Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

NELA-GT-2018: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles

Jeppe Nørregaard,[†] Benjamin D. Horne,[*] Sibel Adalı[*]
Technical University of Denmark[†], Rensselaer Polytechnic Institute[*]
jepno@dtu.dk, horneb@rpi.edu, adalis@rpi.edu

Abstract

In this paper, we present a dataset of 713k articles collected between 02/2018-11/2018. These articles are collected directly from 194 news and media outlets including mainstream, hyper-partisan, and conspiracy sources. We incorporate ground truth ratings of the sources from 8 different assessment sites covering multiple dimensions of veracity, including reliability, bias, transparency, adherence to journalistic standards, and consumer trust. The NELA-GT-2018 dataset can be found at https://doi.org/10.7910/DVN/ULHLCB.

1 Introduction

One of the main gaps in the study of misinformation is finding broad labelled datasets, which this data set aims to fill. There are a number of published misinformation datasets with ground truth, but they are often small, event specific, engagement specific, or incomplete. As a result, they are not sufficient for answering a wide-range of research questions.

First, for many studies, particularly those involving machine learning methods, a large dataset with ground truth labels is necessary. Article-level ground truth (i.e. true/false) for such datasets can be infeasible, as fact-checking requires experts conducting a slow and labor-intensive process. Furthermore, the slow speed of fact-checking makes datasets quickly out-of-date. One solution that has been proposed to mitigate problems with article level labels is to use higher level labels, such as source reliability over an extended period of time (Horne et al. 2018; Baly et al. 2018).

Secondly, fact-checkers tend to concentrate their efforts on articles that receive a lot of attention, making datasets with fact-checked labels engagement-driven. Engagement-driven news datasets (for example those based on social media mentions), are very useful in engagement-driven studies, but may not provide a complete picture of attention to malicious news sources. For example, The Drudge Report, a site known for spreading mixed-veracity information, is 41st in United States in terms of the amount of Internet traffic, making it a highly influential source. Readers spend a long time on the site, averaging 25 minutes with about 11 clicks pages per visit. However, readers only reach the site using social-media links 4% of the time, while 83% of the time they reach it through direct links[1]. As a result, we argue that there is a need for datasets collected independent of social media in order to understand the full impact of and tactics used by misleading and hyper-partisan news producers.

Lastly, news, particularly state-sponsored propaganda, can misinform through methods other than explicitly fabricated claims (Zannettou et al. 2018). Hence, fact-checking labels may not capture all types of misinformation. This leads to labeling mechanisms that account for other factors, such as whether the sources have bias in their reporting or how much they adhere to journalistic standards. Therefore, we argue that datasets should contain multiple types of ground truth at the source-level in order to perform complete studies of misinformation.

The dataset presented in this paper is and engagement-independent collection of news articles with multiple types of source-level ground truths. Our dataset contains 713,534 articles from 194 news outlets collected between 01/02/2018-30/11/2018. These articles are collected directly from each news producers' websites, independent of social media. We corroborate ground truth labels from eight different assessment sites covering multiple dimensions of veracity, including reliability, bias, transparency, and consumer trust. The dataset sources are from both mainstream media and alternative media across multiple countries. The dataset can be found at https://doi.org/10.7910/DVN/ULHLCB. In this paper, we outline dataset collection, ground-truth corroboration, and provide a few use cases.

2 Related Work

There are many recent news datasets focused on misinformation, each with different focus in labelling. Labels include various dimensions of reliability and various dimensions of bias. **BuzzfeedNews**[2] is a small dataset of news articles that had high Facebook engagement during the 2016 U.S. Presidential Election. The dataset contains 1627 articles that are fact-checked by 5 Buzzfeed journalists. The dataset labels include if the article is false or true, along with the political leaning of the source that produced the article. **FakeNewsCorpus**[3] is a dataset containing nearly 10M articles labeled

[1] source: similarweb.com, consulted on 13/01/2019
[2] github.com/BuzzFeedNews/2016-10-facebook-fact-check
[3] github.com/several27/FakeNewsCorpus

using opensources.co. OpenSources is a list of sources labeled by experts. These labels include 13 different labels related to the reliability of the source. **FakeNewsNet** is a collection of datasets containing news articles and tweets. The dataset includes rich metadata including social features and spatiotemporal information (Shu et al. 2018). While this dataset is described in a paper on arxiv.com, to the best of our knowledge, the data has not been completely released to the public at this time [4].

Many other misinformation datasets have focused on individual claims rather than complete news articles. While claims can be extracted from news articles, most of these datasets use claims made on social media or by political figures in speeches. **LIAR** is a fake claim benchmark dataset that has 12.8K fact-check short statements from politifact.com (Wang 2017). The claims in the dataset are from social media posts and political speeches.**CREDBANK** is a dataset of 60M tweets between 2015 and 2016. Each tweet is associated to a news event and is labeled with credibility by Amazon Mechanical Turkers (Mitra and Gilbert 2015). Again, this dataset only contains claims/tweets, not complete news articles. **PHEME** is a dataset of 330 tweet threads annotated by journalist. Each tweet is associated with a news story (Zubiaga et al. 2016). **FacebookHoax** is a dataset containing 15K Facebook posts about science news. The posts are labeled as "hoax" or "non-hoax" and come from 32 different Facebook pages (Tacchini et al. 2017). These datasets are highly related to the smaller tweet credibility datasets created in the last decade (Castillo, Mendoza, and Poblete 2011).

There are also several recent unlabelled news datasets, which are much larger than most of the labeled datasets. **NELA2017** is a political news article dataset that contains 136K articles from 92 media sources in 2017 (Horne, Khedr, and Adalı 2018). The dataset includes sources from mainstream, hyper-partisan, conspiracy, and satire media sources. Along with the news articles, the dataset includes a rich set of natural language features on each news article, and the corresponding Facebook engagement statistics. The dataset contains nearly all of the articles published by the 92 sources during the 7 month period. **GDELT** is an open database of event-based news articles with temporal and location features. It is said to be one of the most comprehensive event-based news datasets. However, GDELT does not explicitly contain maliciously fake or hyper-partisan news sources, needed for misinformation studies.

While all of these datasets are useful, there are several limitations we address with the dataset presented int his paper:

1. Small number of sources and articles - With the exception of FakeNewsCorpus and the NELA2017 dataset, the current publicly available datasets are either small in the number of media sources they contain, small in the number of articles, or both. Furthermore, many of the larger datasets do not contain multiple types of sources. In comparison to FakeNewsCorpus, our dataset covers a wider range of news, in particular more mainstream news. In addition, our dataset is collected over a longer and more consistent period of time, where as the many of alternative news sources in FakeNewsCorpus no longer exists and the time frame of FakeNewsCorpus is unknown.
2. Engagement-driven - The majority of the current datasets, both for news articles and claims, contain only data has been highly engaged with on social media or has received attention from fact-checking organizations. While understanding the engagement of misinformation is an important task, engagement driven news datasets fail to show the complete picture of misinforming news. Both malicious fake news producers and hyper-partisan media produce hundreds, sometimes thousands of articles in a year, most of which are never seen on social media or fact-checkers. Questions about when fake news tactics work or do not work remain unanswered.
3. Lack of ground truth labels - All of the current large-scale news article datasets do not have any form of labeling for misinformation research, with exception of FakeNewsCorpus. While some contain a mix of reliable and unreliable sources, it is not necessarily clear to what extent each source is reliable or what dimensions of credibility should be used to assess the sources. For example, a news article can spread misinformation (or disinformation) in many ways other than false statements. A news article may use partially false information, decontextualized information, or information misrepresented by hyper-partisan language. For both machine learning and comparative studies, having well defined labels about multiple dimensions of veracity is important in understand what signals a machine learning model is learning or why discovered patterns exist in news data.

Thus, our goal with the NELA-GT-2018 dataset is to create a large, veracity-labeled news article dataset that in independent of social media engagement and specific events.

3 Dataset Creation

We created this dataset, with the following steps:

1. We gathered a wide variety of news sources from varying levels of veracity, including many well-studied misinforming sources and other less well-known sources.
2. We scraped article data from the gathered sources' RSS feeds twice a day for 10 months in 2018.
3. We combine and corroborated source-level veracity labels from 8 independent assessments, some of which are used in the misinformation literature, others that are not. These labels provide multiple and complementary ground truth allowing for many different ways to characterize the sources.

Through this process, we provide **713,534 articles** from **194 news and media producers**. Along with these articles, we provide multiple **labels from 8 independent assessments** for each source. The final set of article data is arranged in an sqlite data, with date, source, title, and cleaned text content for each article. The labels are provided in CSV format, with rows being sources and columns being each label gathered from all the assessment sites. The set of labels

[4] github.com/KaiDMML/FakeNewsNet

can also be found in Table 3 and Table 4. Specifics on the file-formats can be found in the documentation given with the dataset. We describe the collection process and ground truth in detail below.

News Article Data

To collect our dataset, we scraped the RSS feeds of each source twice a day starting on 02/02/2018 using the Python libraries feedparser and goose. Our starting point for source selection was mainstream outlets and alternative sources that are mentioned in other studies or high profile cases of false news coverage. An initial subset of 92 sources was available in NELA2017 dataset (Horne, Khedr, and Adalı 2018), which already covered a wide array of media types. We then continued to expand this source set using the same criteria, as well as by automated Google searches to find other outlets that published similar articles as those already in our dataset. Specifically, we queried the Google Search API with the titles of the news articles that were previously collected. If a news source that was not in our source collection list appeared in the top 10 pages of the Google search, we added it to our source collection list. Note, we do not include small local news sources or sources that did not have operational RSS feeds, which significantly reduces the size of the expected source set. Furthermore, this Google expansion process was ran in July 2018, which caused a large increase in unlabeled news sources, as shown in Figure 1.

By the end of the collection process (30/11/2018) we had 713K articles from 194 news and media producers. These sources come from a variety of countries, but are all articles are in English. In Tables 3, 4, and 5 we write the date of the first scraped article from each source. After these dates, we have near complete data from the respective sources RSS-feeds. In Figure 1 we show the number of articles collected over time.

Ground Truth Data

A number of organizations and platforms have developed methods for assessing reliability and bias of news sources. These organizations come from both the research community and from practitioner communities. While each of these organizations and platforms provide useful assessments on their own, each uses different criteria and methods to make their assessments, and most of these assessments cover relatively few sources. Thus, in order to create a large, centralized set of veracity labels, we collected ground truth (GT) data from eight different sites, which all attempt to assess the reliability and/or the bias of news.

These assessment sites are:

1. NewsGuard
2. Pew Research Center
3. Wikipedia
4. OpenSources
5. Media Bias/Fact Check (MBFC)
6. AllSides
7. BuzzFeed News
8. Politifact

We gather data from all these sites, using html-scraping and GUI-automation, and combine their labels to create a centralized set of veracity ground truth labels. Of the 194 sources in our data set, 154 sources have GT labels from at least one of the assessment sites, while the remaining 40 sources remain unlabelled. Tables 3 and 4 show the combined labels, while Table 5 lists the sources where no label information was found. Table 2 provide a detailed described of each assessment and Table 1 lists urls for the assessment sites.

NewsGuard uses a group of trained journalists to assess credibility and transparency of news websites. They emphasizes the use of trained people rather than algorithms to determine credibility of sources. They allows respective news outlets comment on their verdict before publishing it. They provide extensions for major browsers to inform users of the credibility of the sites they visit. They also display icons on search results in search engines like Google and Duck Duck Go. Their analysis produces 9 granular, binary labels for each site, with assigned point scores that sums to 100. Based on the sum of points the sites get an overall label for credibility - green for good score, red for bad score. Three additional overall labels exist for satire, user-produced content and sites with unfinished analysis. Table 2 describes the granular labels. NewsGuard is transparent about their methodology and publish a policy for ethics and conflicts of interest. Their full staff is listed with names online and their ratings are free.

Pew Research Center published an article entitled "Political Polarization & Media Habits" which analysed trust in specific news sources by liberals and conservatives. This analysis used 5 groups of people, ranging from liberals to conservatives, and each group provided a rating of how much they trust each source. The ratings are aggregated to show whether readers with different political leanings predominantly trust or distrust a specific source. We provide this trust label for each source and political leaning, as a label for congruency between bias a readership (rather than a fact-checking label).

Wikipedia published a list of fake news websites, which they define as sites that "*intentionally, but not necessarily solely, publish hoaxes and disinformation for purposes other than news satire*". The page has more than 500 edits, 162 cited references and has been in existence since 18/11/2016. There is no information on how the sites were selected, but for each source there are references to other sites which has reported their bad behaviour. We provide a fake-news tag for sources on the list.

Open Sources describes itself as a "*curated resource for assessing online information sources, available for public use*" and its analysis are done by its own team of experts. The criteria is published online in detail. This list has also been used in several academic studies (Horne and Adali 2017; Horne et al. 2018; Baly et al. 2018). Unfortunately, last repository commit was 2 years ago and many of the labeled sources no longer exist. The site provides a list of sources with 1-3 tags per source (See Table 2).

Media Bias/Fact Check is a platform that analyzes news

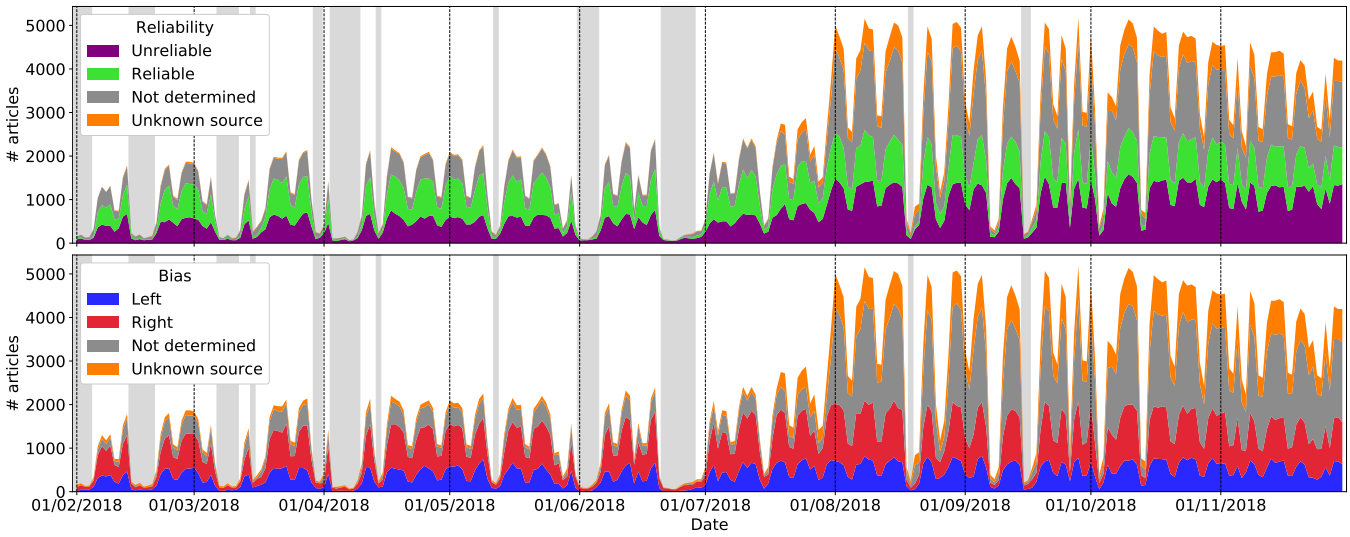

Figure 1: Number of articles in the dataset over time. For each source, we compute an aggregated reliability and bias rating, and label all articles in the source with this rating for illustration purposes. The two stack-plots contain the same datapoints, but dissected with these two distinct aggregated labels. If the aggregated label is uncertain we label the articles with gray. Grey-shaded vertical regions are marks where unusually little data were collected due to some problem with data-scraping or potentially low activity. The increase in the number of data points around the 01/08/2018 is caused by the addition of new sources to the collection.

NewsGuard	newsguardtech.com
Pew Research Center	journalism.org/2014/10/21/political-polarization-media-habits
Wikipedia	en.wikipedia.org/wiki/List_of_fake_news_websites
Open Sources	opensources.co
Media Bias/Fact Check	mediabiasfactcheck.com
Allsides	allsides.com
PolitiFact	politifact.com
BuzzFeed News	buzzfeednews.com/article/craigsilverman/inside-the-partisan-fight-for-your-news-feed
Alexa Analysis top sites	s3.amazonaws.com/alexa-static/top-1m.csv.zip

Table 1: Links for online resources.

sources to determine their credibility, as well as to "*educate the public on media bias and deceptive news practices*". The site publishes the names of its editorial team and only accepts outside information from individuals who have accepted International Fact-Checking Network's code of principles. According to its published methodology, the site numerically evaluates each news outlet in 4 categories; *biased wording/headlines*, *factual/sourcing*, *story choices* and *political affiliation*, and uses the mean of these for a final verdict. As of January 2019, we were unfortunately not able to find the numerical categories for the sources. We were able to find a *factual reporting* label, which is derived from the previously mentioned scores. Many sources also had descriptive labels, some of which were related to reliability and some of which were related to bias. All these labels are described in Table 2.

Allsides takes a very idealistic approach to assessing bias of sites and is mainly data-driven. They emphasize that news are inherently biased, that a mixed news "diet" is the true goal for newsreaders and that bias can be hidden and unconscious. This site creates data through a set of methods, each of which are noted for the sources. It conducts blind surveys on material in the public as well as in an editorial board, use third party data and assessment, conducts internal research on sources if needed, and also has a community feedback function for all bias assessments. In the community feedback, users can vote to agree or disagree with Allsides assessment of a source. They note that the community feedback is not normalized with respect to bias, and should more be used as a flag for their own use on whether their assessments are off and needs updating. We include their bias label and feedback numbers (votes agreeing and votes disagreeing) for each source. The feedback number are not shown in the paper, but can be found in the dataset.

BuzzFeed News published an article "Inside The Partisan Fight For Your News Feed" on 08/08/2017 which describes a study conducted by them on how partisan websites and Facebook pages have been created in increasing numbers. They publish an associated dataset with news sources and their political leaning (left and right), which we include.

PolitiFact is a well-known fact-checking organization which investigates claims and evaluates the truthfulness of those claims. The statements can be from any public per-

son or simply rumours that gain enough attention. PolitiFact's data is very different from the rest of our labelling sites, as their assessment is on article/statement level and not source level. They also aggregate the statements and their labels for the sources that published the statements. We have counted the types of statements coming from each source, which could be used to indicate their truthfulness. However the data is not well normalized, as some sites have many noted statements, while some have none, due to the origin of the statements and the amount of attention each source has.

Amazon's Alexa provides a ranking of nearly all websites based on frequency of visits, to which they provide free access to the top 1M. We include the position of the sources in this rating in the dataset based on our access to Alexa on 13th of January 2019. Note, this data comes from the free portion of Alexa's data, not the paid portion. Furthermore, these rankings will change over time.

4 Use Cases

There are many threads of misinformation research that this dataset can benefit. We argue that our dataset can especially benefit automated news veracity methods, which need large labelled datasets, and qualitative studies that focus on the tactics used by malicious and hyper-partisan news producers. We discuss a few examples below.

Distant Supervised Learning

Much research in news has been focused on automated methods for detecting misinformation (Kumar and Shah 2018). For machine learning systems, this analysis generally requires article-level labelling (i.e. false/bias labels of individual articles). One problem with this approach is that labelling individual articles requires a lot of resources and is often times not possible. For many machine learning algorithms the minimum requirement of labelled samples is in the thousands. Furthermore, verifying articles will commonly require considerable time from an expert. A second problem is that the verification of statements in articles can require a lot of time. This can make available labelled articles outdated for analyzing contemporary articles, due to shifts in topics and news cycle.

An alternative approach to creating labels is through distant supervision (or weak supervision), where labels are created at the source-level and used as proxies for article-level labels. One advantage of the approach is that it reduces the workload of labelling. Additionally, labels are known instantaneously for articles from known sources allowing real time update of parameters and analysis of news. This approach has been shown promising in recent misinformation detection work (Horne et al. 2018; Baly et al. 2018). The NELA-GT-2018 dataset can be used out-of-the-box for this type of machine learning study.

Semi-Supervised Learning

Another commonly debated issue in misinformation research is handling new articles from mixed-veracity (partial truths, benign or malicious) sources or handling articles from newly emerging sources during events (such as elections). One potential way to address these problems is using semi-supervised learning, in which these uncertain veracity news sources are included as unlabelled data. This approach can improve stability and increase the working domain for automated systems. In fact, it has been shown that, with some assumptions, semi-supervised approaches can improve performance over fully supervised approaches, where unlabelled samples enables classifiers to reduce risk exponentially with the number of labelled samples (Castelli and Cover 1996). Depending on the problem, this dataset provides consistent labels of 100+ sources, verified by multiple assessment sites. Remaining sources are either completely unknown, or are sparsely labelled, but can be utilized with semi-supervised methods.

Mixed-Method Studies

There are unanswered research questions about the tactics used by news producers publishing false, misleading, or propaganda news. These questions cannot be answered through machine learning studies, but rather require mix-method assessments in order to be answered. For example, recent work has focused on content sharing by alternative media sources (Starbird et al. 2018). This work sheds light on the tactics employed by state-sponsored news to create alternative narratives around an event, but can continue to be improved with data that is more complete and independent of social media. Other question include: how do false news producers change with events? Do they keep consistent ideologies? or do they adapt with the given event? Many of these potential tactics are unknown. This dataset provides news over many major events, which can be easily extracted for specific studies. For qualitative researchers, the data can provide a "head-start" on exploring the data, as the veracity of each source is known.

5 Conclusion

In this paper, we present a labelled news dataset for the study of misinformation. We argue that the research community lacks large labelled datasets for use in both mixed-method and machine learning studies. To address this need, we provide a large dataset of news articles (713K articles), collected over many sources (194), over a long period of time 02/2018-11/2018. The articles are independent of engagement from online communities, and reflect the publish patterns of the news producers. We have furthermore gathered labels for these sources from 8 different assessment sites, each of which seeks to assess the reliability and bias of sources and claims. Combined they provide a detailed and near-complete labelling of sources, which can be used for predictive analysis and qualitative studies of the news landscape.

A Appendix

Section		Description	(NewsGuard points)	Coloring
NewsGuard	1.	Does not repeatedly publish false content	(22.0)	
	2.	Gathers and presents information responsibly	(18.0)	
	3.	Regularly corrects or clarifies errors	(12.5)	
	4.	Handles the difference between news and opinion responsibly	(12.5)	
	5.	Avoids deceptive headlines	(10.0)	
	6.	Website discloses ownership and financing	(7.5)	
	7.	Clearly labels advertising	(7.5)	
	8.	Reveals who's in charge, including any possible conflicts of interest	(5.0)	
	9.	Provides information about content creators	(5.0)	
	10.	Aggregated `score` computed from 1-9		-
	11.	Column 10 thresholded at 60 points		
Pew Research Center	12.	Trust from consistently-liberals		
	13.	Trust from mostly-liberals		
	14.	Trust from mixed groups		
	15.	Trust from mostly-conservatives		
	16.	Trust from consistently-conservatives		
	17.	Aggregated trust from 12-16		
Wikipedia	18.	Existence of source on Wikipedia's list of fake news sources		
Open Sources	19.	Marked reliable		
	20.	Marked blog		
	21.	Marked clickbait		
	22.	Marked rumor		
	23.	Marked fake		
	24.	Marked unreliable		
	25.	Marked biased		
	26.	Marked conspiracy		
	27.	Marked hate speech		
	28.	Marked junk science		
	29.	Marked political		
	30.	Marked satire		
	31.	Marked state news		
Media Bias / Fact Check	32.	Factual reporting from 5 (good) down to 1 (bad)		
	33.	Special label; conspiracy, pseudoscience or questionable source (purple), and satire (orange)		
	34.	Political leaning / bias from left to right.		
Allsides	35.	Political leaning / bias		
BuzzFeed	36.	Political leaning / bias, but only left and right		
PolitiFact	37.	Has brought story labelled as "pants on Fire!"		
	38.	Has brought story labelled as false		
	39.	Has brought story labelled as mostly false		
	40.	Has brought story labelled as half-true		
	41.	Has brought story labelled as mostly true		
	42.	Has brought story labelled as true		
Alexa Ranking		The Alexa ranking of the source.		Numerical
# Articles		The number of articles collected from the source.		Numerical
First Observed		The date of first articles collected from the source.		dd-mm-yyyy

Table 2: Details of the information for sources found in Tables 3, 4 and 5. We generally use green-to-purple for good-to-poor reliability/credibility, with grey as inconclusive. For bias we use blue-to-red for left-to-right bias, with grey as unbiased. Orange is used for special cases. In NewsGuard data it represents missing information, in Open Sources it marks auxiliary labels and for Media Bias / Fact Check it marks satire.

Table 3: Labelling of first part of sources.

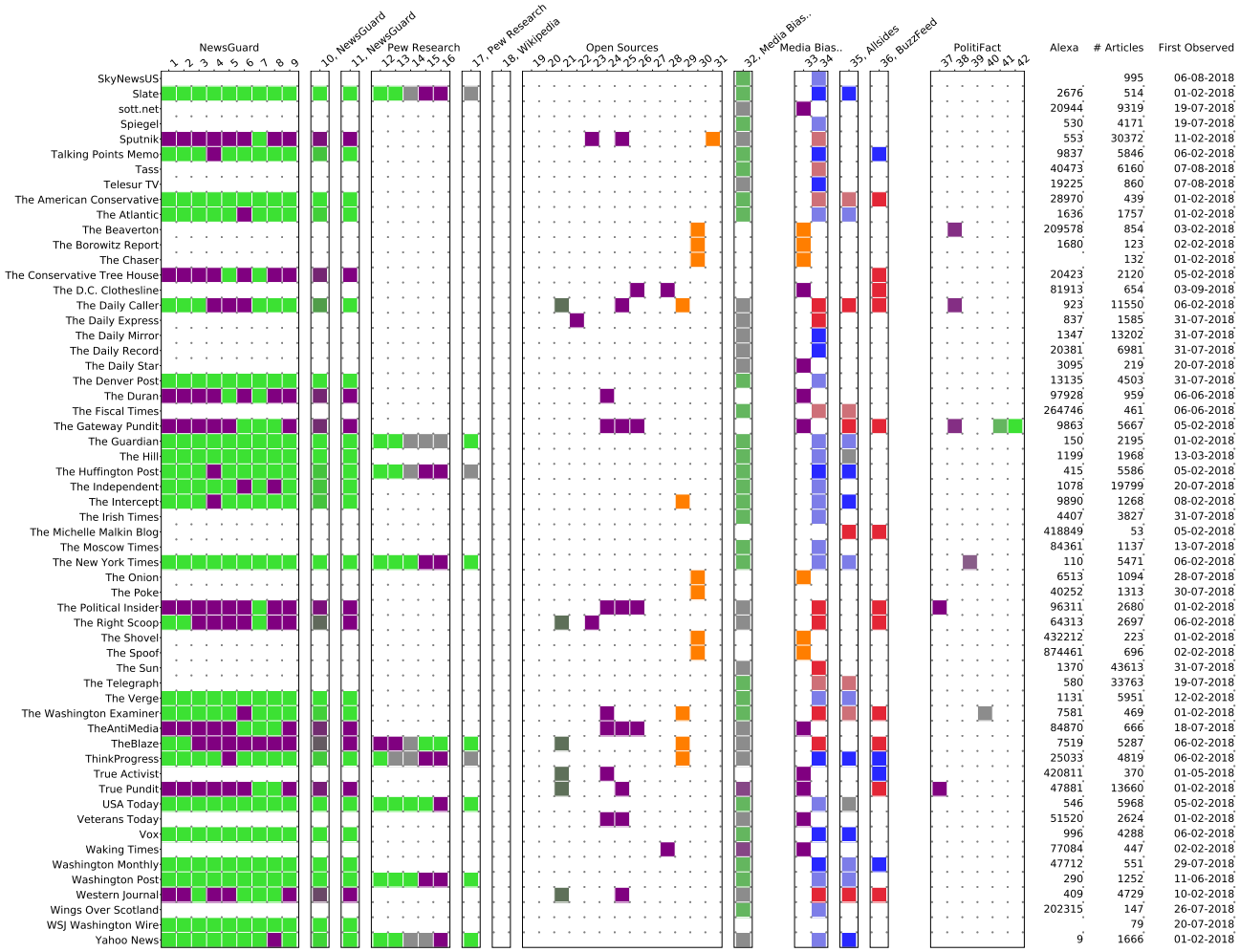

Table 4: Labelling of second part of sources.

Source	Alexa	# Articles	First Observed
Anonymous Conservative		616	09-02-2018
BBC UK		5504	30-07-2018
Channel 4 UK	2817	888	30-07-2018
Common Dreams		27	21-03-2018
Conservative Home	304146	2248	11-02-2018
Conservative Tribune		2353	06-02-2018
Crikey	827664	391	27-07-2018
Delaware Liberal		1132	09-02-2018
Dick Morris Blog	157827	400	07-02-2018
Fort Russ	75353	1090	18-07-2018
Freedom-Bunker		2229	18-07-2018
Hit and Run		3441	09-02-2018
Hullabaloo Blog	126769	958	28-07-2018
Informnapalm	281115	32	20-07-2018
JewWorldOrder		1521	19-07-2018
LabourList	221981	430	30-07-2018
Liberal Democrat Voice	206720	573	26-07-2018
Losercom		10	02-10-2018
Mail	1383	8461	19-07-2018
Mint Press News		1707	09-02-2018

Source	Alexa	# Articles	First Observed
Newsnet Scotland		35	22-07-2018
Newswars	68363	4275	13-08-2018
OSCE	136945	636	06-06-2018
Politicalite		737	30-07-2018
Politicscouk		341	01-02-2018
Prepare For Change	121860	11	28-11-2018
Slugger OToole	309300	303	26-07-2018
The Daily Blog		457	01-02-2018
The Daily Echo	55841	3329	30-07-2018
The Guardian UK		16947	20-07-2018
The Huffington Post UK	11216	5855	31-07-2018
The Inquisitr		2467	02-02-2018
The Manchester Evening News	7335	8447	31-07-2018
The Week UK	33604	2207	31-07-2018
Trump Times		86	21-09-2018
Unian	10908	3312	18-07-2018
Window on Eurasia Blog	495303	840	15-07-2018
Wizbang		58	05-08-2018
rferl	31069	2318	19-07-2018
theRussophileorg		31842	06-08-2018

Table 5: Sources with no labels found.

References

Baly, R.; Karadzhov, G.; Alexandrov, D.; Glass, J.; and Nakov, P. 2018. Predicting factuality of reporting and bias of news media sources. In *EMNLP 2018*.

Castelli, V., and Cover, T. M. 1996. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *Ieee Transactions on Information Theory* 42(6):2102–2117.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of WWW*, 675–684. ACM.

Horne, B. D., and Adali, S. 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. In *NECO Workshop 2017*.

Horne, B. D.; Dron, W.; Khedr, S.; and Adalı, S. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *WWW 2018*.

Horne, B. D.; Khedr, S.; and Adalı, S. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *ICWSM*.

Kumar, S., and Shah, N. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.

Mitra, T., and Gilbert, E. 2015. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, 258–267.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Starbird, K.; Arif, A.; Wilson, T.; Van Koevering, K.; Yefimova, K.; and Scarnecchia, D. 2018. Ecosystem or echosystem? exploring content sharing across alternative media domains.

Tacchini, E.; Ballarin, G.; Della Vedova, M. L.; Moret, S.; and de Alfaro, L. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

Wang, W. Y. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Zannettou, S.; Sirivianos, M.; Blackburn, J.; and Kourtellis, N. 2018. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *arXiv preprint arXiv:1804.03461*.

Zubiaga, A.; Liakata, M.; Procter, R.; Hoi, G. W. S.; and Tolmie, P. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one* 11(3):e0150989.