A Survey on Multimodal Disinformation Detection

Firoj Alam[1], **Stefano Cresci**[2], **Tanmoy Chakraborty**[3*], **Fabrizio Silvestri**[4], **Dimitar Dimitrov**[5], **Giovanni Da San Martino**[6], **Shaden Shaar**[1], **Hamed Firooz**[7], **Preslav Nakov**[8]

[1]**Qatar Computing Research Institute, HBKU**, [2]**IIT-CNR, Pisa**, [3]**IIT Delhi**, [4]**Sapienza University of Rome, Italy**, [5]**Sofia University**, [6]**University of Padova, Italy**, [7]**Facebook AI**, [8]**Mohamed bin Zayed University of Artificial Intelligence**

{fialam, sshaar}@hbku.edu.qa, s.cresci@iit.cnr.it, tanchak@ee.iitd.ac.in, fsilvestri@diag.uniroma1.it
mitko.bg.ss@gmail.com, dasan@math.unipd.it,mhfirooz@fb.com, preslav.nakov@mbzuai.ac.ae

Abstract

Recent years have witnessed the proliferation of offensive content online such as fake news, propaganda, misinformation, and disinformation. While initially this was mostly about textual content, over time images and videos gained popularity, as they are much easier to consume, attract more attention, and spread further than text. As a result, researchers started leveraging different modalities and combinations thereof to tackle online multimodal offensive content. In this study, we offer a survey on the state-of-the-art on *multimodal disinformation detection* covering various combinations of modalities: text, images, speech, video, social media network structure, and temporal information. Moreover, while some studies focused on *factuality*, others investigated how *harmful* the content is. While these two components in the definition of disinformation – (*i*) factuality, and (*ii*) harmfulness –, are equally important, they are typically studied in isolation. Thus, we argue for the need to tackle disinformation detection by taking into account multiple modalities as well as both factuality and harmfulness, in the same framework. Finally, we discuss current challenges and future research directions.

1 Introduction

The proliferation of online social media has encouraged individuals to freely express their opinions and emotions. On one hand, the freedom of speech has led to a massive growth of online content which, if systematically mined, can be used for citizen journalism, public awareness, political campaigning, etc. On the other hand, its misuse has given rise to the proliferation of hostility online (Brooke, 2019; Joksimovic et al., 2019), resulting in offensive content in the form of fake news, hate speech (Schmidt and Wiegand, 2017a; Davidson et al., 2017), propaganda (Da San Martino et al., 2019), cyberbullying (Van Hee et al., 2015), etc. Indeed, researchers have argued that this situation has set the dawn of the Post-Truth Era, dominated by emotions and "alternative facts" (Lewandowsky et al., 2017; Cooke, 2018; Nakov and Da San Martino, 2020). More recently, with the emergence of the COVID-19 pandemic, a new blending of medical and political false information has given rise to the first global infodemic (Paka et al., 2021; Zarocostas, 2020; Patwa et al., 2021).[1]

The term "fake news" is commonly used, although it is very generic, and misleads people to focus only on veracity. That is why international organizations such as the UN, WHO, EU, and NATO prefer the term ***disinformation*** (Ireton and Posetti, 2018), which refers to information that is (*i*) *fake* and also (*ii*) spreads *deliberately to deceive and harm* others. The latter aspect of the disinformation (i.e., harmfulness) is often ignored, but it is equally important. A related term is *misinformation*, which also refers to the spreading of false content, but lacks the underlying intention to do harm. This is illustrated by the definitions of these notions by First Draft (Ireton and Posetti, 2018) where ***misinformation*** is defined as "*unintentional mistakes such as inaccurate photo captions, dates, statistics, translations, or when satire is taken seriously*", while ***disinformation*** is "*fabricated or deliberately manipulated text/speech/visual context, and also intentionally created conspiracy theories or rumors*".

In our survey, we will focus on disinformation, and we will study both the factuality and harmfulness aspects of the problem, with focus on different modalities. Note that there are posts that can be harmful but factually true or non-factual but harmful (e.g., hate speech); our study also covers some related work on them. The term *factuality* refers to automatically evaluating the solidity of the reporting/social media statements in terms of facts and

[*] Work done while T. Chakraborty was at IIIT-Delhi, India.

[1] https://www.who.int/health-topics/infodemic

Proceedings of the 29th International Conference on Computational Linguistics, pages 6625–6643
October 12–17, 2022.

figures (Ireton and Posetti, 2018). The *harmfulness* or *harmful content* typically refers to "*anything online which causes a person distress or harm*".[2] Figure 2, in Appendix, gives examples of such content. Alam et al. (2021) addressed both aspects of disinformation using social media content related to the COVID-19 infodemic. They demonstrated a correlation between factuality and harmfulness, which varies across languages even in the same country, e.g., for Arabic, 56% of the false content was harmful, while for English, it was 24%.

Disinformation often spreads as text. However, Internet and social media allow the use of different modalities, which can make a disinformation message attractive as well as impactful, e.g., a meme or a video is much easier to consume, attracts much more attention, is perceived as more credible (Hameleers et al., 2020), spreads further than simple text (Zannettou et al., 2018), and can be weaponized (Olsen, 2018).

Notably, multimodality remains under-explored in disinformation detection. Bozarth and Budak (2020) performed a meta-review of 23 fake news models and the data modality they leveraged, and found that 91.3% used text, 47.8% looked into social media network structure, 26% relied on temporal data, and only a handful made use of images or videos. Moreover, while there has been research to detect whether an image or a video has been manipulated, the attempt is less in a truly multimodal setting (Pérez-Rosas et al., 2015; Tan et al., 2020; Zhang et al., 2022; Song et al., 2021; Giachanou et al., 2020; Denaux and Gomez-Perez, 2020).

Here we survey research on multimodal disinformation detection covering various combinations of modalities: text, images, speech, video, social media network structure, and temporal information. The data sources include social media (e.g., Twitter), news, video (e.g., courtroom trials), and TV shows. We further argue for the need to cover multiple modalities in the same framework, while taking both factuality and harmfulness into account.

While there have been a number of surveys on "fake news" (Shu et al., 2017; Kumar and Shah, 2018; Cardoso Durier da Silva et al., 2019; Zhou and Zafarani, 2020), misinformation (Islam et al., 2020), fact-checking (Thorne and Vlachos, 2018; Kotonya and Toni, 2020a), truth discovery (Li et al., 2016), rumour detection (Bondielli and Marcelloni, 2019), harmful memes (Sharma et al., 2022) and propaganda detection (Da San Martino et al., 2020), none of them had multimodality as the main focus. Moreover, they targeted either factuality (most surveys above), or harmfulness (the latter survey), but not both. Here, we aim to bridge this gap. Therefore, in the present survey, we analyze the literature covering various aspects of multimodality (text, image, speech, video, network, and temporal), with a focus on the two aspects of disinformation: factuality and harmfulness, as shown in Figure 1.

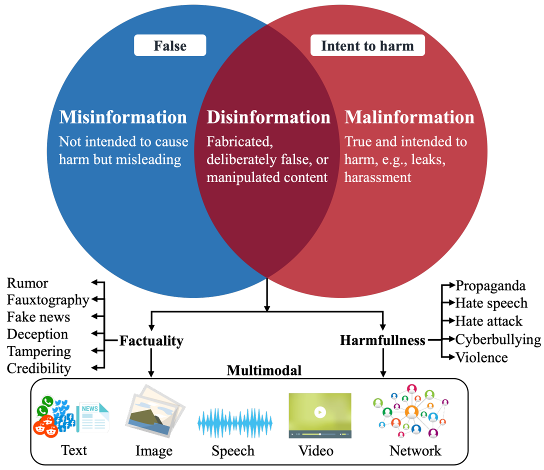

Figure 1: Our vision of multimodality to interact with harmfulness and factuality in this survey.

2 Multimodal Factuality Prediction

In this section, we focus on the first aspect of disinformation – *factuality*. Automatic detection of factual claims is important to debunk the spread of misleading information, as it is crucial to detect the factuality of statements that can mislead people. A large body of work has been devoted to fact-checking textual claims but such claims are often expressed and disseminated together with other modalities such as images, speech, and video, and are further propagated through social networks. We summarize relevant studies in Table 1.

2.1 Text

Due to the availability of large amounts of textual content, research on the text modality is comparatively richer than for other modalities. Notable work in this direction covers fake news spread on social media (Vosoughi et al., 2018b), fake news and fact-checking on news media (Rashkin et al., 2017), fact-checking such as fact-checked URL recommendation model (Vo and Lee, 2018) to reduce the spread, fact-checking with stance detection (Baly et al., 2018b), factuality of media outlets

[2] https://swgfl.org.uk/services/report-harmful-content/

Ref.	Task	Modality T	I	V	N	S	Data/Source	Anno.	Lang	Method
(Baly et al., 2020)	Bias, factuality	✓				✓	MBFC	M	En	SVM, BERT
(Dinkov et al., 2019)	Bias	✓				✓	MBFC	M	En	MM deep learning architecture
(Baly et al., 2018a)	Bias, factuality	✓					MBFC	M	En	SVM
(Shao et al., 2018)	Credibility∗	✓					Articles and tweets	M	En	Statistical analysis
(Sen et al., 2020)	Deception			✓			CTD: 121 videos	M	En	RF, SVM and NN classifiers
(Soldner et al., 2019)	Deception			✓			TV Show	M	En	RF
(Volkova et al., 2019)	Deception	✓	✓				Twitter; T1: 2,485, T2-T3: 56,691, T4: 496,929	M	En	Feature fusion with AdaBoost/NN
(Krishnamurthy et al., 2018)	Deception			✓			CTD: 121 videos	M	En	MLP
(Kaya and Karpov, 2016)	Deception					✓	CSC: 25 videos	M	En	PLS/ELM based model
(Levitan et al., 2016)	Deception					✓	CSC: 25 videos	M	En	SMO, Bagging, Dagging, BN and NB
(Pérez-Rosas et al., 2015)	Deception			✓			CTD: 121 videos	M	En	DT and RF
(Hirschberg et al., 2005)	Deception					✓	CSC: 25 videos	M	En	Rule-based classifier
(Kazemi et al., 2021)	Facuality	✓					FEVER	M	En	Deep Q-learning network
(Atanasova et al., 2020)	Factuality	✓					Liar-Plus	M	En	DistilBERT
(Sathe et al., 2020)	Facuality	✓					WikiFactCheck	M	En	SVM, Decomposable attention model
(Shaar et al., 2020)	Facuality	✓					Political debates	M	En	Learning-to-rank approach, BM25, BERT, RoBERTa, sentence-BERT
(Kopev et al., 2019)	Facuality	✓		✓			Political debates	M	En	MM fusion: concatenation
(Vo and Lee, 2018)	Facuality	✓					Fact-checked tweets from: Snopes.com, Politifact.com, FactCheck.org, OpenSecrets.org, TruthOrfiction.com and Hoax-slayer.net	M	En	BPRMF , MF, CoFactor, CTR, proposed a joint model
(Baly et al., 2018b)	Facuality	✓					Claims from Verify and Reuters	M	Ar	Gradient boosting, multilayer perceptron, softmax layer, end-to-end memory network
(Rashkin et al., 2017)	Facuality	✓					Politifact	M	En	LSTM, MaxEnt, NB
(Nguyen et al., 2020)	Fake news	✓			✓		PHEME , Twitter (snopes.com), Weibo, FakeNewsNet	M	En	Graphical social context
(Nakamura et al., 2020)	Fake news	✓	✓				Reddit: 1m posts	DS	En	MM fusion
(Shu et al., 2020)	Fake news	✓			✓		PolitiFact and GossipCop	M	En	GNB, DT, LR, and RF
(Shu et al., 2019)	Fake news	✓			✓		BuzzFeed and PolitiFact	M	En	LR, NB, DT, XGBoost, AdaBoost, and GB
(Vosoughi et al., 2018a)	Fake news	✓			✓		Twitter: 126,000 posts	M	En	Statistical analysis, Topic modeling
(Liu and Wu, 2018)	Fake news	✓			✓		Weibo: 4,664 (Ma et al., 2016), Twitter15: 1,490 (Ma et al., 2017), Twitter16: 818 (Ma et al., 2017)	M	En	DT, SVM, GRU, RF, RNN, CNN
(Rashkin et al., 2017)	Fake news	✓					Gigaword corpus, articles from seven unreliable news sites	M	En	MaxEnt
(Boididou et al., 2016)	Fake		✓	✓			Social media	M	En	-
(Gupta et al., 2013)	Fake news				✓		Twitter: 16,117 tweets	M	En	DT on balanced dataset, NB
(Wang et al., 2021)	Fauxtography	✓	✓				Twitter, 4chan, and Reddit	M	En	Analytical
(Zhang et al., 2018)	Fauxtography	✓	✓				Reddit: 91, Twitter: 390	M	En	Feature fusion with XGBoost
(Heller et al., 2018)	Image tampering∗∗		✓				Reddit: 102,028 images	A	-	-
(Garimella and Eckles, 2020)	Misinformation∗∗		✓				WhatsApp: 2,500 images	M	-	-
(Zannettou et al., 2018)	Memes propagation	✓	✓				Twitter, Reddit, 4chan, and Gab	DS	-	Memes analysis
(Vosoughi et al., 2017)	Rumor	✓			✓		Twitter: 113 false and 96 true	M	En	Temporal, propagation linguistic, and user credibility features
(Kwon et al., 2017)	Rumor	✓			✓		Twitter, snopes.com, and urban-legends.about.com	M	En	RF

Table 1: Summary of the most relevant works on factuality, covering different modalities and tasks. **T:Text**, **I: Image**, **V:Video**, **N:Network**, **S:Speech**. CTD: Courtroom trial dataset, CSC: Columbia/SRI/Colorado Corpus. Anno.: Annotation, M: manual annotation; DS: distant supervision. MM: Multimodal, SVM: Support Vector Machine, RF: Random Forest, DT: Decision Tree; NN: Neural Network, MLP: Multi-layer Perceptron, PLS: Partial Least Squares regression; ELM: Extreme Learning Machines, NB: Naïve Bayes, BN: BayesNet, BPRMF: Bayesian Personalized Ranking Matrix Factorization, , MF: Matrix Factorization, CTR: Collaborative Filtering Regression, GNB: Gaussian Naive Bayes; LR: Logistic Regression; GB: Gradient Boosting, GRU: Gated Recurrent Units, RNN: Recurrent Neural Networks, CNN: Convolutional Neural Networks. ∗ Also include botometer features. T1-T4 represents different tasks. ∗∗ dataset only.

(Baly et al., 2020, 2018a), generating justifications for verdicts on claims (Atanasova et al., 2020), and fact-checking claims from Wikipedia (Sathe et al., 2020). There have also been recent efforts for fact-checking from political debates (Shaar et al., 2020, 2022, 2021; Nakov et al., 2022b,a), fact-checking with evidence reasoning (Si et al., 2021; Jiang et al., 2021; Wan et al., 2021) and fact-checking by claim matching (Kazemi et al., 2021). Given that there have been surveys on the text modality for fake news/disinformation detection and fact-checking, here we will not go into more detail about the individual studies.

2.2 Image

Text with visual content (e.g., images) in social media is more prominent as it is more intuitive; thus, it is easier to consume, it spreads faster, it gets 18% more clicks, 89% more likes, and 150% more retweets (Zhang et al., 2018). Due to the growing number of claims disseminated with images, in the current literature, there have been various studies that address the visual content with text for predicting misleading information (Volkova et al., 2019), fake images (Gupta et al., 2013), images shared with misinformation in political groups (Garimella and Eckles, 2020), and fauxtography (Zhang et al., 2018; Wang et al., 2021). Some of these studies attempt to understand how two different modalities are used. Their analyses show that the extension of text with images increases the effectiveness of misleading content. Gupta et al. (2013) highlighted the role of Twitter to spread fake images. This study reports that 86% tweets spreading fake images are retweets. Garimella and Eckles (2020) manually annotated a sample of 2,500 images collected from public WhatsApp groups, and labeled them as *misinformation*, *not misinformation*, *misinformation already fact-checked*, and *unclear*; however, experiments were conducted with binary labels: misinformation vs. not-misinformation. The authors found that violent and graphic images spread faster. Nakamura et al. (2020) developed a multimodal dataset containing 1M posts including text, images, metadata, and comments collected from Reddit. The dataset was labeled with 2, 3, and 6-ways labels. Volkova et al. (2019) proposed models for detecting misleading information using images and text.

Fauxtography is defined as *"visual images, especially news photographs, which convey a questionable (or outright false) sense of the events they seem to depict"* (Cooper, 2007). It is also commonly used in social media in different forms such as a fake image with false claims, a true image with false claims. Zhang et al. (2018) defined that *"a post is a fauxtography if the image of the post (i) directly supports a false claim, or (ii) conveys misinformation of a true claim."* An example is shown in Figure 2 (in Appendix A). Zhang et al. (2018) developed FauxBuster to detect fauxtographic social media content, which uses social media comments in addition to the content in the images and the texts. Zlatkova et al. (2019) investigated the factuality of claims with respect to images and compared the performance of different feature groups between text and images. Wang et al. (2021) analyzed fauxtography images in social media posts and found that posts with doctored images increase user engagement in the form of re-shares, likes, and comments, specifically in Twitter and Reddit. They pointed out that doctored images are often used as memes to mislead or as a means of satire, and that they have a 'clickbait' power to drive engagement.

2.3 Speech/Audio

There have been attempts to use acoustic signals to predict the factuality of claims in political debates (Kopev et al., 2019; Shaar et al., 2020), left-center-right bias in YouTube channels (Dinkov et al., 2019), and deception in speech (Hirschberg et al., 2005). Kopev et al. (2019) found that the acoustic signal helps in improving the performance compared to using only textual and metadata features. Similarly, Dinkov et al. (2019) reported that the use of speech signal improves the performance of the system for detecting the political bias (i.e., left, center, right) of Youtube channels. Moreover, a large body of work was done on deception detection using the acoustic signal. Hirschberg et al. (2005) created the Columbia-SRI-Colorado (CSC) corpus by eliciting within-speaker deceptive and non-deceptive speech. Their experiments consist of the use of acoustic, prosodic, and a variety of lexical features including 68 LIWC categories, filled pauses, and paralinguistic information (e.g., speaker information, gender, field-pause). Using the same corpus, an evaluation campaign was organized, where different multimodal approaches were proposed, such as fusion of different acoustic, prosodic, lexical, and phonotactics representations (Levitan et al., 2016; Kaya and Karpov, 2016).

2.4 Video

In addition to textual, imagery, and speech content, the information in video plays an important role in capturing cues of deceptive behavior. Such cues in videos (e.g., facial expression, gestures) have been investigated in several studies for deception detection. Pérez-Rosas et al. (2015) developed a real-life courtroom trial dataset, which includes 61 deceptive and 60 truthful videos. They explored the use of n-gram features from transcripts and non-verbal features (i.e., facial expressions, eyebrows, eyes, mouth openness, mouth lips, and head movements, hand gestures) to classify liars and truth-tellers. Krishnamurthy et al. (2018) explored textual, speech, and visual features for deception detection. They used a 3D CNN to extract visual features from each frame, spatio-temporal features, and facial expressions such as smile, fear, or stress. Soldner et al. (2019) developed a multimodal deception dataset using TV shows and experimented with textual, visual and dialog features.

2.5 Network and Temporal Information

The rationale for leveraging network information stems from early work (Shao et al., 2018; Vosoughi et al., 2018a) that showed that propagation and interaction networks of fake news are deeper and wider than those of real news. Vosoughi et al. (2018a) further found that fake information spreads faster than factual one, thus advocating for the use of temporal information.

Propagation networks can be homogeneous or heterogeneous (e.g., encompassing news articles, publishers, users, and posts) and they can be analyzed at different scales (e.g., node-level, ego-level, triad-level, community-level and the overall network, as shown in Figure 3, in Appendix) (Zhou and Zafarani, 2019). Shu et al. (2020) tackled the fake news classification task by proposing an approach based on hierarchical propagation networks. At both micro- and macro-scale, they extracted and jointly considered network features, temporal features, and linguistic features. Experiments on PolitiFact and GossipCop datasets revealed that temporal features have maximum contribution, followed by network and linguistic features. Shu et al. (2019) provided one of the most thorough multimodal frameworks for fake news classification. Their experimental results suggest that social context (i.e., network-derived) features are more informative than news content ones.

Vosoughi et al. (2017) proposed Rumor Gauge, a system that jointly exploits temporal and propagation features, in conjunction with linguistic and user credibility features, for checking the veracity of rumors. In particular, Rumor Gauge leverages text, and network propagation. The temporal modality does not directly provide features, but is instead considered by recomputing all other features at regular time steps, thus yielding multiple time series. Results by Vosoughi et al. (2017) and Kwon et al. (2017) also demonstrated that the contribution of the different data modalities change over time.

To mitigate the "cold start" problem of propagation-based early detection of fake news, Liu and Wu (2018) proposed an approach that is primarily based on user and temporal information. First, they built a propagation path of each news as a time series of user representations. The time series for a given news only contains the ordered representations of those users that shared such news. Then, they learned two vector representations of each propagation path via GRUs and CNNs, respectively. Zannettou et al. (2018) analyzed different aspects of memes, such as how they evolve and propagate in different mainstream and fringe web communities, and variants of memes that propagate. Finally, Nguyen et al. (2020) proposed Factual News Graph (FANG) to exploit the social structure and the engagement patterns of users for fake news detection.

3 Multimodal Harmful Content Detection

In this section, we focus on the second aspect of disinformation: *harmfulness*. It is essential to filter or to flag online harmful content. The harmful content includes *child abuse material*, *violent and extreme content*, *hate speech*, *graphic content*, *sexual content*, and *spam content* (Banko et al., 2020).[3] In recent years, the ability to recognize harmful content within online communities has received a lot of attention by researchers (Pramanick et al., 2021a,b) and policymakers that aim to keep users safe in the digital world. Studies in this direction include detecting harmful contents in network science (Ribeiro et al., 2018), natural language processing (Waseem et al., 2017; Schmidt and Wiegand, 2017b; Fortuna and Nunes, 2018) and computer vision (Yang et al., 2019a; Vijayaraghavan et al., 2021; Gomez et al., 2020; Dimitrov et al., 2021b). In Table 2, we provide a list of relevant work addressing different types of harmful content, modalities, source of data, annotation approach, language of the content and the methods.

3.1 Text

In the past few years there has been significant research effort on detecting harmful content (e.g., hate speech) from social media posts (Van Hee et al., 2015; Waseem and Hovy, 2016; Waseem et al., 2017; Schmidt and Wiegand, 2017b). Waseem and Hovy (2016) developed a dataset of hate speech consisting of 16K tweets, and reported a baseline results using char- and word- ngrams and a logistic regression classifier. (Davidson et al., 2017) distinguished between hate speech, and offensive language. They developed a dataset of ∼24K labeled tweets with categories such as hate speech, offensive language and neither. Qian et al. (2018) took a different approach to classic hate speech classification. Instead of binary classes, they proposed 13 fine-grained hate categories such as nationalist, anti-immigrant, racist skinhead, among others, providing a dataset of tweets collected from 40 hate groups. Ribeiro et al. (2018) proposed an approach to find hateful users on Twitter. Mathew et al. (2019) analyzed 341K

[3] https://swgfl.org.uk/services/report-harmful-content/

Ref	Task	Modality					Data/Source	Anno.	Lang	Method
		T	I	V	N	S				
(Nizzoli et al., 2021)	CIB				✓		Twitter: 1.1m users, 11m tweets	DS	En	Statistical and similarity analysis
Weber and Neumann (2020)	CIB				✓		Twitter	–	En	Statistical and network analysis
(Wang et al., 2020)	Cyberbullying	✓	✓				Posts: Vine (970), Instagram (2,218)	M	En	SVM, NB, LR, RF, LSTM, CNN
(Soni and Singh, 2018)	Cyberbullying	✓	✓			✓	Vine videos	M	En	KNN, SVM, LR, RF, GNB
(Dadvar and Eckert, 2018)	Cyberbullying	✓					Youtube 54k posts	M	En	LSTM, BiLSTM, CNN
Hosseinmardi et al. (2015)	Cyberbullying	✓	✓		✓		Instagram	M	En	SVM
(Rafiq et al., 2015)	Cyberbullying	✓			✓		Vine videos	M	En	NB, AdaBoost, DT and RF
(Van Hee et al., 2015)	Cyberbulling	✓					Ask.fm: 85k QA pairs	M	Nl	SVM
(Chatzakou et al., 2019)	Cyberbullying, Cyberaggression	✓			✓		Twitter: 1,303 users, 9,484 tweets	M	En	NB, RF, AdaBoost, Ensemble, NN
(Liang et al., 2017)	Gunshots					✓	Videos: freesound.com, Youtube; Test: CSV, TRECVID Gunshot, UrbanSound Gunshot	DS	En	Localized self-paced reranking
(Mariconti et al., 2019)	Hate attacks	✓	✓			✓	Youtube videos (428)	M	En	Ensemble, CNN, RNN
(Kiela et al., 2020)	Hate speech	✓	✓				FB: Hateful Memes Challenge	M	En	Late fusion, Concat BERT, MMBT, ViLBERT, VisualBERT
(Das et al., 2020)	Hate speech	✓	✓				FB: Hateful Memes Challenge	M	En	VisualBERT, MM fusion
(Gomez et al., 2020)	Hate speech	✓	✓				Twitter: MMHS150K	M	En	Inception v3, LSTM, and MM fusion
(Yang et al., 2019a)	Hate speech	✓	✓				FB: train+dev 378k, test 53k	M	En	Fusion: text + image embedding
(Waseem and Hovy, 2016)	Hate speech	✓					Twitter: 16,914 tweets	M	En	LR
(Davidson et al., 2017)	Hate speech	✓					Twitter: 24,802 tweets	M	En	LR, SVM, NB, DT, RF
(Qian et al., 2018)	Hate speech	✓					Twitter: 40 accounts, 3.5m tweets	DS	En	LR, SVM, Char-CNN, BiLSTM, HCVAE
(Ribeiro et al., 2018)	Hate speech	✓			✓		Twitter: 4,972 users	M	En	GradBoost, AdaBoost, GraphSage
(Mathew et al., 2019)	Hate speech	✓			✓		Gab: 21m posts by 341k users	DS	En	Lexicon based filtering, DeGroot
Dimitrov et al. (2021b)	Propaganda	✓	✓				FB: SemEval-2021 task 6: 950 Facebook memes	M	En	MM fusion, MM joint representation
(Vijayaraghavan et al., 2021)	Hate speech	✓			✓		In-house developed and curated datasets	M	En	MM late fusion, LR, SVM, CNN, BiGRU, BiLSTM
(Constantin et al., 2020)	Violence		✓	✓		✓	VSD96: Hollywood, Youtube	M	En	MM Early fusion; SVM, HMM, GMM, Bayesian, MLP, QDA, PLDA, CNN, KNN, unsupervised, hybrid
(Acar et al., 2013)	Violence			✓		✓	MediaEval VSD	M	En	SVM (mid-level audio + low-level visual)
(Giannakopoulos, 2009)	Violence					✓	Movies	M	-	BN, kNN

Table 2: Summary of the most relevant works on harmful content. **T:Text**, **I: Image**, **V:Video**, **N:Network**, **S:Speech**, Anno.: Annotation, CIB: Coordinated Inauthentic Behavior, QA: Question-answer, CSV: Real-life Conflict Scene Videos, VSD: Violent Scene Detection. Nl: Dutch. KNN: k-Nearest Neighbors, LSTM: Long Short-Term Memory, BiLSTM: Bidirectional LSTM, MMBT: MultiModal BiTransformers, HCVAE: Hierarchical Conditional Variational Autoencoder, QDA: Quadratic Discriminant Analysis, PLDA: Probabilistic Linear Discriminant Analysis.

users and 21M posts collected from Gab to understand the diffusion dynamics of hateful content. Their findings suggest that the posts from hateful user diffuse faster, wider, and have a greater outreach compared to the posts from non-hateful ones.

3.2 Image

Among different types of harmful content, cyberbullying is one of the major growing problems, significantly affecting teens. Hosseinmardi et al. (2015) investigated Instagram images and their associated comments for detecting cyberbullying and online harassment. They developed a manually labeled dataset using CrowdFlower (which is now Appen), where they followed standard procedures for the annotation: using annotation guidelines, qualification tests, gold standard evaluation and quality control criteria such as minimum annotation time. The annotated dataset consists of 998 media sessions (images and their associated comments). A key finding of this study is that a large fraction of the annotated posts (48%) with a high percentage of negative words have not been labeled as cyberbullying. To train and to evaluate the model, the authors used n-grams from text, meta-data (e.g., the number of followers, followees, likes, and shared media), and image categories as features and experimented with Naïve Bayes and SVM classifiers. Their study suggests that combining multiple modalities helps to improve the performance of the SVM classifier.

Hate speech is another important problem that spreads over social media. The "Hateful Memes Challenge" is an important milestone to advance the research on this topic and the tasks is to detect hateful memes (Kiela et al., 2020). Das et al. (2020) proposed different approaches for hatefulness detection in memes such as (*i*) extract the caption and include this information with the multimodal model, (*ii*) use sentiment as an additional feature with multimodal representations. For hate speech detection, Yang et al. (2019a) explored different fusion techniques such as concatenation, bilinear, gated summation, and attention, and reported that combining the text with image embedding boosted the performance in all cases. Vijayaraghavan et al. (2021) proposed methods for interpreting multimodal hate speech detection models, where the modalities consist of text and socio-cultural information rather than images. Concurrently, Gomez et al. (2020) introduced a larger dataset of 150K tweets for multimodal hate speech detection, consisting of six labels.

Propaganda is another topic that has been explored in multimodal settings. Seo (2014) showed how Twitter was used as a propaganda tool during the 2012 Gaza conflict to build international support for each side of the conflict. Dimitrov et al. (2021b) addressed the detection of persuasion techniques in memes. Their analysis of the dataset showed that while propaganda is not always factually false or harmful, most memes are used to damage the reputation of a person or a group of people. Dimitrov et al. (2021a) highlighted the importance of both modalities for detecting fine-grained propaganda techniques, with VisualBERT yielding 19% improvement compared to using the image modality only (with ResNet-152), and 11% improvement compared to using the text modality only (with BERT). Similar observations were made by (Kiela et al., 2020) for hateful meme detection. Glenski et al. (2019) explored multilingual multimodal content and categorizes disinformation, propaganda, conspiracy, hoax, and clickbait.

3.3 Speech/Audio

Cues in spoken content can represent harmful behaviors and those cues can be used to automatically detect such content. Due to the lack of data, studies using the speech-only modality are comparatively lower than other modalities even though it plays a major role in many contexts. For example, for detecting violent content such as screaming and gunshots, the speech modality can play an important role, which other modalities might not be able to offer. This is important as most often user-generated contents are posted on newspapers or their social media accounts without verifying the content of the post, which can have serious consequences (Harkin et al., 2012; Rauchfleisch et al., 2017).

Giannakopoulos (2009) studied the audio segmentation approaches for segmenting violent (e.g., gunshots, screams) and non-violent (e.g., music, speech) content in movies. The studies related to violent content detection using acoustic features also include (Acar et al., 2013), where the focus was on finding violent content in movies.

Liang et al. (2017) proposed Localized Self-Paced Reranking (LSPaR) for detecting gunshots and explosion in videos using acoustic features. Soni and Singh (2018) investigated audio, visual and textual features for cyberbullying detection. Their findings suggest that audio and visual features are associated with the occurrence of cyberbullying, and both these features complement textual features.

3.4 Video

There are multiple studies on detecting cyberbullying in video-based social networks such as Vine (Rafiq et al., 2015) and YouTube (Dadvar and Eckert, 2018). These studies show that although the percentage of cyberbullying in video sessions is quite low, automatic detection of these types of content is very challenging. Wang et al. (2020) used textual, visual, and other meta-information to detect social media posts with bullying topics. Their proposed method was evaluated on publicly available multimodal cyberbullying datasets. Abd Kadir et al. (2016) investigated the relationship between emotion and propaganda techniques in Youtube videos. Their findings suggest that propaganda techniques in Youtube videos affect emotional responses. Content (e.g., Youtube videos) can also be attacked by hateful users via posting hateful comments through a coordinated effort. Mariconti et al. (2019) investigated whether a video is likely to be attacked using different modalities such as metadata, audio transcripts, and thumbnails.

There has been a recent interest from different government agencies to stop the spread of violent content. Constantin et al. (2020) developed a multimodal dataset, which consists of more than 96 hours of Hollywood and YouTube videos and high variability of content. Their study suggests that multimodal approaches with audio and images perform better.

3.5 Network and Temporal Information

The use of network data for predicting factuality was motivated by results showing different propagation patterns for fake vs. real content. Such results are lacking for harmful content. However, the intention to harm in social media is often pursued via coordinated actions, for instance, by groups of users (e.g., social bots and trolls (Cresci, 2020)) that target certain people or minorities. These collaborative harmful actions, perpetrated to increase the efficacy of the harm, are best addressed using network analysis to detect likely coordinated harmful campaigns. Chatzakou et al. (2019) focused on detecting cyberbullying and cyberaggression by training machine learning models for detecting: (*i*) bullies, (*ii*) aggressors, (*iii*) spammers, and (*iv*) normal users on Twitter. To solve these tasks, they leveraged a combination of 38 features extracted from user profiles, the textual content of their posts, and network information (e.g., user degree and centrality measures in the social graph). Orthogonal and in synergy with respect to the detection of disinformation, scholars have recently focused on the novel task of detecting Coordinated Inauthentic Behavior (CIB) (Nizzoli et al., 2021). *CIB is defined as coordinated activities that aim to mislead and manipulate others.*[4] Detecting CIB typically involves analyzing both interaction networks to detect suspicious coordination, as well as the coordinated users and the content they shared to detect inauthentic users and harmful content (Nizzoli et al., 2021, 2020; Pacheco et al., 2021). Given the importance of coordination in CIB, the analysis typically starts from the available network data by applying community detection algorithms, and subsequently moving to the analysis of textual data.

4 Modeling Techniques

In this section, we discuss modeling techniques for both factuality and harmfulness. To combine multiple modalities, there have been several approaches: (i) *early-fusion*, where low-level features from different modalities are learned, fused, and fed into a single prediction model (Jin et al., 2017b; Yang et al., 2018; Zhang et al., 2019; Singhal et al., 2019; Zhou et al., 2020; Kang et al., 2020); (ii) *late-fusion*, where unimodal decisions are fused with some mechanisms such as averaging and voting (Agrawal et al., 2017; Qi et al., 2019), and (iii) *hybrid-fusion*, where a subset of learned features are passed to the final classifier (early-fusion), and the remaining modalities are fed to the classifier later (late-fusion) (Jin et al., 2017a). Within these fusion strategies, the learning setup can also be divided into *unsupervised*, *semi-supervised*, *supervised* and *self-supervised* methods.

Dimitrov et al. (2021b) investigated different fusion strategies (e.g., *early*- and *late*-fusion and *self-supervised* models) for propaganda detection using VisualBERT (Li et al., 2019), MMBT (Kiela et al., 2019), and ViLBERT (Lu et al., 2019). Their findings suggest that self-supervised joint learning models, such as MMBT, ViLBERT, and VisualBERT perform better in increasing order, respectively, compared to the other fusion methods. As a part of "Hateful Memes Challenge" to classify hateful vs. non-hateful memes, several such models have been investigated by Kiela et al. (2020).

Attempts to design *unsupervised* models are limited. Müller-Budack et al. (2020) introduced Cross-modal Consistency Verification Tool (CCVT) to check the coherence between images and associated texts. Yang et al. (2019b) defined trust of news and credibility of users who spread the news and used Bayesian learning to iteratively update these quantities. News with low trustworthiness is returned as fake news. Gangireddy et al. (2020) proposed GTUT, a graph-based approach that exploits the underlying bipartite network of users and news articles to detect the dense communities of fake news and fraud users.

Due to the scarcity of labeled data, a few studies attempted to design *semi-supervised* methods by leveraging an ample amount of unlabelled data. Helmstetter and Paulheim (2018); Gravanis et al. (2019) presented weak-supervision and Guacho et al. (2018) presented a tensor-decomposition semi-supervised method for fake content detection. Dong et al. (2020) developed a deep semi-supervised model via two-path learning (one path uses a limited labeled data, the other path explores the unlabelled data) for timely fake news detection. Paka et al. (2021) presented, Cross-SEAN, a cross-stitch semi-supervised end-to-end neural attention model for COVID-19 fake news detection.

Within a *supervised* learning setup, two other types of learning method have also been explored for disinformation detection such as *adversarial learning* and *autoencoder based*. *Adversarial learning* models for fake news detection include EANN (Wang et al., 2018), an event adversarial neural network to detect emerging and time-critical fake news, and SAME (Cui et al., 2019), a sentiment-aware multimodal embedding method which leverages multiple modalities with the sentiment expressed by readers in their comments.

5 Major Challenges

Recently, several initiatives were undertaken by major companies and government entities to combat disinformation in social media (DIGI, 2021).[5] However, automatic detection of misleading and harmful content poses a number of challenges as discussed below and in Appendix (Section D).

[4] https://medium.com/1st-draft/how-to-improve-our-analysis-of-coordinated-inauthentic-behavior-a4ec62ce9bff

[5] For example, http://digi.org.au/disinformation-code/

Models Combining Multiple Modalities. The major challenge is to devise a mechanism to combine multiple modalities in a systematic way so that one modality complements the others. Current state-of-the-art primarily adopts early and late fusion, which are limited and do not always yield strong results (Dimitrov et al., 2021a). Very recently, jointly trained multimodal transformer-based models (e.g., ViLBERT (Lu et al., 2019), Visual BERT (Lin et al., 2014) and Multimodal Bitransformers (MMBT) (Kiela et al., 2019)) have shown strong potential (Dimitrov et al., 2021b,a; Kiela et al., 2020). However, such models are trained considering only two modalities (textual and visual), while fact-checking or disinformation-related content consists of more than two modalities e.g., text, speech, video, network, etc. (Baly et al., 2020). Hence, there is a room for improvement in developing multimodal models that involve additional, and potentially more than two modalities. Another important problem is cross-modal inconsistency in social media content, as shown in Figure 2(c), which poses a challenge in a multimodal setting (Tan et al., 2020).

Datasets. One of the major challenges when working with such diverse modalities, i.e., text, image, speech, video, and network, is to get access to an appropriate dataset, and moreover to one that considers both factuality and harmfulness. Furthermore, there is a need to integrate data from multiple platforms (e.g., news, posts from Twitter, Reddit and Instagram) as different data sources present different styles and focus on different topics.

6 Future Directions

Based on the aforementioned challenges, we forecast the following research directions:

Explainability. Model interpretation remains largely unexplored. This can be addressed in future studies to understand the general capability of the models. Providing evidence of why certain claims are false is also important. There has been work in this direction such as TabFact (Chen et al., 2020) and FEVER (Hanselowski et al., 2018). However, such approaches rely on existing knowledge bases (e.g., Wikipedia) and may fail for a new problem such as disinformation about COVID-19. It is also important to understand what models learn, e.g., lexical or semantic concepts or a set of neurons may learn one aspect better than the others. Moreover, while current studies on explainable fact-checking focus on explaining the predictions, very few focus on model explanations (Kotonya and Toni, 2020b).

Beyond Content and Network Signals. State-of-the-art methods for multimodal factuality prediction and harmful content detection are primarily based on content signals and network structure. However, the information in these signals is limited and does not include personal preferences or cultural aspects. In the future, we envision multimodal techniques for disinformation detection that would go beyond content and network and would include signals like common sense and information about the user. Moreover, multimodal models will become larger with more heterogeneous signals as input, and they would be pre-trained on a wider variety of tasks to shelter both aspects of disinformation: factuality and harmfulness.

Knowledge-based Method. The use of knowledge-based approaches to check the factuality of claims based on what has been checked before could be ideal solutions as some claims are often repeated by politicians. Current approaches in this direction are limited and this can be explored further by creating a common repository of previously fact-checked claims and harmful content. Relevant studies in this direction include detecting previously fact-checked claims (Shaar et al., 2020), studying the role of context at the sentence level (Shaar et al., 2022) or at the document level (Shaar et al., 2021), and claim matching across languages (Kazemi et al., 2021).

7 Conclusion

We surveyed the state-of-the-art in multimodal disinformation detection based on prior work on different modalities, focusing on disinformation, i.e., information that is both false and intents to do harm. We covered the major research topics of factuality and disinformation. Our survey brought several interesting research challenges for multimodal disinformation detection, such as combining various modalities, which are often not aligned and are in different representations (e.g., text vs. speech vs. network structure), and the lack of such datasets to foster future research. In addition to highlighting the challenges, we also pointed to several research directions. While doing so, we argued for the need to tackle disinformation detection by taking into account multiple modalities as well as both factuality and harmfulness in the same framework.

Acknowledgments

This research is part of the Tanbih mega-project, developed at the Qatar Computing Research Institute, HBKU, which aims to limit the impact of "fake news," propaganda, and media bias by making users aware of what they are reading, thus promoting media literacy and critical thinking.

Limitations

We might not have covered *all* relevant work that falls under the topics of factuality and harmfulness across different modalities.

References

Shamsiah Abd Kadir, Anitawati Lokman, and T. Tsuchiya. 2016. Emotion and techniques of propaganda in youtube videos. *Indian journal of science and technology*, Vol (9):1–8.

Esra Acar, Frank Hopfgartner, and Sahin Albayrak. 2013. Violence detection in hollywood movies by the fusion of visual and mid-level audio cues. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, page 717–720, New York, NY, USA. Association for Computing Machinery.

Taruna Agrawal, Rahul Gupta, and Shrikanth Narayanan. 2017. Multimodal detection of fake social media use through a fusion of classification and pairwise ranking systems. In *25th European Signal Processing Conference (EUSIPCO)*, pages 1045–1049. IEEE.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of EMNLP 2021*, EMNLP '21. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 7352–7364, Online. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 3364–3374, Online. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, ACL '18, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137, Online. Association for Computational Linguistics.

Christina Boididou, Symeon Papadopoulos, Stuart E. Middleton, Giulia Boato, Duc-Tien Dang-Nguyen, and Michael Riegler. 2016. Verifying multimedia use at MediaEval 2016. *MediaEval*.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55.

Lia Bozarth and Ceren Budak. 2020. Toward a better performance evaluation framework for fake news classification. In *Proceedings of the international AAAI conference on web and social media*, volume 14 of *AAAI '20*, pages 60–71.

Sian Brooke. 2019. "condescending, rude, assholes": Framing gender and hostility on Stack Overflow. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 172–180, Florence, Italy. Association for Computational Linguistics.

Fernando Cardoso Durier da Silva, Rafael Vieira, and Ana Cristina Garcia. 2019. Can machines learn to detect fake news? A survey focused on social media. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, volume 6, pages 2763–2770.

Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali, and Nicolas Kourtellis. 2019. Detecting cyberbullying and cyberaggression in social media. *ACM Transactions on the Web (TWEB)*, 13(3):1–51.

Jing Chen, Chenhui Wang, Kejun Wang, Chaoqun Yin, Cong Zhao, Tao Xu, Xinyi Zhang, Ziqiang Huang, Meichen Liu, and Tao Yang. 2021. Heu emotion: a large-scale database for multimodal emotion recognition in the wild. *Neural Computing and Applications*, pages 1–17.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

M. G. Constantin, L. D. Stefan, B. Ionescu, C. Demarty, M. Sjoberg, M. Schedl, and G. Gravier. 2020. Affect in multimedia: Benchmarking violent scenes detection. *IEEE Transactions on Affective Computing*, pages 1–1.

Nicole A Cooke. 2018. *Fake news and alternative facts: Information literacy in a post-truth era*. American Library Association.

Stephen D Cooper. 2007. A concise history of the fauxtography blogstorm in the 2006 lebanon war. *American Communication Journal*, 9(2).

Stefano Cresci. 2020. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83.

Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. SAME: sentiment-aware multi-modal embedding for detecting fake news. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 41–48. ACM.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4826–4832.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, EMNLP-IJCNLP '19, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Maral Dadvar and Kai Eckert. 2018. Cyberbullying detection in social networks using deep learning based models; a reproducibility study. *arXiv:1812.08046*.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv:2012.14891*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11 of *AAAI '17*.

Ronald Denaux and Jose Manuel Gomez-Perez. 2020. Linked credibility reviews for explainable misinformation detection. In *International Semantic Web Conference*, pages 147–163. Springer.

Digital Industry Group Inc DIGI. 2021. Australian code of practice on disinformation and misinformation. Online.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, pages 6603–6617, Online. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b. Task 6 at SemEval-2021: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation*, SemEval '21, Bangkok, Thailand. Association for Computational Linguistics.

Yoan Dinkov, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Predicting the leading political ideology of youtube channels using acoustic, textual, and metadata information. In *Interspeech*, pages 501–505.

Xishuang Dong, Uboho Victor, and Lijun Qian. 2020. Two-path deep semisupervised learning for timely fake news detection. *IEEE Transactions on Computational Social Systems*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4).

Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Siva Charan Reddy Gangireddy, Cheng Long, and Tanmoy Chakraborty. 2020. Unsupervised fake news detection: A graph-based approach. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 75–83.

Kiran Garimella and Dean Eckles. 2020. Images and misinformation in political groups: Evidence from whatsapp in india. *Harvard Kennedy School Misinformation Review*.

Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal multi-image fake news detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654. IEEE.

Theodoros Giannakopoulos. 2009. Study and application of acoustic information for the detection of harmful content, and fusion with visual information. *University of Athens*.

Maria Glenski, Ellyn Ayton, Josh Mendoza, and Svitlana Volkova. 2019. Multilingual multimodal digital deception detection and disinformation spread across social platforms. *arXiv:1909.05838*.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1470–1478.

Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. 2019. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications*, 128:201–213.

Gisel Bastidas Guacho, Sara Abdali, Neil Shah, and Evangelos E. Papalexakis. 2018. Semi-supervised content-based detection of misinformation via tensor embeddings. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 322–325. IEEE Computer Society.

Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking Sandy: Characterizing and identifying fake images on Twitter during hurricane Sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736.

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

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Juliette Harkin, Kevin Anderson, Libby Morgan, and Briar Smith. 2012. Deciphering user-generated content in transitional societies. *Philadelphia, PA: Center for Global Communication Studies, Annenberg School for Communication, University of Pennsylvania*.

Silvan Heller, Luca Rossetto, and Heiko Schuldt. 2018. The ps-battles dataset-an image collection for image manipulation detection. *arXiv preprint arXiv:1804.04866*.

Stefan Helmstetter and Heiko Paulheim. 2018. Weakly supervised learning for fake news detection on twitter. In *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*, pages 274–277. IEEE Computer Society.

Julia Hirschberg, Stefan Benus, Jason Brenier, Frank Enos, Sarah Hoffman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, Bryan L. Pellom, Elizabeth Shriberg, and Andreas Stolcke. 2005. Distinguishing deceptive from non-deceptive speech. In *Interspeech*, pages 1833–1836.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the Instagram social network. *arXiv:1503.03909*.

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

Md Rafiqul Islam, Shaowu Liu, Xianzhi Wang, and Guandong Xu. 2020. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *SNAM*, 10(1):1–20.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, ACL-IJCNLP '21, pages 402–410, Online. Association for Computational Linguistics.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017a. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017b. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, 19(3):598–608.

Srecko Joksimovic, Ryan S. Baker, Jaclyn Ocumpaugh, Juan Miguel L. Andres, Ivan Tot, Elle Yuan Wang, and Shane Dawson. 2019. Automated identification of verbally abusive behaviors in online discussions. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 36–45, Florence, Italy. Association for Computational Linguistics.

SeongKu Kang, Junyoung Hwang, and Hwanjo Yu. 2020. Multi-modal component embedding for fake news detection. In *14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–6. IEEE.

Heysem Kaya and Alexey A Karpov. 2016. Fusing acoustic feature representations for computational paralinguistics tasks. In *Interspeech*, volume 2016, pages 2046–2050.

Ashkan Kazemi, Kiran Garimella, Devin Gaffney, and Scott Hale. 2021. Claim matching beyond English to scale global fact-checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), ACL-IJCNLP '21, pages 4504–4517, Online. Association for Computational Linguistics.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *arXiv:1909.02950*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

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

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

Gangeshwar Krishnamurthy, Navonil Majumder, Soujanya Poria, and Erik Cambria. 2018. A deep learning approach for multimodal deception detection. *arXiv:1803.00344*.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv:1804.08559*.

Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. 2017. Rumor detection over varying time windows. *PLoS ONE*, 12(1).

Sarah Ita Levitan, Guozhen An, Min Ma, Rivka Levitan, Andrew Rosenberg, and Julia Hirschberg. 2016. Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception detection. In *Interspeech*, pages 2006–2010.

Stephan Lewandowsky, Ullrich K.H. Ecker, and John Cook. 2017. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv:1908.03557*.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. In *SIGKDD*, volume 17, pages 1–16.

Junwei Liang, Lu Jiang, and Alexander G. Hauptmann. 2017. Temporal localization of audio events for conflict monitoring in social media. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 1597–1601. IEEE.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. 2021. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*.

Yang Liu and Yi-fang Brook Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 354–361. AAAI Press.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2016, pages 3818–3824.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do":

Proactive detection of YouTube videos targeted by coordinated hate attacks. *Proceedings of the ACM on Human-Computer Interaction*, 3:1–21.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. 2015. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. *arXiv:2003.10421*.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6149–6157.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, Javier Beltrán, Michael Wiegand, Melanie Siegel, and Juliane Köhler. 2022a. Overview of the CLEF-2022 CheckThat! lab on fighting the COVID-19 infodemic and fake news detection. In *Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization*, CLEF '2022, Bologna, Italy.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, and Javier Beltrán. 2022b. The CLEF-2022 CheckThat! Lab on fighting the covid-19 infodemic and fake news detection. In *Advances in Information Retrieval*, pages 416–428, Cham. Springer International Publishing.

Preslav Nakov and Giovanni Da San Martino. 2020. Fact-checking, fake news, propaganda, and media bias: Truth seeking in the post-truth era. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–19, Online. Association for Computational Linguistics.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: leveraging social context for fake news detection using graph representation. In *The 29th ACM International Conference on Information and Knowledge Management*, pages 1165–1174. ACM.

Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. 2021. Coordinated behavior on social media in 2019 uk general election. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 443–454.

Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, Maurizio Tesconi, and Emilio Ferrara. 2020. Charting the landscape of online cryptocurrency manipulation. *IEEE Access*, 8:113230–113245.

Deidre Olsen. 2018. How memes are being weaponized for political propaganda. *Salon. Salon. com, February*, 24.

Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2021. Uncovering coordinated networks on social media: Methods and case studies. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 455–466.

William Scott Paka, Rachit Bansal, Abhay Kaushik, Shubhashis Sengupta, and Tanmoy Chakraborty. 2021. Cross-SEAN: A cross-stitch semi-supervised neural attention model for covid-19 fake news detection. *Applied Soft Computing*, 107:107393.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 21–29, Cham. Springer International Publishing.

Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *ICMI*, pages 59–66, New York, NY, USA. ACM.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

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

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Hierarchical CVAE for fine-grained hate speech classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3550–3559, Brussels, Belgium. Association for Computational Linguistics.

Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in vine. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 617–622, Paris, France. ACM.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Adrian Rauchfleisch, Xenia Artho, Julia Metag, Senja Post, and Mike S. Schäfer. 2017. How journalists verify user-generated content during terrorist crises. analyzing twitter communication during the brussels attacks. *Social Media + Society*, 3(3):2056305117717888.

Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on Twitter. In *Twelfth international AAAI conference on web and social media*, volume 12.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from Wikipedia. In *Proceedings of the 12th Language Resources and Evaluation Conference*, ACL '20, pages 6874–6882, Marseille, France. European Language Resources Association.

Anna Schmidt and Michael Wiegand. 2017a. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017b. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Umut Mehmet Sen, Veronica Perez-Rosas, Berrin Yanikoglu, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2020. Multimodal deception detection using real-life trial data. *IEEE Transactions on Affective Computing*.

Hyunjin Seo. 2014. Visual propaganda in the age of social media: An empirical analysis of Twitter images during the 2012 Israeli–Hamas conflict. *Visual Communication Quarterly*, 21(3):150–161.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2022. The role of context in detecting previously fact-checked claims. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1619–1631, Seattle, United States. Association for Computational Linguistics.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2021. Assisting the human fact-checkers: Detecting all previously fact-checked claims in a document. *arXiv preprint arXiv:2109.07410*.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 3607–3618, Online. Association for Computational Linguistics.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5597–5606. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2020. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Kai Shu, Suhang Wang, and Huan Liu. 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 2019, Melbourne, VIC, Australia, February 11-15, 2019*, pages 312–320. ACM.

Jiasheng Si, Deyu Zhou, Tongzhe Li, Xingyu Shi, and Yulan He. 2021. Topic-aware evidence reasoning and stance-aware aggregation for fact verification. In *Proceedings of the 59th Annual Meeting of the*

Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1612–1622, Online. Association for Computational Linguistics.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. SpotFake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1768–1777, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1):102437.

Xiaoyu Song, Hong Chen, Qing Wang, Yunqiang Chen, Mengxiao Tian, and Hui Tang. 2019. A review of audio-visual fusion with machine learning. In *Journal of Physics: Conference Series*, volume 1237, page 022144. IOP Publishing.

Devin Soni and Vivek K Singh. 2018. See no evil, hear no evil: Audio-visual-textual cyberbullying detection. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–26.

Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul. 2021. Recent advances and trends in multimodal deep learning: A review. *arXiv: 2105.11087*.

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

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

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. 2021. Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*.

Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking URL recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 275–284. ACM.

Svitlana Volkova, Ellyn Ayton, Dustin L Arendt, Zhuanyi Huang, and Brian Hutchinson. 2019. Explaining multimodal deceptive news prediction models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 659–662.

Soroush Vosoughi, Mostafa 'Neo' Mohsenvand, and Deb Roy. 2017. Rumor gauge: Predicting the veracity of rumors on Twitter. *ACM transactions on knowledge discovery from data (TKDD)*, 11(4):1–36.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018a. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018b. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Hai Wan, Haicheng Chen, Jianfeng Du, Weilin Luo, and Rongzhen Ye. 2021. A DQN-based approach to finding precise evidences for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL-IJCNLP '21, pages 1030–1039, Online. Association for Computational Linguistics.

Kaige Wang, Qingyu Xiong, Chao Wu, Min Gao, and Yang Yu. 2020. Multi-modal cyberbullying detection on social networks. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

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

Yuping Wang, Fatemeh Tahmasbi, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, David Magerman, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the use of fauxtography on social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):776–786.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In

Proceedings of the First Workshop on Abusive Language Online, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Derek Weber and Frank Neumann. 2020. Who's in the gang? revealing coordinating communities in social media. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 89–93. IEEE.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019a. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 11–18, Florence, Italy. Association for Computational Linguistics.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. 2019b. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5644–5651.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. 2018. TI-CNN: Convolutional neural networks for fake news detection. *arXiv:1806.00749*.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202.

John Zarocostas. 2020. How to fight an infodemic. *The lancet*, 395(10225):676.

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

Guobiao Zhang, Anastasia Giachanou, and Paolo Rosso. 2022. Scenefnd: Multimodal fake news detection by modelling scene context information. *Journal of Information Science*, page 01655515221087683.

Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1942–1951.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 354–367. Springer.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter*, 21(2):48–60.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. 2021. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, pages 1–26.

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

Appendix

A Examples of Factuality and Harmful Content

In Figure 2, we provide examples textual and visual content that are harmful and false, true image with false claim, and harmful meme.

B Modeling Techniques

Figure 4 shows various multimodal approaches that have been proposed in the literature.

C Lessons Learned

1. A lot of progress has been made on the problem, but the two components in the definition of disinformation (falseness and harmfulness) have been considered mostly in isolation. We argue that there is a need for tight integration of the factuality and the intentional harmfulness into the same detection model. These two aspects have been addressed together in (Alam et al., 2021), which shows that 56% of Arabic false content is also harmful. From Tables 1 and 2, we observe that most multimodal datasets cover just 2–3 modalities, which combine some approaches depicted in Figure 4. Moreover, no multimodal dataset

[6] https://www.snopes.com/fact-check/abe-lincoln-racist-protest-sign/

Figure 2: Examples of textual and visual contents that show (a) *fauxotographic content* (which is both harmful and false), [6] (b) harmful content promoting bad cure (text-only, and false), (c) true image with a false claim about it (malicious), and (d) harmful content, where the text and the image collectively appeal to fear.

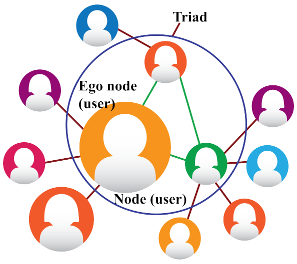

Figure 3: Example of social network with users. **Node:** A node can be a users or a spreader. **Ego:** "Ego" is an individual "focal" node (central user) and the nodes that are directly connected to it are called "alters/spreaders." **Triad:** It (a set of three connected users) is the most basic subgraph of the network. **Community:** A community structure refers to the occurrence of groups of nodes in a network that are more densely connected internally than with the rest of the network.

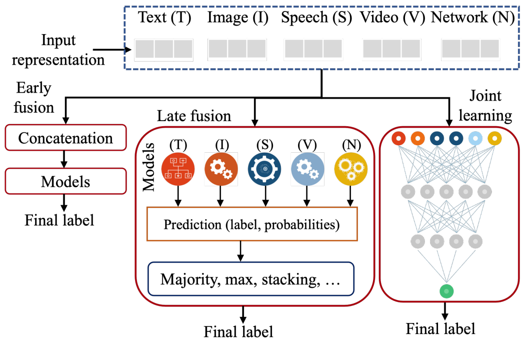

Figure 4: Multimodal approaches, including early and late fusion, and joint modal learning. The hybrid approach (combining early and late fusion) is not shown.

looks at both aspects of disinformation: factuality and harmfulness. While Alam et al. (2021) did address both aspects, they only covered the text modality.

2. In the early phase of (dis)information spreading, user and content features are those that provide the highest contribution for detecting factuality. Indeed, at that time, a few interactions with content are available and the propagation network is small and sparse. As information spreads, the contribution of content-derived features remains constant, while propagation-derived features become richer and more informative. In summary, early prediction of factuality and veracity must necessarily rely heavily on users and content – be it text, image, audio or video. Instead, analyses carried out at later times benefit more from network and temporal data. In the past decade, research on multimodality has shown its potential in several fields, which include audio-visual fusion (Mroueh et al., 2015; Zhu et al., 2021; Song et al., 2019), emotion recognition (Chen et al., 2021), image and video captioning (Liu et al., 2021), multimedia retrieval and visual question answering (Summaira et al., 2021). For factuality, Baly et al. (2020) showed that combining different modalities such as text, speech, and metadata yields improved performance compared to using individual modalities. Similar phenomena have been observed for other tasks such as hateful memes (Kiela et al., 2020), and propaganda detection (Dimitrov et al., 2021b).

D More Challenges

1. **Contextualization.** Existing methods of disinformation detection are mostly non-contextualized, i.e., the broader context of a news article in terms of the responses of the readers and how the users perceive them are not captured. We argue that the response

thread under a news, the underlying social network among users, the propagation dynamics of the news and its mentions across social media need suitable integration to better capture the overall perspective on the news.

2. **Meta Information.** Along with the news and the context, other information such as the authenticity of the news, the credibility of the authors of the news, the factuality of the news also play an important role for disinformation detection. Moreover, detecting whether the disinformation attack is a coordinated effort or an individual activity would also help understanding its severity.

3. **Bias, Region, and Cultural Awareness.** The performance of most of the existing systems is limited to the underlying dataset, particularly to the demography and the underlying cultural aspects. For instance, a model trained on an Indian political dataset may not generalize well to a US health-related dataset (Fortuna et al., 2021).

4. **Disinformation on Evolving Topics.** Often, claims or harmful content are disseminated based on the current event; information about COVID-19 and vaccines are examples of such use cases. Existing models might fail on such use cases, and thus zero-shot or few-shot learning might be an important future avenue to explore.

5. **Transparent and Accountable Models.** The detection models should be designed in a way that their outcomes are unbiased and more accountable to ethical considerations. The models for disinformation detection should present the outcome in such a way that a practitioner can interpret it and understand why a piece of information is flagged as disinformation, what is the related real news based on which the judgment was made, and which part of the information was counterfeit. There is also a lack of datasets containing disinformation with explanations and the corresponding real information.

6. **Fine-grained Detection.** Existing disinformation detection models are mostly binary classifiers: given a piece of news, they aim to detect whether it is a disinformation or not. Such binary signals might be enough in certain cases. However, in many other cases, more fine-grained labels can help to make a better decision. For example, whether a social media post is fake or genuine can help fact-checkers, but having more fine-grained information such as true, satire/parody, misleading, manipulated, false connection, or imposter content can be even more helpful (Nakamura et al., 2020). Therefore, rather than a binary classification, one could cast the problem as a multi-class classification task or even an ordinal regression, or just a regression task. This would also help prioritize disinformation for reactive measurements.