zeng et al. [92] proposes to capture the correlations between text and image by a VAE-based multi-modal feature fusion method. In another work, Wang et al. propose Event Adversarial Neural Networks (EANN) [83], an end-to-end framework which can derive event-invariant features and thus benefit the detection of fake news on newly arrived events. It consists of three main components: a multi-modal feature extractor, the fake news detector, and the event discriminator. The multi-modal feature extractor is responsible for extracting the textual and visual features from posts. It cooperates with the fake news detector to learn the discriminating representation of news articles. The role of event discriminator is to remove the event-specific features and keep shared features among the events.

In another work [84], Wang et al. propose MetaFEND framework which is able to detect fake news on emergent events with a few verified posts using an event adaption strategy. MetaFEND framework has two stages: event adaption and detection. In event adaption stage, the model adapts to specific events and then in the detection stage, the event-specific parameter is leveraged to detect fake news on a given event. Although MetaFEND does not apply a generative architecture, it leverages an auxiliary network to learn event-specific parameter set to improve efficiency of fake news detector.

The last example is a work [72] by Silva et.al where they propose a cross-domain framework using text and propagation network. The proposed model consists of two components: an unsupervised domain embedding learning; and a supervised domain-agnostic news classification. The unsupervised domain embedding exploits text and propagation network to represent a news domain with a low-dimensional vector. The classification model represents each news record as a vector using the textual content and the propagation network. Then, the model maps this representation into two different subspaces such that one preserves the domain-specific information. Later on, these two components are integrated to identify fake news while exploiting domain-specific and cross-domain knowledge in the news records.

3.2.4 Graph neural network architectures. In resent years, Graph Neural Networks (GNNs) have been successfully exploited for fake news detection [9, 19, 78], thereby they have caught researchers' attentions for multi-modal misinformation detection task as well. In this category of deep learning solutions, article content e.g., text, image etc. are represented by graphs and then graph neural networks are used to extract the semantic-level features.

For instance, Wang et al. construct a graph for each social media post based on the point-wise mutual information (PMI) score of pairs of words, extracted objects in visual content and knowledge concept through knowledge distillation. Then, utilize a Knowledge-driven Multi-modal Graph Convolutional Network (KMGCN) which extracts the multi-modal representation of each post through Graph convolutional networks [85].

Another GCN based model is GAME-ON [22] which represents each news with uni-modal visual and textual graphs and then project them into a common space. To capture multi-modal representations, Game-on applies a graph attention layer on a multi-modal graph generated out of modality graphs.

3.2.5 Cross-modal discordance-aware architectures. In the previous categories we discussed above, deep learning models are used to fuse different modalities in order to obtain discriminating representations. However, in this category, deep learning architectures are customized based on some identified discordances between modalities. The intuition is that: fabrication of either modality will lead to dissonance between the modalities and results in misrepresented, misinterpreted and misleading news. Therefore, there are subtle cross-modal discordance clues that could be identified and learned by customized architectures.

For instance, in many cases, fake news propagators use irrelevant modalities e.g., image, video, audio etc., for false statements to attract readers' attention. Thus, the similarity of text to other modalities e.g., image, audio etc., is a cue for measuring the credibility of a news article.

With that said, Zhou et al. [95], propose SAFE, a Similarity-Aware Multi-Modal Fake News Detection framework by defining the relevance between news textual and visual information using a modified cosine similarity.

Similarly, Giachanou et al., propose a multi-image system that combines textual, visual and semantic information [25]. The semantic representation refers to the text-image similarity calculated using the cosine similarity between the title and image tags embeddings.

In another work, Singhal et.al. [74] develop an inter-modality discordance based fake news detector which learns discriminating features and employs a modified version of contrastive loss that explores the inter-modality discordance. Xue et al. [90], propose a Multi-modal Consistency Neural Network (MCNN) which utilizes a similarity measurement module that measures the similarity of multi-modal data to detect the possible mismatches between the image and text. Lastly, Biamby et al. [10], leverage a Contrastive Language-Image Pre-Training (CLIP) model [58], to jointly learn image/text representation to detect Image-Text inconsistencies in Tweets. Instead of concatenating vector representation, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples.

On video-based platforms such as YouTube videos, typically different producers use different title and description, as users and subscribers express their opinions in different writing styles.

Having this clue in mind, Choi et.al. propose a framework to identify fake content on YouTube [16]. They propose to use domain knowledge and "hit-likes" of comments to create the comments embedding which is effective in detecting fake news videos. They encode Multi-modal features i.e., image and text and detect differences between title, description or video and user's comments.

In another work [63], Shang et.al. develop TikTec, a multi-modal misinformation detection framework that explicitly exploits the captions to accurately capture the key information from the unreliable video content. This framework learns the composed misinformation that is jointly conveyed by the visual and audio content. TikTec consists of four major components. A Caption-guided Visual Representation Learning (CVRL) module which identify the misinformation-related visual features from each sampled video frame, An Acoustic-aware Speech Representation Learning (ASRL) module that jointly learns the misleading semantic information that is deeply embedded in the unstructured and casual audio tracks and the Visual-speech Co-attentive Information Fusion (VCIF) module which captures the multiview composed information jointly embedded in the heterogeneous visual and audio contents of the video. Finally, the Supervised Misleading Video Detection (SMVD) module identifies misleading COVID-19 videos.

Table 1. A summary of the existing deep learning solutions.

Paper	Concatenation	Attention-based	Generative Networks	Graph Neural Networks	Cross-modal Discordance-aware
[75]	✓	✗	✗	✗	✗
[64]	✓	✗	✗	✗	✗
[7]	✓	✗	✗	✗	✗
[55]	✓	✗	✗	✗	✗
[60]	✓	✗	✗	✗	✗
[59]	✓	✗	✗	✗	✗
[38]	✓	✓	✗	✗	✗
[50]	✓	✓	✗	✗	✓
[57]	✓	✓	✗	✗	✗
[52]	✓	✓	✗	✗	✗
[62]	✓	✓	✗	✗	✗
[46]	✗	✓	✗	✗	✗
[40]	✓	✓	✗	✗	✗
[87]	✓	✓	✗	✗	✗
[76]	✓	✓	✗	✗	✗
[81]	✓	✓	✗	✗	✗
[36]	✓	✓	✗	✗	✗
[83]	✓	✗	✓	✗	✗
[42]	✓	✗	✓	✗	✗
[92]	✓	✓	✓	✗	✗
[37]	✓	✓	✓	✗	✗
[72]	✓	✗	✓	✗	✗
[93]	✓	✗	✓	✗	✗
[84]	✓	✓	✓	✗	✗
[85]	✗	✗	✗	✓	✗
[22]	✗	✓	✗	✓	✗
[95]	✗	✗	✗	✗	✓
[25]	✓	✓	✗	✗	✓
[90]	✗	✓	✗	✗	✓
[63]	✓	✓	✗	✗	✓
[74]	✓	✓	✗	✗	✓
[10]	✗	✗	✗	✗	✓
[16]	✓	✓	✗	✗	✓

A summary of the aforementioned deep learning based works is demonstrated in Table 1. It is worth mentioning that many of the state-of-the-art solutions utilize a hybrid of deep learning solutions.

4 MULTI-MODAL DATA STUDY

Data accusation and preparation is the most important building block of a machine learning pipeline. Machine learning models leverage training data to continuously improve themselves over time. Thus, sufficient good quality, and in most cases annotated data is extremely crucial for these models to operate effectively. With that said, in this section we introduce and compare some of the existing multi-modal datasets for feke news detection task. Later on we will discuss some of the limitations of these datasets.

Image-Verification-Corpus. [9] is an evolving dataset containing 17,806 fake and real posts with images shared on Twitter. This dataset is created as an open corpus of tweets containing images that may be used for assessing online image verification approaches (based on tweet texts and user features), as well as building classifier for new content. Fake and real images of this dataset have been annotated by online sources that evaluate the credibility of the images and the events they are associated with [11].

Fakeddit. [10] is a dataset collected from Reddit, a social news and discussion website where users can post submissions on various subreddits. Fakeddit consists of over 1 million submissions from 22 different subreddits spanning over a decade with the earliest submission being from 3/19/2008, and the most recent submission being from 10/24/2019. These subreddits are posted on highly active and popular pages by over 300,000 users. Fakeddit consists of submission titles, images, user comments, and submission metadata including score, username of the author, subreddit source, sourced domain, number of comments, and up-vote to down-vote ratio. Approximately 64% of the samples have both text and image data [53]. Samples of this dataset are annotated with 2-way, 3-way and 6-way labels including true, satire/paroday, misleading content, manipulated content, false connection and imposter content.

NewsBag. comprises 200,000 real news and 15,000 fake articles. The real training articles have been collected from the Wall Street Journal and the fake ones from The Onion website[11]which publishes satirical content. However, the samples of the test set are collected from different websites i.e., TheRealNews[12] and ThePoke[13]. The rational behind using different sources of news for the training and test sets is to observe how well the models could be generalized to unseen data samples. The NewsBag dataset is a highly imbalanced dataset. Thus, to tackle this issue, NewsBag ++ is

[9]https://githubhelp.com/MKLab-ITI/image-verification-corpus
[10]https://github.com/entitize/Fakeddit
[11]https://www.theonion.com/
[12]https://therealnews.com/
[13]https://www.thepoke.co.uk/

also released which is the augmented training version of the `NewsBag` dataset and contains 200,000 real and 389,000 fake news articles. Another weakness of `NewsBag` dataset is that it does not have any social context information such as spreader information, sharing trends and reactions such as user comments and engagements [39].

MM−COVID. [14] is a multi-lingual and multi-dimensional COVID-19 fake news data repository. This dataset comprises 3981 fake news and 7192 trustworthy information in 6 different languages i.e., English, Spanish, Portuguese, Hindi, French and Italian. `MM−COVID` consists of visual, textual and social context information e.g., users and networks information [49]. Samples of this dataset are annotated by Snopes[15] and Poynter[16] crowed source domains where the experts and journalists evaluate and fact check news content and annotate contents as either fake or real. While Snopes is an independent publication which mainly contains English content, Poynter is an international fact-checking network (IFCN) which unites 96 different fact-checking agencies such as PolitiFact[17] in 40 languages.

ReCOVery. [18] contains 2,029 news articles that have been shared on social media, most of which (2,017 samples) have both textual and visual information for multi-modal studies. `ReCOVery` is imbalanced in news class i.e., the proportion of real vs. fake articles is around 2:1. The number of users who spread real news (78,659) and users sharing fake articles (17,323) is greater than the total number of users included in the dataset (93,761). In this dataset the assumption is that users can engage in spreading both real and fake news articles. Samples of this dataset are annotated by two factchecking resources: NewsGuard[19] and Media Bias/Fact Check (MBFC)[20] which is a website that evaluates factual accuracy and political bias of news media. MBFC labels each news media as one of six factual-accuracy levels based on the fact-checking results of the previously published news articles. Samples of `ReCOVery` are collected from 60 news domains, from which 22 are the sources of reliable news articles (e.g., National Public Radio[21] and Reuters[22]) and the remaining 38 are sources to collect unreliable news articles (e.g., Human Are Free[23] and Natural News[24] [94].

[14] https://github.com/bigheiniu/MM-COVID
[15] 2www.snopes.com
[16] www.poynter.org/coronavirusfactsalliance/
[17] https://www.politifact.com/
[18] https://github.com/apurvamulay/ReCOVery
[19] https://www.newsguardtech.com/
[20] https://mediabiasfactcheck.com/
[21] https://www.npr.org/
[22] 2https://www.reuters.com
[23] 3http://humansarefree.com/
[24] www.https://www.naturalnews.com/

CoAID. [25] Covid-19 heAlthcare mIsinformation Dataset or `CoAID` is a diverse COVID-19 healthcare misinformation dataset, including fake news on websites and social platforms, along with users' social engagement about the news. It includes 5,216 news, 296,752 related user engagements, 926 social platform posts about COVID-19, and ground truth labels. The publishing dates of the collected information range from December 1, 2019 to September 1, 2020. In total, 204 fake news articles, 3,565 true news articles, 28 fake claims and 454 true claims are collected. Real news articles are crawled from 9 reliable media outlets that have been cross-checked as reliable e.g., National Institutes of Health (NIH)[26] and CDC[27] Fake news are retrieved from several fact-checking websites, such as PolitiFact, Health Feedback[28] etc [18].

MMCoVaR. is a Multi-modal COVID-19 Vaccine Focused Data Repository (MMCoVaR). Articles of this dataset are annotated using two news website source checking and the tweets are fact checked based on stance detection approach. `MMCoVaR` comprises 2,593 articles issued by 80 publishers and shared between 02/16/2020 05/08/2021 and 24,184 Twitter posts collected between 04/17/2021 to 05/08/2021. Samples of this dataset are annotated by Media Bias Chart and Media Bias/Fact Check (MBFC) and classified into two levels of credibility: reliable and unreliable. Thus, articles are labeled as either credible or unreliable and tweets are annotated as reliable, inconclusive or reliable [14]. It is worth mentioning that textual, visual and social context information are available for the news articles.

N24News. [29] is a multi-modal dataset extracted from the New York Times articles published from 2010 to 2020. Each news belongs to one of 24 different categories e.g., science, arts etc. The dataset comprises up to 3,000 samples of real news for each category. Totally, 60,000 news articles are collected. Each article sample contains category tag, headline, abstract, article body, image, and corresponding image caption. This dataset is randomly split into training/validation/testing sets in the ratio of 8:1:1. [86]. The main weakness of this dataset is that it does not have any fake samples and all of the real samples are collected from a single source i.e., The New York Times.

MuMiN. [30] Large-Scale Multilingual Multi-modal Fact-Checked Misinformation Social Network Dataset (MuMin) comprises 21 million tweets belonging to 26 thousand Twitter threads, each of which has been linked to 13 thousand fact-checked claims in 41 different languages. `MuMiN` is available in three large, medium and small versions with largest one consisting 10,920 articles and

[25] https://github.com/cuilimeng/CoAID
[26] 2https://www.nih.gov/news-events/news-releases
[27] https://www.cdc.gov/coronavirus/2019-ncov/whats-new-all.html
[28] https://healthfeedback.org/
[29] https://github.com/billywzh717/N24News
[30] https://github.com/MuMiN-dataset/mumin-build

6,573 images. In this dataset, if the claim is "mostly true", it is labeled as factual. When the claim is deemed "half true" or "half false" it is labeled as misinformation, with the justification that a statement containing a significant part of false information should be considered as a misleading content. When there is no clear verdict then the verdict is labelled as other. [54].

A summary and a side by side comparison of the aforementioned datasets are illustrated in Fig. 4 to 5 and Table 2. As demonstrated, most of these datasets are small in size, are annotated with binary labels, extracted from limited resources such as Twitter, and only contain limited modalities i.e., text and image.

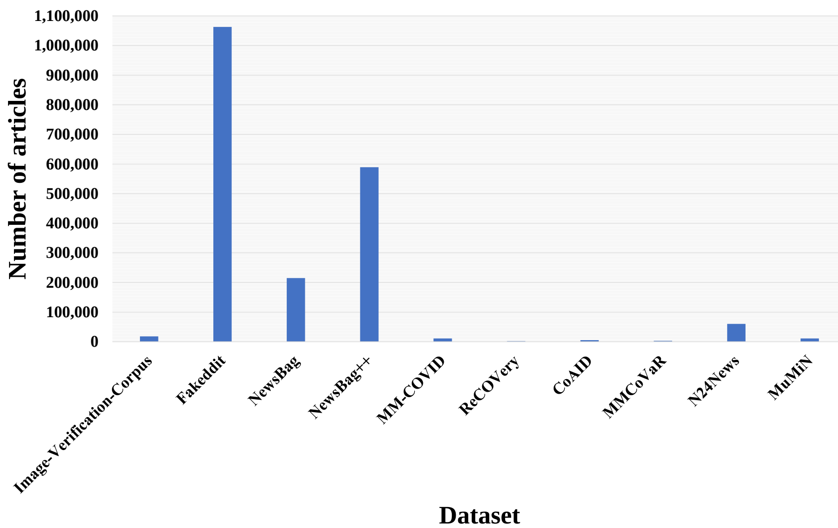
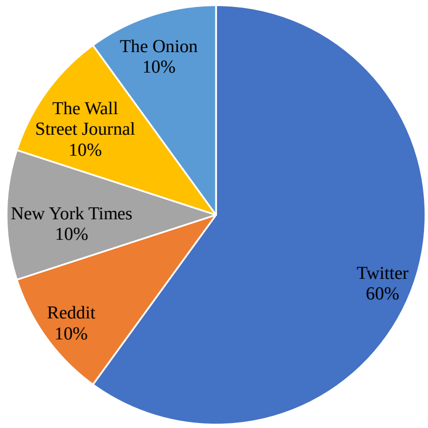
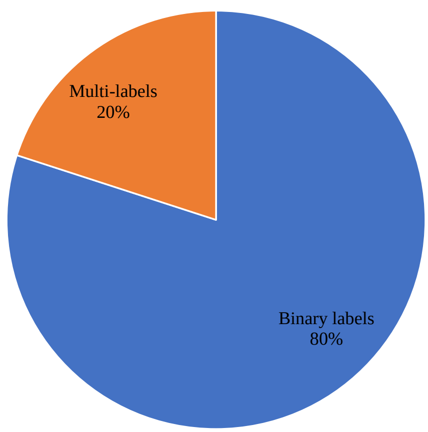


Fig. 4. Number of news articles by dataset.



(a) Source of dataset



(b) Type of ground truth

Fig. 5. Statistics of the existing datasets.

Table 2. Statistics of multi-modal databases for fake news detection

Dataset	Total Samples	# classes	Modalities	Source	Details
image-verification-corpus [11]	17,806	2	image,text	Twitter	
Fakeddit [53]	1,063,106	2,3,6	image,text	Reddit	682,996 samples are multi-modal
NewsBag [39]	215,000	2	image, text	TTrain: Wall Street&Onion.Test: TheRealNews & ThePoke	This dataset is highly imbalanced. There are only 15,000 fake samples.
NewsBag++ [39]	589,000	2	image,text	Train: Wall Street&Onion. Test: TheRealNews & ThePoke	Same as NewsBag but fake samples are synthetic samples created by augmentation techniques
MM-COVID [49]	11,173	2	image,text,social context	Twitter	3,981 fake samples and 7,192 real samples
ReCOVery [94]	2,029	2	text,image	Twitter	Imbalanced with ratio of 2:1 real vs.fake
CoAID [18]	5,216	2	image,text	Twitter	Consists of 296,752 user engagements (926 social platforms)
MMCoVaR [14]	2,593 articles & 24,184 tweets	2	image,text,social context	Twitter	Tweets as labeled as reliable,inconclusive and unreliable
N24News [86]	60,000	24	image,text	New York Times	All samples are real from 24 different categories
MuMiN [54]	10,920	3	image,text	Twitter	Consists of 10,920 articles and 6,573 images.

5 CHALLENGES AND OPPORTUNITIES IN MULTI-MODAL MISINFORMATION DETECTION

Recent studies on multi-modal learning have made significant contributions to the field of multi-modal fake news detection. However, there are still weaknesses and shortcomings which recognizing them opens the door to new opportunities not only in fake news detection study, but also in multi-modal field in general. In this section, we discuss challenges, shortcomings and opportunities in multi-modal fake news detection. We provide non-exhaustive lists of challenges and shortcomings for each direction of multi-modal misinformation study. A summary of all these challenges is demonstrated in Fig. 6.

5.1 Data related challenges

This category refers to the weaknesses of current multi-modal datasets for misinformation detection. We briefly discussed some of these weaknesses in the multi-modal data study section. An itemized list of such limitations and shortcomings is as follows:

- **Lack of large and comprehensive datasets** As illustrated in Fig. 4, most of the existing datasets are small in size, and sometimes highly imbalanced in terms of fake to real ratio.
- **Lack of cross-lingual datastes** Almost all social media platforms are multi-lingual environments where users share information in multiple languages. Although misinformation spreads in multiple languages, a vast majority of the existing detests are mono-lingual i.e., they only provide English content. Therefore, there is a serious lack of non-English contents and annotations.
- **Limited modalities** As discussed earlier, most of the existing multi-modal datasets only provide image and text modalities. Thus, neglect useful information that are conveyed by other modalities such as video, audio etc. The necessity of providing more modalities comes into light more when we consider popular social media such as YouTube, TikToK and Clubhouse that are mainly video or audio based platforms.
- **Bias in event-specific datasets** Many of the existing datasets are created for specific events such as COVID-19 crises, thereby do not cover a variety of events and topics and as a result of this they might not be sufficient to train models to detect fake news in other contexts.
- **Binary and domain level ground truth** Most of the existing datasets provide us with binary and domain level ground truth for well-known outlets such as Onion or New York times. In addition, they often do not provide any information about reasons of mis-informativeness e.g., cross-modal discordance, false connection, imposter content etc.
- **Subjective annotations and inconsistency of labels** As discussed in the data study section, different datasets use different crowd-source and fact-checking agencies, thereby articles

are annotated subjectively with different labels across different datasets. Thus, it is very hard to analyze, compare and interpret results on different datasets.

5.2 Feature related challenges

This category comprises shortcomings related to cross-modal feature identification and extraction in the multi-modal fake news detection pipeline. Some of the most important weaknesses in the current feature based study are:

- **Insufficiency of cross-modal cues** Although researchers have proposed some multi-modal cues, most of the existing models naively fuse image based features with textual features as a supplement. There exist fewer works that leverage explainable cross-modal cues other than image and text combination. However, there are still plenty of useful multi-modal cues which are mainly being neglected by researchers.
- **Ineffective cross-modal embeddings** As mentioned earlier, the majority of the existing approaches only fuse embeddings with simple operations such as concatenation of the representations, thereby fail to build an effective and non-noisy cross-modal embedding. Such architectures fail in many cases, as the resulted cross-modal embedding consists of unuseful or irrelevant parts which may result in noisy representations.
- **Lack of language-independent features** The majority of existing work on misinformation leverage text features that are highly dependent on dataset languages which is mostly English. Identifying language-independent features, is an effective way to cope with mono-lingual datasets.

5.3 Model related challenges

This category refers to the shortcomings of current machine learning solutions in detecting misinformation in multi-modal environments. The following is a non-exhaustive list of existing shortcomings:

- **Inexplicability of current models** A majority of the existing models do not provide any explicable information about the regions of interest, common patterns of inconsistencies among modalities and type of misinformation e.g., manipulation, exaggeration etc. While some recent works attempt to use attention-based techniques to overcome the problem of ineffective multi-modal embedding and provide some interpretability, most of them usually follow a trial and error approach like masking to find relevant sections to attend to. However, interpretable and explainable AI is crucial in building trust and confidence as well as fairness and transparency which is being mostly neglected.
- **Non-transferable models to unseen events** Most of the existing models are designed in such a way that they extract and learn event-specific features e.g., COVID-19, election etc.

Thus, they are most likely biased toward specific events and as a result, non-transferable to unseen and emerging events. For this very reason, building models that learn general features and separate them from the nontransferable event-specific features would be extremely useful.

- **Unscalibility of current models** Considering the expensive and complicated structures of deep networks and the fact that most of the existing multi-modal models leverage multiple deep networks (one for each modality), they are not scalable if the number of modalities increases. Moreover, many of the existing models require heavy computing resources and need large volume of memory storage and processing units. Therefore, scalablity of proposed models should be taken into account while developing new architectures.
- **Vulnerabilities against adversarial attacks** Malicious adversaries continuously try to fool the misinformation detection models. This is especially feasible when the underlying model's techniques and cues are revealed to the attacker, such as when the attacker can probe the model. As a result, many of the detection techniques become dated in a short period of time. Thus, there is a need to create detection models that are resistant to manipulation.

5.4 Opportunities in multi-modal misinformation detection study

Considering the challenges and shortcomings in multi-modal misinformation detection we discussed above, we propose opportunities in furthering research in this field. In what follows we discuss these opportunities by each direction of multi-modal misinformation detection study.

5.4.1 Opportunities in multi-modal data study. Considering the data related challenges we discussed earlier, we propose the following data related avenues:

- **Comprehensive multi-modal and multi-lingual datasets** As we discussed in detail, one important gap in the misinformation detection study is the lack of a comprehensive multi-modal dataset which needs to be addressed in the future. Multi-modal misinformation detection requires large, multi-lingual, multi-source datasets that cover a variety of modalities, web resources, events etc. and provide fine-grained ground truth for the samples.
- **Standardized annotation strategy** As mentioned above, current datasts annotate articles by different fact-checking agencies and as a result the labels are in many cases subjective. Having a standard agreement across all datasets on how to label articles makes the cross-dataset comparison and analysis much easier.

5.4.2 Opportunities in multi-modal feature study. Based on feature related challenges we discussed in previous section, we propose the following research opportunities to overcome some of the existing challenges in multi-modal feature study:

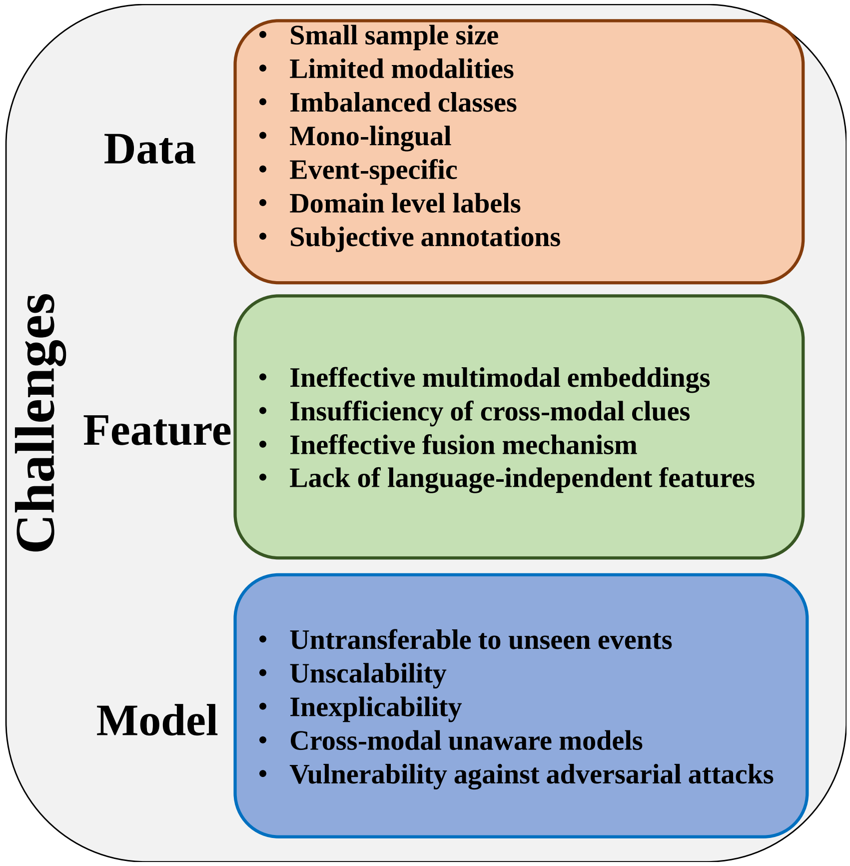


Fig. 6. A summary of the challenges in multi-modal misinformation detection.

- **Identifying cross-modal clues** As mentioned earlier, cross-modal cues are currently limited to a handful of trivial clues such as similarity of text and image. Identifying subtler, yet neglected cues not only helps in development of discordance-aware models, but also could be helpful in recognizing vulnerabilities of the serving platforms which is a part and parcel of adversarial learning.
- **Developing efficient fusion mechanism** As discussed before, many of the existing solutions leverage naive fusion mechanisms such as concatenation which may result in inefficient and noisy multi-modal representations. Therefore, another fruitful avenue of research lies in the study and development of more efficient fusion techniques to produce information richer representations.
- **Identifying language-independent features to cope with mono-lingual datasets** As discussed above, the majority of existing datasets are mono-lingual, thereby are not sufficient

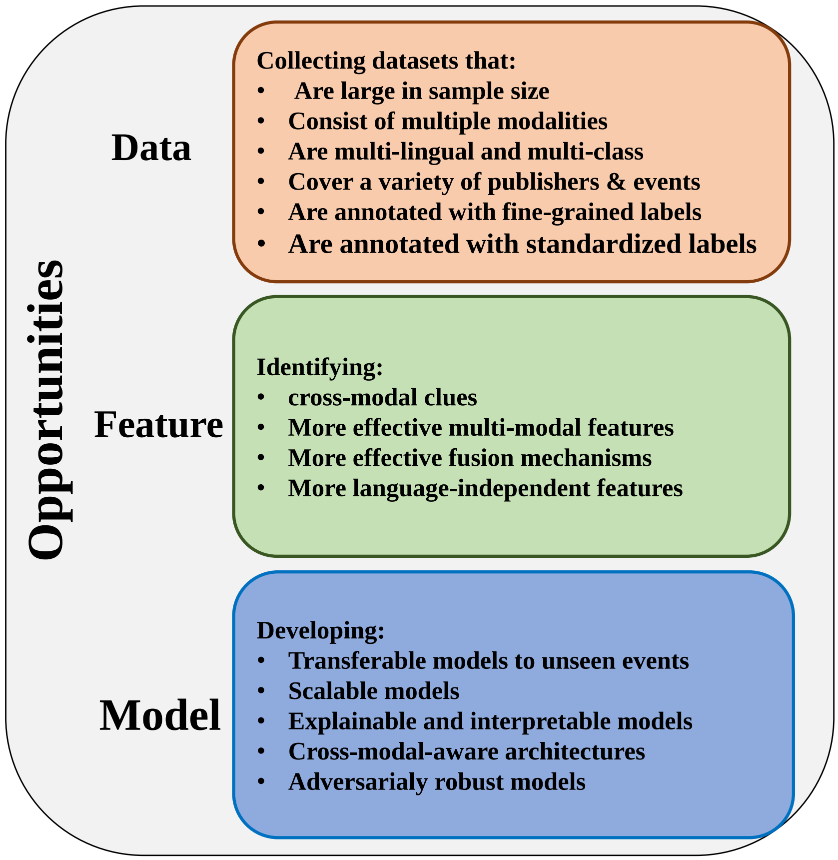


Fig. 7. A summary of the opportunities in multi-modal misinformation detection.

enough to train models for non-English tasks. One way to compensate for the lack of multi-lingual datasets is to use language-independent features [80]. Identifying such features, specially in multi-modal environment where there are more features and aspects, would be highly effective to cope with mono-lingual datasets.

5.5 Opportunities in multi-modal model study

In what follows, we propose some unexplored research avenues to tackle some of the existing model related challenges in multi-modal misnformation detection.

- **Developing cross-modal discordance-aware architectures** As described earlier, most of the existing works, either blindly merge modalities or take a trial and error approach to attend to the relevant modalities. Implementing discordance-aware models not only results in information richer representations, but also may be useful in making attention based techniques more efficient.
- **Adversarial learning in multi-modal misinformation detection** Although there are existing generative-based architectures, adversarial study of multi-modal misinformation detection has been mostly neglected. In order to make the detection models more adversarialy robust, it is of utmost importance to dedicate time and effort to the study and development of generative and adversarial learning techniques.
- **Interpretability of multi-modal models** Development of explainable frameworks to help better understand and interpret predictions made by multi-modal detection models is another opportunity in multi-modal misinformation detection. Explicablity can be very useful for related tasks such as predictability of model, fairness and bias, and adversarial learning.
- **Transferable models to unseen events** As mentioned earlier, except a few works, most of the existing models are designed for specific events and as a result, ineffective for emerging ones. Since misinformation spreads during a variety of events, developing general and transferable models is extremely crucial.
- **Development of scalable models** Another opportunity is to develop models that are more efficient in terms of time and resources and do not become intolerably complicated while increasing the number of fused modalities.

A summary of the aforementioned opportunities in all three directions of multi-modal misinformation detection is demonstrated in Fig. 7.

6 CONCLUSIONS

In this work, we study existing works on multi-modal misinformation detection, analyze their strengths and weaknesses and offer new opportunities for advancing research in this field. First, we introduce some of the widespread misinformation categories and commonly used cross-modal clues for detecting them and then we discuss different fusion mechanisms to merge involved modalities in such cross-modal clues. Next, we categorize existing solutions into two main categories of classic machine learning and deep learning solutions and then break each one of them down based on the utilized techniques. In addition, we introduce and compare existing datasets on multi-modal misinformation detection and identify some of the weaknesses of these datastes. Moreover, we discuss some of the existing challenges in multi-modal fake news detection study by categorizing them into data, feature and model based shortcomings. Finally, we propose new directions and research avenues to address each one of those shortcomings.

7 ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant number 2127309 to the Computing Research Associate for the CIFellows Project. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of funding agencies.

REFERENCES

[1] Sara Abdali, Rutuja Gurav, Siddharth Menon, Daniel Fonseca, Negin Entezari, Neil Shah, and Evangelos E. Papalexakis. 2021. Identifying Misinformation from Website Screenshots. *Proceedings of the International AAAI Conference on Web and Social Media* 15, 1 (May 2021), 2–13. https://ojs.aaai.org/index.php/ICWSM/article/view/18036

[2] Sara Abdali, Neil Shah, and Evangelos E. Papalexakis. 2020. HiJoD: Semi-Supervised Multi-aspect Detection of Misinformation using Hierarchical Joint Decomposition. In *ECML/PKDD*.

[3] Sara Abdali, Neil Shah, and Evangelos E. Papalexakis. 2021. KNH: Multi-View Modeling with K-Nearest Hyperplanes Graph for Misinformation Detection. *CoRR* abs/2102.07857 (2021). arXiv:2102.07857 https://arxiv.org/abs/2102.07857

[4] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: Visual Question Answering. *Int. J. Comput. Vision* 123, 1 (may 2017), 4–31. https://doi.org/10.1007/s11263-016-0966-6

[5] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2021. A Survey on Multimodal Disinformation Detection.

[6] Santiago Alonso-Bartolome and Isabel Segura-Bedmar. 2021. Multimodal Fake News Detection. *CoRR* abs/2112.04831 (2021). arXiv:2112.04831 https://arxiv.org/abs/2112.04831

[7] Santiago Alonso-Bartolome and Isabel Segura-Bedmar. 2021. Multimodal Fake News Detection. arXiv:2112.04831 [cs.CL]

[8] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and M. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems* 16 (2010), 345–379.

[9] Adrien Benamira, Benjamin Devillers, Etienne Lesot, Ayush K. Ray, Manal Saadi, and Fragkiskos D. Malliaros. 2019. Semi-Supervised Learning and Graph Neural Networks for Fake News Detection *(ASONAM '19)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3341161.3342958

[10] Giscard Biamby, Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Twitter-COMMs: Detecting Climate, COVID, and Military Multimodal Misinformation. arXiv:2112.08594 [cs.CV]

[11] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval* 7, 1 (2018), 71–86. https://doi.org/10.1007/s13735-017-0143-x

[12] Said Boulahia, Abdenour Amamra, Mohamed Madi, and Said Daikh. 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications* 32 (11 2021). https://doi.org/10.1007/s00138-021-01249-8

[13] Yegui Cai, George O. M. Yee, Yuan Xiang Gu, and Chung-Horng Lung. 2020. Threats to Online Advertising and Countermeasures: A Technical Survey. *Digital Threats: Research and Practice* 1, 2, Article 11 (may 2020), 27 pages. https://doi.org/10.1145/3374136

[14] Mingxuan Chen, Xinqiao Chu, and K. P. Subbalakshmi. 2021. MMCoVaR: Multimodal COVID-19 Vaccine Focused Data Repository for Fake News Detection and a Baseline Architecture for Classification *(ASONAM '21)*. Association for Computing Machinery, New York, NY, USA, 31–38. https://doi.org/10.1145/3487351.3488346

[15] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR* abs/1504.00325 (2015). arXiv:1504.00325 http://arxiv.org/abs/1504.00325

[16] Hyewon Choi and Youngjoong Ko. 2022. Effective fake news video detection using domain knowledge and multimodal data fusion on youtube. *Pattern Recognition Letters* 154 (2022), 44–52. https://doi.org/10.1016/j.patrec.2022.01.007

[17] Anshika Choudhary and Anuja Arora. 2021. ImageFake: An Ensemble Convolution Models Driven Approach for Image Based Fake News Detection. In *2021 7th International Conference on Signal Processing and Communication (ICSC)*. 182–187. https://doi.org/10.1109/ICSC53193.2021.9673192

[18] Limeng Cui and Dongwon Lee. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. arXiv:2006.00885 [cs.SI]

[19] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation *(KDD '20)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3394486.3403092

[20] Fernando Cardoso Durier da Silva, Rafael Vieira, and Ana Cristina Bicharra Garcia. 2019. Can Machines Learn to Detect Fake News? A Survey Focused on Social Media. In *HICSS*.

[21] Dipto Das. 2019. *A MULTIMODAL APPROACH TO SARCASM DETECTION ON SOCIAL MEDIA*. Ph. D. Dissertation.

[22] Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. 2022. GAME-ON: Graph Attention Network based Multimodal Fusion for Fake News Detection.

[23] Yuan Dong, Shan Gao, Kun Tao, Jiqing Liu, and Haila Wang. 2014. Performance Evaluation of Early and Late Fusion Methods for Generic Semantics Indexing. *Pattern Anal. Appl.* 17, 1 (feb 2014), 37–50. https://doi.org/10.1007/s10044-013-0336-8

[24] Ignazio Gallo, Gianmarco Ria, Nicola Landro, and Riccardo La Grassa. 2020. Image and Text fusion for UPMC Food-101 using BERT and CNNs. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. 1–6. https://doi.org/10.1109/IVCNZ51579.2020.9290622

[25] Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal Multi-image Fake News Detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. 647–654. https://doi.org/10.1109/DSAA49011.2020.00091

[26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 6325–6334.

[27] Gisel Bastidas Guacho, Sara Abdali, Neil Shah, and Evangelos E. Papalexakis. 2018. Semi-supervised Content-Based Detection of Misinformation via Tensor Embeddings. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 322–325. https://doi.org/10.1109/ASONAM.2018.8508241

[28] Danna Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chin Lin, Kristen Grauman, Jiebo Luo, and Jeffrey Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. (02 2018).

[29] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. *Captioning Images Taken by People Who Are Blind*. 417–434. https://doi.org/10.1007/978-3-030-58520-4_25

[30] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan. 2021. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems* 117 (2021), 47–58. https://doi.org/10.1016/j.future.2020.11.022

[31] Stefan Helmstetter and Heiko Paulheim. 2018. Weakly Supervised Learning for Fake News Detection on Twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

274–277. https://doi.org/10.1109/ASONAM.2018.8508520

[32] Benjamin D. Horne and Sibel Adali. 2017. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. *CoRR* abs/1703.09398 (2017). arXiv:1703.09398 http://arxiv.org/abs/1703.09398

[33] Drew Hudson and Christopher Manning. 2019. GQA: a new dataset for compositional question answering over real-world images.

[34] Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei Efros. 2018. Fighting Fake News: Image Splice Detection via Learned Self-Consistency.

[35] Md Rafiqul Islam, Shaowu Liu, Wang Xianzhi, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining* 10 (12 2020). https://doi.org/10.1007/s13278-020-00696-x

[36] Vidit Jain, Rohit Kaliyar, Anurag Goswami, Pratik Narang, and Yashvardhan Sharma. 2022. AENeT: an attention-enabled neural architecture for fake news detection using contextual features. *Neural Computing and Applications* 34 (01 2022). https://doi.org/10.1007/s00521-021-06450-4

[37] Ramji Jaiswal, Upendra Pratap Singh, and Krishna Pratap Singh. 2021. Fake News Detection Using BERT-VGG19 Multimodal Variational Autoencoder. In *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. 1–5. https://doi.org/10.1109/UPCON52273.2021.96676

[38] 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* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 795–816. https://doi.org/10.1145/3123266.3123454

[39] S. Jindal, R. Sood, Richa Singh, Mayank Vatsa, and T. Chakraborty. 2020. NewsBag: A multimodal benchmark dataset for fake news detection. , 138- 145 pages.

[40] Quanliang Jing, Di Yao, Xinxin Fan, Baoli Wang, Haining Tan, Xiangpeng Bu, and Jingping Bi. 2021. TRANSFAKE: Multi-task Transformer for Multimodal Enhanced Fake News Detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533433

[41] S. Wang K. Shu, A. Sliva and H. Liu. 2019. Beyond news contents: the role of social context for fake news detection. In *WSDM '19 Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 312–320.

[42] 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* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2915–2921. https://doi.org/10.1145/3308558.3313552

[43] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *CoRR* abs/2005.04790 (2020). arXiv:2005.04790 https://arxiv.org/abs/2005.04790

[44] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision* 123, 1 (may 2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

[45] Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559* (2018).

[46] Rina Kumari and Asif Ekbal. 2021. AMFB: Attention based multimodal Factorized Bilinear Pooling for multimodal Fake News Detection. *Expert Systems with Applications* 184 (2021), 115412. https://doi.org/10.1016/j.eswa.2021.115412

[47] Dana Lahat, T. Adalı, and Christian Jutten. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* 103 (2015), 1449–1477.

[48] Lily Li, Or Levi, Pedram Hosseini, and David A. Broniatowski. 2020. A Multi-Modal Method for Satire Detection using Textual and Visual Cues. *ArXiv* abs/2010.06671 (2020).

[49] Yichuan Li, Bohan Jiang, Kai Shu, and Huan Liu. 2020. MM-COVID: A Multilingual and Multimodal Data Repository for Combating COVID-19 Disinformation. arXiv:2011.04088 [cs.SI]

[50] Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 505–514. https://doi.org/10.18653/v1/2020.acl-main.48

[51] Héctor P. Martínez and Georgios N. Yannakakis. 2014. Deep Multimodal Fusion: Combining Discrete Events and Continuous Signals. In *Proceedings of the 16th International Conference on Multimodal Interaction* (Istanbul, Turkey) *(ICMI '14)*. Association for Computing Machinery, New York, NY, USA, 34–41. https://doi.org/10.1145/2663204.2663236

[52] Nicola Messina, Fabrizio Falchi, Claudio Gennaro, and Giuseppe Amato. 2021. AIMH at SemEval-2021 Task 6: Multimodal Classification Using an Ensemble of Transformer Models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics, Online, 1020–1026. https://doi.org/10.18653/v1/2021.semeval-1.140

[53] 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*. European Language Resources Association, Marseille, France, 6149–6157. https://aclanthology.org/2020.lrec-1.755

[54] Dan Saattrup Nielsen and Ryan McConville. 2022. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. *arxiv* (2022). https://arxiv.org/abs/2202.11684

[55] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. *Improving Fake News Detection by Using an Entity-Enhanced Framework to Fuse Diverse Multimodal Clues*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3474085.3481548

[56] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting Multi-domain Visual Information for Fake News Detection. 518–527. https://doi.org/10.1109/ICDM.2019.00062

[57] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. *Hierarchical Multi-Modal Contextual Attention Network for Fake News Detection*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/3404835.3462871

[58] 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 *ICML*.

[59] Chahat Raj and Priyanka Meel. 2022. ARCNN framework for multimodal infodemic detection. *Neural Networks* 146 (2022), 36–68. https://doi.org/10.1016/j.neunet.2021.11.006

[60] Saed Rezayi, Saber Soleymani, Hamid R. Arabnia, and Sheng Li. 2021. Socially Aware Multimodal Deep Neural Networks for Fake News Classification. In *2021 IEEE 4th International Conference on Multimedia Information*

Processing and Retrieval (MIPR). 253–259. https://doi.org/10.1109/MIPR51284.2021.00048

[61] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. *CoRR* abs/1901.08971 (2019). arXiv:1901.08971 http://arxiv.org/abs/1901.08971

[62] Tanmay Sachan, Nikhil Pinnaparaju, Manish Gupta, and Vasudeva Varma. 2021. SCATE: Shared Cross Attention Transformer Encoders for Multimodal Fake News Detection. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Virtual Event, Netherlands) *(ASONAM '21)*. Association for Computing Machinery, New York, NY, USA, 399–406. https://doi.org/10.1145/3487351.3490965

[63] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. 899–908. https://doi.org/10.1109/BigData52589.2021.9671928

[64] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. DEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 395–405. https://doi.org/10.1145/3292500.3330935

[65] Kai Shu, Ahmed Hassan, Susan Dumais, and Huan Liu. 2020. Detecting Fake News With Weak Social Supervision. *IEEE Intelligent Systems* PP (05 2020), 1–1. https://doi.org/10.1109/MIS.2020.2997781

[66] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2019. Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation. arXiv:1903.09196 [cs.SI]

[67] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.* 19 (2017).

[68] Kai Shu, Suhang Wang, and Huan Liu. 2018. Understanding User Profiles on Social Media for Fake News Detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 430–435. https://doi.org/10.1109/MIPR.2018.00092

[69] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan, Scott Ruston, and Huan Liu. 2020. Leveraging Multi-Source Weak Social Supervision for Early Detection of Fake News.

[70] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019. The Role of User Profiles for Fake News Detection *(ASONAM '19)*. Association for Computing Machinery, New York, NY, USA, 436–439. https://doi.org/10.1145/3341161.3342927

[71] Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management* 58, 5 (2021), 102618. https://doi.org/10.1016/j.ipm.2021.102618

[72] Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021. Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multi-modal Data. (02 2021).

[73] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models that can Read.

[74] Shivangi Singhal, Mudit Dhawan, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2021. Inter-Modality Discordance for Multimodal Fake News Detection. In *ACM Multimedia Asia* (Gold Coast, Australia) *(MMAsia '21)*. Association for Computing Machinery, New York, NY, USA, Article 33, 7 pages. https://doi.org/10.1145/3469877.3490614

[75] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. In *2019 IEEE Fifth International Conference on*

Multimedia Big Data (BigMM). 39–47. https://doi.org/10.1109/BigMM.2019.00-44

[76] Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Inf. Process. Manag.* 58 (2021), 102437.

[77] Chenguang Song, Kai Shu, and Bin Wu. 2021. Temporally evolving graph neural network for fake news detection. *Information Processing & Management* 58, 6 (2021), 102712. https://doi.org/10.1016/j.ipm.2021.102712

[78] Chenguang Song, Kai Shu, and Bin Wu. 2021. Temporally evolving graph neural network for fake news detection. *Information Processing & Management* 58, 6 (nov 2021), 102712. https://doi.org/10.1016/j.ipm.2021.102712

[79] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion* 64 (2020), 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

[80] Inna Vogel and Meghana Meghana. 2020. Detecting Fake News Spreaders on Twitter from a Multilingual Perspective. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (2020), 599–606.

[81] Jingzi Wang, Hongyan Mao, and Hongwei Li. 2022. FMFN: Fine-Grained Multimodal Fusion Networks for Fake News Detection. *Applied Sciences* 12, 3 (2022). https://doi.org/10.3390/app12031093

[82] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. 2015. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*. 1–6. https://doi.org/10.1109/ICMEW.2015.7169757

[83] 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 *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 9 pages. https://doi.org/10.1145/3219819.3219903

[84] Yaqing Wang, Fenglong Ma, Haoyu Wang, Kishlay Jha, and Jing Gao. 2021. Multimodal Emergent Fake News Detection via Meta Neural Process Networks. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021).

[85] Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. *Fake News Detection via Knowledge-Driven Multimodal Graph Convolutional Networks*. Association for Computing Machinery, New York, NY, USA, 540–547. https://doi.org/10.1145/3372278.3390713

[86] Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2021. N24News: A New Dataset for Multimodal News Classification. arXiv:2108.13327 [cs.CL]

[87] Zuhui Wang, Zhaozheng Yin, and Young Anna Argyris. 2021. Detecting Medical Misinformation on Social Media Using Multimodal Deep Learning. *IEEE Journal of Biomedical and Health Informatics* 25, 6 (2021), 2193–2203. https://doi.org/10.1109/JBHI.2020.3037027

[88] Liang Wu, Jundong Li, Xia Hu, and Huan Liu. 2017. Gleaning wisdom from the past: Early detection of emerging rumors in social media. In *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 99–107.

[89] Liang Wu and Huan Liu. 2018. Tracing Fake-News Footprints: Characterizing Social Media Messages by How They Propagate. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (2018).

[90] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manag.* 58 (2021), 102610.

[91] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78. https://doi.org/10.1162/tacl_a_00166

[92] Jiangfeng Zeng, Yin Zhang, and Xiao Ma. 2020. Fake news detection for epidemic emergencies via deep correlations between text and images. *Sustainable Cities and Society* 66 (12 2020), 102652. https://doi.org/10.1016/j.scs.2020.102652

[93] Honghao Zhou, Tinghuai Ma, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. 2022. MDMN: Multi-task and Domain Adaptation based Multi-modal Network for early rumor detection. *Expert Systems with Applications* 195 (2022), 116517. https://doi.org/10.1016/j.eswa.2022.116517

[94] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3205–3212. https://doi.org/10.1145/3340531.3412880

[95] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-Modal Fake News Detection. *ArXiv* abs/2003.04981 (2020).

[96] Xinyi Zhou and Reza Zafarani. 2019. Network-Based Fake News Detection: A Pattern-Driven Approach. *SIGKDD Explor. Newsl.* 21, 2 (nov 2019), 48–60. https://doi.org/10.1145/3373464.3373473

[97] Zhi-Hua Zhou. 2017. A Brief Introduction to Weakly Supervised Learning. *National Science Review* 5 (08 2017). https://doi.org/10.1093/nsr/nwx106