Multi-modal Fake News Detection

Tanmoy Chakraborty

Abstract The primary motivation behind the spread of fake news is to convince the readers to believe false information related to certain events or entities. Human cognition tends to consume news more when it is visually depicted through multimedia content than just plain text. Fake news spreaders leverage this cognitive state to prepare false information in such a way that it looks attractive in the first place. Therefore, multi-modal representation of fake news has become highly popular. This chapter presents a thorough survey of the recent approaches to detect multi-modal fake news spreading on various social media platforms. To this end, we present a list of challenges and opportunities in detecting multi-modal fake news. We further provide a set of publicly available datasets, which is often used to design multi-modal fake news detection models. We then describe the proposed methods by categorizing them through a taxonomy.

Keywords Multi-modal fake news · Multimedia · Microblogs · Supervised methods · Unsupervised methods

1 Introduction

A new article usually gains more visibility when it is accompanied by attractive visuals—images, videos, etc. Human psychology often relates the multi-modal content more to an individual's daily life than a textual content. Therefore, it is not surprising that fraudulent content creators often take advantage of such human cognition of biased multi-modal/multimedia content consumption to design catchy fake news in order to increase overall visibility and reach. Studies revealed that tweets with images receive 18% more clicks, 89% more likes, and 150% more retweets than those without images.[1] Moreover, visual component is frequently

[1] https://www.invid-project.eu/tools-and-services/invid-verification-plugin/.

Deepak P et al., *Data Science for Fake News*, The Information Retrieval Series 42, https://doi.org/10.1007/978-3-030-62696-9_3

considered as a proof of the trustworthiness of the story in our common sense. This is another reason for a multimedia story to attract a large audience.[2] The dissemination of multi-modal fake news is thus even more detrimental than usual unimodal (text only or image only) fake news. Note that a fake image without any caption or description may not have the storytelling capability of a fake image with textual content. For example, an image depicting "a black person is beaten by several white persons" may not be that attractive if it is not accompanied by the associated story, such as where the incident happened (say, New York City) and what was the reason behind the incident (say, the black person challenged the state authorities). Such stories also lead to communal hatred, regional riot, etc. In this chapter, we will cover several recent research that deal with fake news detection by leveraging "multi-modal" or "multimedia" content.

Note that we refrain ourselves from discussing image/video forensics such as forgery, doctoring, or tampering detection [3] as well as fake news detection methods, which leverage only images or videos in isolation. Readers are encouraged to read notable studies in this direction, such as Gupta et al. [18], which made an effort to understand the temporal, social reputation, and influence patterns for the spreading of fake images on microblogs, and Angiani et al. [2], which proposed a supervised method for image-based hoax detection, etc.

Another body of research deals with image repurposing detection, where the task is to detect visual content that is real (not manipulated) but is published together with a false caption about the depicted event. These studies attempt to measure the semantic integrity of images and their corresponding captions using reference resources or knowledge bases [16, 20, 21, 52]. We also purposefully skip them in this chapter because they fall under the study of image caption generation. Captions are often not considered as equivalent news, tweets, or posts. Moreover, these models mostly look at the manipulation of image metadata such as image creation date, owner, location, etc., which are often not publicly available with social media content and online news articles.

We strictly confine our discussion to methods that consider *at least text and visual content* of an article/post for fake news detection. Also note that a "news" can be a social media post such as a "tweet" or an article in a newspaper or blog. Figure 1 shows an example of the type of fake news considered in this chapter.

In 2015, a workshop, called MediaEval,[3] was organized as a satellite event of Interspeech conference,[4] where one of the competitions was "Verifying Multimedia

[2] https://www.businesswire.com/news/home/20190204005613/en/Visual-SearchWins-Text-Consumers%E2%80%99-Trusted-Information.

[3] http://www.multimediaeval.org/mediaeval2015/.

[4] http://interspeech2015.org/.

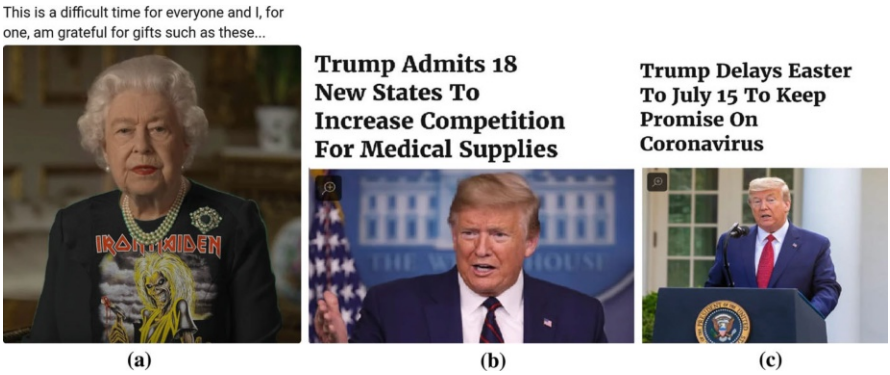

Fig. 1 Three examples of multi-modal fake news. Example (**a**) was picked up from a recent video of Queen Elizabeth regarding the current situation of coronavirus in the United Kingdom. The Queen's dress in the photo was modified, and the caption was changed to create the fake news [61]. Example (**b**) was picked up from a speech of Trump, and the above fake caption was attached to it, which states that the number of states in the United States of America was increased by 18, making the total number of states to 68 due to the current situation of coronavirus [46]. Example (**c**) is another example where the coronavirus situation has been used. A photo of a speech of Trump attached with the fake caption that states that Easter has been postponed to a future date. This is a reference to the other events around the world, which are being postponed due to coronavirus to avoid mass gatherings. Since Easter is a festival, its date cannot be changed [47]

Use (New in 2015!)." The organizers defined the following task:

"Given a tweet and the accompanying multimedia item (image or video) from an event that has the profile to be of interest in the international news, return a binary decision representing verification of whether the multimedia item reflects the reality of the event in the way purported by the tweet."

As a part of the task, the organizers released the MediaEval dataset,[5] which contained ∼400 images used in about ∼20K different tweets in the context of ∼10 events (Hurricane Sandy, Boston Marathon bombings, etc.). This dataset is considered as one of the first multi-modal fake news datasets and has been used extensively for evaluating different models (see Table 3). Three competing teams were shortlisted to present their systems [5]: Middleton [37], Jin et al. [22], and Boididou et al. [6], which achieved 0.83, 0.92, and 0.91 F1-scores, respectively. This was followed by another recent competition hosted jointly by the Institute of Computing Technology, Chinese Academy of Sciences and the Beijing Academy of Artificial Intelligence (BAAI) Research Institute, called MCG-FNews19,[6] where three different tasks were given related to fake news: False News Text Detection, False News Image Detection, and False Multi-modal News Detection.

[5] https://github.com/MKLab-ITI/image-verification-corpus.

[6] https://biendata.com/competition/falsenews/.

Two recent studies are worth mentioning: (i) Volkova et al. [63] explained the multi-modal deceptive news detection models by studying their behavior on a curated Twitter dataset. The authors categorized deceptive news into six classes and defined them: disinformation, propaganda, hoaxes, conspiracies, clickbait, and satire. They empirically showed that although text-only models outperform image-only models, combining both image and text modalities with lexical features performs even better. The authors also developed ErrFILTER,[7] an online interactive tool that allows users to explain model prediction by characterizing text and image traits of suspicious news content and analyzing patterns of errors made by the various models. (ii) Glenski et al. [14] performed fake news detection on a dataset comprising $7M$ posts in a variety of languages—Russian, English, Spanish, German, French, Arabic, Ukrainian, Portuguese, Italian, and unknown. Using a simple framework consisting of user network extractor and text and image feature extractors, they achieved 0.76 F1-score.

Li et al. [33] surveyed various datasets and methods for rumor detection. Cao et al. [10] defined fake news as follows and presented a survey on multi-modal approaches:

Definition 1 "A piece of fake news is a news post that shares multimedia content that does not faithfully represent the event that it refers to."

In this chapter, we start by discussing the major challenges faced by the multi-modal fake news detection models (Sect. 2). Section 3 introduces relevant multi-modal datasets that are often used for fake news detection. Section 4 presents the overview of the tools and techniques used for multi-modal fake news detection, which are further elaborated in Sects. 5–10. Section 11 concludes the chapter with possible future directions.

2 Challenges and Opportunities

The major challenges faced by multi-modal fake news detection methods can be divided into the following categories, based on which the existing methods can be differentiated:

– *Scarcity of Data*: Most of the publicly available datasets are small as human annotation is extremely costly and time consuming. Even if someone manages to employ multiple human annotators, it is extremely challenging to annotate a news as fake or real without knowing its context. For example, an expert in the social media domain may not be able to annotate news related to healthcare.

[7] https://github.com/pnnl/errfilter.

- *Class Imbalance*: The number of instances labelled as "fake" should be significantly smaller than that of the "real" category, thanks to the current online media that are mostly reliable and trustworthy. Therefore, most of the models face difficulties in handling highly skewed classes.
- *Capturing Multiple Modalities*: How to efficiently capture multiple modalities present in a news article is a challenge. Most of the methods extract features from different modalities independently and fuse them to obtain a combined representation of the article. Such methods usually fail to capture the dependency between modalities in the final representation.
- *Novel Fake News*: Fraudulent content creators are continuously adopting intelligent obfuscation strategies to evade quick detection of their story. Therefore, a model trained on an outdated dataset may not be able to spot the newly invented fake news articles.
- *Early Detection*: The effect of a highly damaging fake news may be detrimental to the society. Therefore, it is essential to adopt a strategy to detect fake stories immediately upon their publication. A model that takes into account time-dependent features, such as the number of shares/retweets and the underlying user network properties, may not be able to fulfill this requirement.
- *Explainability*: An additional challenge is to understand why a news is marked as "fake," explaining the root cause and answering the "why" and "how" of the method. Most of the existing methods fail to explain their results.
- *Generalizability*: A model may suffer from three types of problems: (i) *Domain Adaptation*: if it is trained on a healthcare-related fake news dataset, it may not perform equally well on social media posts; (ii) *Entity-Type Adaptation*: if it is trained on short texts such as tweets, it may not be able to generalize well on long news such as blogs or full-length news articles; and (iii) *Geo-location Adaptation*: if it is trained on a news dataset related to the US presidential election, it may not be able to perform well on Indian general elections (as the major sociological issue in the West is "black vs. white," on which the fake stories are often written, whereas in India, it is "Hindu vs. Muslim").

These challenges open up a tremendous opportunity to the research community to solve this problem in an efficient way in terms of both scalability and accuracy.

3 Multi-modal Fake News Datasets

In this section, we briefly describe some of the popular multi-modal fake news datasets. Table 1 presents a brief statistics of the datasets along with the link to obtain them. The datasets are broadly divided into two categories—datasets containing microblog posts (tweets, Weibo posts, Reddit posts, etc.) and datasets containing full-length news articles.

Table 1 Summary of the datasets used by various approaches for multi-modal fake news detection. Datasets are arranged in chronological order of the year of publication. The last column indicates the link where the source code is available

Dataset	Brief description	Entity	Size	Year	Link
MediaEval [5]	Tweets related to events or places along with images	Tweet	Training: ∼5K genuine, ∼7K fake, Test: ∼1.2K genuine, ∼2.5K fake	2015	[7]
Weibo-JIN [23]	Fake tweets from the official rumor busting system of Sina Weibo[a], genuine tweets verified by Xinhua News Agency	Tweet	50,287 tweets (19,762 of them have images attached), 25,953 images, 42,441 distinct users	2016	N.A.[b]
Weibo-att [24]	Human verified false rumor posts, genuine tweets verified by Xinhua News	Tweet	Training (rumor, 3749; real, 3783), Test (rumor, 1000; real, 996)	2017	[25]
Twitter [35]	Keywords were extracted from 530 rumors obtained from snopes.com, and tweets were scraped by queries using the keywords	Tweet	498 events each for rumor and non-rumors, 491,229 users, 1,101,985 tweets	2016	[36]
TI-News [66]	Real news obtained from authoritative news sites (*NYT*, *Washington Post*, etc.) and fake news collected by B.S. Detector[c]	News	8074 real and 11,941 fake news articles	2016	[51]
PHEME [71]	Tweets collected based on the newsworthy events identified by the journalists; news was marked as rumors/real by human annotators	Tweet	1972 rumor and 3830 real tweets	2017	[72]
PolitiFact [55]	A fact-checking website authenticating claims by elected officials. It contains news content, corresponding images, users' retweets/replies, and news profile (source, publisher, and keywords)	News, tweets, replies, users	624 real and 432 fake news, 558,937 users, and 552,698 replies	2018	[56]
Gossip Cop [55]	A fact-checking website for celebrity reporting investigating the credibility of entertainment stories	News, tweets, replies, users	16,817 real and 5323 fake news, 1,390,131 users, 379,996 replies	2018	[56]
Tampered News [40]	News articles written in English and published in 2014 across different domains (sports, politics, etc.); entities are tampered automatically using a tampering technique	News	72,561 news articles	2020	[42]

News400 [40]	News articles crawled from popular German news websites (faz.net, haz.de, and sued-deutsche.de); each news has at least one image and text	News	4000 news articles	2020	[41]
BuzzFeed News [58]	Facebook provided a dataset consisting of political orientation and portal label	News	181 fake and 757 real news articles	2016	[58]
NewsBag [26]	Real (fake) news collected from *The Wall Street Journal* (The Onion); human experts annotated the news	News	200,000 real and 15,000 fake news articles	2020	N.A.[d]
NewBag++ [26]	Augmented version of NewsBag where the size of two classes is balanced	News	200,000 real and 389,000 fake news articles	2020	N.A.[d]
NewsBag Test [26]	Real (fake) news articles scrapped from The Real News (The Poke)	News	11,000 real and 18,000 fake news articles	2020	N.A.[d]
Fakeddit [43]	1M submissions from 22 different subreddits were collected (March 19, 2008–October 24, 2019) and passed through several filtering steps; two-way, three-way, and six-way classification labels	Reddit posts and their metadata	628,501 fake and 527,049 real posts, 682,996 multi-modal posts	2020	[44]

N.A.: Not available

[a] http://service.account.weibo.com/

[b] Authors did not respond to our email

[c] https://github.com/selfagency/bs-detector

[d] Authors informed us that the dataset will be shared upon request

3.1 Fake Microblog Datasets

Each sample of these datasets is relatively small. Two widely used datasets in this category are MediaEval and Weibo-att. Along with them, we also describe a few other datasets that are often being used to detect fake news.

1. **MediaEval:** The dataset was collected as a part of the *Verifying Multimedia Use* task of MediaEval 2015 [5]. It contains tweets related to events or places along with images. A tweet was annotated as "genuine" if the associated image corresponds to the event that the text of the tweet points to; otherwise, it was marked as "fake." Overall, there are 400 images that are used in about ∼20K different tweets in the context of ∼10 events (Hurricane Sandy, Boston Marathon bombings, etc.).
2. **Weibo-JIN:** Jin et al. [23] collected tweets related to diverse events from Weibo. Instead of human annotation, the ground-truth was prepared based on the authenticity of the news sources. Specifically, fake news events were collected from the official rumor busting system of Sina Weibo, and real events were gathered from a hot news detection system of Xinhua News Agency, the official and most authoritative news agency in China, as the main source. From 146 event-related news articles, keywords were extracted based on which tweets were collected from Weibo. This dataset is larger than that of MediaEval.
3. **Weibo-att:** Jin et al. [24] collected false rumors posted from May 2012 to January 2016 from the official rumor debunking system of Weibo. The real tweets were collected from Xinhua News Agency, an authoritative news agency in China. This is one of the highly used datasets in multi-modal fake news detection.
4. **Twitter:** Ma et al. [35] collected 778 verified rumor and real events during March–December 2015 from www.snopes.com. Upon extracting the keywords and iteratively refining them, composite queries were fired on Twitter API. Non-rumor events were collected from some existing datasets [11, 30].
5. **PHEME**: Zubiaga et al. [71] collected this dataset by emulating the scenario in which a journalist is following a story. They hired few expert journalists and kept getting information about the new events. Upon receiving information about a new event, the crawler immediately started collecting tweets related to the event. After preprocessing, the remaining tweets were annotated by the experts based on whether there was any evidence about the trustworthiness of the fact expressed in the tweet or any authoritative source was found. The collected tweets were related to five events—Ferguson unrest, Ottawa shooting, Charlie Hebdo shooting, Sydney siege, and Germanwings plane crash.
6. **Fakeddit**: Nakamura et al. [43] collected 1M submissions from 22 different subreddits posted between March 19, 2008, and October 24, 2019. The dataset contains the title of the submission, images, comments made by the users, other user information, scores, upvote and downvote counts, etc. Around 64% of text comments have accompanying images. Initial quality assessment was done based on the metadata information such as the ratio of upvotes and downvotes, users'

karma score, etc. Second-level assessment was conducted by the experts. A series of preprocessing steps were followed to clean up the subreddit posts before entering the annotation process. The annotation was done in three levels—*two-way*, whether a sample is real or fake; *three-way*, whether a sample is completely real or it is fake and contains text that is true or the sample is fake with false text; and *six-way*, whether a sample is real, satire/parody, misleading content, imposter content, false connection, and manipulated content.

3.2 Fake News Datasets

Each sample of these datasets is relatively large and contains a full-length article. Two widely used datasets in this category are PolitiFact and Gossip Cop [55]. Along with these datasets, we also discuss some other datasets of this type that are often used for fake news detection.

1. **TI-News**: Yang et al. [66] created a collection of news from Megan Risdal and Kaggle, containing 11,941 fake and 8074 real news articles. We call this dataset TI-News. The real news articles were related to well-known authoritative sites such as *The New York Times*, *The Washington Post*, etc. Along with the text and image information, each sample contains the author of the news and the website where it was posted.
2. **PolitiFact and Gossip Cop**: Shu et al. [55] utilized two fact-checking websites, namely, PolitiFact[8] and Gossip Cop.[9] The former accommodates news related to politics, and the latter contains fact-checking stories related to films and entertainment. The ground-truth labels were provided by their expert teams. True news were collected from E! Online,[10] which is a well-known trusted media website for publishing entertainment news pieces. Social contexts were collected by searching Twitter API with the titles of the news articles. Users' responses were also collected for every post. Along with these, spatiotemporal information such as locations (if explicitly provided by the users), timestamps of user engagement, replies, likes, retweets, etc. enriched the dataset.[11]
3. **TamperedNews**: Müller-Budack et al. [40] collected an existing dataset, called BreakingNews [50], which covers $100K$ news related to different domains (sports, politics, healthcare, etc.). They further designed a tampering mechanism such as random replacement of named entities to synthetically generate fake news.

[8] https://www.politifact.com/.

[9] https://www.gossipcop.com/.

[10] https://www.eonline.com/.

[11] PolitiFact and Gossip Cop are combined in FakeNewsNet dataset [54, 55].

4. **News400**: In order to evaluate their model on cross-language datasets, Müller-Budack et al. [40] further created News400, a repository containing news articles from three popular German news websites (faz.net, haz.de, and sueddeutsche.de). The news were published during August 2018–January 2019 and were related to four topics—politics, economy, sports, and travel. Similar tampering mechanism was applied to obtain fake news.
5. **NewsBag**: Jindal et al. [26] created the largest dataset of multi-modal fake news articles. Real and fake news were collected from *The Wall Street Journal* and The Onion,[12] respectively. Several human experts were asked to verify 15,000 articles as fake. However, the number of fake articles was much lesser than the real articles. To make a balanced dataset, the authors further created NewsBag++, comprising $200K$ real and $389K$ fake news by running a data augmentation method on NewsBag. They also created NewsBag Test, a separate dataset for testing the models. This dataset contains $11K$ real news collected from The Real News[13] and $18K$ fake news collected from The Poke.[14]

4 State-of-the-Art Models

Most of the existing models are supervised and follow *fusion technique*—low-level features are extracted from different modalities (text, image, etc.) and combined using various fusion mechanisms, based on which existing models can be divided into three broad categories: early fusion, late fusion, and hybrid fusion [34]. Let $\mathbf{v_m}$ be the low-level feature representation of modality m, and there are M modalities in a post. Semicolon (;) is used to indicate concatenation operation. The three fusion techniques are defined below:

– *Early Fusion*: Low-level features from different modalities are combined (generally through concatenation), and a joint representation is created from the combined features. Next, a single model is trained to learn the correlation and interactions between low-level features of each modality. Let h be the single model and p be the final prediction. Then,

$$p = h([\mathbf{v}_1; \mathbf{v}_2; \cdots, \mathbf{v}_m; \cdots; \mathbf{v}_M])$$

.

– *Late Fusion*: From different modalities, unimodal decisions are obtained using other models. These decisions are then fused with some mechanism (such as averaging, voting, or a learned model). Let h_m be the model for mth modality,

[12]The Onion publishes satirical articles on both real and fictional events. Link: https://www.theonion.com/.

[13]https://therealnews.com/.

[14]https://www.thepoke.co.uk/.

and F is the mechanism used to fuse the decisions as in the early fusion. Then, the final prediction will be

$$p = F([h_1(\mathbf{v}_1); h_2(\mathbf{v}_2); \cdots; h_m(\mathbf{v}_m); \cdots; h_M(\mathbf{v}_M))$$

– *Hybrid Fusion*: It is a combination of early and late fusion. A subset of features is passed through separate models to obtain the unimodal decisions as in the late fusion. These decisions are combined with the remaining features to obtain a combined representation, which is further passed through a single model for the final decision. Let $n, n+1, \cdots, m-1, m$ be the modalities that follow late fusion. Then, the final prediction will be

$$p = h([h_j(\mathbf{v}_j)]_{n \le j \le m}; [\mathbf{v}_i]_{1 \le i \ne j \le M}])$$

There are some methods that follow unsupervised techniques; some other methods follow advanced neural network techniques such as adversarial learning and variational autoencoder. Table 2 provides a brief summary of the state-of-the-art methods for multi-modal fake news detection, and Fig. 2 shows the taxonomy of the methods. Tables 3 and 4 show a comparative analysis of the methods on four widely used datasets. The following sections elaborate on these methods. The methods in each section are arranged in chronological order of the year of publication.

5 Unsupervised Approach

Müller-Budack et al. [40] introduced the task of cross-model consistency verification in real-world news. The idea is to quantify the coherence between image and text. They proposed the first unsupervised approach for multi-modal fake news detection, which we call CCVT (cross-model consistency verification tool).[15] CCVT links every named entity (person, location, and event) extracted from the text to its corresponding image using some reference image database. Then, the consistency between the texts and images present in the post is measured. CCVT is composed of three major components:

– *Extraction of Textual Entities*: CCVT utilizes spaCy [19] to extract the named entities and link them to the Wikidata [9] knowledge base. To extract the context of the text, sapCy is applied to obtain all nouns (general concepts such as politics, sports, actions, etc.). fastText [8] is used to obtain the embedding of each candidate.
– *Extraction of Visual Features*: Multi-task cascaded convolutional network [69] is used to detect faces from images. The feature vector of each face is extracted using DeepFace [53].

[15] Code is available at https://github.com/TIBHannover/cross-modal_entity_consistency.

Table 2 Summary of the methods used for multi-modal fake news detection. Methods are in chronological order of the year of publication

Method	Approach	Entity	Dataset	Year
JIN [23]	Five types of visual features are extracted and combined with textual features; concatenated feature set is fed to classifiers	Tweet	Weibo-JIN	2016
AGARWAL [1]	Augmentation of classification systems with a learning to rank scheme	Tweet	MediaEval	2017
att-RNN [24]	RNN with attention mechanism to fuse features from text, image, and social context	Tweet	Weibo-att, MediaEval	2017
EANN [64]	Event adversarial neural networks, composed of multi-modal feature extractor, event discriminator, and fake news detector	Tweet	Weibo-att, MediaEval	2018
TI-CNN [66]	Explicit text and image features are extracted and combined with the implicit features obtained from the CNNs and combined for the detection	News	TI-News	2018
MVAE [28]	Multi-modal variational autoencoder that uses a bimodal variational autoencoder coupled with a binary classifier	Tweet	Weibo-att, MediaEval	2019
MVNN [49]	An end-to-end neural network to learn representations of frequency and pixel domains simultaneously and effectively fuse them	Tweet	Weibo-att	2019
MKEMN [68]	Multi-modal knowledge-aware network to obtain text, visual, and external knowledge, and an event memory network to capture event-invariant feature	Tweet	Twitter, PHEME	2019
SAME [12]	Triplet (news publisher, user, and news) extraction followed by adversarial learning for detecting a semantic correlation between different modalities and finally incorporation of users' sentiment	News	PolitiFact, Gossip Cop	2019
SpotFake [59]	A concatenation of BERT-based text embedding and VGG-19-based image embedding	Tweet	Weibo-att, MediaEval	2019
SpotFake+ [60]	A transfer learning-based approach by combining XLNet and VGG-19 modules	News	PolitiFact, Gossip Cop	2020
CCVT [40]	An unsupervised approach that measures the consistency of image and text to detect fake news	News	Tampered News, News400	2020
MCE [27]	After obtaining the embedding from each modality, a combined representation is learned to score each news based on its magnitude and consistency	News	MediaEval, BuzzFeed News	2020
SAFE [70]	A fusion model is used to obtain a joint representation of news; two representations are compared to measure their similarity; both of them are combined to obtain final loss	News	PolitiFact, Gossip Cop	2020

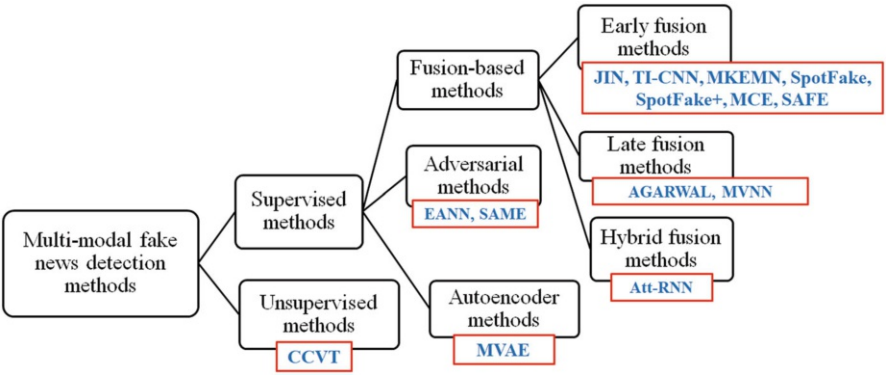

Fig. 2 Taxonomy of the multi-modal fake news detection models with respect to the techniques used for the detection

Table 3 Performance of the multi-modal fake news detection methods, which were evaluated on two popular microblog datasets—MediaEval and Weibo-att. The accuracy corresponding to the best setting of each model was taken from the original paper

Model	MediaEval				Weibo-att			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
JIN	0.898	–	0.835	–	–	–	–	–
att-RNN	0.682	0.78	0.615	0.689	0.788	0.862	0.686	0.764
EANN	0.715	0.822	0.638	0.719	0.827	0.847	0.812	0.829
MVAE	0.745	0.801	0.719	0.758	—	0.689	0.777	0.730
MVNN	–	–	–	–	0.846	0.809	0.857	0.832
MVNN+att-RNN	–	–	–	–	0.901	0.911	0.901	0.906
MVNN+EANN	–	–	–	–	0.897	0.930	0.872	0.900
MVNN+MVAE	–	–	–	–	0.891	0.896	0.898	0.897
SpotFake	0.777	0.751	0.900	0.820	0.892	0.902	0.964	0.932
MCE	0.967	0.875	0.976	0.923	–	–	–	–

Table 4 Performance of the multi-modal fake news detection methods, which were evaluated on two popular news datasets—PolitiFact and Gossip Cop. The accuracy corresponding to the best setting of each model was taken from the original paper

Model	PolitiFact				Gossip Cop			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
SAME	–	–	–	0.772	–	–	–	0.804
SpotFake+	0.846	–	–	–	0.856	–	–	–
SAFE	0.874	0.889	0.903	0.896	0.838	0.857	0.937	0.895

– *Verification of Shared Cross-Model Entities*: The scene contexts extracted from images and texts are compared. First, for each named entity, a set of k images is retrieved from Google/Bing search engine. Second, a denoising step is executed to remove irrelevant images from the set. It is followed by a clustering technique, and the mean feature vector corresponding to the majority cluster serves as the

representative of the queried person. Finally, the feature vectors of all faces in the image are compared to the vector of each person in the text. Similarly, the consistency of locations and events is measured.

CCVT was evaluated on the TamperedNews and News400 datasets to show its efficacy compared to other baselines.

6 Early Fusion Approaches

6.1 JIN

Jin et al. [23] proposed JIN,[16] an early fusion approach to separate fake and real *events* (instead of detecting fake tweets/news). An event is composed of a set of tweets containing certain keywords, which indicate a real incident. The authors observed that given the same number of tweets in events, real events tend to contain more images than fake events. Their major contribution was to come up with five novel visual features:

- *Visual Clarity Score* (VCS): The intuition behind this score is that if a set of images (corresponding to an event) is distinct from the entire collection, then the event is likely to be genuine. First, the local descriptor of each image is extracted. Second, all descriptors are quantized to form a visual word vocabulary. Third, each image is represented by a bag-of-words model. Fourth, two language models are calculated—one from the event and the other from the entire collection. Finally, the "clarity score" is defined as the Kullback–Leibler divergence between two language models.
- *Visual Coherence Score* (VCoS): It measures how coherent images in a certain event are. GIST-based global image descriptor [45] is extracted from each image within an event, and an average similarity of all pairs of images within the event is computed.
- *Visual Similarity Distribution Histogram* (VSDH): For each event, the inter-image similarity is measured between all pairs of images based on VCoS. The similarity scores are divided into 10 bins. For each bin, the normalized number of elements indicates the entry of the feature. Ten features (corresponding to ten bins) are obtained after this step.
- *Visual Diversity Score* (VDS): For every event, images are ranked based on the popularity on social media. For each image, the average dissimilarity score (1-VCS) is then calculated between the image and all the other images ranked higher than the given image. The final VDS score is the average of the VDS scores of all the images in the event.

[16] If there is no explicit name of the method mentioned in the original paper, we use the name of the first author to denote the method.

Text Content
Count of Message, Average Word/Character Length, Fraction of Question/Exclamation Mark, Fraction of Multi Question/Exclamation Mark Ratio, Fraction of First/Second/Third Pronouns, Fraction of URL/@/#, Count of Distinct URL/@/#, Fraction of Popular URL/@/#, Count of Distinct People/Location/Organization, Fraction of People/Location/Organization, Fraction of Popular People/Location/Organization, Average Sentiment Score, Fraction of Positive/Negative Tweets.

User
Count of Distinct Users, Fraction of Popular Users, Average Followers/Followees/Posted Tweets, Fraction of Verified User/Organization.

Propagation
Size of Max Subtree, Average Likes, Average Degree/Non-zero Degree.

Fig. 3 Set of statistical features used by the JIN model

– *Visual Clustering Score* (VCIS): Each image is represented by the bag-of-word model as in VCS. For every event, images are placed in a Euclidean space, and a hierarchical agglomerative clustering (single-link strategy) is used to detect the number of clusters, which constitutes the feature of the event.

JIN also considers 42 statistical features, broadly divided into 3 categories—text content, user, and propagation based (as shown in Fig. 3).

Four classifiers, namely, SVM, Logistic Regression, KStar, and Random Forest, were run on the Weibo-JIN dataset, among which Random Forest was reported to be the best model considering both non-image- and image-based features, achieving 0.83 F1-score.

6.2 TI-CNN

Yang et al. [66] mentioned that the lexical diversity and cognition of the deceivers are totally different from true tellers. Images play a major role in fake news detection. For instance, a fake image is often of low resolution and not correlated with the text. The authors proposed TI-CNN (Text and Image information-based Convolutional Neural Network), which takes explicit user-defined features and implicit CNN-based features and gets trained on both texts and images.

TI-CNN is composed of two major components:

– *Text Feature Extractor*: Several features (such as the length of the news, number of question marks, exclamation, capital letters, etc.) are explicitly extracted from the text and passed through a Fully Connected Layer (FCL). Latent textual features are extracted using a Convolutional Neural Network (CNN). Both of them are concatenated to obtain a combined textual representation.

– *Image Feature Extractor*: Several image features (such as the number of faces, resolution of the image, etc.) are extracted and combined with the latent features obtained from another CNN.

Both these features are further combined and passed through a FCL for the final detection.

TI-CNN outperformed various unimodal classifiers with 0.921 F1-score on the TI-News dataset.

6.3 MKEMN

Zhang et al. [68] argued that along with the text and multimedia, one should also consider the rich knowledge information present in the existing rumor texts, which might often be used for rumor verification. Their proposed method, MKEMM (Multi-modal Knowledge-aware Event Memory Network), utilizes the multi-modal knowledge-aware network to obtain a shared representation of text, existing knowledge, and images (see Fig. 4 for the framework). An Event Memory Network (EMN) is used to obtain event-independent features as suggested in EANN [64] (see Sect. 9). MKEMM attempts to detect whether a claim is a rumor or not, where a claim comprises a sequence of correlated posts with timestamp associated with each post. Two major components of MKEMN are discussed below:

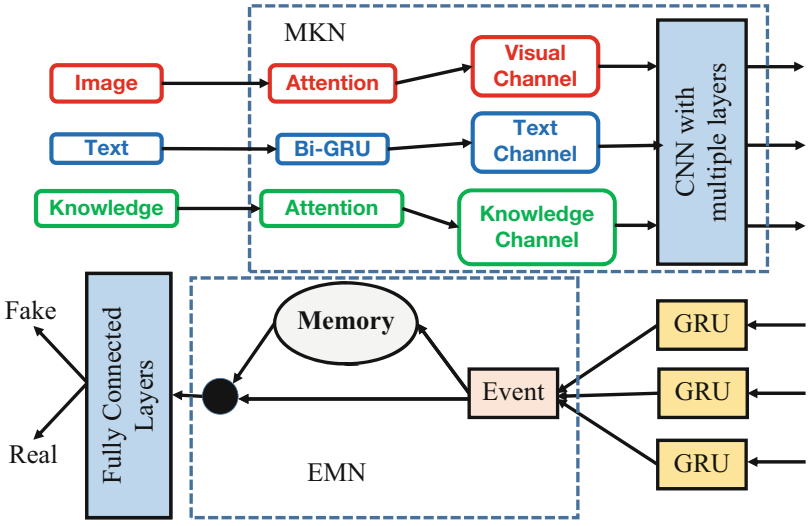

Fig. 4 A simplified visualization of the MKEMN architecture. *Filled circle* indicates concatenation operation

– *Multi-modal Knowledge-aware Network* (MKN): To capture four signals from a post $p = \{w_1, w_2, \cdots, w_n\}$ into the final embedding, four separate modules are designed—(i) *Text Encoder*, which takes a short text and uses a Bi-GRU to obtain a text embedding h_t; (ii) *Knowledge Encoder*, which first extracts entities from a post, then acquires concept information for each entity from existing knowledge graphs [65] and taxonomies [39], and finally obtains a concept knowledge vector k_t for each entity using an attention mechanism; (iii) *Visual Encoder*, which uses VGG-19 [57] to obtain the initial visual representation. A word-guided visual attention model is incorporated, which takes VGG-19 features and Bi-GRU embedding and projects regions that correspond to the highly relevant words to obtain a visual embedding v_t; and (iv) *Multi-modal knowledge-aware CNN*, which, instead of directly concatenating h_t, k_t, and v_t, uses two continuous transformation functions $\mathcal{H}_k(.)$ and $\mathcal{H}_v(.)$ to map k_t and v_t, respectively, to the word space keeping their semantic relation. Finally, a combined representation is obtained as $G = \begin{pmatrix} h_i \\ \mathcal{H}_k(k_i) \\ \mathcal{H}_v(v_i) \end{pmatrix}_{1 \le i \le n}^{3 \times n \times d}$. Afterward, multiple layers with different filters are applied to obtain the final representation of the post.
– *Event Memory Network* (EMN): To obtain event-independent features, EMN first generates an event representation x by passing the MKN embedding of posts related to the event through GRUs and feeding their outputs to a memory, which measures how dissimilar a query event is with respect to the previous events. The output of the memory network is concatenated with x to generate the new event representation X.

The final classification is performed by a deep neural network classifier $z = \mathcal{D}(X)$ using cross-entropy loss.

MKEMM achieved 0.870 and 0.814 F1-scores on the Twitter [35, 36] and PHEME [71] datasets, respectively, and outperformed six baselines including EANN.

6.4 SpotFake and SpotFake+

Singhal et al. [59] argued that existing (adversarial) models [64] are heavily dependent on the secondary tasks performed by the discriminator. An inappropriate choice of the secondary task may deteriorate the performance by up to 10%. The authors proposed SpotFake (Spotting Fake News), a multi-modal early fusion approach to combine texts and images.

– *Textual Feature Extractor*: SpotFake uses BERT [13] to obtain the embeddings of words, which are further concatenated to form the embedding of a sentence.
– *Visual Feature Extractor*: A pretrained VGG-19 model is adopted, and the output of the second last layer is passed through a FCL.

– *Multi-modal Fusion*: The outputs of the above two extractors are concatenated to obtain the final representation.

While comparing with nine baselines including att-RNN [24] (see Sect. 8), EANN [64], and MVAE [28] (see Sect. 10), SpotFake turned out to be outperforming others with 0.82 and 0.932 F1-scores on the Weibo-att and MediaEval datasets, respectively.

Singhal et al. [60] further extended SpotFake to a transfer learning framework and proposed SpotFake+.[17] It leverages a pretrained language transformer (XLNet [67]) and a pretrained ImageNet model (VGG-19) for feature extraction. The authors claimed that SpotFake+ is the first multi-modal approach that performs fake news detection on *full-length articles*. On the PolitiFact and Gossip Cop datasets, SpotFake+ achieved 0.846 and 0.854 F1-scores, respectively, outperforming four baselines including EANN, MVAE, and SpotFake.

6.5 MCE

Kang et al. [27] proposed MCE (Multi-modal Component Embedding) that focuses on the reliability of various multi-modal components and the relationship among them. A vector representation is learned for each modality whose magnitude and direction indicate "reliability" and "consistency." A news will have overall high magnitude if the sum of its component magnitudes is high and all of them are closely aligned (high consistency). MCE learns a latent space such that the magnitude of the real news would be higher than that of fake news. Text-CNN [29] and VGG-19 are used to extract textual and visual features, respectively. For event-related features, multilayer perceptron is used. The final representation of a news is the sum of the representation of its individual components.

MCE was reported to outperform three baselines with 0.9234 and 0.5915 F1-scores, respectively, on the MediaEval and BuzzFeed News datasets.

6.6 SAFE

Zhou et al. [70] also argued to measure the consistency between two modalities and hypothesized that fake news articles tend to contain uncorrelated/dissimilar text and image modalities. Their proposed model SAFE (Similarity-Aware FakE news detection method) attempts to combine the representations of two modalities along with their dissimilarities in an end-to-end framework, which is composed of three components.

[17] https://github.com/shiivangii/SpotFakePlus.

- *Multi-modal Feature Extraction*: Similar to MCE, Text-CNN is used for textual embedding F_t. However, for visual feature extraction, unlike other methods that directly apply pretrained VGG-19, SAFE first uses a pretrained image2sentence model [62] to obtain the initial embeddings that are further fed to a similar Text-CNN framework with an additional FCL to obtain the final visual embedding F_v.
- *Modal-Independent Fake News Prediction*: Two different representations are further concatenated to obtain the final representation, which is passed through a FCL with cross-entropy loss $\mathcal{L}_p$.
- *Cross-Modal Similarity Extraction*: This component independently assumes that texts and images are dissimilar in the case of fake news; thus, a loss can also be computed between the ground-truth and the similarity between two modalities. The similarity between F_t and F_v is computed using a modified cosine similarity measure as follows:

$$\mathcal{M}_s(F_t, F_v) = \frac{F_t F_v + \|F_t\|\|F_v\|}{2\|F_t\|\|F_v\|}$$

  The loss function calculated in this step assumes that news formed by dissimilar texts and images is more likely to be fake and thus is defined as follows:

$$\mathcal{L}_s = y \log(1 - \mathcal{M}_s(F_t, F_v)) + (1 - y) \log \mathcal{M}_s(F_t, F_v)$$

  where $y = 1$ if the article is fake, 0 otherwise.
- *Model Integration and Joint Learning*: The model is jointly trained by combining both the losses: $\mathcal{L} = \alpha\mathcal{L}_p + \beta\mathcal{L}_s$, where α and β balance their corresponding components.

SAFE outperformed seven baselines including att-RNN and models obtained by dropping each modality in isolation from SAFE, by achieving 0.896 and 0.895 F1-scores on the PolitiFact and Gossip Cop datasets.

7 Late Fusion Approaches

7.1 AGARWAL

Agrawal et al. [1] detected fake multimedia tweets containing texts and images. They defined fake news as follows:

Definition 2 "A multimedia news is fake if the multimedia content (image/video) is unrelated to the texts."

The authors proposed a fusion technique (we call it AGARWAL) that concatenates the output of a ranking method with the other features of the tweet entities and

Content features		User Features	
length of tweet	num of words	num of friends	num of followers
contains question mark	contains exclamation mark	follower-friend ratio	num of times listed
num of question marks	num of exclamation marks	user has a URL	user is a verified user
contains happy emoticon	contains sad emoticon	num of tweets	
contains 1st order pronoun	contains 2nd order pronoun		
contains 3rd order pronoun	num of uppercase characters		
num of negative senti words	num of positive senti words		
num of mentions	num of hashtags		
num of URLs	num of retweets		

Fig. 5 Content and user features used by [1, 4] to characterize a tweet entity

feeds the concatenated features into a classifier. The other features of a tweet entity can be broadly categorized into three classes as follows:

– *Image-Based Features*: These features are often used to identify if an image is doctored [5, 17, 32]. The intuition is that a multimedia fake news is generally associated with doctored image(s). The used features are as follows:

  - Probability map of the aligned double JPEG compression
  - Probability map of the nonaligned double JPEG compression
  - Potential primary quantization steps for the first six DCT (discrete cosine transform) coefficients of the aligned double JPEG compression
  - Potential primary quantization steps for the first six DCT coefficients of the nonaligned double JPEG compression
  - Block artifact grid
  - Photo-response nonuniformity.

– *Twitter Content and User-Based Features*: These features (as shown in Fig. 5) are taken from Boididou et al. [4] to capture the social status of users who post the news and the lexicographic properties of tweet texts.
– *Tweet-Based Features*: Doc2vec [31] embedding method is trained on the Sentiment140 corpus [15] to obtain the vector representation of the text. The authors showed that document embedding outperforms n-gram-based features.

Various traditional classifiers (such as SVM, deep neural network, and logistic regression) were trained along with a ranking model. The ranking model was trained in such a way that it prefers genuine tweets more than fake tweets. The ranking model produces a score, which was further used as a feature along with the other features mentioned before. AGARWAL achieved 83.5% unweighted average recall in detecting fake multimedia tweets.

7.2 MVNN

Qi et al. [49] classified fake images into two categories: *tampered images* that have been modified digitally, and *misleading images* that are not modified, but content-wise they are misleading (outdated images used for current events, images taken in one country are used for another country, etc.). They defined fake news as follows:

Definition 3 "Fake news is a post that is intentionally and verifiably false. A fake-news image is an image that is attached to a fake news."

The authors proposed MVNN (Multi-domain Visual Neural Network) that combines frequency and pixel information for fake news detection. It is composed of three modules:

- *Frequency Domain Sub-network*: Discrete cosine transformer (DCT) is used to transfer images from pixel domain to frequency domain. A CNN (three convolutional blocks and a FCL) is used to process the output of DCT and return the final feature representation l_o.
- *Pixel Domain Sub-network*: This module is used to extract the visual features of the input image at the semantic level. A multi-branch CNN network is used to extract multiple levels of features, and a bidirectional GRU (Bi-GRU) network is utilized to model the sequential dependencies between features. The proposed CNN model is composed of four blocks, each having a 3×3 and a 1×1 convolution layer and a max-pooling layer. One CNN block feeds its input to the next CNN block. Furthermore, the outputs of all CNN blocks are fed to a Bi-GRU to obtain a strong dependency between features. The composite representation obtained from GRU is denoted by $L = \{l_1, l_2, l_3, l_4\}$, where l_i is the output of the ith GRU unit.
- *Fusion Sub-network*: All features extracted so far may not contribute equally. For instance, misleading images may not have gone through tampering; therefore, semantic features are more effective than pixel-level features. Fusion sub-network introduces an attention mechanism to weigh individual features.

Finally, the weighted feature vector is passed through a FCL (with cross-entropy loss) to make the final prediction.

Note that MVNN only considers image-related features for fake news detection.[18] It was compared with four baselines, and 0.832 F1-score was reported on the Weibo-att dataset. Furthermore, while the visual feature extraction module of att-RNN (Sect. 8), EANN (Sect. 9), and MVAE (Sect. 10) was replaced by MVNN, it improves the performance of the original methods. The highest accuracy was obtained with att-RNN+MVNN with 0.906 F1-score (see Fig. 3 for a comparative analysis).

[18] Although we avoid any method that solely uses image features for fake news detection, we intentionally add MVNN as it has widely been used as a baseline by other multi-modal fake news detection models. Moreover, it shows significant performance gain when being incorporated into the existing methods (see Table 3).

8 Hybrid Fusion Approach

Jin et al. [24] proposed att-RNN, a multi-modal deep fusion model to leverage multiple modalities present in the tweets (see Fig. 6 for the schematic diagram of att-RNN). It captures the intrinsic relations among three modalities—text, multimedia content (image), and social context (metadata of the tweets). The model intrinsically captures the coherence between these three modalities. The authors hypothesized that images would have certain correlations with text or social context in genuine tweets.

A tweet is represented as a tuple $I = \{T, S, V\}$, where T is the text of the tweet, S is its social context (hashtag topic, mentions and retweets, emotion, sentimental polarity, etc.), and V is the visual content. The model extracts features from each of these modalities to obtain a combined representation. The model follows three steps:

- *Step 1*: The text $T = \{T_1, T_2, \cdots, T_n\}$ and the social context S are fused using an RNN to obtain a joint representation as follows. A pretrained Word2Vec [38] is used to obtain the embedding R_{T_i} of each word T_i in the tweet. The social context vector R_S is passed through a FCL to match the dimension of R_{T_i} and to obtain $R_{S'} = W_{sf} R_S$, where W_{sf} is the weight matrix of a FCL. Next, for each time step (word), an LSTM cell takes $[R_{T_i}; R_{S'}]$ as an input, and the final representation R_{TS} is obtained by averaging the output neurons of all LSTM cells.
- *Step 2*: A visual representation R_V is obtained using deep CNN. The authors used the standard VGG-19 network in the initial layer and added back to back two 512-dimensional FCLs to obtain R_V. In order to capture the correlation between the text/social context and image, a visual attention mechanism is incorporated. From every time step (word) in Step 1, the output hidden state h_i of LSTM is passed through two FCLs (the first FCL with ReLU and the second FCL with softmax function) to obtain the attention vector A_n (of the same dimension as that of R_V). The output of this step is an attention vector $R_{V'}$.

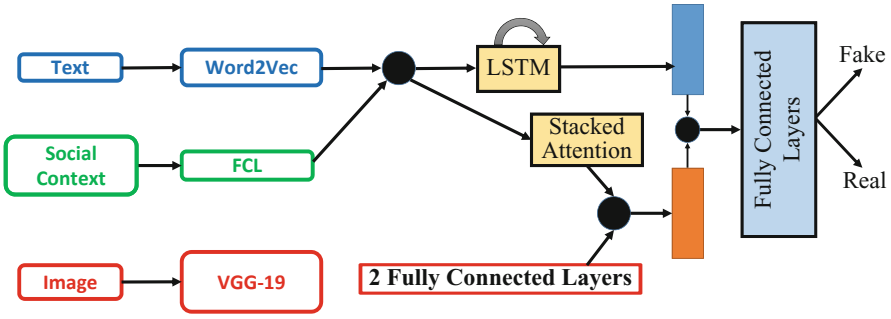

Fig. 6 Schematic diagram describing the architecture of att-RNN. *Filled circle* indicates concatenation operation

– *Step 3*: A combined representation for each tweet is obtained by concatenating R_{TS} and $R_{V'}$: $R_I = [R_{TS}; R_{V'}]$, which is fed to a softmax layer with cross-entropy loss.

The proposed method was evaluated on two datasets—Weibo-att and MediaEval; it achieved 0.764 and 0.689 F1-scores, respectively, for two datasets and outperformed seven baselines (including different variants of att-RNN).

9 Adversarial Model

Wang et al. [64] argued that most of the existing approaches tend to detect event-specific fake news; therefore, they fail miserably in detecting fake news on newly emerged and time-critical events (novel fake news). The proposed method EANN (Event Adversarial Neural Networks) attempts to overcome this problem by learning an event-independent feature representation of every tweet using an adversarial network (see Fig. 7). It consists of three components:

1. *Multi-modal Feature Extractor* (MEF): Text-CNN is used to encode tweet texts. A pretrained vector embedding is used to initialize each word. Multiple filers with various sizes are applied to extract textual features with different granularity. Following this, a FCL is used to ensure the same dimension of the text representation with that of the image representation (discussed below).
   For image-level feature extraction, the same architecture as proposed by [24] was adopted. These two features are then concatenated to form a multi-modal feature R_F.
2. *Fake News Detector* (FND): Given the multi-modal feature R_F, this module uses a FCL with softmax to predict if a post is real or fake. The cross-entropy loss is used to calculate the detection loss L_d.
3. *Event Discriminator* (ED): Given the multi-modal feature R_F, this module uses two FCLs to classify posts into one of the K events. Cross-entropy loss L_e is

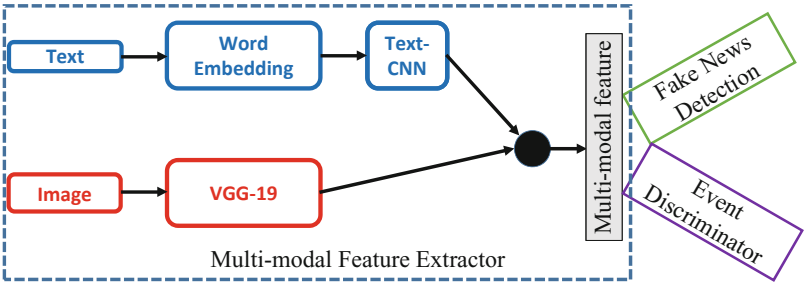

Fig. 7 Schematic diagram describing the architecture of EANN. *Filled circle* indicates concatenation operation

calculated to estimate the dissimilarities between the representations of different events—large loss indicates a similar distribution of the representations of events, which in turn ensures that the resultant representation is event-invariant.

Finally, the model integrator combines the two losses as follows: $L = L_d - \lambda L_e$, where λ balances between two losses. The combined loss ensures that MEF tries to fool ED to achieve event-invariant representations by maximizing $L_e(.)$, whereas ED tries to identify each event by minimizing $L_e(.)$.

On two datasets, namely, MediaEval and Weibo-att, ENVV outperforms six baselines including att-RNN with 0.719 and 0.829 F1-scores, respectively.

9.1 SAME

Cui et al. [12] argued that along with multiple modalities, the views of readers expressed on a particular post also play an important role to detect whether the post is fake or not. Users' viewpoints can be captured by the comments left for the post. The authors statistically validated that users tend to express more sentiment polarity on the comments related to fake news than real news. The proposed model, named SAME (Sentiment-Aware Multi-modal Embedding), consists of three components:

- *Feature Extractor*: To generate the embedding of images, texts, and user profiles, three different networks are designed—a pretrained VGG-19 is used to extract image feature, a pretrained Glove [48] embedding followed by a multilayer perceptron is used to extract text feature, and a two-layer multilayer perceptron is used to extract user profile (represented by a vector of discrete values such as topics) feature. These features are passed through the adversarial network (discussed below) before integrating using a FCL with three hidden units.
- *Adversarial Learning*: In order to bridge the gap between three modalities, an adversarial network is designed. It consists of two modality discriminators for image and profile features—one takes image and text features, and the other takes profile and text features, to discriminate whether the feature corresponds to the image or the profile. Here, the feature extractor acts as a generator.
- *Fake News Detector*: A FCL with cross-entropy loss is used to discriminate a news as fake or real.

SAME achieved 0.772 and 0.804 (macro) F1-scores while comparing with six baselines including EANN on the PolitiFact and Gossip Cop [55] datasets.

10 Autoencoder Model

Qi et al. [49] argued that existing methods [24, 64] do not have any explicit objective function to discover correlations across the modalities. The authors proposed MVAE (Multi-modal Variational Autoencoder) that consists of three modules (see Fig. 8):

– *Encoder*: Two sub-modules are used for encoding texts and images. The encoder architecture is similar to MEF in EANN [64]. Here, instead of using a CNN, the authors used stacked bidirectional LSTM units (Bi-LSTMs). Upon obtaining the embeddings of words from a pretrained word embedding model, the embedding vectors are passed through two Bi-LSTMs, followed by a FCL to get the textual embedding R_T.
The visual encoder is the same as the image-level feature extractor in MEF of EANN, except in this case where two FCLs are used to pass the VGG-19 feature, which outputs a visual embedding R_V.
The concatenated representation $[R_T; R_V]$ is passed through another FCL to obtain two vectors μ and σ, indicating the mean and variance, respectively, of the distribution of the shared representation. The final output of the encoder is a linear combination of μ and σ as follows: $R_m = \mu + \epsilon\sigma$, where ϵ is a random variable sampled from a Gaussian distribution.
– *Decoder*: The decoder module is just the reverse of the encoder. It also has two sub-modules—one for text and the other for image. These sub-modules try to reconstruct the original data from the sampled multi-modal representation. The text decoder takes R_m and passes it through a FCL followed by the stacked Bi-LSTMs to obtain the original text. Similarly, the image decoder passes R_m through two FCLs to reconstruct the image.

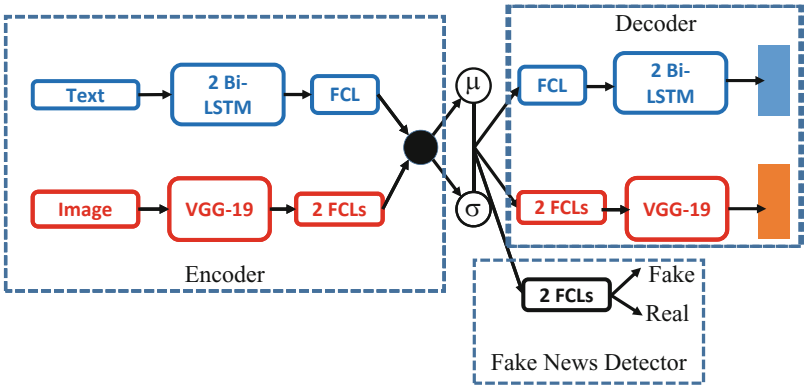

Fig. 8 Schematic diagram showing the flow in the MVAE model. *Filled circle* indicates concatenation operation

– *Fake News Detector*: The shared representation R_m is passed through two FCLs that minimize the cross-entropy loss for a binary classification.

The proposed VAE model and the fake news detector are trained jointly, and the combined loss is minimized in an end-to-end setting.

MVAE was evaluated on two datasets, Weibo-att and MediaEval, and compared with six baselines, including different variants of the original model, att-RNN and EANN. EANN outperforms all the baselines with 0.730 (MediaEval) and 0.837 (Weibo-att) F1-scores.

11 Summary of the Chapter

This chapter presented the current research on multi-modal fake news detection. We introduced various challenges that the existing methods deal with, which further open up opportunities for further research. We also summarized major datasets that are being used for multi-modal fake news detection. While summarizing the methods, we observed that

– Most of the methods adopted multi-modal fusion techniques, and feature-level fusion was incorporated at different positions of the architecture.
– MVNN as an image feature extractor turned out to be highly efficient, improving the performance of most of the methods significantly (Table 3).
– MAVE, although presents a completely different model paradigm, does not seem to be as effective as other fusion-based models.
– BERT-based embedding for text representation shows significant improvement in SpotFake.

We observed that there is still a scarcity of research on multi-modal approaches for large texts such as full-length news articles, blogs, etc. We also noticed that most of the methods have not been shown to be generalized across datasets of diverse domains. Model explainability is the other property that has not been addressed in any of the studies. Other modalities such as videos and audios should also be considered for fake news detection as these modalities are even more powerful and can easily communicate the story to the society.

Acknowledgment The author would like to acknowledge the support of Sarah Masud in writing the chapter.

References

1. Agrawal, T., Gupta, R., Narayanan, S.: Multimodal detection of fake social media use through a fusion of classification and pairwise ranking systems. In: 2017 25th European Signal Processing Conference (EUSIPCO), pp. 1045–1049. IEEE, New York (2017)

2. Angiani, G., Balba, G.J., Fornacciari, P., Lombardo, G., Mordonini, M., Tomaiuolo, M.: Image-based hoax detection. In: Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good, pp. 159–164 (2018)
3. Arun Anoop, M.: Image forgery and its detection: a survey. In: 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1–9 (2015)
4. Boididou, C., Papadopoulos, S., Kompatsiaris, Y., Schifferes, S., Newman, N.: Challenges of computational verification in social multimedia. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 743–748 (2014)
5. Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Riegler, M., Kompatsiaris, Y., et al.: Verifying multimedia use at mediaeval 2015. MediaEval **3**(3), 7 (2015a)
6. Boididou, C., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Kompatsiaris, Y.: The certh-unitn participation@ verifying multimedia use 2015. In: MediaEval (2015b)
7. Boididou, C., Papadopoulos, S., Kompatsiaris, Y., Schifferes, S., Newman, N.: Challenges of computational verification in social multimedia (2015). http://www.multimediaeval.org/mediaeval2015/verifyingmultimediause/index.html
8. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. Assoc. Comput. Linguist. **5**, 135–146 (2017)
9. Brank, J., Leban, G., Grobelnik, M.: Semantic annotation of documents based on wikipedia concepts. Informatica **42**(1), 23–32 (2018)
10. Cao, J., Qi, P., Sheng, Q., Yang, T., Guo, J., Li, J.: Exploring the role of visual content in fake news detection (2020). Preprint. arXiv:200305096
11. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684 (2011)
12. Cui, L., Wang, S., Lee, D.: Same: sentiment-aware multi-modal embedding for detecting fake news. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 41–48 (2019)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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, vol. 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN (2019). https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
14. Glenski, M., Ayton, E., Mendoza, J., Volkova, S.: Multilingual multimodal digital deception detection and disinformation spread across social platforms (2019). Preprint. arXiv:190905838
15. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N project report, Stanford 1(12) (2009)
16. Goebel, M., Flenner, A., Nataraj, L., Manjunath, B.: Deep learning methods for event verification and image repurposing detection. Electron. Imag. **2019**(5), 530–531 (2019)
17. Goljan, M., Fridrich, J., Chen, M.: Defending against fingerprint-copy attack in sensor-based camera identification. IEEE Trans. Inf. Foren. Secur. **6**(1), 227–236 (2010)
18. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 729–736 (2013)
19. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings. In: Convolutional Neural Networks and Incremental Parsing (2017)
20. Jaiswal, A., Sabir, E., AbdAlmageed, W., Natarajan, P.: Multimedia semantic integrity assessment using joint embedding of images and text. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 1465–1471 (2017)
21. Jaiswal, A., Wu, Y., AbdAlmageed, W., Masi, I., Natarajan, P.: AIRD: adversarial learning framework for image repurposing detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11330–11339 (2019)

22. Jin, Z., Cao, J., Zhang, Y., Zhang, Y.: MCG-ICT at mediaeval 2015: verifying multimedia use with a two-level classification model. In: MediaEval (2015)
23. Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q.: Novel visual and statistical image features for microblogs news verification. IEEE Trans. Multimedia **19**(3), 598–608 (2016)
24. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM international conference on Multimedia, pp. 795–816 (2017)
25. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Weibo dataset: Multimodal fusion with recurrent neural networks for rumor detection on microblogs (2017). http://mcg.ict.ac.cn/mm2017dataset.html
26. Jindal, S., Sood, R., Singh, R., Vatsa, M., Chakraborty, T.: Newsbag: a benchmark multimodal dataset for fake news detection. In: Proceedings of the Workshop on Artificial Intelligence Safety, co-located with 34th AAAI Conference on Artificial Intelligence, SafeAI@AAAI 2020, New York City, NY, February 7, 2020, CEUR Workshop Proceedings, AAAI, vol. 2560, pp. 138–145 (2020)
27. Kang, S., Hwang, J., Yu, H.: Multi-modal component embedding for fake news detection. In: 2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM), pp. 1–6. IEEE, New York (2020)
28. Khattar, D., Goud, J.S., Gupta, M., Varma, V.: MVAE: multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference, pp. 2915–2921 (2019)
29. Kim, Y.: Convolutional neural networks for sentence classification (2014). Preprint. arXiv:14085882
30. Kwon, S., Cha, M., Jung, K., Chen, W., Wang, Y.: Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th International Conference on Data Mining, pp. 1103–1108. IEEE, New York (2013)
31. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196 (2014)
32. Li, W., Yuan, Y., Yu, N.: Passive detection of doctored jpeg image via block artifact grid extraction. Signal Process. **89**(9), 1821–1829 (2009)
33. Li, Q., Zhang, Q., Si, L., Liu, Y.: Rumor detection on social media: datasets, methods and opportunities (2019). Preprint. arXiv:191107199
34. Liu, K., Li, Y., Xu, N., Natarajan, P.: Learn to combine modalities in multimodal deep learning (2018). Preprint. arXiv:180511730
35. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M.: Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16) (2016)
36. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.-F., Cha, M.: Twitter dataset (2016). http://alt.qcri.org/wgao/data/rumdect.zip
37. Middleton, S.: Extracting attributed verification and debunking reports from social media: mediaeval-2015 trust and credibility analysis of image and video (2015)
38. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013). Preprint. arXiv:13013781
39. Miller, G.A.: Wordnet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
40. Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., Ewerth, R.: Multimodal analytics for real-world news using measures of cross-modal entity consistency (2020a). Preprint. arXiv:200310421
41. Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., Ewerth, R.: News400 dataset (2020b). https://github.com/TIBHannover/cross-modal_entity_consistency
42. Müller-Budack, E., Theiner, J., Diering, S., Idahl, M., Ewerth, R.: Tamperednews dataset (2020c). https://github.com/TIBHannover/cross-modal_entity_consistency
43. Nakamura, K., Levy, S., Wang, W.Y.: r/Fakeddit: a new multimodal benchmark dataset for fine-grained fake news detection (2019). Preprint. arXiv:191103854
44. Nakamura, K., Levy, S., Wang, W.Y.: Fakeddit dataset (2020). https://github.com/entitize/fakeddit

45. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**(3), 145–175 (2001)
46. ONION T: Trump admits 18 new states to increase competition for medical supplies (2020a). https://www.thepoke.co.uk/2020/04/08/queens-dress-perfect-green-screen-hilariously-exploited/
47. ONION T: Trump delays Easter to July 15 to keep promise on coronavirus (2020b). https://politics.theonion.com/trump-delays-easter-to-july-15-to-keep-promise-on-coron-1842566559
48. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
49. Qi, P., Cao, J., Yang, T., Guo, J., Li, J.: Exploiting multi-domain visual information for fake news detection (2019). Preprint. arXiv:190804472
50. Ramisa, A., Yan, F., Moreno-Noguer, F., Mikolajczyk, K.: Breakingnews: article annotation by image and text processing. IEEE Trans. Patt. Anal. Mach. Intell. **40**(5), 1072–1085 (2017)
51. Risdal, M.: Ti-news dataset (2016). https://www.kaggle.com/mrisdal/fake-news
52. Sabir, E., AbdAlmageed, W., Wu, Y., Natarajan, P.: Deep multimodal image-repurposing detection. In: Proceedings of the 26th ACM international conference on Multimedia, pp. 1337–1345 (2018)
53. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
54. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet dataset (2018a). https://github.com/KaiDMML/FakeNewsNet
55. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: a data repository with news content, social context and dynamic information for studying fake news on social media (2018b). Preprint. arXiv:180901286
56. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Politifact dataset (2018c). https://github.com/KaiDMML/FakeNewsNet
57. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). Preprint. arXiv:14091556
58. Singer-Vine, J.: Buzzfeednews dataset (2016). https://github.com/BuzzFeedNews/2016-10-facebook-fact-check
59. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: Spotfake: a multi-modal framework for fake news detection. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). IEEE, New York, pp. 39–47 (2019)
60. Singhal, S., Kabra, A., Sharma, M., Shah, R.R., Chakraborty, T., Kumaraguru, P.: Spotfake+: a multimodal framework for fake news detection via transfer learning. In: AAAI (Student Abstract), pp. 1–2 (2020)
61. The POKE: The queen's dress made the perfect green screen and people hilariously exploited it (2020). https://politics.theonion.com/trump-admits-18-new-states-to-increase-competition-for-1842708962
62. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans. Patt. Anal. Mach. Intell. **39**(4), 652–663 (2016)
63. Volkova, S., Ayton, E., Arendt, D.L., Huang, Z., Hutchinson, B.: Explaining multimodal deceptive news prediction models. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, pp. 659–662 (2019)
64. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J.: 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, pp. 849–857 (2018)

65. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: a probabilistic taxonomy for text understanding. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 481–492 (2012)
66. Yang, Y., Zheng, L., Zhang, J., Cui, Q., Li, Z., Yu, P.S.: Ti-CNN: Convolutional neural networks for fake news detection (2018). Preprint. arXiv:180600749
67. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: generalized autoregressive pretraining for language understanding (2019). http://arxiv.org/abs/1906.08237. Cite arxiv:1906.08237. Comment: Pretrained models and code are available at https://github.com/zihangdai/xlnet
68. Zhang, H., Fang, Q., Qian, S., Xu, C.: Multi-modal knowledge-aware event memory network for social media rumor detection. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 1942–1951 (2019)
69. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
70. Zhou, X., Wu, J., Zafarani, R.: Safe: similarity-aware multi-modal fake news detection (2020). Preprint. arXiv:200304981
71. Zubiaga, A., Liakata, M., Procter, R.: Exploiting context for rumour detection in social media. In: Ciampaglia, G.L., Mashhadi, A., Yasseri, T. (eds.) Social Informatics. Springer International Publishing, Cham, pp. 109–123 (2017a)
72. Zubiaga, A., Liakata, M., Procter, R.: Pheme dataset (2017b). https://github.com/azubiaga/pheme-twitterconversation-collection