GREENER: Graph Neural Networks for News Media Profiling

Anonymous ACL submission

Abstract

We study the problem of profiling news media on the Web with respect to their factuality of reporting and bias. This is an important but under-studied problem related to disinformation and "fake news" detection, but it addresses the issue at a coarser granularity compared to looking at an individual article or an individual claim. This is useful as it allows to profile entire media outlets in advance. Unlike previous work, which has focused primarily on text (*e.g.*, on the text of the articles published by the target website, or on the textual description in their social media profiles or in Wikipedia), here our main focus is on modeling the similarity between media outlets based on the overlap of their audience. This is motivated by homophily considerations, *i.e.*, the tendency of people to have connections to people with similar interests, which we extend to media, hypothesizing that similar types of media would be read by similar kinds of users. In particular, we propose GREENER (GRaph nEural nEtwork for News mEdia pRofiling), a model that builds a graph of inter-media connections based on their audience overlap, and then uses graph neural networks to represent each medium. We find that such representations on their own, or when augmented with representations for articles, and from Twitter, YouTube, Facebook, and Wikipedia are quite useful for predicting the factuality and the bias of news media outlets, yielding state-of-the-art results on four datasets for the two tasks.

1 Introduction

The problem of news media profiling with respect to their factuality of reporting and political bias is important but under-studied. It is related to disinformation and "fake news" detection, but it is of different granularity compared to looking at an individual article or at an individual claim. This kind of profiling can be done by professional fact-checkers, who inspect the articles and the multimedia material published by the target news outlet.

However, doing this automatically while solely relying on text features is a very challenging task as previous work has shown (Baly et al., 2018, 2020). It gets even more complicated when considering news sources where only limited amount of content is available for evaluation. Therefore, not only is there a need to more thoroughly characterize news media, but there is also a need to be able to do so in a predictive fashion using limited information.

A crucial consideration is the need to complement the textual representation with other elements of a news medium that may serve as reliable indicators of its factuality of reporting and bias. These may relate to multimedia creation and curation processes (Jin et al., 2016; **?**), to its underlying infrastructure and technological components used to serve its content (Fairbanks et al., 2018; Castelo et al., 2019; Hounsel et al., 2020), and, more critically, to characteristics of its audience (Baly et al., 2020; Chen and Freire, 2020).

Here, we explore ways to augment the textual representations from the articles published by a target news medium by introducing new information sources that relate to media audience homophily, audience engagement, and media popularity. In particular, we propose the GREENER (GRaph nEural nEtwork for News mEdia pRofiling) model, which builds graph neural networks that model the audience overlap between websites, which we further complement with other state-of-the-art representations. Our contributions are as follows:

- We propose a novel model, based on graph neural networks that models the audience overlap between media in order to predict the factuality and the bias of entire news outlets.

- We show that the information in our graph is complementary to other information sources such as the text of the articles by the target news outlet, as well as to information from Twitter, Youtube, Facebook, and Wikipedia.

- We report sizable improvements over the state of the art on four standard datasets and for two tasks: predicting the factuality of reporting and the bias of news outlets.

- We release the code, the data, the processed features, and the representations used in our experiments (https://anonymous/).

2 Related Work

Previous work on automating the process of characterizing news sites based on the factuality of their reporting and on their political bias has mainly focused on analysis of the textual content of the respective website (Afroz et al., 2012; Rubin et al., 2015; Rashkin et al., 2017; Potthast et al., 2018; Baly et al., 2018; Pérez-Rosas et al., 2019). Although style-based analysis of the text can help reveal the intent of an article, it cannot ultimately evaluate the authenticity and the objectivity of the claims stated in that article. In fact, as demonstrated by the results in (Baly et al., 2020) on a manually fact-checked and categorized dataset, state-of-the-art textual representations can only achieve a prediction accuracy around 70% for factuality and 80% for bias. Thus, several approaches have been proposed to supplement the content-level analysis with other contextual and relational information available about the target news outlet.

Multimedia has been an important element of conveying news and information by all news media. Due to its prevalence, tampering detection and identification of processing related traces in photos and videos have long been a focus of study (Sencar and Memon, 2013). The fact that multimedia editors of a news site follow a workflow when creating, acquiring, editing, and curating content for their pages makes it possible to characterize a website based on multimedia content. Therefore, visual features are increasingly being explored and used to predict the factuality of reporting of news media (Jin et al., 2016; Huh et al., 2018; Khattar et al., 2019; Zlatkova et al., 2019; Qi et al., 2019; Singhal et al., 2019).

Beyond textual and visual features, news sites also exhibit distinct characteristics in the way they set up their infrastructure to serve content. To detect low-factuality news sites, it was proposed to use features that relate to network, web design, and data elements of the target website. At the network level, it was shown that a website's domain, certificate, and hosting properties can serve as reliable identifiers (Hounsel et al., 2020). Concerning the web design aspect, several features capturing the pattern of elements that govern the structure and the style of a web page have been also used (Castelo et al., 2019). Finally, at the data level, shared content among web sites and mutually linked sites were used to identify similar sites (Fairbanks et al., 2018). Overall, a major advantage of using infrastructure features is their content-agnostic nature.

Another set of features used to estimate the factuality and the bias of a news source is based on audience characteristics following the homophily principle, which simply states that similar individuals interact with each other at a higher rate than with dissimilar ones. In the context of social media platforms, several approaches were proposed to infer the similarity between news media through obtaining and comparing descriptive characteristics of the followers of a news medium (Baly et al., 2020) and by profiling how these followers respond to the content of the target news medium in their comments and with their posting and sharing behavior (Wong et al., 2013; Chen and Freire, 2020). In this regard, a more reliable indicator for similarity of news sites is how much the followers of different news media overlap (Darwish et al., 2020).

Ultimately, these features were all obtained from disparate data sources and are all complementary in nature. Therefore, a more accurate characterization of the news reporting practice of a given news medium can be achieved by deploying more comprehensive heterogeneous learning approaches. To this objective, in this work, we propose to use graph neural networks to model the audience homophily relations based on audience overlap and engagement statistics from Alexa. In order to provide a more holistic view, our representation is also coupled with state-of-the-art textual representations extracted from media articles, as well as on other audience characteristics proposed in the context of social media platforms.

3 Method

To characterize the similarity between news media in terms of their factuality of reporting and political bias, we mainly rely on audience overlap, which is based on the idea that if a group of visitors have a common interest in some websites, then those websites must be similar in some respect. With this idea, we create an undirected Web audience overlap graph, where nodes represent news media sites and edges indicate that that two news sites have an overlapping set of visitors, as well as the degree of overlap. The graph is created using a seed list of news sites for which factuality and bias ratings are manually annotated by professional fact-checkers. This initial graph only captures the relation between websites due to visitors that are interested in a pair of sites, and it cannot represent indirect relations where visitors might have common taste in their news consumption, but do not necessarily visit the same websites.

In order to also identify such connections between news sites, we iteratively expand the graph by adding new neighboring nodes for a more comprehensive representation of the audience overlap, which is discussed in detail in section 3.2. The graph is further enhanced by incorporating user engagement statistics as node attributes in order to model the relation between a site and its visitors better. We then use graph neural networks to encode these relations and to obtain node embeddings representing different categories of news sites. We further combine these embeddings with textual representations from articles from each news website.

3.1 Data Sources

3.1.1 Alexa Metrics

Alexa is a web traffic analysis company that produces statistics about the browsing behavior of Internet users. These statistics are computed over a rolling three-month window; they are updated daily, and are either obtained directly from sites that choose to install a tracking script on their web pages or are estimated from a sample of data generated by millions of users using browser extensions and plug-ins related to Alexa.[1] Figure 1 shows a sample Alexa page providing web traffic and domain statistics for the website wsj.com.

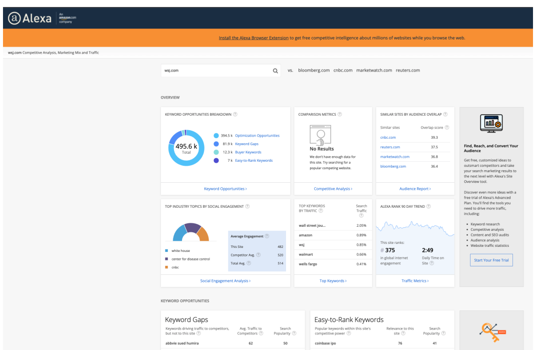

Figure 1: Alexa Rank information for *wsj.com*.

[1] www.alexa.com/find-similar-sites

We used the Alexa Audience Overlap Tool to extract statistics, which we used to build our Web audience overlap graph: links and node attributes.

Audience Overlap: This includes a list of websites most similar to the target. Alexa calculates the similarity between two websites based on shared visitors and overlap in the keywords used in their webpages. For each pair of overlapping sites, a score is computed to quantify the degree of overlap. Preliminary analysis of Alexa Rank has shown that a highly factual site, such as reuters.com, has sizable audience overlap with other factual sites. Similarly, a low-factuality website such as *infowars.com*, shares audience with other low-factulity websites. The audience homophily also holds for political bias, e.g., *foxnews.com* and *cnn.com* share audience primarily with other right- and left-leaning websites, respectively.

Figure 2 shows the overlapping websites for wsj.com, where we can see its homophily with other high-factuality websites. A similar pattern is observed for bias, where left/right-leaning websites overlap with other left/right-leaning websites.

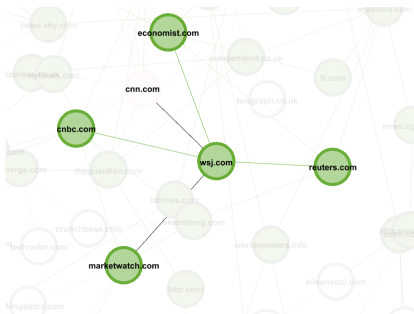

Figure 2: Audience overlap graph for *The Wall Street Journal*, showing that most of its neighboring nodes have the same factuality label: *high*.

Traffic Rank: A site's rank is a measure of its popularity, which is computed based on the number of unique users that visit it and on the total number of URL requests they made on a single day. Page views corresponding to different URL requests are counted separately only if they are 30 minutes apart from each other. We logarithmically scale this rank for a more compact representation.

Sites Linking In: This is the number of websites in the Common Crawl corpus that link to a given website. The list excludes links placed to influence search engine rankings of the linked page.

Bounce Rate: Bounce rate is an engagement statistic showing the level of interest visitors have in the content of a website. It is measured as the percentage of visits that consist of a single pageview, *i.e.*, when the visitor does not click on any of the links on the landing page.

Daily Pageviews per Visitor: This is the average number of pages viewed (or refreshed) by visitors.

Daily Time on Site: This is another engagement statistics, which shows the average time, in minutes and seconds, that a visitor spends on a target website each day. We convert it to seconds.

Binarized Alexa Metrics: Among the above-described Alexa site metrics, *Sites Linking In* produces a list of websites through analysis of web crawled data. Therefore, the completeness of the list depends on the crawling coverage. The last three metrics, (*i.e.*, daily page views, bounce rate, and daily time on site) measure the level of user engagement with the website. If users bounce at a higher rate, do not stay very long, or only view a few pages, they are likely less interested in that website. Hence, the reliability of these three metrics depends on the size of the sample of users that was used for the measurements. Due to these limitations, not all sites have such corresponding metrics calculated by AlexaRank: Table 6 shows statistics about the overall availability of these metrics for websites in the two datasets. Therefore, as a more crude measure of site popularity and engagement, we also use the binary versions of these four metrics as features showing whether Alexa was able to provide these metrics for the target website. These are given in rows 8–11 of Table 6.

3.1.2 Supplementary Sources

News Articles and Wikipedia: Previous work on the task used either GloVe (Baly et al., 2018) or fine-tuned BERT encodings (Baly et al., 2020) of the news articles, and averaged these encodings across articles by the website to obtain a textual representation for the website/domain. Similarly, GloVe and pre-trained BERT were used to get encodings for the Wikipedia descriptions of media. Thus, we also used articles and Wikipedia descriptions to obtain site-level textual representations. For the EMNLP-2018 Bias and Factuality tasks, we used the averaged GloVe encodings of the articles present on the website. For the ACL-2020 Bias and Factuality tasks, we used sentence encoders based on RoBERTa (Reimers and Gurevych, 2019) to encode the text, *i.e.*, the articles or Wikipedia descriptions. For news media without a Wikipedia page, we used a vector of zeroes. We refer to these textual representations as *Articles* and *Wikipedia*.

Audience Characteristics: In addition to modeling the similarity between news media in terms of the overlap of their audience and of quantifying the level of engagement between a medium and its followers, we also obtained an audience-centric representation for each medium, by considering the users of social media platforms that have interest in the content created by these news sources. For this purpose, we considered three features that were reported to perform well in characterization of followers of a news medium (Baly et al., 2020).

The first feature is based on how Twitter users following the account of the medium self-describe in their publicly accessible Twitter profiles. For each medium, this is obtained by encoding the biographic descriptions of 5,000 English-speaking Twitter followers, using BERT and obtaining an average representation. The second feature involves how audience of the medium's YouTube channel respond to each video in terms of the number of comments, views, likes and dislikes; by averaging these statistics over all videos, another medium-level representation is generated. The last feature includes audience estimates from Facebook's advertising platform which is used to obtain demographic information for the audience interested in each medium; this data is used to obtain the audience distribution over the political spectrum, the distribution is then divided into five categories, and each medium is labeled accordingly. These three features are hereinafter referred to as *Twitter*, *YouTube*, and *Facebook* audience representations.

3.2 Audience Overlap Graph Construction

When queried with a target news site's address, the Alexa *siteinfo*[2] tool returns a list of 4-5 sites that are most similar to the queried website based on audience overlap. For example, for `wsj.com`, we obtain the following list of similar websites and similarity scores: `marketwatch.com` 39.4, `cnbc.com` 39.4, `bloomberg.com` 35.9, `reuters.com` 34.5. We use these pairs of websites and overlap scores to build the edges of our graph, as shown in Figure 2. Given a set of websites, we repeatedly query for each website and we grow our graph by adding new nodes and edges. The resulting graph, obtained after performing this task for every site in our dataset, is referred to as level 0 audience overlap graph.

For richer and denser representations, we then

[2] http://www.alexa.com/siteinfo

expand our overlap graph to higher levels. For this, we repeat the aforementioned steps of connecting website nodes according to audience overlap for the new websites identified during building the level-0 overlap graph, which were not initially in our seed list of websites. This yields to level-1 overlap graph as displayed in Figure 3, where the distinction between low-factuality and high-factuality nodes can be clearly observed. The same procedure is repeated until obtaining level-4 graphs.

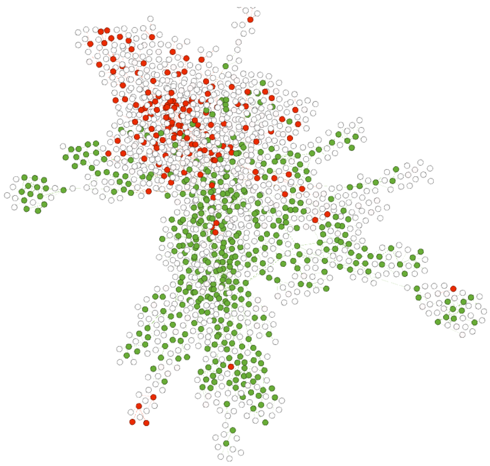

Figure 3: Bird's eye view of our overlap graph. Nodes represent news sites and colors code site factuality: red corresponds to low-factuality, green to high-factuality, and white to mixed factuality and unknown sites.

3.3 Graph Embeddings

In recent years, Graph Neural Networks (GNNs) have been extensively used to model dependencies and relations between entities and for representation learning to map graph nodes to low-dimensional dense representations. To get a representation for news source nodes in our overlap graphs, we used node2vec (Grover and Leskovec, 2016), one of the earliest GNN frameworks. The model is inspired by word2vec (Mikolov et al., 2013), but instead of using sequences of words and optimizing the proximity loss, sequences for graph are generated by sampling random walks of a fixed maximum length for each node. These sequences of random walks are then used with a skip-gram model, just as with word2vec, to learn representations for the nodes. We obtain a 512-dimensional vector representation for each (website) node in our graph; we will refer to these representations as *graph embeddings* throughout the paper.

EMNLP-2018				ACL-2020			
Political Bias		Factuality		Political Bias		Factuality	
Left	189	High	256	Left	243	High	162
Centre	564	Mixed	268	Centre	272	Mixed	249
Right	313	Low	542	Right	349	Low	453

Table 1: Label distribution for the two datasets.

4 Experiments and Evaluation

Datasets To evaluate our system, we use two datasets from previous work: (Baly et al., 2018) and (Baly et al., 2020). We will refer to them as *EMNLP-2018 dataset* and *ACL-2020 dataset*, respectively. Both datasets contain lists of media domains along with their bias and factuality labels from Media Bias/Fact Check,[3] which is an independent journalism outlet. Factuality is modeled on a three-point scale, *i.e.*, *high*, *mixed*, and *low*. Originally, political bias was modeled on a seven-point scale, but previous work has merged the fringe labels together and converted it into a three-point scale, *i.e.*, *left*, *centre*, and *right*. Table 1 shows the label distribution of the two datasets.

Experimental Setup For the EMNLP-2018 dataset, we used our graph embeddings, and GloVe (Pennington et al., 2014) representations for the articles. For the ACL-2020 dataset, we used average RoBERTa sentence representations of the articles and of the Wikipedia descriptions along with graph embeddings, as well as representations based on information from Twitter, YouTube, and Facebook from the repository of (Baly et al., 2020). In our repository we've documented every package version so everyone can replicate our results.

For comparability, we kept our experimental setup identical to the previous work, with the only change being our new representations. We used five-fold cross-validation to train and to evaluate an SVM model using different representations. We performed grid search to tune the values of the hyper-parameters of our SVM model with an RBF kernel. As the datasets for both years and for both tasks are imbalanced, we optimized macro-F1 using grid search. We evaluated our model on the remaining unseen fold, and we report both macro-F1 score and accuracy.

For studying the efficacy of our system, we compare the results of EMNLP-2018 dataset to the best previous overall models and with models using

[3]http://mediabiasfactcheck.com/

#	Model	F1	Acc.
1	Majority class baseline	22.47	50.84
	Previous work: (Baly et al., 2018)		
2	Articles (GloVe)	58.02	64.35
3	Best overall model (Articles + Twitter + Wikipedia + URL analysis + Alexa Rank)	***59.91***	***65.48***
	Our results		
4	Graph embeddings	60.60	68.19
5	Graph embeddings + AlexaMetrics	60.42	67.73
6	Graph embeddings + Articles (early fusion)	62.25	68.11
7	Graph embeddings + Articles (late fusion)	65.28	72.33
8	Graph embeddings + Articles + AlexaMetrics (late fusion)	**65.88**	**72.51**

Table 2: Factuality prediction on the EMNLP-2018 data. In lines 6–8, we use article representation from line 2.

#	Model	F1	Acc.
1	Majority class baseline	22.93	52.43
	Previous work: (Baly et al., 2020)		
2	Best "Who Read It" model	42.48	58.76
3	Articles (BERT)	61.46	67.94
4	Best overall model (Articles + Twitter + YouTube)	***67.25***	***71.52***
	Our results		
5	Graph embeddings	59.70	67.20
6	Graph embeddings + AlexaMetrics	59.55	66.01
7	Articles (RoBERTa)	61.06	66.94
8	Graph embeddings + Articles (early fusion)	65.59	70.20
9	Graph embeddings + Articles (late fusion)	62.26	67.87
10	Graphs embeddings + Articles + Twitter + YouTube + Facebook (early fusion)	64.34	69.73
11	Graphs embeddings + Articles + Twitter + YouTube + Facebook (late fusion)	68.05	72.76
12	Graphs embeddings + Articles + Twitter + YouTube + Facebook + AlexaMetrics (late fusion)	**69.67**	**73.69**

Table 3: Factuality prediction on ACL-2020 data. In lines 8–12, we use the article representation from line 7; in lines 10–12, we use the representations for Twitter, YouTube, and Facebook from the GitHub of (Baly et al., 2020).

only textual representations (which was also the best-performing single feature). As our audience overlap graph falls under the *Who Read It* category of features in (Baly et al., 2020), for the 2020 tasks, in addition to the best previous model and the best model using textual representations, we also compare to the best *Who Read It* model.

We used two strategies to combine representations: early fusion and late fusion. In early fusion, we train a single classifier using a concatenation of the representations. In late fusion, we train separate classifiers for each type of representation, and then we train an ensemble by averaging the posterior probabilities obtained by each model.

In the case of late fusion, our models learn different weights, which ensures that more attention is paid to the probabilities produced by better models.

We used Nvidia's K80 GPUs to train the graph embeddings and to obtain the RoBERTa encodings, both of which took around 30 minutes. The neural network training and inference phases were both carried out on the CPU.

Factuality Prediction Table 2 shows our results for the EMNLP-2018 Factuality Task. We can see that our graph embeddings (row 4) outperform the Articles representations (row 2) and the best result from previous work (row 3), which combines representations from several sources. When our graph representations are used together with the Articles representation, we improve the best previous result by +5.37 macro-F1 points absolute (row 7). This also confirms that graph embeddings are complementary to the textual representations. Adding Alexa Metrics (*Has Daily Pageviews per Visitor*) in our system yields an additional improvement of +0.50 macro-F1 points absolute (row 8).

Table 3 shows our results on the ACL-2020 Factuality dataset and task. Here our graph embeddings and Articles representation (rows 5-6) perform comparable to the best text representation from previous work (row 3), which used fine-tuned BERT. When our graph embeddings are used together with Articles representations (rows 7-8), we outperform previous Article representations (row

#	Model	F1	Acc.
1	Majority class baseline	22.61	51.33
	Previous work: (Baly et al., 2018)		
2	Articles (GloVe; our rerun)	61.64	68.01
3	Best overall model (Articles + Wikipedia + URL analysis + Alexa Rank)	***63.27***	***69.89***
	Our results		
4	Graph embeddings	67.64	73.55
5	Graph embeddings + AlexaMetrics	66.22	72.89
6	Graphs embeddings + Articles (early fusion)	68.52	73.55
7	Graphs embeddings + Articles (late fusion)	70.79	75.61
8	Graphs embeddings + Articles + AlexaMetrics (late fusion)	**72.18**	**76.17**

Table 4: Bias prediction on EMNLP-2018 data. In lines 6–8, we use the article representation from line 2.

#	Model	F1	Acc.
1	Majority Class	19.18	40.39
	Previous work: (Baly et al., 2020)		
2	Articles (BERT)	79.34	79.75
3	Best “Who Read it” model	65.12	66.44
4	Best overall model (Articles + Wikipedia + Twitter + YouTube)	***84.77***	***85.29***
	Our results		
5	Graph embeddings	75.70	76.95
6	Graph embeddings + AlexaMetrics	73.80	74.97
7	Articles (RoBERTa)	79.75	80.21
8	Graphs embeddings + Articles (early fusion)	78.48	79.05
9	Graphs embeddings + Articles (late fusion)	82.60	83.24
10	Graphs embeddings + Articles + Wikipedia + Twitter + YouTube (early fusion)	78.53	79.16
11	Graphs embeddings + Articles + Wikipedia + Twitter + YouTube (late fusion)	85.72	86.15
12	Graphs embeddings + Articles + Wikipedia + Twitter + YouTube + AlexaMetrics (late fusion)	**86.15**	**86.50**

Table 5: Bias Prediction on ACL-2020 data. In lines 8–12, we use the article representation from line 7; in lines 10–12, we use the representations for Wikipedia, Twitter, and YouTube from the GitHub of (Baly et al., 2020).

3). Comparing the graph embeddings with other audience characteristics (the *Who Read It* category of features), we can see that the discrimination power inherent to the audience overlap feature is much higher (by +17.22 macro-F1 points absolute) than that of the latter features. We outperform the previous best published result (row 4) when we incorporate graph embeddings and textual representations with Twitter, YouTube, and Facebook features (row 11). Finally, adding Alexa Metrics (*Has Bounce Rate* and *Has Daily Time on Site*) increases the macro-F1 score by +0.43 points absolute (row 12).

Bias Prediction Table 4 shows evaluation results on the EMNLP-2018 Bias task. Here, we observe that our graph embeddings (row 4) alone outperform the previous best overall model (row 3). Our graph embeddings when combined with Articles representations (row 7) yield a substantial increase of +7.52 macro-F1 points absolute over the best previous result. The results in rows 2, 4, and 7 demonstrate that graph embeddings and Articles representations are complementary. Finally, row 8 demonstrates that adding Alexa Metrics (*Has Daily Pageviews per Visitor* and *Has Daily Time on Site*) to the system in row 7 further improves the performance by +1.39 macro-F1 points absolute.

Table 5 shows the results for the ACL-2020 Bias task. Our RoBERTa-based Articles representations (row 7) perform better than previous Articles representations (row 2), which used fine-tuned BERT. Similarly to the ACL-2020 Factuality task, our graph embeddings (row 5) here too outperform the best result from the *Who Read It* feature group (row 3) by +10.58 macro-F1 points absolute. Then, graph embeddings, when combined with our Articles representation (row 9) perform comparably to the previous best overall result (row 4). When we also use Wikipedia, Twitter and YouTube representations (row 11), we improve the previous best result for the task (row 4). The results further improve in row 12 when adding Alexa Metrics (*Has Daily Time on Site*): an increase of +0.43 macro-F1 points absolute compared to row 11.

5 Discussion

Other Features Tested Alexa Site Info maintains a wide array of audience centric statistics for the websites. Apart from audience overlap, we also experimented with other features: *Alexa Rank*, *Total Sites Linking In*, *Daily Page Views per Visitor*, *Bounce Rate*, *Average Daily Time per Visitor*. Table 6 shows that these features performed better than the majority class baselines, they are not very strong. Note that most of these features were heavily unpopulated for a substantial part of our website dataset, which could be the reason for their mediocre performance. Regardless, site popularity and engagement metrics are potentially very useful for bias and factuality prediction. In fact, as our results show, even their binarized versions are helpful, even on top a very a strong system.

Inductive Graph Embeddings Inductive Graph Representations are a topic of great research interest in Deep Learning right now, and they have been recently used in the misinformation and disinformation domain as well (Nguyen et al., 2020). The main advantage of Inductive Graph Representations is that, in case of addition of new nodes to the graph, the resulting representations can be generated without recomputation. This saves time and computational resources for retraining these embeddings. In particular, we tried GraphSAGE (Hamilton et al., 2018) and Attri2Vec (Zhang et al., 2019) representations of nodes with Alexa Features as attributes, but due to their absence for most of the nodes, we could not achieve much improvement.

Different Levels Our preliminary experiments have shown that, as we use embeddings from higher level graphs, performance improves. Table 7 shows our results on incremental levels of graphs on the EMNLP-2018 factuality dataset. We can notice a jump of +15.40 macro-F1 points absolute when going from a level-0 to a level-4 graph. This increase in performance can be attributed to the addition of more nodes and denser connections between them in the graph, which enhances our graph embeddings. After these preliminary results, we decided to use level 4 embeddings as our overlap graph embeddings in all our experiments.

***Who Read It* vs. *What Was Written* Features** With the introduction of graph embeddings in the *Who Read It* feature category, we narrowed the gap between *What Was written* and *Who Read It* features, as reported in (Baly et al., 2020).

#	Model	% Pop.	F1	Acc.
1	Majority class baseline	–	22.47	50.84
2	Alexa Rank (reciprocal)	99.92	22.46	50.75
3	Alexa Rank (logarithm)	99.92	44.81	55.07
4	Total Sites Linking In	94.98	45.28	55.72
5	Bounce Rate	31.09	44.70	55.25
6	Average Daily Time	36.27	44.13	56.10
7	Daily Pageviews	61.08	44.93	56.85
8	Has Total Sites Linking In	94.98	23.03	50.94
9	Has Bounce Rate	31.09	42.70	59.38
10	Has Average Daily Time	36.27	42.50	59.47
11	Has Daily Pageviews	61.08	37.19	56.10
12	Combination of 3–7	–	48.14	57.50
13	Combination of 8–11	–	43.08	59.19

Table 6: Factuality prediction on the EMNLP-2018 data using different statistics from Alexa. Line 2 shows a result from (Baly et al., 2018). Line 12 combines lines 3–7, and line 13 combines lines 8–11. For missing values, we take the mean value of the feature.

Model	Nodes	Edges	F1	Acc.
Majority	–	–	22.47	50.84
level 0	1,062	4,837	45.20	57.50
level 1	4,238	20,335	55.80	64.70
level 2	11,867	57,320	56.78	65.01
level 3	30,889	149,110	57.70	66.10
level 4	78,429	377,260	60.60	68.19

Table 7: Ablation study: factuality prediction on the EMNLP-2018 data using graph embeddings from graphs of different levels of expansion.

6 Conclusion and Future Work

We studied the problem of media profiling with respect to their factuality of reporting and bias. Motivated by homophily considerations, we built a graph of inter-media connections based on the audience overlap for the target pair of news media, and then we used graph neural networks to come up with representations for each medium. We found that such representations, especially when augmented with Alexa Metrics and additional information sources from Twitter, Facebook, YouTube, and Wikipedia, are quite useful, yielding state-of-the-art results on four standard datasets for predicting the factuality and the bias of news media.

In future work, we plan to experiment with other kinds of graph neural networks. We further want to integrate additional information sources.

Ethics and Broader Impact

Data Collection and Limitations We collected the data for our graph using the Alexa Audience Overlap Tool.[4] Although obtained Alexa statistics provide an extensive view of audience overlap across media sites, it is not comprehensive as they are only limited to top-five sites for each query. Further, sites with fewer audience are likely to be more prone to measurement error, therefore inferring factuality and bias ratings of those sites is more challenging.

Biases There might be biases in our gold labels from Media Bias/Fact Check, as in some judgments for factuality and bias might be subjective. These biases, in turn, will likely be exacerbated by the supervised models trained on them (Waseem et al., 2020). This is beyond our control, as are the potential biases in pre-trained large-scale transformers such as BERT and RoBERTa, which we use in our experiments.

Intended Use and Potential Misuse Our models can enable analysis of entire news outlets, which could be of interest to fact-checkers, journalists, social media platforms, and policymakers. Yet, they could also be misused for malicious attacks like targeting specific parts of the audience with misinformation news. We, therefore, ask researchers to exercise caution.

Environmental Impact We would also like to warn that the use of large-scale Transformers requires a lot of computations and the use of GPUs/TPUs for training, which contributes to global warming (Strubell et al., 2019). This is a bit less of an issue in our case, as we do not train such models from scratch; rather, we fine-tune them on relatively small datasets. Moreover, running on a CPU for inference, once the model is fine-tuned, is perfectly feasible, and CPUs contribute much less to global warming.

References

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *2012 IEEE Symposium on Security and Privacy*, pages 461–475. IEEE.

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

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

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

Zhouhan Chen and Juliana Freire. 2020. Proactive discovery of fake news domains from real-time social media feeds. In *Companion Proceedings of the Web Conference 2020*, pages 584–592.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 141–152.

James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. 2018. Credibility assessment in the news: do we need to read. In *Proc. of the MIS2 Workshop held in conjuction with 11th Int'l Conf. on Web Search and Data Mining*, pages 799–800.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive representation learning on large graphs.

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

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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

[4] `http://alexa.com/marketing-stack/audience-overlap-tool`

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

V Pérez-Rosas, B Kleinberg, A Lefevre, and R Mihalcea. 2019. Automatic detection of fake news. Association for Computational Linguistics.

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

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

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Victoria L Rubin, Niall J Conroy, and Yimin Chen. 2015. Towards news verification: Deception detection methods for news discourse. In *Hawaii International Conference on System Sciences*, pages 5–8.

Husrev T. Sencar and Nasir Memon, editors. 2013. *Digital Image Forensics*. Springer, New York, NY.

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

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. 2020. Disembodied machine learning: On the illusion of objectivity in NLP.

Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2013. Quantifying political leaning from tweets and retweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2019. Attributed network embedding via subspace discovery.

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

Appendix

Figures 4–7 show examples of Alexa Rank Audience Overlap statistics for *reuters.com*, *foxnews.com*, *cnn.com*, and *infowars.com*. We can see that a highly factual site, such as reuters.com, has sizable audience overlap with other factual sites. Similarly, a low-factuality website such as *infowars.com*, shares audience with other low-factuality websites. The audience homophily also holds for political bias as can be seen in cases of *foxnews.com* and *cnn.com*.

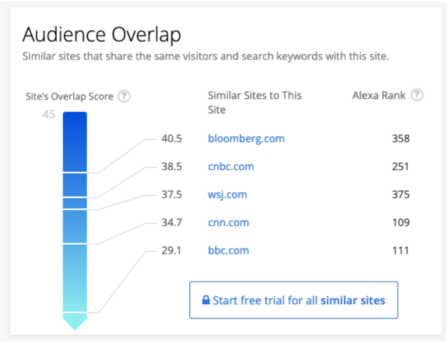

Figure 4: Alexa Rank audience overlap for *reuters.com*.

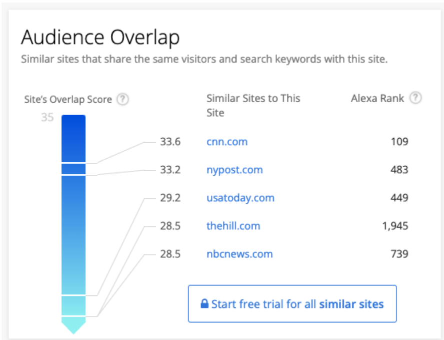

Figure 5: Alexa Rank audience overlap for *foxnews.com*.

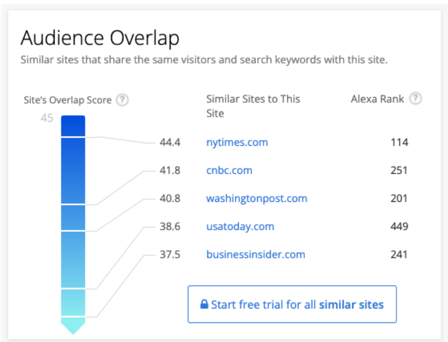

Figure 6: Alexa Rank audience overlap for *cnn.com*.

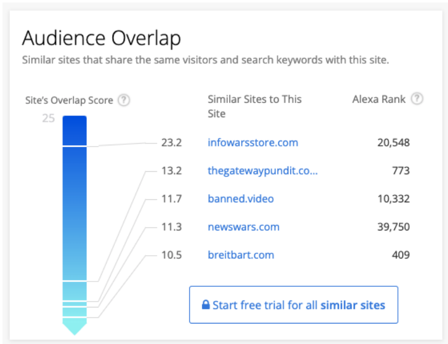

Figure 7: Alexa Rank audience overlap for *infowars.com*.