Social Network Analysis and Mining (2023) 13:101
https://doi.org/10.1007/s13278-023-01104-w

ORIGINAL ARTICLE

Multimodal fake news detection on social media: a survey of deep learning techniques

Carmela Comito[1] · Luciano Caroprese[2] · Ester Zumpano[3]

Received: 28 February 2023 / Revised: 9 May 2023 / Accepted: 14 July 2023 / Published online: 1 August 2023

Abstract
The escalation of false information related to the massive use of social media has become a challenging problem, and significant is the effort of the research community in providing effective solutions to detecting it. Fake news are spreading for decades, but with the rise of social media, the nature of misinformation has evolved from text-based modality to visual modalities, such as images, audio, and video. Therefore, the identification of media-rich fake news requires an approach that exploits and effectively combines the information acquired from different multimodal categories. Multimodality is a key approach to improving fake news detection, but effective solutions supporting it are still poorly explored. More specifically, many different works exist that investigate if a text, an image, or a video is fake or not, but effective research on a real multimodal setting, 'fusing' the different modalities with their different structure and dimension is still an open problem. The paper is a focused survey concerning a very specific topic which is the use of deep learning (DL) methods for multimodal fake news detection on social media. The survey provides, for each work surveyed, a description of some relevant features such as the DL method used, the type of analysed data, and the fusion strategy adopted. The paper also highlights the main limitations of the current state of the art and draws some future directions to address open questions and challenges, including explainability and effective cross-domain fake news detection strategies.

Keywords Fake news · Deep learning · Social media

1 Introduction

The world is highly connected and ideas easily spread in it. Moreover, the easy access to social media platforms has greatly increased so that offering the possibility to produce and share information, ideas, and emotions in different forms such as text, video, audio, and images. The freedom to share and access content without cost and supervision has surely positive implications, but it has also led to the consequent spread of low-quality news and false news, referred to as *fake news*. Inaccurate and fake information is often intentionally posted online by malicious users in order to manipulate public emotions, influence people's thoughts and actions, damage a group or a community, generate confusion, and gain profits through misinformation.

Fake news is misleading and difficult to catch by humans, but also by artificial intelligence (AI) algorithms, as often it combines both false and real information. The propagation of false information through social media has negative effects on many different aspects of social life. Let's think, for example, of the spread of fake news related to COVID-19 like the one related to chloroquine drug overdose as an effective treatment to fight the epidemic. As another example scenario, some studies support the claim that US Presidential election in 2016 has been influenced by the spread of fake news on many different social media platforms (Bovet and Makse 2019).

The term fake news can be declined in different forms: (i) *misinformation* describes fake content produced without a specific reason; (ii) *disinformation* describes fake

✉ Luciano Caroprese
luciano.caroprese@unich.it

Carmela Comito
carmela.comito@icar.cnr.it

Ester Zumpano
e.zumpano@dimes.unical.it

[1] ICAR-CNR, Rende, Italy

[2] DIMES, University of Calabria, Arcavacata, Italy

[3] INGEO, University G. d'Annunzio of Chieti-Pescara, Pescara, Italy

content produced for a specific reason; (iii) *malinformation* describes fake content that is deliberately produced to harm others.

The widespread diffusion of fake news on social media is a challenging problem. The research community is devoting great attention to the topic, putting in place important efforts to provide effective fake news detection solutions.

Early works on the fake news detection topic just rely on textual content. Anyhow, even if it is undoubtedly necessary to analyse news content in order to obtain a good indicator for detecting misinformation, it is clear that only textual analysis is not sufficient. Post and online articles contain not only textual information but also audio, images, and video, and misinformation can, therefore, spread through different modalities. Many different sophisticated tools exist to produce fake images or fake videos so that attracting users' attention and thus be shared. The term *deepfakes* refers to the use of deep learning tools to create manipulated images and video that spread easily and quickly with respect to text (Zannettou et al. 2018; Hameleers et al. 2020; Li and Xie 2020).

Revealing a fake image involves an accurate analysis of the features related to the image, its associated caption, and the relationship between the image and the caption. Revealing a fake video implies, among others, a detailed analysis of the features related to the images, the sounds, and the narrative associated to the video.

Thus far, besides textual information, it is important to exploit and correctly combine information acquired from images and audio in order to detect fake news. Multimodality is the real key point to properly address the misinformation detection challenge. However, the results obtained by the research community are not yet very effective. More specifically, many different works exist that investigate if a text, an image, or a video is fake or not, but effective research on a real multimodal setting, 'fusing' the different modalities with their different structure and dimension, also including the news propagation network, is still an open problem.

This paper surveys the recent literature covering various aspects of multimodality, like the news propagation network, text, image, audio and video, and discusses the fusion strategies proposed in the literature to merge the different modalities for fake news detection.

1.1 Scope of the survey and contributions

The goal of this review is to offer a complete overview of deep learning techniques for multimodal fake news detection on social media. The proposal investigates and discusses an extensive collection of papers published in recent years with the purpose of highlighting how deep learning can help to fight fake news. In particular, the review has been conducted by trying to answer the following questions: (i) Which are the deep learning methods used to detect fake news? (ii) What is the effectiveness of such methodologies?

The paper explores these questions, with the following contributions:

- It is a focused survey exploring the specific topic of multimodal fake news detection with the lens of deep learning techniques. In fact, even if there are several useful surveys on fake news detection (Shu et al. 2017; Kumar and Shah 2018; da Silva et al. 2019), only a few of them focus on multimodal strategies (Alam et al. 2022; Abdali 2022; Hangloo and Arora 2022) and even a smaller number of them is restricted to the use of deep learning methods (Hangloo and Arora 2022). Therefore, the final purpose of this survey is to undertake a complete analysis of multimodal fake news detection by considering only the recent advancements in artificial intelligence brought by deep neural networks based solutions.
- It provides, for each work, an analysis of the rationale behind the approach, highlighting some relevant features such as the method used, the type of analysed data, the fusion strategy adopted, and the results achieved.
- It summarizes the main research contributions related to the role of deep learning for multimodal fake news detection on social media by reporting in Table 1 the main characteristics of state-of-the-art methods, to guide the reader through the key results of the relevant primary research outcomes.
- It discusses the main limitations of the current approaches and the challenges that remain to be addressed by future research works, including explainability and effective use of cross-domain fake news detection strategies.

2 Multimodal fake news detection on social media

Misinformation frequently emerges as textual content. The Internet and social media, however, enable the use of several modalities, which can make a misinformation message intriguing in addition to detrimental. For instance, a meme or a video is much simpler to digest, gets a great deal more interest, and disseminates farther than basic text. In this section, we first introduce the multimodal fake news detection problem in the social media setting and after that, we report the main differences with already published surveys on this topic.

2.1 Problem formulation and key concepts

The *multimodal fake news detection* problem refers to the classification of news with respect to its adherence to real

Table 1 Summary of methods

Publication	Method	Data types	Datasets	Data fusion	Cross-domain
SpotFake (Singhal et al. 2021)	BERT, CNN (VGG-19)	Textual, Images	MediaEval, Weibo A	Concatenation	No
EXMULF (Amri et al. 2021)	LDA, BERT, CNN	Textual, Images	Twitter, Weibo A	Concatenation	No
Alonso-Bartolome and Segura-Bedmar (2021)	CNN	Textual, Images	Fakeddit	Concatenation	No
EM-FEND (Qi et al. 2021)	BERT, CNN (VGG-19), Entity detector, OCR model	Textual, Images	Yang dataset, Weibo A	Co-attention transformer	No
EANN (Wang et al. 2018)	CNN, VGG-19	Textual, Images	MediaEval, Weibo A	Concatenation	Event classifier with Adversarial NN
MetaFEND (Wang et al. 2021)	Meta-learning, CNN, VGG-19	Textual, Images	MediaEval, Weibo A	Concatenation, Attention Mechanism	Meta-learning, Neural Process Network
MVAE (Khattar et al. 2019)	Bi-LSTM, VGG-19	Textual, Images	MediaEval, Weibo A	Variational Autoencoder	No
SAFE (Zhou et al. 2020b)	CNN, Word embedding, Image2sentence	Textual, Images	PolitiFact, GossipCop	Concatenation and multi-loss	No
MCNN (Xue et al. 2021)	BERT, Bi-GRU, ResNet50, Attention	Textual, Images	MCG-FNeWS, PolitiFact, MediaEval, Yang dataset	Attention, multi-loss	No
att-RNN (Jin et al. 2017)	Bi-LSTM, Word embedding, VGG-19	Textual, Images, Social	Weibo A, MediaEval	Concatenation, Neuron-level attention	No
MKEMN (Zhang et al. 2019)	Bi-GRU, Word Embedding, Concept Embedding, VGG-19	Textual, Knowledge (semantics)	Ma dataset, PHEME	Attention, multi-channel CNN	Event memory network
BDANN (Zhang et al. 2020)	BERT, CNN (VGG-19)	Textual, Images	MediaEval, Weibo A	Concatenation	Domain Adaptation with Adversarial NN Layer)
AIFN (Wu and Rao 2020)	BERT, Bi-LSTM, GAIN	Textual, Emoticons	MediaEval, Weibo A	Concatenation	Semantic-level fusion, self-attention networks (SFSN)
CARN-MCN (Song et al. 2021)	Word Embedding, VGG-19, self-attention residual network	Textual, Comments, Images	MediaEval, Weibo A, Weibo B	Cross-modal Attention Residual (CARN), Multi-channel CNN (MCN)	No
Silva et al. (2021)	BERT, supervised network representation learning, FF-NN	Textual, Propagation Network	PolitiFact, GossipCop, CoAID	Concatenation	Unsupervised Multimodal Domain Discovery, domain-specific and Cross-domain embedding
REAL-FND (Mosallanezhad et al. 2022)	BERT, RNN Hierarchical Attention Network, FF-NN	Textual, Comments, User-news interactions	PolitiFact, GossipCop	Concatenation, FF-Neural Network	Reinforcement Learning-based Domain Adaptation
Rezayi et al. (2021)	GloVe Embedding, LSTM, FF-NN	Textual, Social, Network	PHEME, Volkova dataset	Concatenation, FF-Neural Network	No
dEFEND (Shu et al. 2019a)	RNN, GRU, co-Attention	Textual, Comments	PolitiFact, GossipCop	Concatenation, Sentiment-comment co-attention	No
SCATE (Sachan et al. 2021)	BERT, VGG-19, Dot product Transformer Attention	Textual, Images	MediaEval, Weibo A, Weibo B	Bilinear Pooling, Dot product Attention	No

Table 1 (continued)

Publication	Method	Data types	Datasets	Data fusion	Cross-domain
AMFB (Kumari and Ekbal 2021)	Attention-based stacked Bi-LSTM, Attention-based multi-level CNN-RNN	Textual, Images	MediaEval, Weibo A	Multimodal Factorized Bilinear Pooling	No
TRANSFAKE (Jing et al. 2021)	BERT, Faster-RCNN, Sequence position embedding, Segment embedding	Textual, Comments, Images	PolitiFact, GossipCop, Weibo A	Multi-layer bidirectional Transformer	No
GCAN (Lu and Li 2020)	GCN, GRU, CNN	Textual, User features, Social, Propagation Network	Ma dataset	Dual co-attention mechanism, pooling	No
FauxBuster (Zhang et al. 2018)	deep autoencoding	Textual, Images	Reddit, Twitter	Concatenation	No
Krishnamurthy et al. (2018)	CNN, 3D-CNN, OpenSMILE	Textual, Images, Audio, Video	Dataset by Pérez-Rosas et al. (2015)	Concatenation, Hadamard	No
Giachanou et al. (2020)	BERT, VGG-19, LSTM	Textual, Images	PolitiFact, GossipCop	Concatenation	No
Kaliyar et al. (2020)	LSTM	Textual, User-relationship, Social	Fakeddit, BuzzFeed	Concatenation	No
Kirchknopf et al. (2021)	BERT, CNN, ResNet-v2, ResNet101-v2, Inception-v3	Textual, Visual, Comments, Metadata	Fakeddit	Concatenation	No
SERN (Xie et al. 2021)	BERT, ResNet-152, MLP	Textual, Images	PHEME, Fakeddit	Concatenation	No
DEV (Karimi et al. 2018)	CNN, LSTM	Audio, Video	Dataset by Pérez-Rosas et al. (2015)	Concatenation	No
SAME (Cui et al. 2019)	VGGNet, CNN	Textual, Social Content	PolitiFact, GossipCop	Concatenation	No
Mendels et al. (2017)	lexical BLSTM, MFCC BLSTM, DNN-openSMILE	Audio	Columbia X-Cultural Deception Corpus (CXD)	Concatenation	No
DUAL (Dong et al. 2018)	BGRU, DNN	Textual, Social	LIAR, Buzzfeed News	Concatenation	No
CSI (Ruchansky et al. 2017)	RNN	Textual, Social	Twitter and Weibo	Concatenation	No
Jiang et al. (2019)	CNN, Bi-LSTM	Textual, Social	BuzzFeed, PolitiFact	Late Fusion	No
ConvNet (Raj and Meel 2021)	CCN	Textual, Images	Yang dataset, EMERGENT, MICC-F220		No
TTEC (Hua et al. 2023)	BERT, CCN, Data Augmentation, Contrastive learning	Textual, Images	ReCOvery (Zhou et al. 2020a)	Concatenation	No
UPFD (Dou et al. 2021)	BERT, GNN, NN	Textual, Propagation Graph, Timeline	GossipCop, PolitiFact	Concatenation	No
TriFN (Shu et al. 2019)	Non negative matrix factorization (NMF)	Textual, Users, User-news Interactions, Publisher-news Interactions	BuzzFeed, PolitiFact	Concatenation	No
	No				
SSDL (Mu et al. 2023)	NN, Distillation Knowledge	Textual, Image	NewsCLIPpings	Cross-Modal consistency	No
DGExplain (Shang et al. 2022)	Bi-GRU, LSTM	Textual, Image, User Comments	ReCOVery, MMCoVaR	Cross-Modal consistency	No

Table 1 (continued)

Publication	Method	Data types	Datasets	Data fusion	Cross-domain
TikTec (Shang et al. 2021)	GloVe embedding, Bi-GRU, RCNN	Textual, Audio, Video	Dataset by TikToc	Concatenation	No
Wang et al. (2022)	BERT, WordPiece, MLP	Textual, Video	Dataset by Twitter	Concatenation	No
Mittal et al. (2020)	Siamese Network	Textual, Video, Audio	DFDC, DF-TIMIT	Concatenation	No

facts, carried out by analysing the different parts of their information content, which are usually in different formats. News information formats, also called *modalities* or *features*, are the following:

- *Text* Text is a key part of most news. Text classification is a complex process that requires the analysis of its syntactic, semantic, and stylistic aspects. Furthermore, the text appears in different parts of news: (i) title, (ii) description, (iii) links to other digital content (news, web pages, videos, etc.), (iv) comments from other users.
- *Social features* News spreading on online platforms (social networks, blogs, online newspapers) usually report users *reaction* to the messages posted, sharing, or expressing an appreciation (*like, emotions, comments*).
- *Audio* Audio is often included in videos but can also be self-contained information content. Think for example of podcasts, broadcast networks, radio services, and audio files included in the news.
- *Video* Video content is increasingly included in the news for its high ability to attract the attention of the public. Videos can be extracted from longer-duration sequences (news, documentaries, films, etc.) or captured with mobile devices. The contents of many social media are almost exclusively video-based (e.g. YouTube and TikTok).
- *Images* Images are often included in the news. They can be captured with mobile devices or extracted from video sequences or from other digital content.
- *Users* This category includes information about the user that creates the article, in terms of his/her credibility/reputation, connections in the network, previously created news, and so on.
- *Network and propagation features* The social network context of news refers to the network characteristics and how the news is propagated via social media, and it is an additional criterion for distinguishing fake news from authentic ones. The propagation feature capture information on the propagation of a news, such as the number of replies, and retweets of an article. The propagation graph of news can be represented as a tuple $G = \langle N, E, X \rangle$, where nodes N represent the tweets/retweets of the news and the edges E represent the retweet relationships among them. X is the set of attributes of the nodes.

Accurate classification of the news requires analysing the single modalities and the correlations among them. The models able to process more than one feature to solve the misinformation detection task are called multimodal classifiers. Many architectures for multimodal classifiers have been designed. The first important classification of multimodal classifiers takes into account the fusion mechanisms used to

combine features from different modalities. The main fusion techniques are the following:

- *Early fusion* This technique is sometimes referred to as feature-level fusion and consists of concatenating features from many data modalities at an early stage. This kind of fusion is frequently referred to as intermediate fusion if it is carried out after feature extraction and before classification. Accordingly, most previous work on multimodal disinformation detection embeds each modality into a corresponding vector representation and then concatenates the vectors to obtain a multimodal representation that is used for classification.
- *Late fusion* This technique is also called decision-level fusion and consists in combining the results of the analysis carried out for each data modality separately. In other words, methods like sum, max, average, and weighted average are used to integrate the findings of modality-wise classification. The majority of late fusion solutions employ hand-crafted rules that are subject to human bias and ignore the quirks of the real world.

Early fusion is typically a difficult procedure, whereas late fusion is simpler to carry out. Because decisions made at the semantic level often have the same representation, unlike early fusion where features from various modalities, such as image and text, may have different representations. The fusion of decisions is therefore simpler. The feature level correlation among modalities is not utilized by the late fusion technique, though. In early fusion architectures, there is only one model in charge to process the vector obtained by concatenating the embeddings of the single modalities. Having a view of all modalities, this model can discover correlations among them. Because training is only done once, early fusion has the advantage of requiring less computation time than late fusion, which needs many classifiers for local decisions. However, there are also hybrid strategies that benefit from both early and late fusion techniques. Another important challenge in fake news detection concerns *Cross-domain*. News can come from any domain in the real world, including politics, sports, environment, technology, business, and economics. Although very accurate, a model trained on a single domain dataset may underperform when applied to news from another domain. The reason is that during the training phase, the model learns domain-specific words and patterns that can be less relevant in different domains. A challenging task that might help in the solution of this problem is to accurately choose a subset of domain-invariant attributes (e.g. psychological traits, readability features) from news data. Another strategy is to train the model using data from multiple domains. This is possible because more and more datasets are now available.

2.2 Existing reviews

Although there are numerous worthwhile surveys on the subject of identifying fake news, relatively few of them (Alam et al. 2022; Abdali 2022; Hangloo and Arora 2022) concentrate on multimodal methods. In the following of this section, we briefly describe such existing reviews on multimodal fake news detection and after that, we highlight the main differences with the survey proposed in this paper.

The survey in Alam et al. (2022) provides an overview of the state of the art in multimodal disinformation detection that includes a variety of modalities, including text, photos, speech, video, social media network structure, and temporal data. While both factuality and harmfulness are essential elements in the concept of disinformation, they are frequently examined separately. Differently, the survey in Alam et al. (2022) focuses on disinformation, by studying both the factuality and harmfulness aspects of the problem, with a focus on different modalities.

In Abdali (2022) is reviewed the literature on multimodal misinformation detection, discussed its advantages and disadvantages, and suggested new directions for future research. First, we discuss cross-modal hints and fusion processes. Then, we divide all of the currently available solutions into two primary categories depending on the methodology they use: traditional machine learning and deep learning. The review in Abdali (2022) has as its main aim to assess, categorize, and identify current methodologies as well as the challenges and shortcomings in multimodal misinformation detection.

In Hangloo and Arora (2022) is proposed an overview of modern, cutting-edge methods, techniques and strategies with a focus on multimodal context to address the problem of identifying fake news on social media platforms. Our review focuses primarily on four important factors. First, with a properly defined taxonomy of fake news detection strategies, the study offers a precise definition of fake news and distinctions between several related concepts. We discovered during our research that the multimodal feature of news information has received relatively little attention. Second, a number of widely used deep learning models, frameworks, libraries, and transfer learning techniques have been highlighted, with TensorFlow being the most well known. Third, we've given an overview of different cutting-edge methods for detecting fake news on social media sites utilizing deep learning techniques while taking into account multimodal data.

In this survey we provide an extensive comparative study of numerous research proposals, giving insight into hitherto aspects that have not yet been discussed. In contrast to other

review studies, the primary focus of this work is on various deep learning methodologies, such as transfer learning and pre-trained models utilized for spotting bogus news on social media while taking into account multimodal data. In fact, to the best of our knowledge, this paper represents the first survey focusing only on the recent advancements in deep learning techniques exploited for multimodal fake news detection. Other than this specificity, the proposed survey different from the aforementioned ones (Alam et al. 2022; Abdali 2022; Hangloo and Arora 2022), provides a more detailed view of the individual key approaches proposed in the literature by focusing on the specific deep learning model adopted. Whereas the reviews reported in Alam et al. (2022); Abdali (2022); Hangloo and Arora (2022) provide a general discussion of similar methods omitting specific details of the single proposals.

3 Datasets used for multimodal fake news detection

This section reports a synthetic description of the datasets used to validate the approaches presented in this survey.

A detailed description of a larger collection of datasets can be found in Murayama (2021) that describes 118 datasets related to (i) fake news detection, (ii) fact verification, (iii) satire detection, (iv) news (media) credibility, (iv) check-worthy claims, and (v) claim matching.

- *Weibo* The Weibo dataset is a collection of posts from Sina Weibo, which is a popular Chinese microblogging platform similar to Twitter. The dataset includes a large number of posts, with each post containing various types of content such as text, images, videos, and links. The dataset has been used for various research purposes, including sentiment analysis, topic modeling, and user profiling. It can be useful for researchers studying social media behaviour in China, as well as for those interested in developing machine learning models to analyse social media data. In this paper two versions of this dataset are mentioned: *Weibo A* (Jin et al. 2017) and *Weibo B* (Cao et al. 2019).
- *Fakeddit* Fakeddit (Nakamura et al. 2019) is a large multimodal dataset containing over 1 million entries related to several types of fake news. Each entry includes the attributes: (i) submission title, (ii) image, (iii) comments and additional metadata (i.e. score, number of comments, etc.). The samples go through multiple stages of review and are then labelled using distant supervision into either 2-way, 3-way, or 6-way classification categories.
- *MediaEval* The MediaEval dataset (Boididou et al. 2018) is a valuable resource for researchers and developers who are interested in multimedia retrieval and evaluation. It contains a vast collection of multimedia data, including images, audio files, and video recordings, that have been annotated with associated metadata such as timestamps, geolocation data, and user-generated tags. The dataset is used to support a series of annual MediaEval benchmarking evaluations, which are organized around a set of shared research tasks or challenges. Each year, researchers are invited to develop algorithms and approaches to tackle these challenges, and then submit their results for evaluation against a set of predefined performance metrics. The MediaEval challenges cover a broad range of topics, including audio event detection, multimedia event detection, social media analysis, and multimedia recommendation systems.
- *FakeNewsNet* FakeNewsNet (Shu et al. 2019b) is a publicly available dataset designed for research on fake news detection. The dataset includes various types of information related to the creation and dissemination of fake news, including textual content, images, and social network information. The dataset is composed of two sub-datasets:
  - *Politifact* Contains fact-checking articles from PolitiFact (www.politifact. com), a non-profit organization that evaluates the accuracy of statements made by politicians in the United States (Shu et al. 2017). The news articles in the PolitiFact dataset were published from May 2002 to July 2018.
  - *GossipCop* Contains articles from GossipCop, a website that reports false news about celebrities and entertainment in magazines and on the internet in the United States. It assigns a 0–10 scale to each article depending on its credibility, with 0 indicating that the rumour is wholly untrue or fictive and 10 indicating that the news is 100 per cent factual.

  Each sub-dataset includes both fake and real news articles, and all entries are labelled accordingly. The dataset also includes social network information for some articles, such as the number of retweets and likes on Twitter. FakeNewsNet aims to provide researchers with a standardized and reliable dataset for developing and evaluating fake news detection models. It can be useful for researchers working in the fields of natural language processing, machine learning, and data mining.
- *BuzzFeed* The BuzzFeed News dataset is a collection of news articles and their corresponding metadata from BuzzFeed News (https://www.buzzfeed.com/), a popular news and media website. The dataset contains over 200,000 articles published between 2014 and 2018, covering a wide range of topics including politics, entertainment, and technology. The dataset includes several different types of information for each

article, including the article title, author, publication date, URL, text content, and images. Additionally, the dataset includes social media engagement metrics for each article, such as the number of Facebook likes, comments, and shares. The BuzzFeed dataset has been used for various research purposes, including topic modelling, sentiment analysis, and fake news detection.

- *PHEME* The PHEME dataset (Zubiaga et al. 2016) is a publicly available dataset designed for research on rumour detection and veracity prediction in social media. The dataset includes tweets related to nine different events, such as the Boston Marathon bombing, and the Charlie Hebdo shooting, and is divided into four sub-datasets:
  - *Rumours* Contains tweets that were posted during the events and were classified as either true or false.
  - *Non-rumours* Contains tweets that were posted during the events but were not related to rumours.
  - *Thread structure* Contains information about the structure of tweet threads related to the events, such as the number of tweets in each thread and the number of retweets.
  - *Stance* Contains information about the stance of tweets related to the events, such as whether the tweet supports, denies, or is neutral towards a rumour.

  Each tweet is labelled according to its veracity status (i.e. true or false) and its stance (i.e. support, deny, or neutral). The dataset also includes additional metadata for each tweet, such as the date and time of the tweet, the Twitter user who posted the tweet, and the location of the tweet.
- *LIAR* The LIAR dataset (Wang 2017) is a publicly available dataset designed for research on fact-checking and fake news detection. It contains a collection of statements from various politicians and their corresponding labels, indicating the truthfulness of each statement. The dataset includes the following features:
  - *Statement* The text of the statement made by a politician.
  - *Label* The truthfulness rating of the statement, which is one of six categories: True, Mostly True, Half True, Barely True, False, and Pants on Fire.
  - *Subject* The topic of the statement.
  - *Speaker* The name of the politician who made the statement.
  - *Speaker's Job* The job title of the politician.
  - *State* The state in which the politician holds office.
  - *Party* The political party to which the politician belongs.
  - *Context* Additional context for the statement, such as the location or event where it was made.

  The LIAR dataset includes approximately 12,800 statements made by politicians from different political parties in the United States between 2007 and 2017. The dataset has been used for various research purposes, including fact-checking, fake news detection, and natural language processing. It can be useful for researchers interested in studying the accuracy of politicians' statements and developing machine learning models to detect fake news and misinformation.
- *Yang dataset* The dataset proposed in Yang et al. (2018) contains 20,015 news items, of which 11,941 are fake and 8074 are authentic. The dataset for false news includes text and information collected from over 240 websites. The authentic news has been extracted from the New York Times, Washington Post, and other news sites. The dataset contains multiple pieces of information, such as the title, text, image, author, and website.
- *MCG-FNeWS* MCG-FNeWS is the largest publicly available Chinese fake news dataset. It was released by the Chinese Academy of Sciences' Institute of Computer Technology. It contains news from May 2012 to November 2018 and includes 19,186 non-fake news and 19,258 fake news released on Weibo.
- *Ma dataset* The Twitter dataset presented in Ma et al. (2016) is related to confirmed rumours and non-rumours from www.snopes.com, an online rumour debunking service. 778 reported events are collected during March-December 2015, of which 64% were rumours.
- *Volkova dataset* The Twitter dataset presented in Volkova et al. (2017) has been created by querying Twitter's fire hose during the period March 15 to March 29, 2016—one week before and after the Brussels bombing of March 22, 2016, in relation to 174 suspicious and 252 verified news accounts. The authors collected retweets generated by any user that mentioned one of these accounts and assigned the corresponding label: propagated from *suspicious* or *trusted* news account.
- *CoAID (Covid-19 heAlthcare mIsinformation Dataset)* (Cui and Lee 2020) It is a COVID-19 healthcare misinformation dataset containing fake news extracted from websites and social media platforms, as well as consumers' social media engagement with such information. It includes 5216 news items, 296,752 user engagements, 958 COVID-19 social platform postings, and ground truth labels.
- *CXD (Columbia X-Cultural Deception Corpus)* (http://www.cs.columbia.edu/speech/cxd/index.html) is a collection of within-subject deceptive and non-deceptive

English speech obtained from native speakers of Standard American English and Mandarin Chinese. The dataset includes 134 conversations between 268 people who did not know each other previously, for a study of acoustic, prosodic, and lexical cues to deception.

- *Reddit* The Reddit dataset (https://paperswithcode.com/dataset/reddit) is a graph dataset from Reddit posts collected in September 2014. The node label is the community (*subreddit*) to which the post belongs. In order to build a post-to-post graph, 50 large communities have been sampled by connecting posts if the same user commented on both. The Reddit dataset contains 232,965 posts with an average degree of 492.
- *EMERGENT* The EMERGENT dataset (Ferreira and Vlachos 2016) contains 300 claims and 2595 linked articles. The dataset includes 300 rumoured statements and 2,595 related news articles collected and labelled by journalists as *true*, *fake*, or *unconfirmed*.
- *MICC-F220* The MICC-F220 dataset (Amerini et al. 2011) consists of actual and altered photos with no other type of data. More specifically, this dataset contains 110 fake images and 110 original images.
- *ReCOVery* This dataset (Hua et al. 2023) includes 2029 news related to the COVID-19 pandemic, from January to May 2020. Items in ReCOvery are labelled as *real* or *fake*, according to the credibility of their sources. The dataset also provides multimodal information on the news, including text, image, source, publication time, and writer information.

4 State of the art of deep learning-based multimodal fake news detection

In this section, we review the works in the literature discussing models, methods, and applications of deep learning techniques for multimodal fake news detection. In particular, a detailed analysis of the selected papers is provided throughout the section. For each study in the literature, we extracted the most important features like the method implemented, the data type and size used, the evaluation methods adopted, the accuracy for each method, and the results achieved. For each study, the features are summarized in Table 1.

A significant body of research has been devoted to automatically determining fake news using textual content. Currently, the different approaches for fake news detection always combine text modality with additional modalities such as images, video, audio, and social content used to propagate fake news on social networks. Due to this observation, this section summarizes relevant proposals that report all other modalities but text-only.

4.1 Images

Social media posts combine images and text. Photos flood social media after significant events and contribute to further messing up the boundary between reality and fiction at critical junctures for society. It is well known that social media posts containing photographs and images in general gain a lot more retweets and shares and spread more quickly with respect to those containing just text. Images have high distribution, capture the emotions of people, communicate an understanding of reality and users are frequently susceptible to being deceived. Images associated with a post can be altered or can be just images out of context. It's not new to manipulate photographs for political or personal reasons as well as using photo editing software to alter an image. In addition, the caption associated with an image is of paramount importance and cannot be ignored in fake news detection as it is used to increase the reachability of the post. Therefore, in the analysis of the two modalities: text and images, image captions are crucial to identify clickbait and deceptive captions. The massive diffusion on social media of news disseminated with images let rise to a number of proposals whose goal is using visual and text content to predict fake news. A survey of these approaches is reported in the sequel.

In Hua et al. (2023) an interesting approach for multimodal fake news detection, based on a pre-trained BERT model, a ResNet50 model, a data augmentation strategy, and a contrastive learning process is proposed. The BERT model extracts the features of the text of the news while those of the images are extracted by the ResNet50 model. They are concatenated to obtain the final feature representation of the news. To mitigate the problem of operating with a small training set, it is augmented by back-translating the title and abstract of the news so that obtaining additional news (fake and real) with the same semantics but different structures. In order to capture the interacting information between news on a certain topic, contrastive learning is finally used.

In Amri et al. (2021) EXMULF (EXplainable MULtimodal Fake news detection) is presented, a system able to identify fake news by detecting *discrepancies* between the *topics* covered by its *content* and those that can be extracted from its *image*. The topics related to the news content are extracted by a module based on the LDA (Latent Dirichlet Allocation) technique. The topics associated with the image are extracted by a VGGNet-16 Simonyan and Zisserman (2015) model and by the LDA module that processes the text related to the image (e.g. the caption). The resulting output is passed to a module that measures the similarity between the text and image topics. If the topics are different then the news is classified as fake. Otherwise, the text and the image are passed to the *multimodal detector* module, able to predict the news *veracity*. This module processes the

text with a Bert-based model and the image with a ResNet-based model. Finally, an explanation for the prediction is computed by the *explanatory module*.

In Zhang et al. (2019) is proposed the Multimodal Knowledge-aware Event Memory Network (MKEMN) system, which utilizes the Multimodal Knowledge-aware Network (MKN) and Event Memory Network (EMN) as key components for fake news detection. Specifically, the MKN learns the multimodal representation of the post on social media by extracting textual and visual features and retrieves external knowledge from a real-world knowledge graph to complement the semantic representation of short texts of posts and takes conceptual knowledge as additional evidence to improve event detection. Specifically, it combines word embedding, visual embedding, and knowledge embedding. Afterword embedding, the sentence is projected to a sequence of word vectors and then feed into a Bidirectional GRU to capture the contextual information of the sequence. To extract the event-invariant features, the Event Memory Network (EMN) builds an external memory shared during the whole training process to capture the event-independent latent topics. After that, the event representation is fed into a deep neural network for fake news detection. Extensive experiments on two benchmark datasets, the Twitter dataset by Ma et al. (2016) and PHEME, demonstrate that the method outperforms state-of-the-art methods.

In Singhal et al. (2021) is proposed SpotFake, a multimodal framework for fake news detection, exploiting both the textual and visual features of an article. Specifically, BERT is used to learn text features, while image features are learned from a CNN, VGG-19 pre-trained on ImageNet dataset. All the experiments were performed on two publicly available datasets, MediaEval and Weibo A. Authors stated that the proposed model performs better than the current state of the art.

In Alonso-Bartolome and Segura-Bedmar (2021) authors exploited a CNN that takes as inputs both text and image of an article. The outputs are then concatenated into a single vector. Experimental validation has been carried out on the Fakeddit dataset, using both unimodal and multimodal solutions. Experiments have shown that the multimodal approach achieved the best results, with an accuracy of 87%.

In Qi et al. (2021) is proposed the EM-FEND framework that is based on the extraction of visual entities (such as celebrities and landmarks) to understand the news-related high-level semantics of images. To this purpose, the authors considered a variety of data modalities: text, OCR text, news-related high-level semantics of images, e.g. celebrities and landmarks, visual CNN features of the image, and the embedded text in images as the complementation of the original text. The different features are then concatenated by accounting text-image correlations, mutual enhancement, and entity inconsistency. Authors claimed that extensive experiments demonstrate the superiority of their model compared to the state of the art.

In Wang et al. (2018) the event adversarial neural network (EANN) framework is proposed, which can derive event-invariant features and thus benefit the detection of fake news on newly arrived events. It consists of three main components: the multimodal feature extractor, the fake news detector, and the event discriminator. The multimodal feature extractor is responsible for extracting the textual and visual features from posts. The event discriminator deletes event-specific features while maintaining shared features among events. Experiments are conducted on multimedia datasets collected from Weibo (Weibo A) and Twitter (MediaEval). The results reported in the paper show that EANN can outperform the state-of-the-art methods, and learn transferable feature representations.

In Wang et al. (2021) the authors proposed the MetaFEND framework, which can quickly identify fake news on breaking events with a small number of verified posts. The suggested model specifically combines neural process and meta-learning methods to benefit from each method's advantages. To increase effectiveness by handling categorical information and removing irrelevant messages, a label embedding module and a hard attention method are proposed. Extensive tests are run on multimedia datasets gathered from Weibo (Weibo A) and Twitter (MediaEval). The results of the experiment demonstrate that the MetaFEND model outperforms cutting-edge techniques in its ability to identify fake news on previously unreported events.

In Khattar et al. (2019) is proposed the multimodal variational autoencoder (MVAE) framework, which uses a bimodal variational autoencoder combined with a binary classifier for fake news detection. The model consists of three main components, an encoder, a decoder, and a fake news detector module. The encoder embeds both textual and visual features. To extract features from the textual content, a stacked bi-directional LSTM is used. The input to the visual encoder is the image enclosed in the message. Image features are extracted through a CNN, specifically, the VGG-19 architecture is used and trained over the ImageNet database. Textual and visual features are then concatenated and passed through a fully connected layer to form the shared representation. A detailed experimental evaluation has been performed on two well-known fake news datasets collected from Weibo (Weibo A) and Twitter (MediaEval). The results show that across the two datasets, on average the model outperforms state-of-the-art methods.

In Zhou et al. (2020b) is proposed a Similarity-Aware FakE news detection method (SAFE) which exploits multimodal data, more exactly the textual and visual features from the news. To this purpose, neural networks are used

to extract the textual and visual features, also deriving a similarity among them. The aim of the approach is to classify the news by using either its text or images or the mismatch between the text and images. Experiments have been performed on large-scale real-world data (PolitiFact, GossipCop), showing the effectiveness of the proposed method

In the paper of Xue et al. (2021) is proposed the multimodal consistency neural network (MCNN) tool, which is composed of five modules: the textual feature extraction that exploits BERT, the visual semantic feature extraction, the visual tampering feature extraction, the similarity measurement, and the multimodal fusion module. The visual tampering feature extraction focuses on physical levels feature extraction such as malicious image tampering and recompression by using ResNet. The key aspect of the approach is the similarity measurement module that evaluates the correlation between the text information and the visual one. The different features are then fused by means of attention mechanisms. The framework has been evaluated over 4 Twitter datasets, MCG-FNeWS, PolitiFact, MediaEval, and Yang dataset, showing promising results.

The approach proposed by Jin et al. (2017) fuses features from three modalities, i.e. textual, visual, and social context using an RNN that utilizes an attention mechanism (att-RNN) for feature alignment. Image features are incorporated into the joint features of text and social context, which are obtained with an LSTM network, for enhancing the classification. The neural attention from the outputs of the LSTM is utilized when fusing with the visual features. Extensive experiments are conducted on two multimedia rumour datasets collected from Weibo (Weibo A) and Twitter (MediaEval). Results show the effectiveness of the att-RNN framework.

In Zhang et al. (2020) is proposed the BERT-based domain adaptation neural network (BDANN). BDANN comprises three main components: a multimodal feature extractor, a domain classifier, and a fake news detector. Specifically, the multimodal feature extractor employs the pre-trained BERT base model to extract text features and the pre-trained VGG-19 model to extract image features. The extracted text and image features are then fed to the detector to identify fake news and in the domain classifier to map the multimodal features of different events to the same feature space. The approach implements an adversarial learning mode, exploiting the gradient reversal layer (GRL): the multimodal extractor tends to extract event-invariant features by maximizing the domain classification loss, while the domain classifier tends to discover the event special information from multimodal features by minimizing the domain classification loss. To assess the performance of BDANN, experiments have been performed on two multimedia datasets: MediEval and Weibo A. The experimental results show that BDANN outperforms the state-of-the-art baseline models.

In Song et al. (2021) is proposed a multimodal fake news detection model exploiting text, comments, and images and based on word embedding and convolutional neural network (VGG-19). Precisely, the model is composed of the following components: (1) input embedding layer to obtain the word embedding and image embedding; (2) the cross-modal attention residual (CARN) layer to reinforce the target modality feature representation by selectively extracting information from a different source modality; (3) the self-attention residual network layer to record interactions between different sequence element pairs and transmit original textual information to MCN; (4) the multi-channel convolutional neural network (MCN) to reduce the impact of noise information that may be produced by cross-modal attention residual; and (5) fake news detection module. Experiments have been performed on four real-world datasets: MediaEval, Weibo A, Weibo B. Results show that the model outperforms the state-of-the-art methods and learns more discriminable feature representations.

In Sachan et al. (2021) is proposed Shared Cross Attention Transformer Encoders (SCATE) which exploits deep convolutional neural networks and transformer-based methods to encode image and text information and utilizes cross-modal attention and shared layers for the two modalities. SCATE pays attention to the relevant parts of each modality with reference to the other one, fusing the different modalities through attention mechanisms. A detailed experimental evaluation has been carried out over both Twitter and Weibo datasets.

In Kumari and Ekbal (2021) is proposed the attention-based multimodal factorized bilinear (AMFB) framework for multimodal fake news detection. The framework has been designed with the intention to maximize the correlation between textual and visual information. This framework has four different sub-modules: (i) Attention-Based Stacked Bidirectional Long Short Term Memory (ABSBiLSTM) for textual feature representation, (ii) Attention-Based Multi-level Convolutional Neural Network-Recurrent Neural Network (ABM-CNN-RNN) for visual feature extraction, (iii) multimodal Factorized Bilinear Pooling (MFB) attention mechanism for feature fusion and finally (iv) Multi-Layer Perceptron (MLP) for the classification. Experimental results performed on two real-world datasets show the effectiveness of the approach.

In Jing et al. (2021) is proposed the TRANSFAKE framework that considers different modalities like news content, comments, and images for fake news detection. The textual features are extracted with BERT, while for the visual ones, a Faster-RCNN model is used. TRANSFAKE fuses the different features with a Transformer-based model. It employs

multiple tasks, i.e. rumour score prediction and event classification, as intermediate tasks for extracting useful hidden relationships across various modalities. These intermediate tasks promote each other and encourage TRANSFAKE to make the right decision. Extensive experiments on three real-life datasets demonstrate that TRANSFAKE outperforms state-of-the-art methods.

In Zhang et al. (2018) is proposed FauxBuster, a framework that detects fauxtography analysing the comments posted by users on social media and not. More specifically, FauxBuster is content-free: it does not rely on the actual content of the images and is therefore robust w.r.t. the use of powerful uploaders that modify the text associated with the images. FauxBuster uses deep autoencoding and neural word embedding techniques in order to extract, from the comments posted on social media, a set of relevant signs such as network characteristics, linguistic cues, and metadata. These are then integrated by FauxBuster using a supervised learning framework that is effective in detecting fauxtography on social media. The performances of FauxBuster have been evaluated on the two mainstream social media platforms Reddit and Twitter. More specifically: for the Reddit dataset the number of considered posts is 196, the number of comments is 60,168, and the number of distinct users is 30,702; for the Twitter dataset, the number of considered posts is 721, the number of comments is 1,928,325, and the number of distinct users is 582,281. Results show the proposal is effective (with 25.6% higher F1-score than state-of-the-art image forgery detection baselines) and efficient (reaching 86.1% detection accuracy within one hour of the original post).

In Giachanou et al. (2020) is proposed an interesting multimodal multi-image system in order to perform binary classification of online articles by combining textual, visual, and semantic information; moreover, differently from other approaches, in the case of an article in which more than an image is present, it extracts and combines features extracted from all of them. BERT is used to obtain textual features and a VGG-19 model followed by an LSTM layer and a mean pooling layer is used to obtain visual features. As for the semantic representation, it refers to the text-image similarity that is obtained by applying the cosine similarity between the image tags embeddings and the title, this last is a type of information that is rarely considered in fake news detection. Experimentation is performed using the FakeNewsNet collection. In more detail, from the GossipCop posts of such collection authors collect 2745 fake news and 2714 real news. The proposed multimodal multi-image system outperforms the BERT baseline by 4.19% and SpotFake by 5.39% and achieves an F1-score of 79.55%.

In Xie et al. (2021) is proposed SERN, the Stance Extraction and Reasoning Network to obtain, given a post, its stances representations that are implied in the reply associated with the post itself. Text and images are considered in the proposal and a multimodal representation of these features is performed in order to binary classify fake news. The method works as follows: given a post containing multimodal news, an extractor first constructs stances, i.e. post-reply pairs. Then, BERT is used to extract textual features, and a pre-trained ResNet-152 is used to retrieve visual features. Textual and visual features are therefore concatenated so that obtaining a multimodal feature representation. This last is then the input of a Multi-Layer Perceptron (MLP) that is in charge of performing binary classification of the post. Experimentation demonstrates the proposal outperforms the state-of-the-art baselines on two public datasets: the PHEME dataset and a condensed version of the Fakeddit dataset created by the authors. Results show an accuracy of 96.63 % for Fakeddit and of 76.53% on PHEME.

In Raj and Meel (2021), fake news is detected using a Coupled ConvNet architecture, i.e. a hybrid two-stream convolutional architecture with an Image-CNN module for the visual fake news classification and a Text-CNN module for the textual fake news classification. Input data are firstly preprocessed by using these modules, and then the text stream and image stream are coupled using a late fusion algorithm to feed a CNN.

In Shang et al. (2022) is proposed a generative technique to detect multimodal COVID-19 misinformation. The approach investigates the cross-modal link between the visual and textual content that is intricately woven within the multimodal news content. To these aims is proposed a framework for duo-generative explainable misinformation detection (DGExplain) that efficiently uses user comments and explicitly analyses the cross-modal link between news content in various modalities to identify and explain misinformation in multimodal COVID-19 news items. DGExplain performance has been evaluated on two real-world multimodal COVID-19 news datasets. In terms of the precision of multimodal COVID-19 misinformation detection and the explainability of detection reasons, evaluation results show that DGExplain significantly exceeds state-of-the-art baselines.

In Mu et al. (2023) is proposed a Self-Supervised Distilled Learner (SSDL) to obtain feature representation to identify multimodal misinformation. The learning strategy aims to achieve the following multi-task objectives: (1) task agnostic, which assesses the intra- and inter-mode representational consistencies for improved alignments across related models; and (2) task-specific, which calculates the category-specific multimodal knowledge to allow the classifier to derive more discriminative predictive distributions. In the SSDL method, a Teacher network is used to weakly direct a Student network to imitate a decision pattern similar to that of the Teacher. Using contrastive self-supervised task

agnostic objective and supervised task-specific adjustment in tandem, the Student model is first pre-trained. The Student model is then finetuned using self-supervised knowledge distillation combined with the supervised objective of decision alignment. The authors exploit the dataset NewsCLIPpings, which contains multimodal (i.e. each sample has a text caption accompanied by an image component).

In Kirchknopf et al. (2021) is proposed a method for categorizing fake news in binary form utilizing four different modalities: the textual content of the news, the related comments, the images, and the remaining metadata pertaining to other modalities. This method is experimented over the Fakeddit dataset. The proposed architecture allows aggregating these modalities at different levels and considering different data fusion methods. The best result shows an accuracy of 95.5% and has been obtained by separately pretraining each modality and then training only the fusion and classification layers on top.

Many different proposals exist in the recent literature aiming at detecting fake news by combining visual and textual features. At the moment, these two modalities are still those most frequently analysed. Multimedia datasets mainly originate from Weibo and Twitter (MediaEval), and it is often the case that they just cover these two modalities. As for a general discussion from the above-reviewed proposals, it emerges that BERT is widely and successfully used to analyse text, whereas, for image analysis, CNN-based solutions are commonly used (like VGG-19). Many of the above-reviewed approaches use pre-trained deep neural networks to detect manipulated images. Images are first trained on various neural networks, and then the most accurate model is selected. Some other approaches use techniques to identify digital alterations of the images, e.g. regions reporting a different/unusual compression level. The image captions are used to detect fake images and different text modality models have experimented on the textual information in posts to classify image captions into fake and true. In spite of the consideration that visual content often induces a strong user sentiment impact and that revealing opposite-generated sentiments could quickly facilitate fake image detection just a few works consider the sentiment-related data that images produce. This specific issue has been explored by Cui et al. (2019). Another characteristic that could be relevant in detecting fake news is related to the exploitation of the effectiveness of text-image similarity. This specific issue has been explored by Giachanou et al. (2020); Qi et al. (2021). In addition, it should also be evidenced that metadata and comments to the posts are still scarcely adopted by the state-of-the-art literature, but, combined with user data, they could be profitably used to set the credibility of the user originating the post. This specific issue has been explored in Shu et al. (2019a).

4.2 Audio and video

The analysis of audio and video content in news media plays a crucial role in identifying and countering the dissemination of fake news. Manipulation of audio and video content can be used to create false narratives or distort interpretations of events. Modern AI techniques have advanced to the point where audio and video content can be manipulated in highly sophisticated and realistic ways. The use of Deepfake technology, for example, involves machine learning techniques to produce fake videos that appear genuine and are used to spread disinformation and fake news. By training artificial neural networks, Deepfake technology can replicate a person's facial expressions and features in a video, resulting in a fake video in which the individual appears to say things they never said. The technology can generate fake audio that sounds very convincing and can influence people's perceptions. In addition to these challenges, the emergence of modern social media platforms like TikTok, Instagram, and Twitter has made it simpler than ever to share video content with a vast audience. While this has its advantages, it has also presented new obstacles to detecting and countering fake news. Short video clips can spread false information rapidly and widely, becoming viral within hours. This can be particularly problematic when it comes to political or social issues, where false information can be used to sow discord and undermine trust in institutions. Furthermore, these social media platforms are often driven by complex algorithms that employ AI to recommend content to users. This means that even if a fake news video is flagged as false or misleading, it may still be recommended to users who have previously shown an interest in similar content. However, modern AI techniques can also be used to detect and counteract fake news.

In the sequel of this section, we overview recent proposal for multimodal fake news detection involving audio and video. In Mittal et al. (2020), a learning-based technique for distinguishing between real and deepfake multimedia content is proposed. The system examines the similarity between the audio and visual modalities and additionally, extracts and compares the affective clues that correspond to observed emotion from the two modalities within the same video. The proposal discriminates fake and real information by means of a deep learning network influenced by the Siamese network architecture that extracts the audio and video modalities and then uses a triplet loss function to evaluate the similarity and detect the fake videos. The proposal has been validated over two deepfake benchmark datasets, DeepFake-TIMIT (Korshunov et al. 2018) dataset and DFDC (Dolhansky et al. 2019) dataset and results show an accuracy of 96.6% on DF-TIMIT and 84.4% on DFDC.

The approach in Karimi et al. (2018) studies the problem of deception detection in videos and proposes DEV

(DEceptive Videos) an interesting end-to-end framework that uses a deep learning approach for the automatic extraction of features from video and audio. More specifically, in order to capture the sequential nature of videos, a video file is broken into a set of frames and for each frame, a set of visual and vocal features relevant for deception detection is extracted. This task is performed using CNN networks due to their excellent performance in extracting features and applying different filters to a region of data. Then in order to capture the temporal correlations existing in a sequence of features, the set of CNN outputs feds LSTM networks that are in charge of detecting temporal correlation by effectively propagating information in the given input sequences. In order to perform a good classification, the number of training instances is artificially increased by manipulating the output of LSTMs with a variant of the Large Margin Nearest Neighbour (LMNN) method that works on triplets of videos. The final classification task of deception detection is then performed using the simple k-nearest neighbours method. The proposal has been tested over the real trial dataset in Pérez-Rosas et al. (2015) containing 121 video clips of courtroom trials of which 61 are of deceptive nature and 60 are of truthful nature. Results show that performances greatly increase if audio and video modalities are both used with respect to the case in which they are used separately and achieve an accuracy of 84.16%.

The work by Mendels et al. (2017) is the first to use deep learning approaches for detecting deception. It proposes a series of experiments on the Columbia X-Cultural Deception Corpus (CXD) for detecting deception from speech using lexical and acoustic features. These last are a set of standard features, such as pitch, intensity, spectral, cepstral, duration, and voice quality, that are generally used for many computational para-linguistic tasks, including emotion recognition and deception detection. The selected deep learning models are: a lexical bidirectional long short-term memory (BLSTM) classifier, a Mel-Frequency cepstral coefficients (MFCC) BLSTM classifier, DNN-openSMILE, and a hybrid model achieving the best performance with an F1-score of 63.9%.

In Shang et al. (2021), it is examined the problem of identifying misleading COVID-19 short videos on TikToc in which fake content is jointly expressed in the videos' audio, visual, and textual elements. The two main goals the paper faces are (i) How can the manipulated and altered visual content in TikTok videos be efficiently mined for information? (ii) How can heterogeneous information from several modalities be efficiently aggregated in brief videos? In order to address the aforementioned issues the paper proposes TikTec, a TikTok misinformation detection framework that is in charge of correctly identifying misleading Covid-19 videos in TikTok. In more detail, the proposal firstly creates a caption-guided visual representation learning module that specifically takes advantage of the caption in the audio and visual frames of the video to efficiently extract the essential visual information from the manipulated and edited TikTok videos. Then, in order to successfully predict the link between visual frames and speech content in various modalities and effectively fuse the information inherent in multimodal videos, the proposal develops a visual-speech cognitive information fusion module to address the second problem. TikTec has been assessed on a real-world Covid-19 video dataset obtained from TikTok. The dataset collected using Covid-19-related keywords and hashtags consists of a total of 891 valid TikToc videos, including 226 misleading videos and 665 truthful ones. The evaluation's findings demonstrate that TikTec significantly outperforms cutting-edge baselines in correctly identifying deceptive COVID-19 short videos. Results show TikTec improves accuracy and F1-scores by 6.1 and 4.8%, respectively, above the baseline with the best performance (3DResNet).

In Wang et al. (2022), authors propose techniques to identify false information in social media posts by utilizing text and video modalities. The proposal uses self-supervised learning to develop expressive representations of combined visual and textual data and defines and offer two deep learning novel approaches based on contrastive learning and masked language modeling (MLM) for the detection of semantic inconsistencies in short-form social media video posts. The two proposed methods, evaluated on a dataset consisting of 160,000 video postings gathered from Twitter, beat cutting-edge techniques both on synthetic data produced by randomly switching positive examples and on real-world data on a new manually labelled test set for semantic misleading. More specifically, results show Contrastive Learning outperforms the method in McCrae et al. (2022) by 3% on the accuracy, and MLM performs the best overall, outperforming Contrastive Learning by 5.23% on accuracy.

In Krishnamurthy et al. (2018) is proposed a deep learning approach for deception detection in real-life videos using features from multiple modalities. The first stage of the approach extracts textual, audio, and visual features from each video. More specifically, individual modalities are extracted as follows: (i) A CNN is used to extract features from the transcript of a video. Firstly using a pre-trained Word2Vec model for each word in the transcript is extracted a vector. These vectors are therefore concatenated and used as an input vector to feed the CNN. (ii) A 3D-CNN (Ji et al. 2013) is used to extract features from each image frame, and spatiotemporal features from the whole video that allows improving identification of facial expressions such as smile, fear, or stress; and openSMILE—an open-source toolkit—is used to extract high dimensional features from an audio file (Eyben et al. 2013). The second step consists in fusing the features from individual modalities to map them into a joint

space. To achieve this, different kinds of data fusion techniques have been tested: (i) In the concatenation technique, MLP_C the features from all the modalities are simply concatenated into a single feature vector. (ii) In the Hadamard + Concatenation technique, $MLPH_{H+C}$, the fusion of audio, visual, and, textual features is performed using Hadamard product and the Micro-Expression features are finally just concatenated with the Hadamard product. The evaluation of the proposed deception detection model has been performed using a real-life deception detection dataset by Pérez-Rosas et al. (2015) containing 121 video clips of courtroom trials in which 61 are of deceptive nature and 60 are of truthful nature. Results show that the $MLPH_{H+C}$ proposal outperforms existing baseline techniques (SVM, CNN, Bi-LSTM, BERT) for deception detection achieving an accuracy of 96.14% and a ROC-AUC of 0.9.

As a final consideration, we highlight that the evolution of AI techniques has had both positive and negative effects on the diffusion of fake news. On one hand, AI has made it easier to manipulate audio and video content in more sophisticated ways. On the other hand, AI has also provided useful tools to counteract fake news and protect the truthfulness of news reporting.

4.3 Social and network data

Social context-based approaches, in contrast to content-based approaches, combine elements from user profiles, post contents, and social networks on social media (Shu et al. 2019a). By examining users' postings, comments, tags, retweets, and other interactions with breaking news on social media, social context characteristics show the active user interaction. User characteristics and credibility can be measured by user attributes. Users' social reactions, such as stances, are represented by post features (Jin et al. 2017). By building particular social networks, such as diffusion networks (Ma et al. 2016) or co-occurrence networks (Ruchansky et al. 2017), network properties can be retrieved. The majority of these social context models can be roughly divided into two categories: propagation-based and stance-based. Users' comments, attitudes, or opinions about the news are used by stance-based models to infer the news's validity (Jin et al. 2017; Shu et al. 2019a). According to Shu et al. (2019a), propagation-based models use propagation methods to simulate various information-spread patterns, including interactions between news sources, publishers, and consumers. Research has been focused on challenging issues with fake news identification, including early detection of fake news using adversarial learning (Wang et al. 2018) and user response generation (Qian et al. 2018), semi-supervised detection (Benamira et al. 2020), and unsupervised detection (Yang et al. 2019).

Early research by Vosoughi et al. (2018) demonstrated that the interaction and dissemination networks of fake news are deeper and larger than those of actual news, which provided the justification for using network information. Additionally, Vosoughi et al. (2018) discovered that fake data propagated more quickly than real information, suggesting the utility of temporal information. According to Zhou et al. (2020b), propagation networks can be homogeneous or heterogeneous and can be evaluated at several scales, including the node-level, ego-level, triad-level, community-level, and the overall network.

In Lu and Li (2020) is proposed the GCAN framework, Graph-aware Co-Attention Networks whose main aim is to enable explainable fake news detection on social media. After employing a dual co-attention approach to capture the correlations between user interaction/propagation and tweet content, Lu et al. concatenate representations of user interaction, word representations, and propagation features. The representation of retweet propagation based on user attributes is learned using convolutional and recurrent neural networks. In order to learn the graph-aware representation of user interactions, a graph convolution network is employed to model the potential interactions between users. The ability to understand the correlation between the source tweet and retweet propagation as well as the co-influence between the source tweet and user engagement is provided by the dual co-attention mechanism. The binary prediction is generated based on the learned embeddings. The framework has been evaluated on a real-world dataset, the Ma dataset (Ma et al. 2016). The outcomes show that the novel approach could be successfully applied for fake news detection by exploiting the propagation network.

In Shu et al. (2019a) is proposed the dEFEND (Explainable FakE News Detection) framework for multimodal fake news detection. It consists of four major components: (1) a news content encoder (including word encoder and sentence encoder) component, (2) a user comment encoder component, (3) a sentence-comment co-attention component, and (4) a fake news prediction component. The news content encoder component describes the modelling from the news linguistic features to latent feature space through a hierarchical word- and sentence-level encoding exploiting a bidirectional Gated recurrent unit (GRU). Similar to a word encoder, it utilizes RNNs with GRU units to encode each sentence in news and to model the word sequences in a comment. In particular, the user comment encoder extracts latent features from comments through word-level attention networks. The sentence-comment co-attention component models the mutual influences between the news sentences and user comments for learning feature representations, and the explainability degree of sentences and comments are learned through the attention weights within co-attention learning, giving high weights of representations of news

sentences and comments that are beneficial to fake news detection. Finally, the fake news prediction component concatenates news content and user comment features for fake news classification. A detailed experimental evaluation has been performed on two Twitter datasets, PolitiFact and GossipCop. Results show the effectiveness of dEFEND.

In Shu et al. (2019) is proposed an approach for fake news detection that models the social context of a news dissemination process as a tri-relationship among publishers, news, and users. To this purpose, authors introduced a tri-relationship embedding framework TriFN, that concurrently models publisher-news relationships and user-news interactions for the classification of fake news. It has five main parts: an embedding of news contents, an embedding of users, an embedding of user-news interactions, an embedding of publisher-news relations, and an embedding of semi-supervised classification.

In Cui et al. (2019) is proposed an approach focusing on user comments left for posts and latent sentiments in detecting fake news. The proposal embeds users' latent sentiments into an end-to-end deep embedding framework called SAME (Sentiment-Aware Multimodal Embedding for Detecting Fake News). The approach as a first task uses different networks to reason with multimodality (i.e. news, news publishers, and users) and then introduces an adversarial mechanism to investigate semantic similarity/correlations across the different modalities; finally, it models user sentiments and incorporates them into the proposed approach. SAME has been validated using two real-world datasets, PolitiFact and GossipCop, and results show it outperforms state-of-the-art methods on both datasets.

In Dong et al. (2018) is proposed DUAL a unified framework that combines news content, social content, and their cross information in order to reveal fake news. The framework extracts features using adaptive methods: news content and social content features are learned using an attention-based bidirectional Gated Recurrent Unit (GRU) and the cross information is learned using a deep neural network. The hidden representation of these two features is then combined into an attention matrix in order to learn an attention distribution over the vectors. The framework has been tested on two real-world benchmark datasets: the LIAR dataset and Buzzfeed News and outperforms the state-of-the-art methods and baseline methods.

The approach in Ruchansky et al. (2017) considers three different modalities: the content of an article, the feedback it receives and the source users that promote it. The paper proposes a model called CSI that consists of three modules: Capture, Score, and Integrate. The capture module is based on the response and employs a recurrent neural network to record the temporal pattern of user activity on a particular article. Score module estimates a source of suspiciousness score based on the behaviour of users; the Integrate module integrates the previous two to classify an article as fake or not. Experimental analysis on two real-world social media datasets - Twitter and Weibo. More specifically two microblog datasets are obtained from Twitter (www.twitter.com) and Sina Weibo (weibo.com). The Twitter dataset contains 498 rumours and 494 non-rumours, whereas the Weibo dataset consists of 2313 rumours and 2351 non-rumours (Ma et al. 2016). Results demonstrate that CSI achieves higher accuracy than existing models and extracts meaningful latent representations of both users and articles.

In Jiang et al. (2019) is proposed a strategy to reconstruct the news-user network that enhances the news and user embeddings in the news propagation network and therefore efficiently detects those users who frequently transmit false information. The paper proposes a comprehensive framework for learning both news content and news-user network properties. The paper describes the user-characteristic enhanced model (UCEM), a unified framework created by learning the network of news users and news textual content, respectively. News is treated as a source user, and a homogeneous network is derived by looking at user friendship network and news propagation. Starting from user profiles, the proposal uses AANE to learn user embeddings in friendship networks. The next step of the approach consists in providing a reconstructed news-user network to learn representations for both users and news. Experimental findings on two real-world datasets, namely PolitiFact and BuzzFeed, show how well the suggested approach works. By combining news content and news-user network embeddings, the proposed model determines whether or not the original news is fake and achieves cutting-edge performance.

In Dou et al. (2021) is proposed the framework named User Preference-aware Fake Detection (UPFD) that accounts for user preference and user past behaviour in the fake news detection task. With regard to encoding, different text representation learning techniques (such BERT) are used to represent user historical posts and news information. The propagation graph for each piece of news based on its cascading social media sharing is also utilized by the algorithm as additional information. The vector representations of users and news are employed as node features to integrate different pieces of data, and a Graph Neural Network (GNN) is used to build a joint user engagement embedding. A neural classifier is trained to identify fake news using user engagement embedding and news textual embedding.

In Silva et al. (2021) is proposed a multimodal fake news detection technique for cross-domain news. The approach is able to learn both domain-specific as well as cross-domain features using two independent embedding spaces, domain-specific embedding and cross-domain embedding, which are subsequently used to identify fake news records. In particular, the approach implements an unsupervised multimodal domain discovery. The textual content of the news and the

propagation network are the two considered modalities. BERT is used to obtain the embedding of the textual content, whereas the propagation network-based representation is modelled using an unsupervised network representation learning technique. The features are then fused through a concatenation approach after being further elaborated by FF-Neural Networks. Three datasets are combined, PolitiFact, GossipCop, and CoAID, to produce a cross-domain news dataset and evaluate the effectiveness of the approach. Experiments have shown that the framework outperforms state-of-the-art fake news detection models by as much as 7.55% in F1-score.

In Mosallanezhad et al. (2022) is proposed REinforced Adaptive Learning Fake News Detection (REAL-FND), a multimodal fake news detection approach that adds auxiliary information (such as user comments and user-news interactions) into a novel reinforcement learning-based model. Specifically, the framework encodes news content, user comments, and user-news interactions as representation vectors. News content features are obtained by exploiting BERT; to encode the article's comments is used a Hierarchical Attention Network (HAN), while user-news interactions are elaborated by means of a Feed Forward Neural Network. The vectors of the different features are fused for fake news detection by concatenating the different features into a vector that is then passed to a feed-forward neural network to combine the different information into a single vector. Such a final vector is then exploited for fake news detection and to adapt the domain. Despite being trained in a separate source domain, REAL-FND uses cross-domain and intra-domain knowledge to make it robust in a target domain. Extensive tests on real-world datasets (GossipCop and PolitiFact) show the model's effectiveness, particularly when there is a lack of labelled data in the target domain.

In Rezayi et al. (2021) is proposed an approach that leverages network, textual, and relaying features such as hashtags and URLs, and classifies articles using the concatenation of the feature embeddings. Textual features are obtained by using word embedding to represent each word by a low-dimensional vector and input this to an LSTM to find the contextual embedding of each tweet. Five tweet-level features are considered as relaying features: hashtag count, URL count, retweet count, mention count, and favorite count. For what concerns network features, the framework constructs a network that captures the interactions between users and tweets, creating this way a directed graph of user mentions such that each tweet is connected to a user if their name is mentioned in the tweet text. Using this graph, authors created a one-hot vector of user mentions per tweet. The framework has been evaluated over two datasets, PHEME and Volkova. Results show that the approach is comparable with state-of-the-art performance.

In Kaliyar et al. (2020) is proposed DeepNet, a binary fake news classifier. DeepNet is modelled as a deep neural network that performs its task by considering not only the content of the news shared on social media but also exploits the relationship the user exhibits in the social network. The proposal is built considering the tensor factorization method; therefore, a tensor is in charge of expressing the social context of news articles as a combination of different information related to the news itself, the user, and group with whom the user interacts. DeepNet is structured as follows: it has one embedding layer, three convolutional layers, one LSTM layer, seven dense layers, ReLU for activation, and finally it uses the softmax function in order to perform the binary classification. DeepNet is tested on the Fakeddit and BuzzFeed datasets. Results show an accuracy of 86.4% on the Fakeddit dataset and 95.2% over the BuzzFeed dataset.

In Wu and Rao (2020) are proposed the Adaptive Interaction Fusion Networks (AIFN) enabling cross-interaction fusion among the different types of features for fake news detection. Key elements of the framework are the gated adaptive interaction networks (GAIN) and the semantic-level fusion self-attention networks (SFSN) modules. The GAIN allows capturing adaptively similar semantics and conflicting semantics between posts and comments, whereas the sFSN modules improve semantic correlations and fusion among features. AIFN learns four types of features around posts and comments from the perspectives of words and emotions. It exploits BERT, Bi-LSTM to extract the features that are then concatenated after applying self-attention mechanisms. The approach is able, through the semantic-level fusion and the self-attention networks, to catch cross-domains features. Extensive experiments on two real-world datasets, i.e. MediaEval and PHEME, demonstrate that AIFN achieves the state-of-the-art performance and boosts accuracy by more than 2.05% and 1.90%, respectively.

Existing methods that take advantage of user social interactions just extract features to train classifiers without having a thorough knowledge of these features, making them difficult to comprehend.

Current research on the spread of fake news mostly focuses on evaluating macro-level propagation and conducts an extensive study on employing different propagation network properties for fake news identification. Shu et al. (2020) created a hierarchical propagation network from macro- and micro-levels to bridge this gap and utilize the features from structural, temporal, and linguistic perspectives for fake news identification.

Numerous fields of study are yet possible by accounting network propagation features. First, by researching the hierarchical propagation network structures, which is a prelude to mitigating fake news dissemination, we can learn to forecast whether a user would distribute a fake news piece or not. Second, we can do unsupervised fake news detection by taking use

of the hierarchical structure of propagation networks. Third, we may combine the explicit propagation network properties with deep learning models to improve fake news detection even more.

4.4 Summary of the main features of the state of the art

Table 1 summarizes the main features of the reviewed papers. Specifically, we have a column about the kind of DL model used, and two more columns reporting information about the considered modalities and the used dataset. The last two columns report information about the data fusion technique and the relevant information related to the eventual event-invariant features.

Different DL approaches for multimodal fake news detection have been reviewed. Many of them use different deep neural networks solutions like CNN, LSTM, GAN, RNN, FF-Neural Networks, and Autoencoders. The majority of the approaches consider only two modalities: text and images. Text analysis is the core to identifying fake news, and BERT is widely and successfully used to analyse the text, whereas for image analysis, CNN-based solutions are commonly used (like VGG-19). Very few approaches exploited the news propagation network characteristics as a further modality to enhance the detection task. In particular, one interesting solution combined BERT with supervised network representation learning improving the overall fake news detection process. The majority of the state-of-the-art proposals exploited different datasets collected either from Weibo or Twitter. The datasets were often created ad hoc, e.g. during a specific crisis such as COVID-19 or war, and, as a consequence, it is unfeasible to use such datasets to train models in different domains or even in the same domain but in different countries, if the fake news approach it is not addressing cross-domain. The majority of the approaches use concatenation and attention mechanisms as early fusion techniques, and just a few works adopt different strategies such as variational autoencoder. The cross-domain characteristic is the ability of the fake news detection method to model different application domains by identifying event-invariant features. Unfortunately, only a few of the analysed approaches addressed such a challenging issue, with different interesting solutions. Among such techniques particularly promising are adversarial learning to classify events and thus adapt the domain, meta-learning neural process, event memory network, self-attention mechanisms, reinforcement learning, unsupervised multimodal domain, and cross-domain embedding.

5 Challenges and future directions

Fake news affects both online and offline social communities and different proposals exist in the recent literature investigating at different levels and with different strategies the problem. Multimodal approaches for fake news detection have been proven to be a viable effective approach to address disinformation, however, many are still challenges that remain to be addressed.

- *Datasets* Different multimodal datasets exist, but they are often related to two or a few modalities such as text and images. These datasets have generally small sizes, expose content in just one language, and often are imbalanced either in the fake or real news. An additional issue is that, in order to cope with different styles and different topics, datasets from heterogeneous platforms should be available. Therefore, urgent is a need for real and complete multimodal datasets containing different modalities such as text, images, video audio, social content, and temporal and network propagation features.
- *Finer classification* Existing fake news detection models are mainly binary classifiers that determine whether a piece of news is false or not. This strategy is often not sufficient and a multi-class classification or even a regression task should be used. The final aim should allow enabling prioritized reasoning and consequent strategies in the presence of fake news detection.
- *Scalability* Since deep neural networks are complex and costly to build, and as most existing multimodal models use multiple deep neural networks (one per modality), they are not scalable as the number of modalities grows. Furthermore, many existing models require extensive computing resources, including large amounts of memory storage and processing units. As a result, when developing new architectures, the scalability of proposed models should be considered.
- *Enhancement of basic multimodal classifiers* Many deep learning advanced techniques have been applied in order to improve the performance of multimodal classifiers for misinformation detection. The concatenation of vector representations does not always result in an effective multimodal embedding. Thus, some recent works used the attention mechanism to focus on relevant parts of images or texts. In order to get the most out of embeddings, using an attention mechanism is preferable since it results in richer multimodal representations. One of the most relevant problems when training classifiers using supervised approaches is the lack of labelled and balanced datasets. In order to solve this problem, generative models have been used. They are trained to learn the

patterns that characterize misinformative content and can be used to create synthetic balanced datasets or augment existing ones. As a result of their effectiveness in detecting fake news, Graph Neural Networks (GNNs) are now being studied for their potential to detect misinformation in many media. Users, multimodal content, and relationships among them are modelled by means of a GNN. The learning process derives embeddings for users and multimodal content that can be used by a classifier to detect misinformative content.

- *Source verification and author credibility* Only a limited number of existing approaches evaluate either author's credibility or the veracity of the source of the news article. Those two tasks should be deeply explored in future research. Source credibility is a key point when evaluating fake news as well as author credibility as this last allows an automatic system to retrieve the chain of news authored by the same author or group of authors.
- *Cross-domain* When trained on vast volumes of labelled data on events of interest, deep learning-based models perform well, but when instructed on different events due to domain shift, their performance tends to decline. Because it is challenging to get large-scale labelled information, detecting fake news on emergent events poses substantial challenges for current detection algorithms. Furthermore, including new information from emerging events necessitates either creating an entirely new model from scratch or continuing to refine an existing one, both of which can be difficult, expensive, and unrealistic for use in real-world contexts.
- *Explainability* The explainability of models is largely unexplored. This task is relevant and should be focused on future methods in order to obtain transparent models that provide decisions/suggestions explainable and transparent. More specifically, fake news detection systems have to fulfil some general requirements: they have to provide decisions/suggestions, but also justify how and/or why the provided decisions/suggestions have been given. The justification should be provided by flagging the different pieces of the news with the corresponding truth value (true or false or using a finer granularity) and presented in an easy way. In addition, justification in the output should also include ethical considerations. Concerning this specific task, it should be noticed that, to the best of our knowledge, no dataset exists that contains fake news accompanied by the justification of disinformation.
- *Enhancing the integration of news content features (either text, image or video) with network and propagation features in DL models* Features related to the creator of the news, the characteristics of the network in which the news spreads and the propagation-based information are not fully explored in combination with news content in current DL proposals and are research directions on which the community should investigate more in the next future. A more depth analysis of network and propagation-based features and their fusion with well-known adopted modalities would improve fake information detection.
- *Enhancing the exploitation of emotions expressed in the texts to detect fake news* The use of the emotions extracted from the text combined with additional modalities could enhance the task of fake news detection. The motivation relies on some studies in the literature showing that fake news triggers different emotions in users compared to real news. More specifically (Vosoughi et al. 2018) showed that generally false rumours on Twitter caused followers to react with fear, disgust, and astonishment whereas true rumours caused them to respond with joy, grief, trust, and anticipation. A more in-depth analysis of this specific issue and the embedding of well-known techniques used to catch emotions from a text, such as those based on lexicon or on neural networks could improve fake news detection.
- *Enhancing the exploitation of statistical features to detect fake news* Statistical features can provide a synthetic representation of the relevant information and easily allow to evidence the quantitative distribution patterns that characterize fake and real news. This specific issue could be profitably used to complement the information provided by the different modalities, such as image, video, and audio, as well as social content.

6 Conclusion

The paper provided a rigorous and in-depth survey on a very specific topic related to the use of deep learning for multimodal fake news detection on social media. The paper analysed a large number of deep learning approaches and provided, for each work surveyed, an analysis of the rationale behind the approach, highlighting some relevant features such as the DL method used, the type of data analysed, the datasets used, the fusion strategy adopted and the eventual domain-invariant features. The survey also discusses the main limitations of the current approaches and the challenges that remain to be addressed by future research works including effective use of cross-domain fake news detection strategies.

Acknowledgements This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

Abdali S (2022) Multi-modal misinformation detection: approaches, challenges and opportunities

Alam F, Cresci S, Chakraborty T, Silvestri F, Dimitrov D, Martino GDS, Shaar S, Firooz H, Nakov P (2022) A survey on multimodal disinformation detection. In: Proceedings of the 29th international conference on computational linguistics. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp 6625–6643

Alonso-Bartolome S, Segura-Bedmar I (2021) Multimodal fake news detection. arXiv. https://doi.org/10.48550/ARXIV.2112.04831

Amerini I, Ballan L, Caldelli R, Del Bimbo A, Serra G (2011) A sift-based forensic method for copy-move attack detection and transformation recovery. In: IEEE Transactions on information forensics and security, vol 6, pp 1099–1110. https://doi.org/10.1109/TIFS.2011.2129512

Amri S, Sallami D, Aïmeur E (2021) Exmulf: an explainable multimodal content-based fake news detection system. Springer, Berlin, pp 177–187. https://doi.org/10.1007/978-3-031-08147-7_12

Benamira A, Devillers B, Lesot E, Ray AK, Saadi M, Malliaros FD (2020) Semi-supervised learning and graph neural networks for fake news detection. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining. ASONAM'19, pp 568–569

Boididou C, Papadopoulos S, Zampoglou M, Apostolidis L, Papadopoulou O, Kompatsiaris Y (2018) Detection and visualization of misleading content on twitter. Int J Multimed Info Retr 7:71–86. https://doi.org/10.1007/s13735-017-0143-x

Bovet A, Makse HA (2019) Influence of fake news in twitter during the 2016 US presidential election. Nat Commun. https://doi.org/10.1038/s41467-018-07761-2

Cao J, Sheng Q, Qi P, Zhong L, Wang Y, Zhang X (2019) False news detection on social media. arXiv. https://doi.org/10.48550/ARXIV.1908.10818

Cui L, Lee D (2020) CoAID: COVID-19 healthcare misinformation dataset

Cui L, Wang S, Lee D (2019) SAME: Sentiment-aware multi-modal embedding for detecting fake news. In: 2019 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), pp 41–48. https://doi.org/10.1145/3341161.3342894

da Silva FCD, Vieira R, Garcia ACB (2019) Can machines learn to detect fake news? A survey focused on social media. In: HICSS

Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC (2019) The DeepFake detection challenge (DFDC) preview dataset

Dong M, Yao L, Wang X, Benatallah B, Sheng QZ, Huang H (2018) Dual: A deep unified attention model with latent relation representations for fake news detection. In: Hacid H, Cellary W, Wang H, Paik H-Y, Zhou R (eds) WISE, pp 199–209

Dou Y, Shu K, Xia C, Yu PS, Sun L (2021) User preference-aware fake news detection. arXiv

Eyben F, Weninger F, Groß F, Schuller B (2013) Recent developments in openSMILE, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on multimedia

Ferreira W, Vlachos A (2016) Emergent: a novel data-set for stance classification. https://doi.org/10.18653/v1/N16-1138

Giachanou A, Zhang G, Rosso P (2020) Multimodal multi-image fake news detection. In: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA), pp 647–654. https://doi.org/10.1109/DSAA49011.2020.00091

Hameleers M, Powell TE, Meer TGLAVD, Bos L (2020) A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. Polit Commun 37(2):281–301. https://doi.org/10.1080/10584609.2019.1674979

Hangloo S, Arora B (2022) Combating multimodal fake news on social media: methods, datasets, and future perspective. Multimedia Syst 28:2391–2422

Hua J, Cui X, Li X, Tang K, Zhu P (2023) Multimodal fake news detection through data augmentation-based contrastive learning. Appl Soft Comput 136:110125. https://doi.org/10.1016/j.asoc.2023.110125

Ji S, Xu W, Yang M, Yu K (2013) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231. https://doi.org/10.1109/TPAMI.2012.59

Jiang S, Chen X, Zhang L, Chen S, Liu H (2019) User-characteristic enhanced model for fake news detection in social media. In: Tang J, Kan M, Zhao D, Li S, Zan H (eds) Natural language processing and Chinese computing—8th CCF international conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I. Lecture notes in computer science, vol 11838. Springer, Berlin, pp 634–646

Jin Z, Cao J, Guo H, Zhang Y, Luo J (2017) Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM international conference on multimedia. MM'17. Association for Computing Machinery, New York, pp 795–816

Jing Q, Yao D, Fan X, Wang B, Tan H, Bu X, Bi J (2021) TRANSFAKE: Multi-task transformer for multimodal enhanced fake news detection. In: IJCNN, pp 1–8

Kaliyar RK, Kumar P, Kumar M, Narkhede M, Namboodiri S, Mishra S (2020) Deepnet: an efficient neural network for fake news detection using news-user engagements. In: 2020 5th International conference on computing, communication and security (ICCCS), pp 1–6. https://doi.org/10.1109/ICCCS49678.2020.9277353

Karimi H, Tang J, Li Y (2018) Toward end-to-end deception detection in videos. In: 2018 IEEE international conference on big data (Big Data), pp 1278–1283. https://doi.org/10.1109/BigData.2018.8621909

Khattar D, Goud JS, Gupta M, Varma V (2019) Mvae: Multimodal variational autoencoder for fake news detection. In: The world wide web conference. WWW'19, pp 2915–2921

Kirchknopf A, Slijepčević D, Zeppelzauer M (2021) Multimodal detection of information disorder from social media. In: CBMI Conf., pp 1–4. https://doi.org/10.1109/CBMI50038.2021.9461898

Korshunov P, Marcel S (2018) DeepFakes: a new threat to face recognition? Assessment and detection

Krishnamurthy G, Majumder N, Poria S, Cambria E (2018) A deep learning approach for multimodal deception detection. arXiv. https://doi.org/10.48550/ARXIV.1803.00344

Kumari R, Ekbal A (2021) AMFB: Attention based multimodal factorized bilinear pooling for multimodal fake news detection. Expert Syst Appl 184:115412

Kumar S, Shah N (2018) False information on web and social media: a survey. arXiv. https://doi.org/10.48550/ARXIV.1804.08559

Li Y, Xie Y (2020) Is a picture worth a thousand words? an empirical study of image content and social media engagement. J Mark Res 57(1):1–19. https://doi.org/10.1177/0022243719881113

Lu Y-J, Li C-T (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, pp 505–514

Ma J, Gao W, Mitra P, Kwon S, Jansen BJ, Wong K-F, Cha M (2016) Detecting rumors from microblogs with recurrent neural networks. In: IJCAI'16, pp 3818–3824

McCrae S, Wang K, Zakhor A (2022) Multi-modal semantic inconsistency detection in social media news posts. In: Multimedia modeling. Springer, Berlin, pp 331–343

Mendels G, Levitan SI, Lee K-Z, Hirschberg J (2017) Hybrid acoustic-lexical deep learning approach for deception detection. In: INTERSPEECH

Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) Emotions Don't Lie: an audio-visual deepfake detection method using affective cues

Mosallanezhad A, Karami M, Shu K, Mancenido MV, Liu H (2022) Domain adaptive fake news detection via reinforcement learning. In: Proceedings of the ACM web conference 2022. https://doi.org/10.1145/3485447.3512258

Mu M, Bhattacharjee SD, Yuan J (2023) Self-supervised distilled learning for multi-modal misinformation identification. In: IEEE/CVF Winter conference on applications of computer vision, WACV 2023, Waikoloa, HI, USA, January 2–7, 2023. IEEE, pp 2818–2827

Murayama T (2021) Dataset of fake news detection and fact verification: a survey

Nakamura K, Levy S, Wang WY (2019) r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. arXiv. https://doi.org/10.48550/ARXIV.1911.03854

Pérez-Rosas V, Abouelenien M, Mihalcea R, Burzo M (2015) Deception detection using real-life trial data. In: 2015 ACM on international conference on multimodal interaction, pp 59–66

Qian F, Gong C, Sharma K, Liu Y (2018) Neural user response generator: fake news detection with collective user intelligence. In: Proceedings of the twenty-seventh international joint conference on artificial intelligence, IJCAI-18, pp 3834–3840

Qi P, Cao J, Li X, Liu H, Sheng Q, Mi X, He Q, Lv Y, Guo C, Yu Y (2021) Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues, pp 1212–1220

Raj C, Meel P (2021) Convnet frameworks for multi-modal fake news detection. Appl Intell. https://doi.org/10.1007/s10489-021-02345-y

Rezayi S, Soleymani S, Arabnia HR, Li S (2021) Socially aware multimodal deep neural networks for fake news classification. In: 2021 IEEE 4th international conference on multimedia information processing and retrieval (MIPR), pp 253–259. https://doi.org/10.1109/MIPR51284.2021.00048

Ruchansky N, Seo S, Liu Y (2017) CSI: A hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on conference on information and knowledge management. ACM. https://doi.org/10.1145/3132847.3132877

Sachan T, Pinnaparaju N, Gupta M, Varma V (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. ASONAM'21, pp 399–406

Shang L, Kou Z, Zhang Y, Wang D (2021) A multimodal misinformation detector for COVID-19 short videos on Tiktok. In: 2021 IEEE international conference on big data (big data), pp 899–908. https://doi.org/10.1109/BigData52589.2021.9671928

Shang L, Kou Z, Zhang Y, Wang D (2022) A duo-generative approach to explainable multimodal COVID-19 misinformation detection. In: Proceedings of the ACM web conference 2022. WWW'22, pp 3623–3631

Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. SIGKDD Explor Newsl 19(1):22–36

Shu K, Cui L, Wang S, Lee D, Liu H (2019a) dEFEND: Explainable fake news detection. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'19, pp 395–405

Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2019b) FakeNewsNet: a data repository with news content, social context and spatial temporal information for studying fake news on social media

Shu K, Mahudeswaran D, Wang S, Liu H (2020) Hierarchical propagation networks for fake news detection: Investigation and exploitation. In: Proceedings of the international AAAI conference on web and social media, vol 14, issue 1, pp 626–637

Shu K, Wang S, Liu H (2019) Beyond news contents: The role of social context for fake news detection. In: Proceedings of the twelfth ACM international conference on web search and data mining. WSDM'19, pp 312–320

Silva A, Luo L, Karunasekera S, Leckie C (2021) Embracing domain differences in fake news: cross-domain fake news detection using multi-modal data. In: The thirty-fifth AAAI conference on artificial intelligence (AAAI-21). https://doi.org/10.48550/ARXIV.2102.06314

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition

Singhal S, Dhawan M, Shah RR, Kumaraguru P (2021) Inter-modality discordance for multimodal fake news detection. In: MMAsia

Song C, Ning N, Zhang Y, Wu B (2021) A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf Process Manag 58(1):102437

Volkova S, Shaffer K, Jang JY, Hodas N (2017) Separating facts from fiction: linguistic models to classify suspicious and trusted news posts on Twitter. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers), pp 647–653

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359:1146–1151. https://doi.org/10.1126/science.aap9559

Wang WY (2017) "Liar, Liar Pants on Fire": a new benchmark dataset for fake news detection. arXiv. https://doi.org/10.48550/ARXIV.1705.00648

Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, Su L, Gao J (2018) Eann: Event adversarial neural networks for multi-modal fake news detection. In: KDD, pp 849–857

Wang Y, Ma F, Wang H, Jha K, Gao J (2021) Multimodal emergent fake news detection via meta neural process networks. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery and data mining. KDD'21, pp 3708–3716

Wang K, Chan D, Zhao SZ, Canny J, Zakhor A (2022) Misinformation detection in social media video posts

Wu L, Rao Y (2020) Adaptive interaction fusion networks for fake news detection. In: 24th European conference on artificial intelligence—ECAI 2020

Xie J, Liu S, Liu R, Zhang Y, Zhu Y (2021) SERN: Stance extraction and reasoning network for fake news detection. In: ICASSP, pp 2520–2524. https://doi.org/10.1109/ICASSP39728.2021.9414787

Xue J, Wang Y, Tian Y, Li Y, Shi L, Wei L (2021) Detecting fake news by exploring the consistency of multimodal data. Inf Process Manag 58(5):102610

Yang Y, Zheng L, Zhang J, Cui Q, Li Z, Yu PS (2018) TI-CNN: convolutional neural networks for fake news detection. arXiv

Yang S, Shu K, Wang S, Gu R, Wu F, Liu H (2019) Unsupervised fake news detection on social media: a generative approach. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, issue 01, pp 5644–5651

Zannettou S, Caulfield T, Blackburn J, De Cristofaro E, Sirivianos M, Stringhini G, Suarez-Tangil G (2018) On the origins of memes by means of fringe web communities. arXiv. https://doi.org/10.48550/ARXIV.1805.12512

Zhang DY, Shang L, Geng B, Lai S, Li K, Zhu H, Amin MT, Wang D (2018) FauxBuster: A content-free fauxtography detector using social media comments. In: 2018 IEEE Big Data Conf., pp 891–900. https://doi.org/10.1109/BigData.2018.8622344

Zhang H, Fang Q, Qian S, Xu C (2019) Multi-modal knowledge-aware event memory network for social media rumor detection. In:

Proceedings of the 27th ACM international conference on multimedia. MM'19, pp 1942–1951

Zhang T, Wang D, Chen H, Zeng Z, Guo W, Miao C, Cui L (2020) BDANN: Bert-based domain adaptation neural network for multimodal fake news detection. In: 2020 International joint conference on neural networks (IJCNN), pp 1–8. https://doi.org/10.1109/IJCNN48605.2020.9206973

Zhou X, Mulay A, Ferrara E, Zafarani R (2020a) Recovery: a multimodal repository for COVID-19 news credibility research. In: CIKM'20. Association for Computing Machinery, New York, pp 3205–3212. https://doi.org/10.1145/3340531.3412880

Zhou X, Wu J, Zafarani R (2020b) SAFE: similarity-aware multi-modal fake news detection. arXiv. https://doi.org/10.48550/ARXIV.2003.04981

Zubiaga A, Liakata M, Procter R, Hoi GWS, Tolmie P (2016) Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLOS ONE 11(3):0150989. https://doi.org/10.1371/journal.pone.0150989