Predicting Factuality of Reporting and Bias of News Media Sources

Ramy Baly[1], **Georgi Karadzhov**[3], **Dimitar Alexandrov**[3], **James Glass**[1], **Preslav Nakov**[2]
[1]MIT Computer Science and Artificial Intelligence Laboratory, MA, USA
[2]Qatar Computing Research Institute, HBKU, Qatar;
[3]Sofia University, Bulgaria
{baly, glass}@mit.edu, pnakov@qf.org.qa
{georgi.m.karadjov, Dimityr.Alexandrov}@gmail.com

Abstract

We present a study on predicting the factuality of reporting and bias of news media. While previous work has focused on studying the veracity of claims or documents, here we are interested in characterizing entire news media. These are under-studied but arguably important research problems, both in their own right and as a prior for fact-checking systems. We experiment with a large list of news websites and with a rich set of features derived from (*i*) a sample of articles from the target news medium, (*ii*) its Wikipedia page, (*iii*) its Twitter account, (*iv*) the structure of its URL, and (*v*) information about the Web traffic it attracts. The experimental results show sizable performance gains over the baselines, and confirm the importance of each feature type.

1 Introduction

The rise of social media has democratized content creation and has made it easy for everybody to share and spread information online. On the positive side, this has given rise to citizen journalism, thus enabling much faster dissemination of information compared to what was possible with newspapers, radio, and TV. On the negative side, stripping traditional media from their gate-keeping role has left the public unprotected against the spread of misinformation, which could now travel at breaking-news speed over the same democratic channel. This has given rise to the proliferation of false information that is typically created either (*a*) to attract network traffic and gain financially from showing online advertisements, e.g., as is the case of *clickbait*, or (*b*) to affect individual people's beliefs, and ultimately to influence major events such as political elections (Vosoughi et al., 2018). There are strong indications that false information was weaponized at an unprecedented scale during the 2016 U.S. presidential campaign.

"Fake news", which can be defined as "fabricated information that mimics news media content in form but not in organizational process or intent" (Lazer et al., 2018), became the word of the year in 2017, according to Collins Dictionary. "Fake news" thrive on social media thanks to the mechanism of sharing, which amplifies effect. Moreover, it has been shown that "fake news" spread faster than real news (Vosoughi et al., 2018). As they reach the same user several times, the effect is that they are perceived as more credible, unlike old-fashioned spam that typically dies the moment it reaches its recipients. Naturally, limiting the sharing of "fake news" is a major focus for social media such as Facebook and Twitter.

Additional efforts to combat "fake news" have been led by fact-checking organizations such as Snopes, FactCheck and Politifact, which manually verify claims. Unfortunately, this is inefficient for several reasons. First, manual fact-checking is slow and debunking false information comes too late to have any significant impact. At the same time, automatic fact-checking lags behind in terms of accuracy, and it is generally not trusted by human users. In fact, even when done by reputable fact-checking organizations, debunking does little to convince those who already believe in false information.

A third, and arguably more promising, way to fight "fake news" is to focus on their source. While "fake news" are spreading primarily on social media, they still need a "home", i.e., a website where they would be posted. Thus, if a website is known to have published non-factual information in the past, it is likely to do so in the future. Verifying the reliability of the source of information is one of the basic tools that journalists in traditional media use to verify information. It is also arguably an important prior for fact-checking systems (Popat et al., 2017; Nguyen et al., 2018).

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3528–3539
Brussels, Belgium, October 31 - November 4, 2018.

Fact-checking organizations have been producing lists of unreliable online news sources, but these are incomplete and get outdated quickly. Therefore, there is a need to predict the factuality of reporting for a given online medium automatically, which is the focus of the present work. We further study the bias of the source (left vs. right), as the two problems are inter-connected, e.g., extreme-left and extreme-right websites tend to score low in terms of factual reporting. Our contributions can be summarized as follows:

- We focus on an under-explored but arguably very important problem: predicting the factuality of reporting of a news medium. We further study bias, which is also under-explored.
- We create a new dataset of news media sources, which has annotations for both tasks, and is 1-2 orders of magnitude larger than what was used in previous work. We release the dataset and our code, which should facilitate future research.[1]
- We use a variety of sources such as (*i*) a sample of articles from the target website, (*ii*) its Wikipedia page, (*iii*) its Twitter account, (*iv*) the structure of its URL, and (*v*) information about the Web traffic it has attracted. This combination, as well as some of the sources, are novel for these problems.
- We further perform an ablation study of the impact of the individual (groups of) features.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work. Section 3 describes our method and features. Section 4 presents the data, the experiments, and the evaluation results. Finally, Section 5 concludes with some directions for future work.

2 Related Work

Journalists, online users, and researchers are well-aware of the proliferation of false information, and thus topics such as credibility and fact-checking are becoming increasingly important. For example, the ACM Transactions on Information Systems journal dedicated, in 2016, a special issue on Trust and Veracity of Information in Social Media (Papadopoulos et al., 2016).

[1]The data and the code are at http://github.mit.edu/CSAIL-SLS/News-Media-Reliability/

There have also been some related shared tasks such as the SemEval-2017 task 8 on Rumor Detection (Derczynski et al., 2017), the CLEF-2018 lab on Automatic Identification and Verification of Claims in Political Debates (Atanasova et al., 2018; Barrón-Cedeño et al., 2018; Nakov et al., 2018), and the FEVER-2018 task on Fact Extraction and VERification (Thorne et al., 2018).

The interested reader can learn more about "fake news" from the overview by Shu et al. (2017), which adopted a data mining perspective and focused on social media. Another recent survey was run by Thorne and Vlachos (2018), which took a fact-checking perspective on "fake news" and related problems. Yet another survey was performed by Li et al. (2016), covering truth discovery in general. Moreover, there were two recent articles in *Science*: Lazer et al. (2018) offered a general overview and discussion on the science of "fake news", while Vosoughi et al. (2018) focused on the process of proliferation of true and false news online. In particular, they analyzed 126K stories tweeted by 3M people more than 4.5M times, and confirmed that "fake news" spread much wider than true news.

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.

2.1 Fact-Checking

At the claim-level, fact-checking and rumor detection have been primarily addressed using information extracted from social media, i.e., based on how users comment on the target 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). The Web has 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., 2018).

In both cases, the most important information sources 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?). Other important sources are linguistic expression, meta information, and temporal dynamics.

2.2 Stance Detection

Stance detection has been addressed as a task in its own right, where models have been developed based on data from the Fake News Challenge (Riedel et al., 2017; Thorne et al., 2017; Mohtarami et al., 2018; Hanselowski et al., 2018), or from SemEval-2017 Task 8 (Derczynski et al., 2017; Dungs et al., 2018; Zubiaga et al., 2018). It has also been studied for other languages such as Arabic (Darwish et al., 2017b; Baly et al., 2018).

2.3 Source Reliability Estimation

Unlike stance detection, the problem of source reliability remains largely under-explored. In the case of social media, it concerns modeling the user[2] who posted a particular message/tweet, while in the case of the Web, it is about the trustworthiness of the source (the URL domain, the medium). The latter is our focus in this paper.

In previous 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 manually fact-checked claims, without access to gold labels about the overall medium-level factuality of reporting (Mukherjee and Weikum, 2015; Popat et al., 2016, 2017, 2018). The assumption is that reliable media agree with true claims and disagree with false ones, while for unreliable media it is mostly the other way around. The trustworthiness of Web sources has also been studied from a Data Analytics perspective. For instance, Dong et al. (2015) proposed that a trustworthy source is one that contains very few false facts. In this paper, we follow a different approach by studying the source reliability as a task in its own right, using manual gold annotations specific for the task.

Note that estimating the reliability of a source is important not only when fact-checking a claim (Popat et al., 2017; Nguyen et al., 2018), but it also gives an important prior when solving article-level tasks such as "fake news" and click-bait detection (Brill, 2001; Finberg et al., 2002; Hardalov et al., 2016; Karadzhov et al., 2017a; De Sarkar et al., 2018; Pan et al., 2018; Pérez-Rosas et al., 2018).

[2]User modeling in social media and news community forums has focused on finding malicious users such as opinion manipulation *trolls*, paid (Mihaylov et al., 2015b) or just perceived (Mihaylov et al., 2015a; Mihaylov and Nakov, 2016; Mihaylov et al., 2018; Mihaylova et al., 2018), *sockpuppets* (Maity et al., 2017), *Internet water army* (Chen et al., 2013), and *seminar users* (Darwish et al., 2017a).

2.4 "Fake News" Detection

Most work on "fake news" detection has relied on medium-level labels, which were then assumed to hold for all articles from that source.

Horne and Adali (2017) analyzed three small datasets ranging from a couple of hundred to a few thousand articles from a couple of dozen sources, comparing (*i*) real news vs. (*ii*) "fake news" vs. (*iii*) satire, and found that the latter two have a lot in common across a number of dimensions. They designed a rich set of features that analyze the text of a news article, modeling its complexity, style, and psychological characteristics. They found that "fake news" pack a lot of information in the title (as the focus is on users who do not read beyond the title), and use shorter, simpler, and repetitive content in the body (as writing fake information takes a lot of effort). Thus, they argued that the title and the body should be analyzed separately.

In follow-up work, Horne et al. (2018b) created a large-scale dataset covering 136K articles from 92 sources from `opensources.co`, which they characterize based on 130 features from seven categories: structural, sentiment, engagement, topic-dependent, complexity, bias, and morality. We use this set of features when analyzing news articles.

In yet another follow-up work, Horne et al. (2018a) trained a classifier to predict whether a given news article is coming from a reliable or from an unreliable ("*fake news*" or *conspiracy*)[3] source. Note that they assumed that all news from a given website would share the same reliability class. Such an assumption is fine for training (distant supervision), but we find it problematic for testing, where we believe manual documents-level labels are needed.

Potthast et al. (2018) used 1,627 articles from nine sources, whose factuality has been manually verified by professional journalists from BuzzFeed. They applied stylometric analysis, which was originally designed for authorship verification, to predict factuality (fake vs. real).

Rashkin et al. (2017) focused on the language used by "fake news" and compared the prevalence of several features in articles coming from trusted sources vs. hoaxes vs. satire vs. propaganda. However, their linguistic analysis and their automatic classification were at the article level and they only covered eight news media sources.

[3]We show in parentheses, the labels from `opensources.co` that are used to define a category.

Unlike the above work, (*i*) we perform classification at the news medium level rather than focusing on an individual article. Thus, (*ii*) we use reliable manually-annotated labels as opposed to noisy labels resulting from projecting the category of a news medium to all news articles published by this medium (as most of the work above did).[4] Moreover, (*iii*) we use a much larger set of news sources, namely 1,066, which is 1-2 orders of magnitude larger than what was used in previous work. Furthermore, (*iv*) we use a larger number of features and a wider variety of feature types compared to the above work, including features extracted from knowledge sources that have been largely neglected in the literature so far such as information from Wikipedia and the structure of the medium's URL.

2.5 Media Bias Detection

As we mentioned above, bias was used as a feature for "fake news" detection (Horne et al., 2018b). It has also been the target of classification, e.g., Horne et al. (2018a) predicted whether an article is biased (*political* or *bias*) vs. unbiased. Similarly, Potthast et al. (2018) classified the bias in a target article as (*i*) left vs. right vs. mainstream, or as (*ii*) hyper-partisan vs. mainstream. Finally, Rashkin et al. (2017) studied propaganda, which can be seen as extreme bias. See also a recent position paper (Pitoura et al., 2018) and an overview on bias the Web (Baeza-Yates, 2018).

Unlike the above work, we focus on bias at the medium level rather than at the article level. Moreover, we work with fine-grained labels on an ordinal scale rather then having a binary setup (some work above had three degrees of bias, while we have seven).

3 Method

In order to predict the factuality of reporting and the bias for a given news medium, we collect information from multiple relevant sources, which we use to train a classifier. In particular, we collect a rich set of features derived from (*i*) a sample of articles from the target news medium, (*ii*) its Wikipedia page if it exists, (*iii*) its Twitter account if it exists, (*iv*) the structure of its URL, and (*v*) information about the Web traffic it has attracted. We describe each of these sources below.

[4]Two notable exceptions are (Potthast et al., 2018) and (Pérez-Rosas et al., 2018), who use news articles whose factuality has been manually checked and annotated.

Articles We argue that analysis (textual, syntactic and semantic) of the content of the news articles published by a given target medium should be critical for assessing the factuality of its reporting, as well as of its potential bias. Towards this goal, we borrow a set of 141 features that were previously proposed for detecting "fake news" articles (Horne et al., 2018b), as we have described above. These features are used to analyze the following article characteristics:

- **Structure**: POS tags, linguistic features based on the use of specific words (function words, pronouns, etc.), and features for clickbait title classification from (Chakraborty et al., 2016);
- **Sentiment**: sentiment scores using lexicons (Recasens et al., 2013; Mitchell et al., 2013) and full systems (Hutto and Gilbert, 2014);
- **Engagement**: number of shares, reactions, and comments on Facebook;
- **Topic**: lexicon features to differentiate between science topics and personal concerns;
- **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 as calculated using pre-trained classifiers (Horne et al., 2017);
- **Morality**: features based on the Moral Foundation Theory (Graham et al., 2009) and lexicons (Lin et al., 2017)

Further details are available in (Horne et al., 2018b). For each target medium, we retrieve some articles, then we calculate these features separately for the title and for the body of each article, and finally we average the values of the 141 features over the set of retrieved articles.

Wikipedia We further leverage Wikipedia as an additional source of information that can help predict the factuality of reporting and the bias of a target medium. For example, the absence of a Wikipedia page may indicate that a website is not credible. Also, the content of the page might explicitly mention that a certain website is satirical, left-wing, or has some property related to our task.

Accordingly, we extract the following features:

- *Has Page:* indicates whether the target medium has a Wikipedia page;
- Vector representation for each of the following segments of the Wikipedia page, whenever applicable: *Content*, *Infobox*, *Summary*, *Categories*, and *Table of Contents*. We generate these representations by averaging the word embeddings (pretrained word2vec embeddings) of the corresponding words.

Twitter Given the proliferation of social media, most news media have Twitter accounts, which they use to reach out to more users online. The information that can be extracted from a news medium's Twitter profile can be valuable for our tasks. In particular, we use the following features:

- *Has Account:* Whether the medium has a Twitter account. We check this based on the top results for a search against Google, restricting the domain to `twitter.com`. The idea is that media that publish unreliable information might have no Twitter accounts.
- *Verified:* Whether the account is verified by Twitter. The assumption is that "fake news" media would be less likely to have their Twitter account verified. They might be interested in pushing their content to users via Twitter, but they would also be cautious about revealing who they are (which is required by Twitter to get them verified).
- *Created:* The year the account was created. The idea is that accounts that have been active over a longer period of time are more likely to belong to established media.
- *Has Location:* Whether the account provides information about its location. The idea is that established media are likely to have this public, while "fake news" media may want to hide it.
- *URL Match:* Whether the account includes a URL to the medium, and whether it matches the URL we started the search with. Established media are interested in attracting traffic to their website, while fake media might not. Moreover, some fake accounts mimic genuine media, but have a slightly different domain, e.g., `.com.co` instead of `.com`.
- *Counts*: Statistics about the number of friends, statuses, and favorites. Established media might have higher values for these.
- *Description:* A vector representation generated by averaging the *Google News* embeddings (Mikolov et al., 2013) of all words of the profile description paragraph. These short descriptions might contain an open declaration of partisanship, i.e., left or right political ideology (bias). This could also help predict factuality as extreme partisanship often implies low factuality. In contrast, "fake news" media might just leave this description empty, while high-quality media would want to give some information about who they are.

URL We also collect additional information from the website's URL using character-based modeling and hand-crafted features. URL features are commonly used in phishing website detection systems to identify malicious URLs that aim to mislead users (Ma et al., 2009). As we want to predict a website's factuality, using URL features is justified by the fact that low-quality websites sometimes try to mimic popular news media by using a URL that looks similar to the credible source. We use the following URL-related features:

- *Character-based:* Used to model the URL by representing it in the form of a one-hot vector of character *n*-grams, where $n \in [2, 5]$. Note that these features are not used in the final system as they could not outperform the baseline (when used in isolation).
- *Orthographic:* These features are very effective for detecting phishing websites, as malicious URLs tend to make excessive use of special characters and sections, and ultimately end up being longer. For this work, we use the length of the URL, the number of sections and the excessive use of special characters such as digits, hyphens and dashes. In particular, we identify whether the URL contains digits, dashes or underscores as individual symbols, which were found to be useful as features for detecting phishing URLs (Basnet et al., 2014). We also check whether the URL contains short (less than three symbols) or long sections (more than ten symbols), as a high number of such sections could indicate an irregular URL.

Name	URL	Factuality	Twitter Handle	Wikipedia page
Associated Press	http://apnews.com	*Very High	@apnews	~/wiki/Associated_Press
NBC News	http://www.nbcnews.com/	High	@nbcnews	~/wiki/NBC_News
Russia Insider	http://russia-insider.com	Mixed	@russiainsider	~/wiki/Russia_Insider
Patriots Voice	http://patriotsvoice.info/	Low	@pegidaukgroup	N/A

Table 1: Examples of media with various factuality scores. (*In our experiments, we treat *Very High* as *High*.)

Name	URL	Bias	Twitter Handle	Wikipedia page
Loser.com	http://loser.com	Extreme Left	@Loser_dot_com	~/Loser.com
Die Hard Democrat	http://dieharddemocrat.com/	Left	@democratdiehard	N/A
Democracy 21	http://www.democracy21.org/	Center-Left	@fredwertheimer	~/Democracy_21
Federal Times	http://www.federaltimes.com/	Center	@federaltimes	~/Federal_Times
Gulf News	http://gulfnews.com/	Center-Right	@gulf_news	~/Gulf_News
Fox News	http://www.foxnews.com/	Right	@foxnews	~/Fox_News
Freedom Outpost	http://freedomoutpost.com/	Extreme Right	@FreedomOutpost	N/A

Table 2: Examples of media with various bias scores.

- *Credibility:* Model the website's URL credibility by analyzing whether it (*i*) uses `https://`, (*ii*) resides on a blog-hosting platform such as `blogger.com`, and (*iii*) uses a special top-level domain, e.g., `.gov` is for governmental websites, which are generally credible and unbiased, whereas `.co` is often used to mimic `.com`.

Web Traffic Analyzing the web traffic to the website of the medium might be useful for detecting phishy websites that come and disappear in certain patterns. Here, we only use the reciprocal value of the website's *Alexa Rank*,[5] which is a global ranking for over 30 million websites in terms of the traffic they receive.

We evaluate the above features in Section 4, both individually and as groups, in order to determine which ones are important to predict factuality and bias, and also to identify the ones that are worth further investigation in future work.

4 Experiments and Evaluation

4.1 Data

We use information about news media listed on the Media Bias/Fact Check (MBFC) website,[6] which contains manual annotations and analysis of the factuality of reporting and/or bias for over 2,000 news websites. Our dataset includes 1,066 websites for which *both* bias and factuality labels were explicitly provided, or could be easily inferred (e.g., *satire* is of low factuality).

We model factuality on a 3-point scale (*Low*, *Mixed*, and *High*),[7] and bias on a 7-point scale (*Extreme-Left*, *Left*, *Center-Left*, *Center*, *Center-Right*, *Right*, and *Extreme-Right*).

Some examples from our dataset are presented in Table 1 for factuality of reporting, and in Table 2 for bias. In both tables, we show the names of the media, as well as their corresponding Twitter handles and Wikipedia pages, which we found automatically. Overall, 64% of the websites in our dataset have Wikipedia pages, and 94% have Twitter accounts. In cases of "fake news" sites that try to mimic real ones, e.g., `ABCnews.com.co` is a fake version of `ABCnews.com`, it is possible that our Twitter extractor returns the handle for the real medium. This is where the *URL Match* feature comes handy (see above).

Table 3 provides detailed statistics about the dataset. Note that we have 1-2 orders of magnitude more media sources than what has been used in previous studies, as we already mentioned in Section 2 above.

Factuality		Bias	
Low	256	Extreme-Left	21
Mixed	268	Left	168
High	542	Center-Left	209
		Center	263
		Center-Right	92
		Right	157
		Extreme-Right	156

Table 3: Label distribution (counts) in our dataset.

[5] http://www.alexa.com/

[6] https://mediabiasfactcheck.com

[7] MBFC also uses *Very High* as a label, but due to its very small size, we merged it with *High*.

Source	Feature	Dim.	Factuality				Bias			
			Macro-F_1	**Acc.**	**MAE**	**MAE^M**	**Macro-F_1**	**Acc.**	MAE	MAE^M
Majority Baseline			22.47	50.84	0.73	1.00	5.65	24.67	1.39	1.71
Traffic	*Alexa rank*	1	22.46	50.75	0.73	1.00	7.76	25.70	1.38	1.71
URL	*URL structure*	12	39.30	53.28	0.68	0.81	13.50	23.64	1.65	2.06
Twitter	*created at.*	1	30.72	52.91	0.69	0.92	5.65	24.67	1.39	1.71
	has account	1	30.72	52.91	0.69	0.92	5.65	24.67	1.39	1.71
	verified	1	30.72	52.91	0.69	0.92	5.65	24.67	1.39	1.71
	has location	1	36.73	52.72	0.69	0.82	9.44	24.86	1.54	1.85
	URL match	2	39.98	54.60	0.66	0.72	10.16	25.61	**1.51**	1.97
	description	300	44.79	51.41	0.65	0.70	19.08	25.33	1.73	2.04
	counts	5	46.88	**57.22**	**0.57**	0.66	18.34	24.86	1.62	2.01
	Twitter – All	308	**48.23**	54.78	0.59	**0.64**	**21.38**	**27.77**	1.58	**1.83**
Wikipedia	*has page*	1	43.53	59.10	0.57	0.63	14.33	26.83	1.63	2.14
	table of content	300	43.95	51.04	0.60	0.65	15.10	22.96	1.86	2.25
	categories	300	46.36	53.70	0.65	0.61	25.64	32.16	1.70	2.10
	information box	300	46.39	51.14	0.71	0.65	19.79	26.85	1.68	1.99
	summary	300	51.88	58.91	0.54	0.52	30.02	37.43	1.47	**1.98**
	content	300	55.29	62.10	0.51	0.50	**30.92**	**38.61**	**1.51**	2.01
	Wikipedia – All	301	**55.52**	**62.29**	**0.50**	**0.49**	28.66	35.93	1.51	2.00
Articles	*title*	141	53.20	59.57	0.51	0.58	30.91	37.52	1.29	1.53
	body	141	**58.02**	**64.35**	**0.43**	**0.51**	**36.63**	**41.74**	**1.15**	**1.43**

Table 4: Results for factuality and bias prediction. **Bold** values indicate the best-performing feature type in its family of features, while **underlined** values indicate the best-performing feature type overall.

In order to compute the article-related features, we did the following: (*i*) we crawled 10–100 articles per website (a total of 94,814), (*ii*) we computed a feature vector for each article, and (*iii*) we averaged the feature vectors for the articles from the same website to obtain the final vector of article-related features.

4.2 Experimental Setup

We used the above features in a Support Vector Machine (SVM) classifier, training a separate model for factuality and for bias. We report results for 5-fold cross-validation. We tuned the SVM hyper-parameters, i.e., the cost C, the kernel type, and the kernel width γ, using an internal cross-validation on the training set and optimizing macro-averaged F_1. Generally, the RBF kernel performed better than the linear kernel.

We report accuracy and macro-averaged F_1 score. We also report Mean Average Error (MAE), which is relevant given the ordinal nature of both the factuality and the bias classes, and also MAE^M, which is a variant of MAE that is more robust to class imbalance. See (Baccianella et al., 2009; Rosenthal et al., 2017) for more details about MAE^M vs. MAE.

4.3 Results and Discussion

We present in Table 4 the results of using features from the different sources proposed in Section 3. We start by describing the contribution of each feature type towards factuality and bias.

We can see that the textual features extracted from the ARTICLES yielded the best performance on factuality. They also perform well on bias, being the only type that beats the baseline on MAE. These results indicate the importance of analyzing the contents of the target website. They also show that using the *titles* only is not enough, and that the article *bodies* contain important information that should not be ignored.

Overall, the WIKIPEDIA features are less useful for factuality, and perform reasonably well for bias. The best features from this family are those about the page *content*, which includes a general description of the medium, its history, ideology and other information that can be potentially helpful. Interestingly, the *has page* feature alone yields sizable improvement over the baseline, especially for factuality. This makes sense given that trustworthy websites are more likely to have Wikipedia pages; yet, this feature does not help much for predicting political bias.

Features	Macro-F_1	Acc.	MAE	MAE^M
MAJORITY BASELINE	22.47	50.84	0.73	1.00
FULL	59.91	65.48	0.41	0.44
FULL W/O TRAFFIC	59.90	65.39	0.41	0.43
FULL W/O TWITTER	59.52	65.10	0.41	0.47
FULL W/O URL	57.23	63.32	0.44	0.49
FULL W/O ARTICLES	56.15	63.13	0.46	0.51
FULL W/O WIKIPEDIA	55.93	63.23	0.44	0.52

Table 5: Ablation study for the contribution of each feature type for predicting the factuality of reporting.

Features	7-Way Bias				3-Way Bias			
	Macro-F_1	Acc.	MAE	MAE^M	Macro-F_1	Acc.	MAE	MAE^M
MAJORITY BASELINE	5.65	24.67	1.39	1.71	22.61	51.33	0.49	0.67
FULL	37.50	39.87	1.25	1.55	61.31	68.86	0.39	0.53
FULL W/O TRAFFIC	37.49	39.84	1.25	1.55	61.30	68.86	0.38	0.53
FULL W/O TWITTER	36.88	39.49	1.20	1.38	63.27	69.89	0.38	0.50
FULL W/O URL	36.60	39.68	1.24	1.48	60.93	68.11	0.40	0.53
FULL W/O WIKIPEDIA	34.75	37.62	1.33	1.58	59.92	66.89	0.41	0.54
FULL W/O ARTICLES	29.95	36.96	1.40	1.85	53.67	62.48	0.47	0.62

Table 6: Ablation study for the contribution of each feature type for predicting media bias.

The TWITTER features perform moderately for factuality and poorly for bias. This is not surprising, as we normally may not be able to tell much about the political ideology of a website just by looking at its Twitter profile (not its tweets!) unless something is mentioned in its *description*, which turns out to perform better than the rest of the features from this family. We can see that the *has twitter* feature is less effective than *has wiki* for factuality, which makes sense given that Twitter is less regulated than Wikipedia. Note that the *counts* features yield reasonable performance, indicating that information about activity (e.g., number of statuses) and social connectivity (e.g., number of followers) is useful. Overall, the TWITTER features seem to complement each other, as their union yields the best performance on factuality.

The URL features are better used for factuality rather than bias prediction. This is mainly due to the nature of these features, which are aimed at detecting phishing websites, as we mentioned in Section 3. Overall, this feature family yields slight improvements, suggesting that it can be useful when used together with other features.

Finally, the *Alexa rank* does not improve over the baseline, which suggests that more sophisticated TRAFFIC-related features might be needed.

4.4 Ablation Study

Finally, we performed an ablation study in order to evaluate the impact of removing one family of features at a time, as compared to the FULL system, which uses all the features. We can see in Tables 5 and 6 that the FULL system achieved the best results for factuality, and the best macro-F_1 for bias, suggesting that the different types of features are largely complementary and capture different aspects that are all important for making a good classification decision.

For factuality, excluding the WIKIPEDIA features yielded the biggest drop in performance. This suggests that they provide information that may not be available in other sources, including the ARTICLES, which achieved better results alone. On the other hand, excluding the TRAFFIC feature had no effect on the model's performance.

For bias, we experimented with classification on both a 7-point and a 3-point scale.[8] Similarly to factuality, the results in Table 6 indicate that WIKIPEDIA offers complementary information that is critical for bias prediction, while TRAFFIC makes virtually no difference.

[8] We performed the following mapping: {*Extreme-Right, Right*}→Right, {*Extreme-Left, Left*}→Left, and {*Center, Right-Center, Left-Center*}→Center

5 Conclusion and Future Work

We have presented a study on predicting factuality of reporting and bias of news media, focusing on characterizing them as a whole. These are under-studied, but arguably important research problems, both in their own right and as a prior for fact-checking systems.

We have created a new dataset of news media sources that has annotations for both tasks and is 1-2 orders of magnitude larger than what was used in previous work. We are releasing the dataset and our code, which should facilitate future research.

We have experimented with a rich set of features derived from the contents of (*i*) a sample of articles from the target news medium, (*ii*) its Wikipedia page, (*iii*) its Twitter account, (*iv*) the structure of its URL, and (*v*) information about the Web traffic it has attracted. This combination, as well as some of the types of features, are novel for this problem.

Our evaluation results have shown that most of these features have a notable impact on performance, with the articles from the target website, its Wikipedia page, and its Twitter account being the most important (in this order). We further performed an ablation study of the impact of the individual types of features for both tasks, which could give general directions for future research.

In future work, we plan to address the task as ordinal regression, and further to model the interdependencies between factuality and bias in a joint model. We are also interested in characterizing the factuality of reporting for media in other languages. Finally, we want to go beyond *left vs. right* bias that is typical of the Western world and to model other kinds of biases that are more relevant for other regions, e.g., *islamist vs. secular* is one such example for the Muslim World.

Acknowledgments

This research was carried out in collaboration between the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) and the Qatar Computing Research Institute (QCRI), HBKU.

We would like to thank Israa Jaradat, Kritika Mishra, Ishita Chopra, Laila El-Beheiry, Tanya Shastri, and Hamdy Mubarak for helping us with the data extraction, cleansing, and preparation.

Finally, we thank the anonymous reviewers for their constructive comments, which have helped us improve this paper.

References

Pepa Atanasova, Lluís Màrquez, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 1: Check-worthiness. In *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France. CEUR-WS.org.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *Proceedings of the 9th IEEE International Conference on Intelligent Systems Design and Applications*, ISDA '09, pages 283–287, Pisa, Italy.

Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM*, 61(6):54–61.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. 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*, NAACL-HLT '18, pages 21–27, New Orleans, LA, USA.

Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims, task 2: Factuality. In *CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, Avignon, France. CEUR-WS.org.

Ram B Basnet, Andrew H Sung, and Quingzhong Liu. 2014. Learning to detect phishing URLs. *International Journal of Research in Engineering and Technology*, 3(6):11–24.

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

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, Boston, MA, USA.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, Hyderabad, India.

Abhijnan Chakraborty, Bhargavi Paranjape, Kakarla Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '16, pages 9–16, San Francisco, CA, USA.

Cheng Chen, Kui Wu, Venkatesh Srinivasan, and Xudong Zhang. 2013. Battling the Internet Water Army: detection of hidden paid posters. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 116–120, Niagara, Canada.

Kareem Darwish, Dimitar Alexandrov, Preslav Nakov, and Yelena Mejova. 2017a. Seminar users in the Arabic Twitter sphere. In *Proceedings of the 9th International Conference on Social Informatics*, SocInfo '17, pages 91–108, Oxford, UK.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017b. Improved stance prediction in a user similarity feature space. In *Proceedings of the Conference on Advances in Social Networks Analysis and Mining*, ASONAM '17, pages 145–148, Sydney, Australia.

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*, COLING '18, pages 3371–3380, Santa Fe, NM, USA.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, pages 60–67, Vancouver, Canada.

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*, COLING '18, pages 3360–3370, Santa Fe, NM, USA.

Howard Finberg, Martha L Stone, and Diane Lynch. 2002. Digital journalism credibility study. *Online News Association. Retrieved November*, 3:2003.

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.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 1859–1874, Santa Fe, NM, USA.

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, Varna, Bulgaria.

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 Horne, Sibel Adali, and Sujoy Sikdar. 2017. Identifying the social signals that drive online discussions: A case study of Reddit communities. In *Proceedings of the 26th IEEE International Conference on Computer Communication and Networks*, ICCCN '17, pages 1–9, Vancouver, Canada.

Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. 2018a. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Proceedings of the The Web Conference*, WWW '18, pages 235–238, Lyon, France.

Benjamin D. Horne, Sara Khedr, and Sibel Adali. 2018b. 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, Stanford, CA, USA.

Clayton 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, Ann Arbor, MI, USA.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017a. We built a fake news & clickbait filter: What happened next will blow your mind! In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, RANLP '17, 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 on Recent Advances in Natural Language Processing*, RANLP '17, pages 344–353, Varna, Bulgaria.

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*, COLING '18, pages 3402–3413, Santa Fe, NM, USA.

David M.J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

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):1–16.

Ying Lin, Joe Hoover, Morteza Dehghani, Marlon Mooijman, and Heng Ji. 2017. Acquiring background knowledge to improve moral value prediction. *arXiv preprint arXiv:1709.05467*.

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 25th International Joint Conference on Artificial Intelligence*, IJCAI '16, pages 3818–3824, New York, NY, USA.

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 on Conference on Information and Knowledge Management*, CIKM '15, pages 1751–1754, Melbourne, Australia.

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*, ACL '17, pages 708–717, Vancouver, Canada.

Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker. 2009. Identifying suspicious URLs: An application of large-scale online learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 681–688, Montreal, Canada.

Suman Kalyan Maity, Aishik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2017. Detection of sockpuppets in social media. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 243–246, Portland, OR, USA.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015a. Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, CoNLL '15, pages 310–314, Beijing, China.

Todor Mihaylov, Ivan Koychev, Georgi Georgiev, and Preslav Nakov. 2015b. Exposing paid opinion manipulation trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '15, pages 443–450, Hissar, Bulgaria.

Todor Mihaylov, Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Georgi Georgiev, and Ivan Koychev. 2018. The dark side of news community forums: Opinion manipulation trolls. *Internet Research*.

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*, ACL '16, pages 399–405, Berlin, Germany.

Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadjov, and James Glass. 2018. Fact checking in community forums. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, AAAI '18, pages 879–886, New Orleans, LA, USA.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 746–751, Atlanta, GA, USA.

Lewis Mitchell, Morgan R Frank, Kameron Decker Harris, 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.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 767–776, New Orleans, LA, USA.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 353–362, Melbourne, Australia.

Preslav Nakov, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Wajdi Zaghouani, Pepa Atanasova, Spas Kyuchukov, and Giovanni Da San Martino. 2018. Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. In *Proceedings of the Ninth International Conference of the CLEF Association: Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Lecture Notes in Computer Science, pages 372–387, Avignon, France. Springer.

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, New Orleans, LA, USA.

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, Monterey, CA, USA.

Symeon Papadopoulos, Kalina Bontcheva, Eva Jaho, Mihai Lupu, and Carlos Castillo. 2016. Overview of the special issue on trust and veracity of information in social media. *ACM Trans. Inf. Syst.*, 34(3):14:1–14:5.

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*, COLING '18, pages 3391–3401, Santa Fe, NM, USA.

Evaggelia Pitoura, Panayiotis Tsaparas, Giorgos Flouris, Irini Fundulaki, Panagiotis Papadakos, Serge Abiteboul, and Gerhard Weikum. 2018. On measuring bias in online information. *SIGMOD Rec.*, 46(4):16–21.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 2173–2178, Indianapolis, IN, USA.

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, pages 1003–1012, Perth, Australia.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A credibility lens for analyzing and explaining misinformation. In *Proceedings of The Web Conference 2018*, WWW '18, pages 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*, ACL '18, pages 231–240, Melbourne, Australia.

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*, EMNLP '17, pages 2931–2937, Copenhagen, Denmark.

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*, ACL '13, pages 1650–1659, Sofia, Bulgaria.

Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. *ArXiv:1707.03264*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, pages 502–518, Vancouver, Canada.

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):22–36.

James Thorne, Mingjie Chen, Giorgos Myrianthous, Jiashu Pu, Xiaoxuan Wang, and Andreas Vlachos. 2017. Fake news stance detection using stacked ensemble of classifiers. In *Proceedings of the EMNLP Workshop on Natural Language Processing meets Journalism*, pages 80–83, Copenhagen, Denmark.

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*, COLING '18, pages 3346–3359, Santa Fe, NM, USA.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 809–819, New Orleans, LA, USA.

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

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Inf. Process. Manage.*, 54(2):273–290.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, 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.