A Survey on Predicting the Factuality and the Bias of News Media

Anonymous ACL submission

Abstract

The present level of proliferation of fake, biased, and propagandistic content online has made it impossible to fact-check every single suspicious claim or article, either manually or automatically. Thus, many researchers are shifting their attention to higher granularity, aiming to profile entire news outlets, which makes it possible to detect likely "fake news" the moment it is published, by simply checking the reliability of its source. Source factuality is also an important element of systems for automatic fact-checking and "fake news" detection, as they need to assess the reliability of the evidence they retrieve online. Political bias detection, which in the Western political landscape is about predicting left-center-right bias, is an equally important topic, which has experienced a similar shift towards profiling entire news outlets. Moreover, there is a clear connection between the two, as highly biased media are less likely to be factual; yet, the two problems have been addressed separately.

In this survey, we review the state of the art on media profiling for factuality and bias, arguing for the need to model them jointly. We further discuss interesting recent advances in using different information sources and modalities, which go beyond the text of the articles the target news outlet has published. Finally, we discuss current challenges and outline future research directions.

1 Introduction

The rise of the Web has made it possible for anybody to create a website and to become a *news medium*. This was a hugely positive development as it elevated freedom of expression to a whole new level, allowing anybody to have their voice heard. With the subsequent rise of social media, anybody could potentially reach out to a vast audience, something that until recently was only possible for major news outlets. One of the consequences was a *trust crisis*: with traditional news media stripped off their gate-keeping role, the society was left unprotected against potential manipulation.

The issue became a general concern in 2016, a year marked by micro-targeted online disinformation at an unprecedented scale in connection to Brexit and the US Presidential election. These developments gave rise to the term "fake news."

In an attempt to solve the trust problem, several initiatives, such as PolitiFact, Snopes, FactCheck, and Full Fact, have been launched to fact-check suspicious claims manually. However, given the scale of the proliferation of false information online, it was unfeasible to fact-check every single suspicious claim, even when this was done automatically, not only for computational reasons but also due to timing. In order to fact-check a claim manually or automatically, it is required to verify the stance of mainstream media concerning that claim and/or the reaction of users on social media. Accumulating this evidence takes time, and delay means more potential sharing of the malicious content. A study has shown that, for some very viral claims, more than 50% of the sharing happens within the first ten minutes after posting the micro-post on social media (Zaman et al., 2014), and thus timing is of utmost importance. Moreover, an extensive recent study has found that "fake news" spreads six times faster and reaches much farther than real news (Vosoughi et al., 2018).

A much more promising alternative is to profile the medium that initially published the news article with a suspicious claim. Since media that have published fake or biased content in the past are more likely to do so in the future, profiling media in advance makes it possible to detect likely "fake news" the moment it is published by simply checking the reliability of its source.

Estimating the reliability of a news source is important for claim fact-checking (Nguyen et al., 2018), and it also gives an important prior when solving article-level tasks such as "fake news" and click-bait detection (Hardalov et al., 2016; Karadzhov et al., 2017a; De Sarkar et al., 2018; Pérez-Rosas et al., 2018; Brill, 2001; Finberg et al., 2002; Pan et al., 2018; Nguyen et al., 2022).

There have been several surveys on fake news (Shu et al., 2017; da Silva et al., 2019; Zhou and Zafarani, 2020), mis/dis-information (Islam et al., 2020; Alam et al., 2022; Hardalov et al., 2022a), fact-checking (Thorne and Vlachos, 2018a; Kotonya and Toni, 2020; Nakov et al., 2021; Guo et al., 2022a), truth discovery (Li et al., 2016), and propaganda detection (Martino et al., 2020). However, they have focused on claims or articles, while here we survey research on profiling entire news outlets for factuality and bias.

2 Factuality

Veracity of information has been studied at different levels: (*i*) claim-level (e.g., *fact-checking*), (*ii*) article-level (e.g., *"fake news" detection*), (*iii*) user-level (e.g., *hunting for trolls*), and (*iv*) medium-level (e.g., *source reliability estimation*). Our primary interest here is in the latter.

At the claim-level, significant effort has been paid to fact-checking and rumor detection using information from social media, i.e., how users reply to the claim (Canini et al., 2011; Castillo et al., 2011; Ma et al., 2015, 2016; Zubiaga et al., 2016; Ma et al., 2017; Dungs et al., 2018; Kochkina et al., 2018; Hardalov et al., 2022b; Nguyen et al., 2022), but there is a need for more comprehensive approaches (Thorne and Vlachos, 2018b; Guo et al., 2022b). A set of web pages and snippets from search engines have also been used as a source of information (Mukherjee and Weikum, 2015; Popat et al., 2016, 2017; Karadzhov et al., 2017b; Mihaylova et al., 2018; Baly et al., 2018b). In either case, the most important information for the claim-level tasks are *stance* (does a tweet or a news article agree or disagree with the claim?) and *source reliability* (do we trust the user who posted the tweet or the medium that published the news article?).

The problem of source reliability remains largely under-explored. In the case of social media and community fora, it concerns modeling the user, e.g., there has been research on finding opinion manipulation *trolls* (Mihaylov and Nakov, 2016), *sockpuppets* (Maity et al., 2017), *Internet water army* (Chen et al., 2013), and *seminar users* (Darwish et al.). In the case of the Web, it is about source trustworthiness (the URL domain, the medium).

In early work, the source reliability of news media has often been estimated automatically based on the general stance of the target medium with respect to known true/false claims, without access to gold labels about the overall medium-level factuality of reporting (Dong et al., 2015; Mukherjee and Weikum, 2015; Popat et al., 2016, 2017, 2018).

More recent work has addressed the task as one on its own right. Baly et al. (2018a) used gold labels from Media Bias/Fact Check, and rich information sources: articles published by the medium, what is said about it on Wikipedia, metadata from its Twitter profile, URL structure, and traffic information. In follow-up work, Baly et al. (2019) used the same representation to jointly predict media factuality and bias on an ordinal scale, using a multi-task ordinal regression setup. Then, Baly et al. (2020b) extended the information sources to include Facebook followers and speech signals from the news medium's channel on YouTube (if any). Hounsel et al. (2020) proposed to use domain, certificate, and hosting information of the website infrastructure. Finally, Panayotov et al. (2022) used audience overlap and graph neural networks.

3 Bias

Compared to factuality, which can be objectively determined by whether a piece of information is true or not, media bias has more complex dimensions. For the last few decades, many scholars have conceptualized media bias in different ways. For instance, a bias can be defined as "imbalance or inequality of coverage rather than as a departure from truth" (Stevenson et al., 1973). A departure from truth, however, can be measured only when the accurate record of the event is available (e.g., trial transcript and reporting).

A different definition, namely "any systematic slant favoring one candidate or ideology over another" (Waldman and Devitt, 1998), is proposed to capture various dimensions rather than coverage imbalance, such as favorability conveyed in visual representations (i.e., news photos). E.g., smiling, speaking at the podium, cheering crowd, and eye-level shots are preferred over frowning, sitting, being alone, and shots from above, respectively.

D'Alessio and Allen (2000) reviewed 59 studies about partisan media bias in presidential elections. They proposed to categorize media bias into the following three types: (*i*) *gatekeeping bias*, where editors and journalists 'select' the stories to report, (*ii*) *coverage bias*, where the amount of news coverage (e.g., the length of newspapers articles, or the time given on television) each party receives is systematically biased to one party at the expense of the other one, and (*iii*) *statement bias*, where news media interject their attitudes or opinions in the news reporting. Groeling (2013) proposed a more relaxed concept of media bias, which is "a portrayal of reality that is significantly and systematically (not randomly) distorted," to take a variety of media bias dimensions into account. In particular, he focused on two main forms of media bias—*selection bias* (i.e., what to cover) and *presentation bias* (i.e., how to cover it)—driven by the choices of newsmakers.

Selection bias or *gatekeeping* bias has been studied in various ways, including qualitative interviews or surveys of journalists and editors about the decision-making process they use to select the stories in their newsroom (Tandoc Jr, 2014). Here, news selection is not necessarily confined to political context. News reporting about any news items can be considered as the unit of analysis.

Data-driven research on selection bias commonly follows three steps: (*i*) collect news articles (for newspapers or online news) or transcripts (for TV news) for a target period, (*ii*) conduct content analysis to find the news coverage of politicians, parties, or events. Optionally, study the tone of the news articles (e.g., negative news are more frequently reported) (Soroka, 2012), and (*iii*) identify systematic biases by comparing news coverage. An exhaustive database of news stories is thus essential for selection bias research. While commercial databases, such as Lexis Nexis, have been widely used (Soroka, 2012; Padgett et al., 2019; Gilens and Hertzman, 2000; Boykoff and Boykoff, 2004), publicly available datasets, such as GDELT, start to get attention (Boudemagh and Moise, 2017; Kwak and An, 2014; Boudemagh and Moise, 2017) and are getting validated by comparing multiple sources (Kwak and An, 2016; Weaver and Bimber, 2008; Kwak and An, 2016).

Presentation bias has been characterized from diverse perspectives, including framing (Entman, 2007), visuals (Barrett and Barrington, 2005), sources (Baum and Groeling, 2008), tone (Soroka, 2012), and more. Particularly, framing bias has been actively studied in many disciplines.

Framing Bias refers to a bias that highlights a certain aspect of an event or an issue more than the others (Entman, 1993). Emphasizing a particular aspect can deliver a distorted view toward the issue even without the use of biased expressions.

Framing biases have been typically studied at issue level (Kim and Johnson, 2022). Researchers collect news articles about an issue or an event, conduct manual content analysis, and build a frame detection model (Baumer et al., 2015). Open-source tools to help the analysis have been proposed (Bhatia et al., 2021; Morstatter et al., 2018). While this approach can characterizes diverse frames, it is not trivial to compare framing across issues.

The Media Frames Corpus (MFC) was proposed to address this limitation (Card et al., 2015). It contains articles annotated with 15 generic frames (including *others*) across three policy issues. Several studies have demonstrated reasonable prediction performance of the general media frames with different datasets (Field et al., 2018; Kwak et al., 2020). These 15 general frames were also used for analyzing political discourse on social media (Johnson et al., 2017). These frames are often customized to a specific issue by adding issue-specific frames (Liu et al., 2019), even though doing so somewhat contradicts the original motivation of general media frames, namely to be able to compare frames across various issues.

News slant was proposed to characterize how framing in news reports favors one side over the other (Entman, 2007). The media-level slant thus could differ across issues (Ganguly et al., 2020).

A variety of methods have been proposed to quantify the extent of news slant in traditional news media by (*i*) linking media outlets to politicians with known political positions, (*ii*) directly analyzing news content, and (*iii*) using shared audience among media outlets. Groseclose and Milyo (2005) assigned an ADA (Americans for Democratic Action) score for each media outlet by investigating co-citations of think-tanks by members of Congress and media outlets. Gentzkow and Shapiro (2010) proposed an ideological slant index of news media in a seminal study. The news slant is measured by the extent of phrases in news coverage that are more frequently used by one political party (i.e., Democratic or Republican) congress members than by another one in the 2005 Congress Record. Their frequency-based approach successfully finds politically charged phrases such as *death tax* or *war on* *terror* by Republicans and associated media and *estate tax* or *war in Iraq* by Democrats and associated media, and they further computed media slant index for 433 newspapers. The choice of words by political party members and news media is considered framing because they purposely highlight some aspect of the issue over other ones.

An et al. (2012) proposed a method to compute media slant scores by measuring distances between media sources by their mutual followers on Twitter (An et al., 2011, 2012). Stefanov et al. (2020) identified the political leanings of media outlets and influential people on Twitter based on their stance on controversial topics. They built clusters of users around core vocal ones based on their behaviour on Twitter such as retweeting, using a procedure proposed in (Darwish et al., 2020).

Left-center-right bias (or left-right bias) was studied based on media-level annotation from specialized online platforms, such as News Guard, AllSides, and Media Bias/Fact Check, where journalists use carefully designed guidelines to make the judgments. Researchers have then trained systems to predict this bias using a variety of information sources such as analyzing the corresponding YouTube channels (Dinkov et al., 2019), and using information from the articles the target news outlet has published, what is there about them in social media and in Wikipedia (Baly et al., 2020b).

There has also been work on predicting the left-center-right bias of articles, which is somewhat relevant here as it can be an element of media-level analysis. Such systems are typically trained using distant supervision, projecting the label from a medium to each article from that medium, which is an easy way to obtain large datasets, needed to train contemporary deep learning models. For example, Kulkarni et al. (2018) used site-level annotations from the AllSides website for political bias detection. The same approach was used to study hyperpartisanship, i.e., extremely one-sided reporting (Potthast et al., 2018), as part SemEval-2019 task 4 on Hyper-partisan News Detection (Kiesel et al., 2019). More recent work has demonstrated the dangers of distant supervision and has introduced a dataset for left-center-right bias with proper manual article-level annotations (Baly et al., 2020a).

4 Joint Modeling

There is a well-known connection between factuality and bias. For example, hyper-partisanship (high bias) is often linked to low trustworthiness (Potthast et al., 2018), e.g., appealing to emotions rather than sticking to the facts, while center media tend to be generally more impartial and also more trustworthy. Moreover, some of the datasets used for the two tasks have media-level annotations for both factuality and bias. Thus, it makes sense to model factuality and bias jointly.

Yet, joint modeling of the two tasks remains severely underexplored. In fact, there has been a single attempt at doing so to date: Baly et al. (2019) proposed a multi-task learning formulation. They further took into account the ordinal nature of the labels for both tasks, noting that classifying an *extreme right* medium as *extreme-left* is a huge error, while classifying it as a *center* is a smaller one, and predicting *right* is an even smaller error. Similarly, predicting a *high-factuality* label for a *low-factuality* medium is a bigger mistake than predicting *mixed factuality*. Thus, they proposed a multi-task ordinal regression model, copula ordinal regression (Walecki et al., 2016), which jointly predicts factuality and bias on ordinal scales. They further used several auxiliary tasks, modeling centrality, hyper-partisanship, as well as left-vs.-right bias on a coarse-grained scale.

This is challenging as it requires understanding the interactions between the two dimensions. Although the relationship between extreme bias and low factuality follows intuition, uncovering the connection between being factual but biased or non-factual but unbiased requires more detailed insights. For news media that exhibit a mixed behavior in both aspects, this poses an even greater difficulty.

5 Basis of Prediction

5.1 Textual Content

5.1.1 Representation

The most natural representation for a source is as a sample of articles it has published, which in turn can be represented using linguistic features or as continuous representations.

Linguistic Features focus on language use, and they have been shown to be useful for detecting fake articles, as well as for predicting the political bias and the factuality of reporting of news media (Horne et al., 2018; Baly et al., 2018a). For example, Horne and Adali (2017) showed that "fake news" pack a lot of information in the title (as many people do not read beyond the title, e.g., in social media), and use shorter, simpler, and repetitive content in the body (as writing fake information takes a lot of effort). Such features can be calculated based on the Linguistic Inquiry and Word Count (LIWC) lexicon and used to distinguish articles from trusted sources vs. hoaxes vs. satire vs. propaganda (pen). They can be also modeled using linguistic markers (Mihaylova et al., 2018) such as *factives* from (Hooper, 1975), *assertives* from (Hooper, 1975), *implicatives* from (Karttunen, 1971), *hedges* from (Hyland, 2005), *Wiki-bias* terms from (Recasens et al., 2013), *subjectivity* cues from (Riloff and Wiebe, 2003), and *sentiment* cues from (Liu et al., 2005). There are 141 such features in the NELA toolkit (Horne et al., 2018):

- **Style**: part-of-speech tags, use of specific words (function words, pronouns, etc.), and features for clickbait title classification;
- **Complexity**: type-token ratio, readability, number of cognitive process words (identifying discrepancy, insight, certainty, etc.);
- **Bias**: features modeling bias using lexicons (Recasens et al., 2013; Mukherjee and Weikum, 2015) and subjectivity, calculated by pre-trained classifiers (Horne et al., 2017);
- **Affect**: sentiment scores from lexicons (Recasens et al., 2013; Mitchell et al., 2013) and full systems (Hutto and Gilbert, 2014);
- **Morality**: features based on the Moral Foundation Theory (Graham et al., 2009) and lexicons (Lin et al., 2018);
- **Event**: features modeling time and location.

Embedding representations: An alternative way to represent an article is to use embedding representations, typically based on large pre-trained language models, such as BERT (Devlin et al., 2019). This can be done without fine-tuning, e.g., by encoding an article (possibly truncated, e.g., BERT can take up to 512 tokens as an input) and then averaging the word representations extracted from the second-to-last layer. Alternatively, one can use pre-trained sentence encoders such as Sentence BERT (Reimers and Gurevych, 2019). Finally, one can obtain representations that are relevant to the target task, e.g., by fine-tuning BERT to predict the label (bias or factuality) of the medium that an article comes from, in the form of distant supervision (Baly et al., 2020b). One issue with distant supervision is that the model can end up learning to detect the source of the target news article instead of predicting its factuality and bias, which can be fixed using adversarial media adaptation and a specially adapted triplet loss (Baly et al., 2020a).

5.1.2 Aggregation

In order to obtain a representation/prediction for an entire medium, there is a need to aggregate the representations/predictions for its articles.

Averaging article-level representations: One could average the representations for all articles to obtain a representation for a medium, which can then be used in a medium-level classifier. Using arithmetic averaging is a good idea as it captures the general trend of articles in a medium, while limiting the impact of outliers. For instance, if a medium is known to align with left-wing ideology, this should not change if it published a few articles that align with right-wing ideology.

Aggregating posterior probabilities: Alternatively, each article can be represented by a $\mathcal{C}$-dimensional vector that corresponds to its posterior probabilities of belonging to each class c_i, $i \in \{1, \ldots, \mathcal{C}\}$ of the given task, whether it is predicting the political bias or the factuality of the target news medium. Finally, these article-level posterior probabilities are averaged in order to aggregate them at the medium level.

5.2 Multimedia Content

Nowadays, almost all news websites heavily rely on multimedia content. This dependence, however, also makes multimedia a very effective means for dispensing an intended, and even manipulated, messages. The increasing availability of automated and AI-powered multimedia editing and synthesis tools, combined with massive computational power, makes such capabilities accessible to everyone.

Given that the multimedia editors of a news site typically follow a defined workflow when creating, acquiring, editing, and curating content for their pages, this pattern adds a crucial dimension to profiling the factuality and the bias of a news source. In fact, questions around the origin and the veracity of photographic images and videos have long been the subject of multimedia forensics research (Sencar and Memon, 2013; Sencar et al., 2022). There has been research on verifying metadata integrity (Yang et al., 2020; Kee et al.; Iuliani et al., 2018; Yang et al., 2020), digital integrity (Cozzolino and Verdoliva, 2018; Korus, 2017; Cozzolino and Verdoliva, 2018), physical integrity (Matern et al., 2020; O'Brien et al., 2012; Iuliani et al., 2017; Matern et al., 2020; Riess et al., 2017; Peng et al., 2017) identification of processing traces (Hadwiger et al., 2019), and discrimination of synthesized (i.e., GAN generated) media (Agarwal et al., 2020; Li et al., 2018; Agarwal et al., 2020; Verdoliva, 2020). However, these capabilities have only been sparsely explored in the context of predicting factuality and bias.

Existing work mainly considered characteristics of images appearing in trustworthy vs. unreliable sources. It was proposed to use visual characteristics (Jin et al., 2016), deep-learning representations (Qi et al., 2019; Khattar et al., 2019; Qi et al., 2019; Singhal et al., 2019), image provenance information from reverse image search (Zlatkova et al., 2019), and self-consistency with respect to metadata (Huh et al., 2018). Overall, multimedia characteristics have a strong potential that is yet to be fully used for news media profiling.

5.3 Audience Homophily

The well-known homophily principle, "birds of a feather flock together," crucially asserts that similar individuals interact with each other at a higher rate. Therefore, audience representation could be another approach to describe a news media outlet whereby an overall, descriptive characteristic of followers of the outlet is obtained. Then, by evaluating the similarity of audience-centric representations with previously categorized news media, its factuality and bias can be inferred.

Ribeiro et al. (2018) used Facebook's targeted advertising tool to infer the ideological leaning of online media based on the political leaning of the users who interacted with these media. An et al. (2012) relied on follow relationships on Twitter to ascertain the ideological leaning of news media and users. Wong et al. (2013) studied retweet behavior to infer the ideological leanings of online media sources and of popular Twitter accounts. Barberá (2015) proposed a model based on the follower relationships to media sources and Twitter personalities to estimate their ideological leaning.

Stefanov et al. (2020) predicted the political leaning of media with respect to a topic by observing the users of which side of the debate on a polarizing topic were sharing content from which media in support of their position. They constructed a user-media graph and then used label propagation and graph neural networks to derive representations for media, which they used for classification. They further aggregated the leanings across several polarizing topics to come up with a left-center-right polarization prediction.

Following a similar approach, (Baly et al., 2020b) considered three social media platforms for audience characterization. On Twitter, they proposed to use self-descriptions in publicly accessible profiles of users following the account of a medium. For each medium, a representation is obtained by encoding the biographic descriptions of Twitter followers and averaging the resulting textual representations. The second characterization involves how the audience of the medium's YouTube channel responds to each video in terms of number of comments, views, likes and dislikes. By averaging these statistics over all videos, a medium-level representation is obtained. The last audience representation is obtained using Facebook's advertising platform, which is used to obtain demographic information for the audience interested in each medium. This data is used to obtain the audience distribution over the political spectrum. The distribution is then divided into five categories to label each medium accordingly: very conservative, conservative, moderate, liberal, and very liberal.

5.4 Infrastructure Characteristics

Beyond textual, visual, and audience features, news sites also exhibit distinct characteristics that relate to the underlying infrastructure and technological components deployed to serve their content online. In this regard, the prediction problem is analogous to a well-studied one in the cybersecurity domain where the goal has been to identify infrastructure characteristics of malicious domains (Anderson et al., 2007; Invernizzi et al., 2014) that are used for malware distribution (Wang et al., 2013; Invernizzi et al., 2014), phishing (Purwanto et al., 2020; James et al., 2013; Mohammad et al., 2012, 2014; Purwanto et al., 2020), online scams (Alrwais et al., 2017; Konte et al., 2009; Hao et al., 2016), and spamming (Anderson et al., 2007; Hao et al., 2009). Since establishing the infrastructure of a news medium involves several decisions with respect to technological aspects, it is plausible to expect that news media with varying IT practices and different levels of access to IT resources will differ in their characteristics.

There has been very little work on network, web design, and data elements of a news website to characterize new sites for factuality and bias. At the network level, (Hounsel et al., 2020) aimed to distinguish disinformation websites vs. authentic websites vs. sites not related to news or politics, and found that features related to a website's domain name, registration, and DNS configuration work best. Concerning the web design aspect, Castelo et al. (2019) introduced a web page classifier based on several features that govern the structure and the style of a page in addition to three categories of linguistic features. Their binary classification results (real vs. fake news) on several datasets showed that the web-markup features consistently perform well and are complementary to linguistic ones.

Finally, at the data level Fairbanks et al. (2018) examined the source of web pages to identify shared data objects, such as mutually linked sites, scripts, and images, across web sites. This information is then used to create a shared data object graph. By comparing the content-level features to the structural properties of the graph, they found that the use of mutually shared objects yields better performance in predicting both factuality and bias for a site, especially for factuality. Overall, a major advantage of using infrastructure features is their content- and audience-agnostic nature. This allows making reliable predictions when only limited textual and visual content is available and without an established audience interest in a news medium.

6 Lessons Learned

Factuality and bias have some commonalities as they exert negative influences on the public by delivering information that is deviated from the truth. Not surprisingly, some news media purposedly take a biased position in the political landscape and appeal to partisan audiences. This trend becomes apparent in recent years mainly because the news industry becomes more and more competitive. Many journalists and editors, however, have concerns about their biases in news selection and reporting and try to be neutral or at least report diverse perspectives of an issue.

As the bias can be conveyed by different means —text, photos, and videos—, media bias can get subtle in many dimensions. Among them, ideological bias is an important conceptualization due to the importance of media bias in the political context. In the US context, the ideological bias could be broadly defined as conservative, center, and liberal. Then, the (ideological) bias prediction task is formulated as predicting whether a given news story, including both text and visual elements, favors one party over the other. Reported results so far show that accurate prediction of this ideological bias of a news medium is a far more easier task than assessing factuality. This is, in fact, not surprising as evaluation of the factuality ultimately depends on the authenticity and the objectivity of the particular claims stated in a news story, essentially requiring verification from other sources and observations.

Although more sophisticated analysis of the text style and multimedia characteristics may be expected to improve the achievable accuracy, it is evident that there is a big need to complement the textual and visual elements of a news medium with others. In this regard, recent studies have demonstrated the potential of audience homophily and the medium's infrastructure characteristics in bridging the existing performance gap. The content-agnostic nature of these characteristics make them useful in the early discovery and categorization of news media even in the absence of sufficient content.

7 Major Challenges

Ordinal scales: While the ideological bias (news slant) is typically modeled as left-center-right, there exists a spectrum within each bias based on bias intensity. A hyperpartisan bias prediction task has been tested to differentiate far-right from right and far-left from left, but it does not model the political bias using an ordinal scale. Difficulties in labeling the bias (i.e., creating ground-truth datasets) by experts or crowdsourcing is a major hurdle for modeling ideological bias as an ordinal variable.

Multimodality: In news reporting, a photo typically gets high attention, and readers can sometimes understand a news storie from news photos only, even without reading the text. Indeed, news text and photos are strongly coupled together and deliver relevant information about news stories to readers. Thus, there should be a benefit from modeling news text and photos together to understand their bias and factuality (Alam et al., 2022), and potential harmfulness (Sharma et al., 2022).

Evaluation granularity: The label of a news medium is inferred from a sample of observations. This can introduce a measurement bias when a news medium does not exhibit the same reporting behavior with all news items it publishes. This is especially the case for media that have a particular stance in only certain issues (Ganguly et al., 2020). Thus, reliable estimation of factuality and bias labels require analyzing a relatively large amount of content covering a range of issues.

Variability in factuality & bias ratings: These ratings are inherently not static and may change over time when a news medium takes corrective action to address issues raised by fact-checkers. Thus, the ground truth needed for building a learning approach varies, triggering the need for re-evaluating the performance of the proposed approaches. Thus, there is a need to take into account the sensitivity of a learning approach to such small but nevertheless inevitable variations.

Dataset size: The datasets for media-level factuality and bias are relatively small, typically of a few hundred examples.They are derived from sites, such as Media Bias/Fact Check and AllSides, where domain experts perform manual analysis.

Annotation vs. modeling: One problem is that human annotators judge the factuality of reporting and the bias of media based on criteria that are not easy to automate or based on information that may not be accessible to automatic systems. For example, if a news outlet is judged to be of mixed factuality based on it having failed just 2-3 fact-checks, for an automatic system to arrive at the same conclusion using the same idea, it would have to select for analysis the exact same articles where the false claims were made.

Data availability: Primarily due to copyright issues, there are only a few publicly available datasets of the full text of news for research purposes. Instead, indexed data (e.g., GDELT dataset[1]) by mentioned actors, events, locations, sources, or tones are available and have been analyzed in many studies. A set of news headlines collected from news websites or aggregated websites (e.g., AllSides) are also shared more actively for research purposes. Considering the importance of social media channels in news dissemination, researchers collect and analyze social media posts of official accounts of news media. As social media posts are relatively more informal than news articles to fit for social media audience (Park et al., 2021), more studies are required for understanding their biases and factuality correctly.

8 Future Forecasting

Support for non-English corpora and different political systems: Most of the studies we review are conducted for English. More research on bias and factuality for other languages thus is expected. Recently, various approaches are proposed to accelerate NLP research for resource-scarce languages, such as multilingual word embeddings. We believe that those efforts help conduct bias and factuality research for non-English corpora. One non-technical issue here is that not all the countries have US-like left-center-right political biases. For example, there might exist a multiparty system in some countries. In that case, understanding relevant political biases should be the first step in media bias research.

Incorporation of video content: TV news accounts for significant portions of the news industry. Also, the presence of news media becomes strong in video-driven social media platforms over time. To get high user engagements, news media outlets upload short video clips curated for social media use, particularly on existing social media. Most previous studies on bias in video news have analyzed their transcripts instead of analyzing video directly. Commercial databases, such as Lexis Nexis, or open-source libraries to create subtitles are used to analyze news transcripts. We expect that more studies on analyzing video contents in an end-to-end manner will be presented to fully understand the bias and factuality of video news.

Bringing practical implications: Since the factuality and the bias of news media largely influence the public, it is crucial to implement working systems, so that readers can benefit from a rich stream of research. Several stand-alone websites, such as Media Bias/Fact Check, AllSides, and Tanbih (Zhang et al., 2019), aim to make media bias and factuality transparent to end-users, thus promoting media literacy. We expect new tools and services to support more media and languages.

9 Conclusion

We reviewed the state of the art on media profiling for factuality and bias, arguing for the need to model them jointly. We further discussed interesting recent advances in exploiting different information sources and different modalities, which go beyond the text of the articles the target news outlet has published. Finally, we discussed current challenges and outlined promising research directions.

[1] https://www.gdeltproject.org/

References

Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. 2020. Detecting deep-fake videos from phoneme-viseme mismatches. In *CVPR*, pages 2814–2822. IEEE.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sumayah Alrwais, Xiaojing Liao, Xianghang Mi, Peng Wang, XiaoFeng Wang, Feng Qian, Raheem Beyah, and Damon McCoy. 2017. Under the shadow of sunshine: Understanding and detecting bulletproof hosting on legitimate service provider networks. In *IEEE SP*, pages 805–823.

Jisun An, Meeyoung Cha, Krishna Gummadi, Jon Crowcroft, and Daniele Quercia. 2012. Visualizing media bias through twitter. In *AAAI ICWSM*, volume 6.

Jisun An, Meeyoung Cha, P. Krishna Gummadi, and Jon Crowcroft. 2011. Media landscape in twitter: A world of new conventions and political diversity. In *AAAI ICWSM*.

David S. Anderson, Chris Fleizach, Stefan Savage, and Geoffrey M. Voelker. 2007. *Spamscatter: Characterizing internet scam hosting infrastructure*. Ph.D. thesis.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020a. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, 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 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020b. 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*, pages 3364–3374, Online. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota.

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)*, pages 21–27, New Orleans, Louisiana.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91.

Andrew W Barrett and Lowell W Barrington. 2005. Bias in newspaper photograph selection. *Political Research Quarterly*, 58(4):609–618.

Matthew A Baum and Tim Groeling. 2008. New media and the polarization of american political discourse. *Political Communication*, 25(4):345–365.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. 2015. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado. Association for Computational Linguistics.

Vibhu Bhatia, Vidya Prasad Akavoor, Sejin Paik, Lei Guo, Mona Jalal, Alyssa Smith, David Assefa Tofu, Edward Edberg Halim, Yimeng Sun, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. OpenFraming: Open-sourced tool for computational framing analysis of multilingual data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–250, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emina Boudemagh and Izabela Moise. 2017. News media coverage of refugees in 2016: a GDELT case study. In *AAAI ICWSM*.

Maxwell T Boykoff and Jules M Boykoff. 2004. Balance as bias: Global warming and the us prestige press. *Global environmental change*, 14(2):125–136.

Ann M Brill. 2001. Online journalists embrace new marketing function. *Newspaper Research Journal*, 22(2):28–40.

Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. In *Proceedings of the IEEE International Conference* *on Privacy, Security, Risk, and Trust, and the IEEE International Conference on Social Computing*, SocialCom/PASSAT '11, pages 1–8.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *WWW Companion*, pages 975–980.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 675–684.

Cheng Chen, Kui Wu, Srinivasan Venkatesh, and Xudong Zhang. 2013. Battling the internet water army: detection of hidden paid posters. In *Advances in Social Networks Analysis and Mining 2013, ASONAM '13, Niagara, ON, Canada - August 25 - 29, 2013*, pages 116–120.

Davide Cozzolino and Luisa Verdoliva. 2018. Camera-based image forgery localization using convolutional neural networks. In *EUSIPCO*, pages 1372–1376.

Fernando Cardoso Durier da Silva, Rafael Vieira, and Ana Cristina Bicharra Garcia. 2019. Can machines learn to detect fake news? a survey focused on social media. In *HICSS*.

Dave D'Alessio and Mike Allen. 2000. Media bias in presidential elections: A meta-analysis. *Journal of communication*, 50(4):133–156.

Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, and Yelena Mejova. Seminar users in the Arabic Twitter sphere. In *SocInfo*.

Kareem Darwish, Peter Stefanov, Michaël J. Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *AAAI ICWSM*, pages 141–152.

Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. 2018. Attending sentences to detect satirical fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3371–3380, Santa Fe, New Mexico, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

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 *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, INTERSPEECH '19.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, 8(9):938–949.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. 2018. Credibility assessment in the news: do we need to read. In *MIS2 Workshop*, pages 799–800.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

H. Finberg et al. 2002. Digital journalism credibility study. *Online News Association. Retrieved November*, 3:2003.

Soumen Ganguly, Juhi Kulshrestha, Jisun An, and Haewoon Kwak. 2020. Empirical evaluation of three common assumptions in building political media bias datasets. In *AAAI ICWSM*.

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71.

Martin Gilens and Craig Hertzman. 2000. Corporate ownership and news bias: Newspaper coverage of the 1996 telecommunications act. *The Journal of Politics*, 62(2):369–386.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16.

Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022a. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022b. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Benjamin Hadwiger, Daniele Baracchi, Alessandro Piva, and Christian Riess. 2019. Towards learned color representations for image splicing detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 8281–8285. IEEE.

S. Hao, A. Kantchelian, B. Miller, V. Paxson, and N. Feamster. 2016. Predator: proactive recognition and elimination of domain abuse at time-of-registration. In *ACM CCS*, pages 1568–1579.

Shuang Hao, Nadeem Ahmed Syed, Nick Feamster, Alexander G Gray, and Sven Krasser. 2009. Detecting spammers with snare: Spatio-temporal network-level automatic reputation engine. In *USENIX Security*, volume 9.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022a. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022b. CrowdChecked: Detecting previously fact-checked claims in social media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, AACL$_{I}JCNLP$ '22, *online*.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2016. In search of credible news. In *Proceedings of the 17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA '16, pages 172–180.

Joan B Hooper. 1975. *On assertive predicates*, volume 4. Academic Press, New York.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *CoRR*, abs/1703.09398.

Benjamin D. Horne, Sibel Adali, and Sujoy Sikdar. 2017. Identifying the social signals that drive online discussions: A case study of Reddit communities. In *IEEE ICCCN*, pages 1–9.

Benjamin D. Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the Twelfth International Conference on Web and Social Media*, ICWSM '18, pages 518–527.

Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. 2020. Identifying disinformation websites using infrastructure features. In *Proceedings of the 10th USENIX Workshop on Free and Open Communications on the Internet*, FOCI '20.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *ECCV*.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media*, ICWSM '14.

Ken Hyland. 2005. *Metadiscourse: Exploring Interaction in Writing*. Bloomsbury Publishing.

Luca Invernizzi, Stanislav Miskovic, Ruben Torres, Christopher Kruegel, Sabyasachi Saha, Giovanni Vigna, Sung-Ju Lee, and Marco Mellia. 2014. Nazca: Detecting malware distribution in large-scale networks. In *NDSS*, volume 14, pages 23–26.

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.

Massimo Iuliani, Marco Fanfani, Carlo Colombo, and Alessandro Piva. 2017. Reliability assessment of principal point estimates for forensic applications. *J. Vis. Comm. Img. Repr.*, 42:65–77.

Massimo Iuliani, Dasara Shullani, Marco Fontani, Saverio Meucci, and Alessandro Piva. 2018. A video forensic framework for the unsupervised analysis of mp4-like file container. *IEEE TIFS*, 14(3):635–645.

J. James, L. Sandhya, and C. Thomas. 2013. Detection of phishing urls using machine learning techniques. In *IEEE ICCC*, pages 304–309.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE TMM*, 19(3):598–608.

Kristen Johnson, Di Jin, and Dan Goldwasser. 2017. Leveraging behavioral and social information for weakly supervised collective classification of political discourse on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 741–752, Vancouver, Canada. Association for Computational Linguistics.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017a. We built a fake news / click bait filter: What happened next will blow your mind! In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 334–343, Varna, Bulgaria.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017b. Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 344–353, Varna, Bulgaria.

Lauri Karttunen. 1971. Implicative verbs. *Language*, 47(2):340–358.

Eric Kee, Micah K. Johnson, and Hany Farid. Digital image authentication from JPEG headers. *IEEE TIFS*.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2915–2921. ACM.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Michelle YoungJin Kim and Kristen Marie Johnson. 2022. CLoSE: Contrastive learning of subframe embeddings for political bias classification of news media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2780–2793, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA.

M. Konte, N. Feamster, and J. Jung. 2009. Dynamics of online scam hosting infrastructure. In *PAM*, pages 219–228.

Pawel Korus. 2017. Digital image integrity–a survey of protection and verification techniques. *Digital Signal Processing*, 71:1–26.

Neema Kotonya and Francesca Toni. 2020. 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.

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium. Association for Computational Linguistics.

Haewoon Kwak and Jisun An. 2014. A first look at global news coverage of disasters by using the gdelt dataset. In *SocInfo*, pages 300–308.

Haewoon Kwak and Jisun An. 2016. Two tales of the world: Comparison of widely used world news datasets GDELT and EventRegistry. In *AAAI ICWSM*.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *ACM WebSci*, page 305–314.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A survey on truth discovery. *SIGKDD Explor. Newsl.*, 17(2).

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. 2018. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *IEEE WIFS*.

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *IEEE/ACM 2018 International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Barcelona, Spain, August 28-31, 2018*, pages 552–559.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 342–351. ACM.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3818–3824.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 1751–1754.

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.

Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2017. Detection of sockpuppets in social media. In *CSCW*, CSCW '17.

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. ijcai.org.

Falko Matern, Christian Riess, and Marc Stamminger. 2020. Gradient-based illumination description for image forgery detection. *IEEE TIFS*, 15:1303–1317.

Todor Mihaylov and Preslav Nakov. 2016. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 399–405, Berlin, Germany.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James R. Glass. 2018. Fact checking in community forums. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pages 5309–5316, New Orleans, Louisiana, USA.

Lewis Mitchell, Kameron Decker Harris, Morgan R. Frank, Peter Sheridan Dodds, and Christopher M. Danforth. 2013. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PloS one*, 8(5):e64417.

R. M. Mohammad, F. Thabtah, and L. McCluskey. 2012. An assessment of features related to phishing websites using an automated technique. In *IEEE ICITST*, pages 492–497.

R. M. Mohammad, F. Thabtah, and L. McCluskey. 2014. Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications*, 25(2).

Morstatter et al. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2):1–18.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 353–362.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, IJCAI '21, pages 4551–4558.

An T. Nguyen, Aditya Kharosekar, Matthew Lease, and Byron C. Wallace. 2018. An interpretable joint graphical model for fact-checking from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, pages 1511–1518.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2022. Fang: Leveraging social context for fake news detection using graph representation. *Commun. ACM*, 65(4):124–132.

O'Brien et al. 2012. Exposing photo manipulation with inconsistent reflections. *ACM Trans. Graph.*, 31(1):4:1–4:11.

Jeremy Padgett, Johanna L Dunaway, and Joshua P Darr. 2019. As seen on tv? how gatekeeping makes the us house seem more extreme. *Journal of Communication*, 69(6):696–719.

Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *Proceedings of the International Semantic Web Conference*, ISWC '18.

Panayot Panayotov, Utsav Shukla, Husrev Taha Sencar, Mohamed Nabeel, and Preslav Nakov. 2022. GREENER: Graph neural networks for news media profiling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP '22, Abu Dhabi, UAE.

Kunwoo Park, Haewoon Kwak, Jisun An, and Sanjay Chawla. 2021. Understanding effects of editing tweets for news sharing by media accounts through a causal inference framework. In *AAAI ICWSM*.

Bo Peng, Wei Wang, Jing Dong, and Tieniu Tan. 2017. Optimized 3d lighting environment estimation for image forgery detection. *IEEE TIFS*, 12(2):479–494.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *CIKM*.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the Web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 1003–1012, Perth, Australia.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 155–158, Lyon, France.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia.

Rizka Purwanto, Arindam Pal, Alan Blair, and Sanjay Jha. 2020. Phishzip: A new compression-based algorithm for detecting phishing websites. In *IEEE CNS*, pages 1–9.

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

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ribeiro et al. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *AAAI ICWSM*, pages 290–299.

Christian Riess, Mathias Unberath, Farzad Naderi, Sven Pfaller, Marc Stamminger, and Elli Angelopoulou. 2017. Handling multiple materials for exposure of digital forgeries using 2-d lighting environments. *MM. Tools Appl.*, 76(4):4747–4764.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112.

Husrev Taha Sencar and Nasir Memon, editors. 2013. *Digital Image Forensics*. Springer.

Sencar et al., editor. 2022. *Multimedia Forensics*. Springer.

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 31st International Joint Conference on Artificial Intelligence*, IJCAI-ECAI '22, Vienna, Austria.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1).

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *IEEE BigMM*, pages 39–47.

Stuart N Soroka. 2012. The gatekeeping function: Distributions of information in media and the real world. *The Journal of Politics*, 74(2):514–528.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

Robert L. Stevenson, Richard A. Eisinger, Barry M. Feinberg, and Alan B. Kotok. 1973. Untwisting the news twisters: A replication of efron's study. *Journalism Quarterly*, 50(2):211–219.

Edson C Tandoc Jr. 2014. Journalism is twerking? how web analytics is changing the process of gatekeeping. *New media & society*, 16(4):559–575.

James Thorne and Andreas Vlachos. 2018a. 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.

James Thorne and Andreas Vlachos. 2018b. Automated fact checking: Task formulations, methods and future directions. *arXiv preprint arXiv:1806.07687*.

Luisa Verdoliva. 2020. Media forensics and deepfakes: an overview. *IEEE J. of STSP*, 14(5):910–932.

S. Vosoughi et al. 2018. The spread of true and false news online. *Science*, 359(6380).

Paul Waldman and James Devitt. 1998. Newspaper photographs and the 1996 presidential election: The question of bias. *Journal Mass Commun Q*, 75(2):302–311.

Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. 2016. Copula ordinal regression for joint estimation of facial action unit intensity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4902–4910.

G. Wang, J. W. Stokes, C. Herley, and D. Felstead. 2013. Detecting malicious landing pages in malware distribution networks. In *IEEE/IFIP DSN*, pages 1–11.

David A Weaver and Bruce Bimber. 2008. Finding news stories: a comparison of searches using LexisNexis and Google News. *Journal Mass Commun Q*, 85(3):515–530.

Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2013. Quantifying political leaning from tweets and retweets. In *AAAI ICWSM*, pages 640–649.

Pengpeng Yang, Daniele Baracchi, Massimo Iuliani, Dasara Shullani, Rongrong Ni, Yao Zhao, and Alessandro Piva. 2020. Efficient video integrity analysis through container characterization. *IEEE J. of STSP*, 14(5):947–954.

Zaman et al. 2014. A bayesian approach for predicting the popularity of tweets. *Ann. Appl. Stat.*, 8(3).

Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James Glass, and Preslav Nakov. 2019. Tanbih: Get to know what you are reading. 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): System Demonstrations*, pages 223–228, Hong Kong, China.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *CSUR*.

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.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, Rob Procter, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29.