Credibility and Transparency of News Sources: Data Collection and Feature Analysis

Ahmet Aker
University of Duisburg-Essen, Duisburg, Germany and
University of Sheffield, Sheffield, England
aker@is.inf.uni-due.de

Vincentius Kevin
University of Duisburg-Essen
Duisburg, Germany
vincentius.kevin@stud.uni-due.de

Kalina Bontcheva
University of Sheffield
Sheffield, England
k.bontcheva@sheffield.ac.uk

Abstract

The ability to discern news sources based on their credibility and transparency is useful for users in making decisions about news consumption. In this paper, we release a dataset of 673 sources with credibility and transparency scores manually assigned. Upon acceptance we will make this dataset publicly available. Furthermore, we compared features which can be computed automatically and measured their correlation with credibility and transparency scores annotated by human experts. Our correlation analysis shows that there are indeed features which highly correlate with the manual judgments.

1 Introduction

The Web has never been as big as it is now. It contains tremendous amount of information represented in form of articles, videos, images, blog and social media posts and many other entries. One of the reasons for this massive growth is that it is not anymore shaped only by few experts or professional people or institutions but by everyone who has access. Although this new style of contribution towards web content has led to immense information richness and diverse views however, it has also brought new challenges. It has stripped the traditional information providers, such as news media, from their gate-keeping role [1] and has left the public in a jungle of web content with varying quality from reliable and true information to misinformation i.e., facts that are not true.

Misinformation is interchangeably used with the terms fake news. Douglas et al. refer to fake news as a "deliberate publication of fictitious information, hoaxes and propaganda" [7], and is similarly defined by others [11]. Furthermore, it is reported that the veracity of the information is highly connected to the publisher, i.e. the source of information [6, 4]. Thus instead of performing judgment on e.g. article level such as performed in [12, 8, 14] there are services to assess the sources publishing online news. NewsGuard[1] is one of such services. NewsGuard analyses manually each news publishing source in terms of credibility and transparency and provides detailed information such as references and reasoning, and the persons accountable behind each analysis. The results are made available to the public via a browser plugin.

In this paper we use NewsGuard to manually collect analyses results of 673 news sources. For each news source we manually record the overall credibility and transparency scores but also detailed information that led to those overall decisions. We plan to make

[1] www.newsguardtech.com

In: A. Aker, D. Albakour, A. Barrón-Cedeño, S. Dori-Hacohen, M. Martinez, J. Stray, S. Tippmann (eds.): Proceedings of the NewsIR'19 Workshop at SIGIR, Paris, France, 25-July-2019, published at http://ceur-ws.org

this dataset freely available.[2] Next,we collect a rich set of well known metrics/features used by e.g. search engines to assess the popularity of a web-site and run correlation analysis between the features and manually assigned NewsGuard scores. Our analysis show that there are features which highly correlate with the NewsGuard scores. This suggests that the manual process done by NewsGuard could be automated.

2 Data Collection

2.1 NewsGuard: Credibility and Transparency Scores

NewsGuard's team manually reviewed thousands of news agencies, which are mostly based in the US, to label them with nine criteria. A news agency is rewarded credibility and transparency scores for each criterion it fulfills. The criteria are listed below.

Credibility criteria:

- Does not repeatedly publish false content (22 points)
- Gathers and presents information responsibly (18 points)
- Regularly corrects or clarifies errors (12.5 points)
- Handles the difference between news and opinion responsibly (12.5 points)
- Avoids deceptive headlines (10 points)

Transparency criteria:

- Website discloses ownership and financing (7.5 points)
- Clearly labels advertising (7.5 points)
- Reveals who's in charge (5 points)
- The site provides the names of content creators, along with either contact information or biographical information (10 points)

The total of credibility and transparency scores is 100 at maximum, and a news website is considered "safe" if it has at least 60 points.

2.2 News Sources

The list of news sources we used were taken from Media Bias Fact Check (MBFC). MBFC aims to categorize sources by political bias. The categories are as follows, with some descriptions (partially) quoted from their website[3].

- Left/Right: "moderately to strongly biased towar" liberal/ conservative causes, may be untrustworthy.
- Left-Center/Right-Center: slight to moderate bias toward liberal/conservative causes.
- Center (Least Biased): minimal bias, most credible media sources.
- Pro-Science: "These sources consist of legitimate science or are evidence based through the use of credible scientific sourcing. ..."
- Conspiracy-Pseudoscience: "Sources in the Conspiracy- Pseudoscience category may publish unverifiable information that is not always supported by evidence. .."
- Questionable Sources: "extreme bias, consistent promotion of propaganda/conspiracies, poor or no sourcing to credible information, a complete lack of transparency and/or is fake news."
- Satire: "... humor, irony, exaggeration, or ridicule to expose and criticize people's stupidity or vices, ... these sources are clear that they are satire and do not attempt to deceive"
- Re-Evaluated Sources: these are sources which have been updated by MBFC. They are duplicates, so this category is removed from our analysis.

We used the sources (in total 2714) from MBFC to run over the NewsGuard (see next Section).

2.3 Collection Procedure

To collect NewsGuard judgments on the sources collected from MBFC we performed a manual process. We installed the NewsGuard as a browser plugin and visited each of the MBFC source. The results shown by the plugin were recorded. For instance for BBC.com, NewsGuard lists the results shown in Figure 1. For this source we recorded the values for the individual labels as well as the overall NewsGuard score (in this case 95). If the results are unavailable because NewsGuard has not analysed the source, the news source is discarded.

We performed this procedure for all 2714 news sources available in the nine categories at the time. NewsGuard scores were available only for 673 of them. Most of the sources in the "Satire" category were unavailable. The scores were found to agree with MBFC's description of each category - in general, least biased and pro-science sources are the most credible ones, while extremely biased and conspiracy/pseudoscience sources can be unreliable. Table 1

[2] https://github.com/ahmetaker/sourceCredibility
[3] mediabiasfactcheck.com

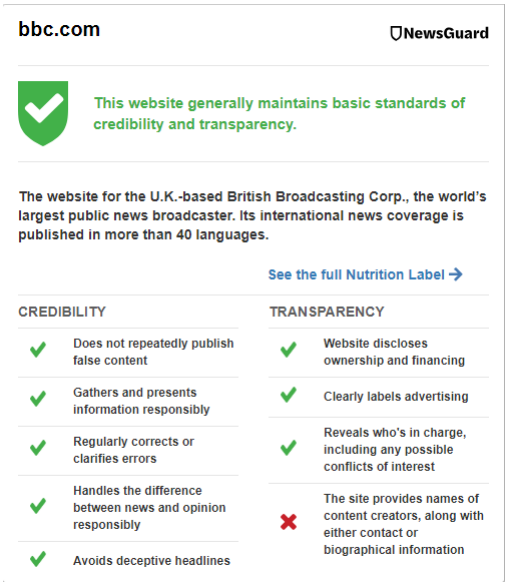

Figure 1: NewsGuard on bbc.com

shows the average score and standard deviation per category. The counts show how many sources are available on NewsGuard out of all that were listed in MBFC.

category	count	$\mu(score)$	$\sigma(score)$	cred.	tran.
Left	85 / 316	77.16	22.25	57.81	19.35
Left Center	185 / 466	94.32	8.11	72.58	21.74
Center	122 / 404	94.29	8.29	72.20	22.09
Right Center	76 / 224	92.01	15.00	70.03	21.97
Right	60 / 269	61.27	26.82	46.02	15.25
Pro Science	27 / 139	93.89	7.51	72.22	21.67
Conspiracy	39 / 287	30.09	27.76	16.88	13.21
Fake News	76 / 478	23.55	17.33	12.93	10.46
Satire*	3 / 131	5.00	4.33	0.00	5.00

Table 1: NewsGuard score per source category and the break down into credibility (max. 75) and transparency (max. 25). The count shows how many news sources are available in NewsGuard out of all sources listed in MBFC. *The satire category is not representative as it has only 3 NewsGuard scores.

3 Correlation Analysis

In the correlation analysis the automatic features are compared to the manually annotated credibility and transparency scores to analyze the correlation and predictive power of the features. We calculated specifically the correlation between each automatic feature against the combined score ($3 \times credibility + transparency$) from NewsGuard[4].

In the followings we outline features we selected as well as the metric used to perform the correlation.

[4]https://www.newsguardtech.com/ratings/rating-process-criteria/

3.1 Automatic Features

3.1.1 CheckPageRank

CheckPageRank[5] (cPR) provides a free online tool which can report page rank score, alexa rank, and a few other domain analysis results for any given website.

The tool does not provide any exact definition or information on how the scores are calculated. However, cPR provides scores which seem to be taken from non-free services such as Moz SEO and Majestic SEO tools. While these tools highly limits usage for free users to ten queries per month and a few queries per day respectively (as of 2019), cPR allows one query every thirty seconds, although it does not provide the full information available in the other tools.

Below is the most likely explanation we found for the features provided by cPR, either because the feature name is self-explanatory or the supposed underlying services give exact or very close scores compared to what is displayed by cPR.

- Google Page Rank: A score from 0 to 10 which estimates the importance of the website based on the quantity and quality of links to it from other websites.

- cPR Score: This is shown visually as one of the most important scores in checkpagerank.net, albeit without any given definition. We presume that 'cPR' simply stands for 'checkPageRank' and cPR score is calculated with a proprietary formula or algorithm.

- Citation Flow and Trust Flow: These two scores are most probably from Majestic[6], an SEO (Search Engine Optimization) tool. According to Majestic's glossary[7], citation flow focuses on the quantity and influential power of links to the website, while trust flow focuses on links from manually reviewed trusted sites. Majestic seems to have crawled over 600 billion URLs by 2014 [13].

- Topic Value: this score also most likely comes from Majestic. Majestic provides a "Topical Trust Flow" score, which, according to their glossary "shows the relative influence [...] in any given topic or category." It is a likely explanation that cPR show only the topic for which the website has the best Topical Trust Flow, since the topic names and value range are exactly the same in cPR and Majestic.

[5]checkpagerank.net
[6]majestic.com
[7]https://majestic.com/help/glossary

- Backlinks: External backlinks mean links from other websites to the subject website. This excludes internal links, which usually exist to let users navigate within the same website.
- Referring domains: this is the number of domains which contains backlink(s) to the subject website.
- EDU and GOV backlinks and domains: Majestic also provides the counts of educational and governmental backlinks and domains.
- Domain Authority and Page Authority: the Moz[8] SEO tool describes these scores as "the ranking potential in search engines based on an algorithmic combination of all link metrics". While MozRank is not used directly by search engines, it is similar and correlated to ranking of major search engines [16]. We tested a few websites and confirmed that cPR shows exactly the same scores as Moz.
- Spam Score: This most likely represents the Moz SEO spam flags explained in their website[9]. The flags represent internal and external features of websites that are indicative of 'spam websites' and have been found to be penalized or banned by Google.
- Alexa Rank: Alexa Rank is described as a popularity measure which "is calculated using a proprietary methodology that combines a site's estimated traffic and visitor engagement over the past three months."[10]
- Alexa Reach Rank: this score is based specifically on the estimated number of people each website is able to reach.
- Indexed URLs: This may be the number of URLs indexed by Google, as is commonly provided in SEO tools, but since there is no information provided, this is only a guess.

3.1.2 Twitter

- Number of followers: the number of users on twitter.com who "subscribes" to the news' Twitter account. Posts made on Twitter will appear on the followers' home screen.
- Listed count: a Twitter user can make lists of users to personally categorize other users. They can keep the list private or publicly visible. Listed count represents the number of public lists in which the Twitter user appears.

[8]moz.com

[9]https://moz.com/blog/spam-score-mozs-new-metric-to-measure-penalization-risk

[10]blog.alexa.com

3.1.3 Facebook

- Page Likes: the number of Facebook users who likes the Facebook page of the news source, by simply clicking on the like button. Likes information is publicly available.
- Page Followers: the number of Facebook users who are following the page, which means any posts by the page will be shown in the users' home screens. By default, when someone likes a page, he automatically follows the page as well. The user can then "Unfollow" while still keeping the "Like". It is also possible to follow a page without liking it.

3.2 Pearson Correlation with Logarithmic Transformation

First, we measured the Pearson correlation [3]. Pearson only measures linear relationships. This means if there is no such relationship Pearson is not a good choice to compute the correlation. However, one way of overcome this limitation is to convert the data to logarithm form. Therefore, we also applied a logarithm (base 10) on the features before calculating the Pearson correlation (with "add one" to avoid math error for the logarithm of zero) to capture the correlations which follow the power law rather than linear.

We expected features such as backlink counts and number of likes in social media to follow the power law, under the assumption that website links and user networks in social media follow the pattern of a scale-free network (preferential attachment) [2].

We also expect behavior of ranking features (e.g. Alexa Rank) to be non-linear. Although it is not necessarily logarithmic, ratio would be a better measure than rank difference. By applying a logarithm kernel, only the ratio is now considered, i.e. the difference between rank 10 and 20 is considered as significant as the difference between ranks 1,000 and 2,000.

3.3 Spearman and Kendall Tau Correlations

Since Pearson correlation only measures linear correlation, we have also computed the Spearman and Kendall Tau correlation scores. This may give a better insight on which variables are more predictive of the news source quality.

Both Spearman [15] and Kendall Tau [9] are rank-based correlation measurement, thus they work well on monotonous correlations. Spearman does not handle tied ranks, which occurs very often in our dataset due to NewsGuard's scoring method. Therefore, Kendall Tau seems to be the better measurement and has been

used to sort the rows in the following table. We have used the tau-b implementation available in scipy [11].

4 Correlation Results

Feature	pearson		spear.	kend.
	linear	log		
GOV Backlinks	*0.031*	0.698	0.656	0.499
GOV Domains	0.201	0.698	0.627	0.473
EDU Backlinks	*0.029*	0.723	0.612	0.454
EDU Domains	0.305	0.723	0.556	0.408
Trust Metric*	0.614	0.662	0.542	0.399
Trust Flow*	0.614	0.662	0.542	0.399
Indexed URLs	*0.019*	0.584	0.537	0.396
Topic Value*	0.589	0.641	0.528	0.387
Ref. Domains	0.227	0.622	0.508	0.367
Google PageRank	0.581	0.575	0.448	0.354
Citation Flow*	0.523	0.538	0.449	0.327
Domain Authority	0.603	0.588	0.445	0.325
cPR Score	0.589	0.584	0.445	0.323
Ext. Backlinks	*0.073*	0.567	0.449	0.322
Page Authority*	0.521	0.524	0.397	0.284
Global Rank	-0.338	-0.427	-0.323	-0.232
Alexa Reach	-0.327	-0.414	-0.313	-0.224
Alexa USA*	-0.379	-0.360	-0.276	-0.197
Facebook Likes	*-0.076*	-0.149	-0.229	-0.163
Twitter Listed	0.131	0.388	0.231	0.162
Twitter Followers	*0.098*	0.327	0.228	0.161
Facebook Follows	*-0.073*	-0.147	-0.225	-0.160
Spam Score	*-0.051*	*0.025*	*0.038*	*0.032*

Table 2: Feature correlation with NewsGuard score: Pearson, Spearman and Kendall tau-b coefficients.

Table 2 shows the correlation scores (Pearson, Spearman, Kendall tau) between each feature and the total score from NewsGuard. Grey values indicate statistically non-significant correlations with $p\_value \geqslant 0.00069$ (using Bonferroni correction, counting both Pearson tests as one).

As expected, applying logarithmic transformation yields big improvements on the Pearson correlation scores. There were six features which have not met our expectation in terms of whether logarithm kernel would improve the linear correlation (marked with a star), even though the differences in these cases are relatively small (< 0.05).

Many of the automatically retrievable features have a significant correlation with the NewsGuard scores. Notably, backlinks and referring domains, especially from government and educational websites, are very good indicators of trustable sources. Trust Metric and Trust Flow also work very well, confirming that seeded network graphs can be useful in practice.

[11] https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.kendalltau.html

One unexpected result is the negative correlation between Facebook likes/follows and NewsGuard scores. This may be caused by the availability of paid "like farms" to get fake likes on the platform, such as BoostLikes and SocialFormula. Even legitimate Facebook ad campaigns can result in significant amounts of such fake likes [5]. However, it requires further analysis of the corresponding Facebook pages to confirm this.

One should note that since the dataset comes from NewsGuard, it is possible for unpopular news sources to be under-represented.

5 Conclusion

In this paper, we release a dataset of 673 sources with credibility and transparency scores manually assigned. The scores come from NewsGuard's plugin. We manually accessed the plugin for 2714 news sources published by Media Bias Fact Check and recorded for those 673 detailed scores about credibility and transparency NewsGuard provides. For the remaining 2042 sources NewsGuard did not have judgments.

We also extracted a rich set of features and performed a correlation analysis. Our results show that there are strong correlations between the NewsGuard scores and features analysed in this work. This indicates that the credibility and transparency scoring could be automated.

In our future work we aim to perform such a step and create a regression model to automatically predict the credibility and transparency scores. This will allow to obtain credibility scores for any source that is so far not judged by NewsGuard. Note since our features are language independent this will allow us to obtain credibility scores for any source reporting in any language. We also plan to use the output of our regression models as information nutrition label within NewsScan[12][10].

Acknowledgements

This work was partially supported by the European Union under grant agreement No. 825297 WeVerify (http://weverify.eu) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2167, Research Training Group "User-Centred Social Media".

References

[1] Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., and Nakov, P. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765* (2018).

[12] www.news-scan.com

[2] Barabási, A.-L., and Pósfai, M. *Network science*. Cambridge University Press, Cambridge, 2016.

[3] Benesty, J., Chen, J., Huang, Y., and Cohen, I. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[4] Burgoon, J. K., and Hale, J. L. The fundamental topoi of relational communication. *Communication Monographs 51*, 3 (1984), 193–214.

[5] De Cristofaro, E., Friedman, A., Jourjon, G., Kaafar, M. A., and Shafiq, M. Z. Paying for likes?: Understanding facebook like fraud using honeypots. In *Proceedings of the 2014 Conference on Internet Measurement Conference* (New York, NY, USA, 2014), IMC '14, ACM, pp. 129–136.

[6] Demchenko, Y., Grosso, P., De Laat, C., and Membrey, P. Addressing big data issues in scientific data infrastructure. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (2013), IEEE, pp. 48–55.

[7] Douglas, K., Ang, C. S., and Deravi, F. Farewell to truth? conspiracy theories and fake news on social media. *The Psychologist* (2017).

[8] Hardalov, M., Koychev, I., and Nakov, P. In search of credible news. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications* (2016), Springer, pp. 172–180.

[9] Kendall, M. G. The treatment of ties in ranking problems. *Biometrika 33*, 3 (1945), 239–251.

[10] Kevin, V., Högden, B., Schwenger, C., Sahan, A., Madan, N., Aggarwal, P., Bangaru, A., Muradov, F., and Aker, A. Information nutrition labels: A plugin for online news evaluation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)* (Brussels, Belgium, Nov. 2018), Association for Computational Linguistics, pp. 28–33.

[11] Klein, D. O., and Wueller, J. R. Fake news: A legal perspective. *Journal of Internet Law 20*, 10 (2017), 6–13.

[12] Markowitz, D. M., and Hancock, J. T. Linguistic traces of a scientific fraud: The case of diederik stapel. *PloS one 9*, 8 (2014), e105937.

[13] Pardeep Sud, M. T. Linked title mentions: A new automated link search candidate. *Scientometrics 101* (2014), 1831–1849.

[14] Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. 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* (2017), pp. 2931–2937.

[15] Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology 15*, 1 (1904), 72–101.

[16] Themistoklis Mavridis, A. L. S. Identifying valid search engine ranking factors in a web 2.0 and web 3.0 context for building efficient seo mechanisms. *Engineering Applications of Artificial Intelligence 41* (2015), 75–91.