# GREENER: Graph Neural Networks for News Media Profiling

**Panayot Panayotov**
Sofia University
p.d.panayotov09@gmail.com

**Utsav Shukla**
TIET

**Husrev T. Sencar**
QCRI, HBKU
hsencar@hbku.edu.qa

**Mohamed Nabeel**
QCRI, HBKU
mnabeel@hbku.edu.qa

**Preslav Nakov**
Mohamed bin Zayed University of Artificial Intelligence
preslav.nakov@mbzuai.ac.ae

## 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 articles published by the target website, or on the textual description in their social media profiles or in Wikipedia), here we focus 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 are quite useful for predicting the factuality and the bias of news media outlets, yielding improvements over state-of-the-art results reported on two datasets. When augmented with conventionally used representations obtained from news articles, Twitter, YouTube, Facebook, and Wikipedia, we improve over previous work by 2.5-27 macro-F1 points absolute for the two tasks and datasets.

## 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; Stefanov et al., 2020). It gets even more complicated when considering news sources where only limited number of examples is available for evaluation. Therefore, not only is there a need to characterize news media more thoroughly, 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 (Jin et al., 2016; Huh et al., 2018), 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; Stefanov et al., 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 can be summarized as follows:

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



*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7470 - 7480
December 7-11, 2022 ©2022 Association for Computational Linguistics

- 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 two 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.[1]

## 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 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., 2018; Baly 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 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 65-71% for factuality and 70-85% for bias, depending on the datasets. Thus, several approaches have been proposed to supplement the content-level analysis with other contextual and relational information 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 et al., 2021). 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 factuality (Jin et al., 2016; Huh et al., 2018; Khattar et al., 2019; Zlatkova et al., 2019; Qi et al., 2019; Singhal et al., 2019; Alam et al., 2022).

[1] http://github.com/Panayot9/News-Media-Peers

Beyond textual and visual features, news web 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 the network, to the web design, and to data elements of the target website. At the network level, it has been 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 also been 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 characteristics of teh audience following the homophily principle, which simply states that similar individuals interact with each other at a higher rate than they do 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., 2016; Chen and Freire, 2020). In this regard, a more reliable indicator for similarity between news web sites is to what extent their followers overlap (Darwish et al., 2020).

Ultimately, these statistics were obtained from disparate data sources and are 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 and heterogeneous learning approaches. With this objective in mind, in the present 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 that 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 relationship 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 relationship between a site and its visitors better. We then use graph neural networks to encode these relations and to obtain node embeddings for news websites. We further combine these embeddings with textual representations from articles from each news website.

### 3.1 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 websites 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.[2] Figure 1 shows an example Alexa page providing web traffic and domain statistics for the website of WSJ.

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

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

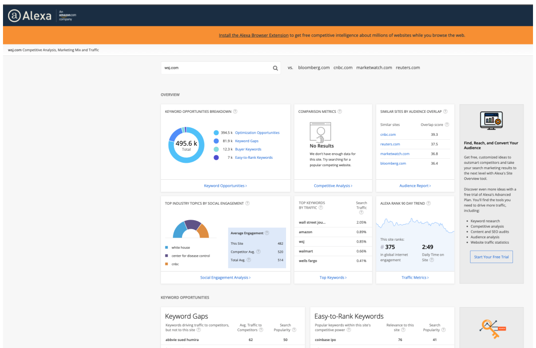

Figure 1: Alexa Rank information for *wsj.com*.

*Audience Overlap:* This includes a list of websites that are 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. Figures 2– 3 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*.

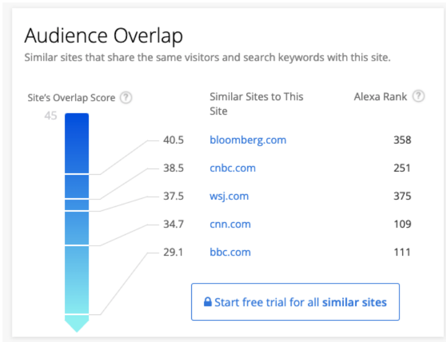


Figure 2: Alexa audience overlap for *reuters.com*.

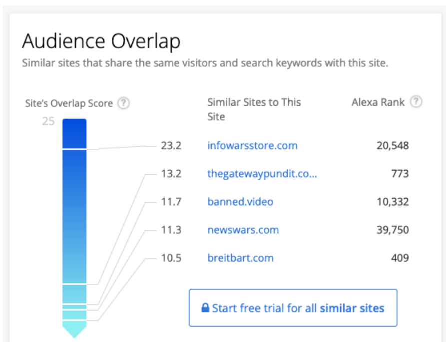


Figure 3: Alexa audience overlap for *infowars.com*.

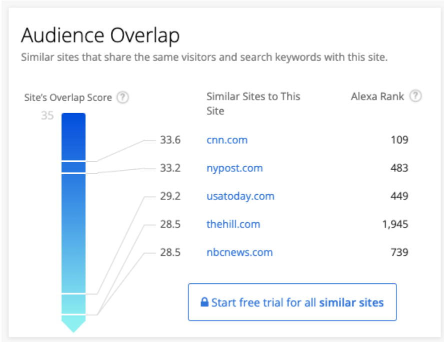


Figure 4: Alexa audience overlap for *foxnews.com*.

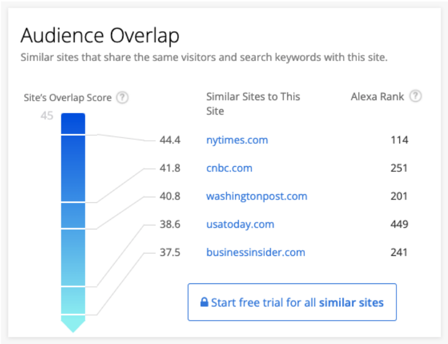


Figure 5: Alexa audience overlap for *cnn.com*.

*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 scale this rank logarithmically 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:* This is an engagement statistics 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 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.

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

*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 target website. If users bounce at a higher rate, do not stay very long, or only view a few pages, this is an indication that 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 websites have such corresponding metrics calculated by Alexa Rank. 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 Rank was able to provide these metrics for the target website. These are given in rows 8–11 of Table 6.

### 3.2 Audience Overlap Graph Construction

When queried with a target news site's address, the Alexa *siteinfo*[3] 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 6. 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.

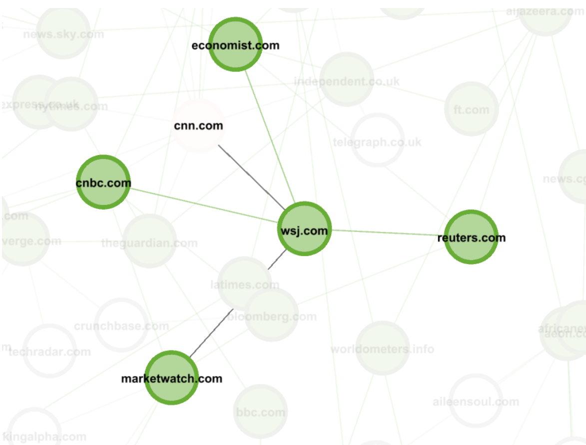


Figure 6: Audience overlap subgraph for WSJ.

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

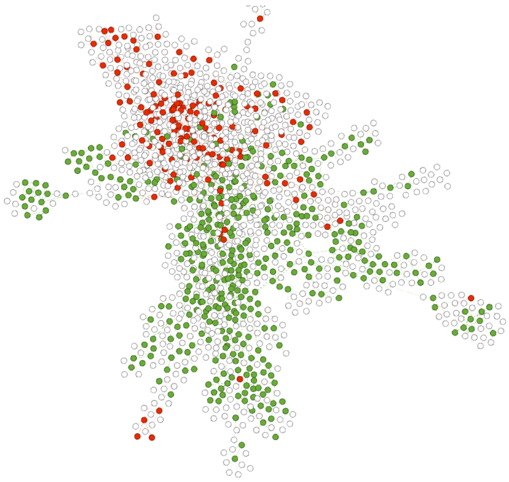

Figure 7: Bird's eye view of our overlap graph. Nodes represent news sites and colors code site factuality: red corresponds to low-factuality, green to high-factuality, and white to mixed factuality and unknown sites.

In order to obtain richer and denser representations, we then 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. Thus, we obtain the level-1 overlap graph as displayed in Figure 7, where the distinction between low-factuality and high-factuality nodes can be clearly observed. We repeat the procedure until level 4; we observed that the performance gain is marginal beyond that level, probably due to the weaker influence of the nodes towards the leaf to the labeled nodes as well as due to decreasing popularity of the domains associated with the leaf nodes.

### 3.3 Representation Learning on Graphs

In recent years, graph learning algorithms have been extensively used to model dependencies and relations between entities and to learn representations that embed graph nodes in a low-dimensional embedding space, which in turn can be used for classification. Naturally, different graph learning algorithms learn different aspects of the nodes in a graph. Thus, we experimented with various ways to obtain representations for the websites of news media in our overlap graphs: in particular, we tried random-walk shallow graph embedding methods such as Node2Vec (Grover and Leskovec, 2016), Graph Convolutional Networks (GCN) (Kipf and Welling, 2017), as well as GraphSAGE (Hamilton et al., 2017).

Node2Vec (Grover and Leskovec, 2016) is one of the earliest graph learning frameworks. The model was inspired by Word2Vec (Mikolov et al., 2013), but instead of using sequences of words and optimizing the proximity loss, it generates sequences for graphs by sampling random walks of a fixed maximum length. These sequences are then used with a skip-gram model, just as with Word2Vec, to learn representations for the nodes, treating nodes the way that words are treated in Word2Vec. While Node2Vec produces embeddings solely based on the graph structure, GCN and GraphSAGE can capture both the connectivity structure of the graph and additional node/edge attributes. In particular, these latter models perform graph convolution operations over the computation graph of each node in the graph. A key difference between GCN and GraphSAGE is how they perform that convolution operation: GCN uses spectral operations, while GraphSAGE relies on spatial operations. Moreover, GCN considers all neighboring nodes, whereas GraphSAGE is flexible to consider only a sampled subset of the neighboring nodes. These differences in the construction give rise to slightly different representations, and below we explore both.

In our experiments, we use the audience overlap graph that we described above, and we further use the Alexa site metrics as potential node features. We impute the missing features by taking the average of the five nearest neighbors. Node2vec takes the graph structure as the input and produces a node embedding for each node in an unsupervised setting. We execute GCN and GraphSAGE under a semi-supervised setting and we use the representations of the last hidden layer in the respective models as the node embeddings. Subsequently, we use the graph embeddings as features to predict the factuality of reporting and the political bias of the news websites corresponding to the nodes.

Using these three graph representation learning algorithms, we obtained low-dimensional embedding representations (512 for Node2Vec, 128 for GCN, and 128 for GraphSAGE) for each node (news website) in our graph. We empirically found that these embedding dimensions for the respective algorithms yielded the best downstream classification performance with a reasonable amount of computing resources. We will refer to these representations as *graph embeddings* throughout the rest of 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

Below, we describe the datasets, the evaluation setup, and the experiments for predicting the factuality and the bias for entire news outlets.

### 4.1 Datasets

We use two datasets from previous work to which we will refer as *EMNLP-2018* (Baly et al., 2018) and *ACL-2020* (Baly et al., 2020). Both datasets contain lists of media domains along with factuality and bias labels from Media Bias/Fact Check, where factuality is modeled on a three-point scale (*high*, *mixed*, and *low*) and so is political bias (*left*, *centre*, and *right*). Table 1 shows the label distribution for the two datasets. In order to allow for direct comparison with previous work, we experiment with the exact same data splits.[4]

### 4.2 Experimental Setup

We evaluated the predictive capability of the node-level representations obtained using the three graph learning models both individually and in combination in a supervised setting. We used five-fold cross-validation to train and to evaluate an SVM model using the node embeddings, and we performed grid search to tune the values of the hyper-parameters of our SVM model with an RBF kernel. As the datasets 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.

When combining the three representations, we adopted a late-fusion strategy. To this end, we trained separate classifiers for each type of representation, and then we trained an ensemble by averaging the posterior probabilities obtained by each model. This allows the ensemble model to learn different weights, thereby ensuring that more attention is paid to the probabilities produced by better models.

[4]https://github.com/ramybaly/News-Media-Reliability

To evaluate the complementary nature of the audience homophily and of the characteristics of the websites themselves, we combined these two sources of information. For the latter, we used the precomputed representations of the news articles and of the Wikipedia descriptions associated with each news medium as well as the Twitter, YouTube, and Facebook audience representations available in the repository of (Baly et al., 2020); see their paper for more detail.

In order to study the efficacy of our models, we further compare the results on the EMNLP-2018 dataset to the best previous overall models and to models using only textual representations, which was the best-performing single feature and included GloVe (Pennington et al., 2014) representations for the articles. 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 (based on average RoBERTa sentence representations), we also compare to the best *Who Read It* model.

We used NVidia's K80 GPUs to train the graph embeddings, which took around 30 minutes, and we performed inference on the CPU. In our repository, we documented every package version for easy replication of our results.

As the datasets have no dedicated validation set, we used five-fold cross-validation on their training partitions to select the values of the hyper-parameters. Eventually, we selected the following values of the hyper-parameters for our models: number of epochs: 1000, number of layers: 4, learning rate: 0.01, weight decay: 0.0005, and dropout: 0.5. For Node2Vec, we further selected the following hyper-parameter values: number of walks: 10, walk length: 100, number of dimensions: 512, return parameter ($p$): 0.5, and in-out parameter ($q$): 2.

### 4.3 Factuality Prediction

Table 2 shows our results for the factuality prediction task on the EMNLP-2018 dataset. In the table, each group of embeddings is referred to by the name of the graph learning algorithm used to generate it. We can see that all three types of graph embeddings (rows 4–6) outperform the Articles representations (row 2) and the best result from previous work (row 3), which combines representations from several information sources.

| # | 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 | Node2Vec | 60.60 | 68.19 |
| 5 | GCN | 72.23 | 75.94 |
| 6 | Supervised GraphSage | 86.04 | 87.55 |
| 7 | Node2Vec+ Supervised GraphSage + GCN (late fusion) | 86.97 | 88.49 |
| 8 | Node2Vec + Supervised GraphSage + GCN + Articles + AlexaMetrics (late fusion) | **87.20** | **88.58** |

Table 2: Factuality prediction on the EMNLP-2018 dataset.

| # | 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 | Node2Vec | 59.70 | 67.20 |
| 6 | GCN | 53.76 | 61.47 |
| 7 | Supervised GraphSage | 56.22 | 63.45 |
| 8 | Node2Vec + Supervised GraphSage + GCN (late fusion) | 63.48 | 69.27 |
| 9 | Node2Vec + Supervised GraphSage + GCN + Articles + Twitter + YouTube + Facebook + AlexaMetrics (late fusion) | **69.61** | **74.27** |

Table 3: Factuality prediction on the ACL-2020 dataset.

As expected, the combination of graph embeddings performs better, improving the macro-F1 score by more than 17 points absolute (row 7). Incorporating our graph representations with a subset of the features yields the best performance (row 8). This provided an additional improvement of +0.23 macro-F1 points absolute.

Table 3 shows our results on the ACL-2020 dataset for the factuality prediction task. Here, our graph embeddings (rows 5–7) perform comparably to the best text representation from previous work (row 3), i.e., the Articles representation obtained using fine-tuned BERT. Comparing the graph embeddings to other audience characteristics (the *Who Read It* category of features), we can see that the discriminative power inherent to audience overlap is much higher, by around 14–17 macro-F1 points absolute. Unlike the EMNLP-2018 dataset, however, none of the graph embeddings outperforms the best model that combines different versions of textual representations (row 4). Moreover, we can see that Node2Vec representations yield more accurate predictions than GCN and GraphSAGE representations on this dataset.

When graph embeddings are combined with the representations for Articles and descriptive characteristics of Twitter, YouTube, and Facebook audiences, the results outperform the best previous result by a margin of +2.36 macro-F1 points absolute (row 9). This result also suggests that graph embeddings are complementary to the textual representations.

### 4.4 Bias Prediction

Table 4 shows the evaluation results for bias detection on the EMNLP-2018 dataset. We observe that among the three graph embeddings (rows 4–7), only Node2Vec outperforms the previous best overall model (row 3). The ensemble classifier's accuracy (row 7) is expectedly very similar to that of the top-performing classifier. Although the graph embeddings do not yield superior performance, their combination with textual and other audience features yields a substantial increase of +9.17 macro-F1 points absolute over the best previous result (row 8). This further confirms the complementarity of audience homophily and textual representations for the bias detection task.

| # | 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 | Node2Vec | 67.64 | 73.55 |
| 5 | GCN | 52.62 | 60.28 |
| 6 | Supervised GraphSage | 52.18 | 64.81 |
| 7 | Node2Vec + Supervised GraphSage + GCN (late fusion) | 65.97 | 73.20 |
| 8 | Node2Vec + GCN + Supervised GraphSage + Articles + AlexaMetrics (late fusion) | **72.44** | **76.98** |

Table 4: Bias prediction on the EMNLP-2018 dataset.

| # | 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 | Node2Vec | 75.70 | 76.95 |
| 6 | GCN | 77.81 | 78.81 |
| 7 | Supervised GraphSage | 88.50 | 88.59 |
| 8 | Node2Vec + GCN + Supervised GraphSage (late fusion) | 89.59 | 89.76 |
| 9 | Node2Vec + GCN + Supervised GraphSage + Articles + Wikipedia + Twitter + YouTube + AlexaMetrics (late fusion) | **91.93** | **92.08** |

Table 5: Bias prediction on the ACL-2020 dataset.

Table 5 shows the corresponding results for the ACL-2020 dataset. In this setting, GraphSage embeddings (row 7) yield much better predictions than the other embeddings (rows 5–6) and the previous best overall model (row 4). The ensemble system is also able to leverage the strengths of the three types of graph embeddings to yield the best performance (row 8). When the graph embeddings capturing audience homophily are combined with other representations (row 9), the improvement is further enhanced by an overall increase of +7.16 macro-F1 points absolute over the previous best result.

## 5 Discussion

**Other Alexa Features** Alexa Site Info maintains a wide array of audience-centric statistics about the websites. Apart from audience overlap, we also experimented with the following features: *Alexa Rank*, *Total Sites Linking In*, *Daily Page Views per Visitor*, *Bounce Rate*, *Average Daily Time per Visitor*. Table 6 shows that, even though these features perform better than the majority class baseline, none of them is particularly strong.

Note that most of these features were heavily unpopulated for a substantial part of the websites in our dataset, which could explain 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 strong system.

**Varying Predictive Power of Graph Learning Methods** Our results show that none of the three graph learning approaches performs consistently better than the rest. Most surprisingly, we observe that Node2Vec yields better results than GCN and GraphSAGE in two settings. We believe an important factor contributing to this result is the sparseness of the node features. As can be seen in Table 6, among all news media websites comprising our audience overlap graph, just three Alexa metrics were available for more than 94% of the websites, and four metrics were available for less than 40% of the websites. Since GNNs' strength stems primarily from their ability to combine graph structure and node information, the missing features likely curtail their performance significantly.

| # | 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 dataset 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 the missing values, we take the mean 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 Node2Vec graph embeddings from graphs of different levels of expansion.

In fact, our earlier experiments performed without imputing the missing features yielded much worse results. Thus, it is plausible to assume that the performance of GNNs will improve in the presence of more discriminant node features.

Moreover, our analysis reveals that features like *Sites Linking In*, *Alexa Rank* and *Daily Time on Site* are more important than *Bounce Rate* and *Daily Pageviews per Visitor* for both tasks. However, there is a slight variation in the order of importance for these tasks. For example, *Alexa Rank* was the most important feature for factuality prediction, whereas *Sites Linking In* was the most important feature for bias prediction. Combining the features with the graph structure helped to improve the performance for both tasks.

**Different Levels in the Graph** Our preliminary experiments showed that, as we use embeddings from higher-level graphs, the performance improves. Table 7 shows our results on graphs of incremental levels for the EMNLP-2018 factuality dataset. We can observe a jump of +15.40 macro-F1 points absolute when going from a level-0 to a level-4 graph. This improvement in the performance can be attributed to the addition of more nodes and of denser connections between them in the graph, which enhances our graph embeddings. Based on 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 down the gap between *What Was written* and *Who Read It* features, as reported in (Baly et al., 2020).

**Alternatives to Alexa Siteinfo** The Alexa Siteinfo service was discontinued in May 2022. While we used the service to obtain information about the audience overlap, we believe that our approach is generic in that it is possible to use alternative SEO data sources such as Ahref, Semrush, Similarweb, or Moz to obtain similar information to construct the audience overlap graph and to extract features for the websites under consideration. As different SEO sources have different coverage of websites, it is also possible to combine multiple such sources to not only address the missing features, but also to increase the number of websites that can be covered, which could lead to the discovery of additional biased or low-factuality websites.

## 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 news 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 sizable improvemets over the state of the art on two standard datasets.

In future work, we plan to experiment with other kinds of graph neural networks. We further want to integrate additional information sources, and to model factuality and bias jointly (Baly et al., 2019).

## Limitations

One limitation is that our work relies on the Alexa website ranking and traffic information to build the input graphs, which is now discontinued. However, we envision that it would be possible to use alternative tools such as Semrush, Ahrefs, and SimilarWeb to build the graph.

Another limitations is that our work excludes isolated nodes (websites) in the constructed graph. Such isolated nodes could occur when a website is either relative new or is not profiled in Alexa due to insufficient traffic. The datasets used in this work have only one such isolated website, but we suggest using non-graph information, as in prior approaches, to classify such websites.

## Ethics and Broader Impact

**Data Collection** We collected the data for our graph using the Alexa Audience Overlap Tool. Although the obtained statistics provide an extensive view of audience overlap across media sites, they are not comprehensive as they are only limited to top-five sites for an input website. Moreover, sites with smaller audiences are likely to be more prone to measurement errors, and thus inferring the factuality and the bias of such websites is more challenging.

**Biases** There might be biases in our gold labels from Media Bias/Fact Check, as some judgments for factuality and bias might be subjective. These biases, in turn, will likely be exacerbated by the supervised models trained on them. This is beyond our control, as are the potential biases in pretrained 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).

## References


Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP '12, pages 461–475, San Francisco, CA, USA.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, COLING '22, pages 6625–6643, Gyeongju, Republic of Korea.

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 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 3364–3374.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 2109–2116, Minneapolis, MN, USA.

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*, WWW '19, pages 975–980, San Fracisco, CA, USA.

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*, WWW '20, 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 of *ICWSM '20*, pages 141–152.

James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. 2018. Credibility assessment in the news: do we need to read. In *Proceedings of the MIS2 Workshop held in conjunction with the 11th International Conference on Web Search and Data Mining*, pages 799–800, Los Angeles, CA, USA.

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*, KDD '16, pages 855–864, San Francisco, CA, USA.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, NIPS '17, pages 1024–1034, Long Beach, CA, USA.

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

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the 15th European Conference on Computer Vision*, volume 11215 of *ECCV '18*, pages 106–124, Munich, Germany.

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.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, pages 2915–2921, San Francisco, CA, USA.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*, ICLR '13, Scottsdale, AZ, USA.

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 '14, pages 1532–1543, Doha, Qatar.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, COLING '18, pages 3391–3401, Santa Fe, NM, USA.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 231–240, Melbourne, Australia.

Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *Proceedings of the 2019 IEEE International Conference on Data Mining*, ICDM '19, pages 518–527.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2931–2937, Copenhagen, Denmark.

Victoria L Rubin, Niall J Conroy, and Yimin Chen. 2015. Towards news verification: Deception detection methods for news discourse. In *Proceedings of the Hawaii International Conference on System Sciences*, HICSS '15, pages 5–8, Kawai, HI, USA.

Husrev T. Sencar, Luisa Verdoliva, and Nasir Memon, editors. 2021. *Multimedia 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 *Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data*, BigMM '19, pages 39–47, Singapore.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 527–537.

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*, ACL '19, pages 3645–3650, Florence, Italy.

Felix Ming Fai Wong, Chee-Wei Tan, Soumya Sen, and Mung Chiang. 2016. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Trans. Knowl. Data Eng.*, 28(8):2158–2172.

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 '19, pages 2099–2108, Hong Kong, China.