

What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context

Ramy Baly¹, Georgi Karadzhov², Jisun An³, Haewoon Kwak³, Yoan Dinkov⁴,
Ahmed Ali³, James Glass¹, Preslav Nakov³

¹MIT Computer Science and Artificial Intelligence Laboratory

²University of Cambridge, ³Qatar Computing Research Institute, HBKU, ⁴Sofia University
{baly, glass}@mit.edu, georgi.karadzhov@cl.cam.ac.uk
{jan, hkwak, amali, pnakov}@hbku.edu.qa, jdinkov@uni-sofia.bg

Abstract

Predicting the political bias and the factuality of reporting of entire news outlets are critical elements of media profiling, which is an understudied but an increasingly important research direction. The present level of proliferation of fake, biased, and propagandistic content online, has made it impossible to fact-check every single suspicious claim, either manually or automatically. Alternatively, we can profile entire news outlets and look for those that are likely to publish fake or biased content. This approach makes it possible to detect likely “fake news” the moment they are published, by simply checking the reliability of their source.

From a practical perspective, political bias and factuality of reporting have a linguistic aspect but also a social context. Here, we study the impact of both, namely (i) *what was written* (i.e., what was published by the target medium, and how it describes itself on Twitter) vs. (ii) *who read it* (i.e., analyzing the readers of the target medium on Facebook, Twitter, and YouTube). We further study (iii) *what was written about the target medium* on Wikipedia. The evaluation results show that *what was written* matters most, and that putting all information sources together yields huge improvements over the current state-of-the-art.

1 Introduction

The rise of the Web has made it possible for anybody to create a website or a blog and to become a *news medium*. Undoubtedly, this was a hugely positive development as it elevated freedom of expression to a whole new level, thus allowing anybody to make their voice heard online. With the subsequent rise of social media, anybody could potentially reach out to a vast audience, something that until recently was only possible for major news outlets.

One of the consequences was a *trust crisis*: with traditional news media stripped off their gate-keeping role, the society was left unprotected against potential manipulation. The issue became a general concern in 2016, a year marked by micro-targeted online disinformation and misinformation at an unprecedented scale, primarily in connection to Brexit and the US Presidential campaign. These developments gave rise to the term “fake news”, which can be defined as “false, often sensational, information disseminated under the guise of news reporting.”¹ It was declared Word of the Year 2016 by Macquarie Dictionary and of Year 2017 by the Collins English Dictionary.

In an attempt to solve the trust problem, several initiatives such as Politifact, Snopes, FactCheck, and Full Fact, have been launched to fact-check suspicious claims manually. However, given the scale of the proliferation of false information online, it became clear that it was unfeasible to fact-check every single suspicious claim, even when this was done automatically, not only due to computational challenges but also due to timing. In order to fact-check a claim, be it manually or automatically, one often needs to verify the stance of mainstream media concerning that claim and the reaction of users on social media. Accumulating this kind of evidence takes time, but time flies very fast, and any delay means more potential sharing of the malicious content on social media. A study has shown that for some very viral claims, more than 50% of the sharing happens within the first ten minutes after posting the micro-post on social media (Zaman et al., 2014), and thus timing is of utmost importance. Moreover, an extensive recent study has found that “fake news” spreads six times faster and reaches much farther than real news (Vosoughi et al., 2018).

¹www.collinsdictionary.com/dictionary/english/fake-news

A much more promising alternative is to focus on the source and to profile the medium that initially published the news article. The idea is that media that have published fake or biased content in the past are more likely to do so in the future. Thus, profiling media in advance makes it possible to detect likely “fake news” the moment it is published, by simply checking the reliability of its source.

From a practical perspective, political bias and factuality of reporting have not only a linguistic aspect but also a social context. Here, we study the impact of both, namely (i) *what was written* (the text of the articles published by the target medium, the text and the audio signal in the videos of its YouTube channel, as well as how the medium self-describes itself on Twitter) vs. (ii) *who read it* (by analyzing the media readers in Facebook, Twitter, and YouTube). We further study (iii) *what was written about the target medium* on Wikipedia.

Our contributions can be summarized as follows:

- We model the leading political ideology (left, center or right bias) and the factuality of reporting (high, mixed, or low) of news media by modeling the textual content of what they publish vs. who reads it in social media (Twitter, Facebook, and YouTube). The latter is novel for these tasks.
- We combine a variety of information sources about the target medium, many of which have not been explored for our tasks, e.g., YouTube video channels, political bias estimates of their Facebook audience, and information from the profiles of the media followers on Twitter.
- We use features from different data modalities: text, metadata, and speech. The latter two are novel for these tasks.
- We achieve sizeable improvements over the current state-of-the-art for both tasks.
- We propose various ensembles to combine the different types of features, achieving further improvements, especially for bias detection.
- We release the data, the features, and the code necessary to replicate our results.

In the rest of this paper, we discuss some related work, followed by a description of our system’s architecture and the information sources we use. Then, we present the dataset, the experimental setup, and the evaluation results. Finally, we conclude with possible directions for future work.

2 Related Work

While leveraging social information and temporal structure to predict the factuality of reporting of a news medium is not new (Canini et al., 2011; Castillo et al., 2011; Ma et al., 2015, 2016; Zubiaga et al., 2016), modeling this at the medium level is a mostly unexplored problem. A popular approach to predict the factuality of a medium is to check the general stance of that medium concerning already fact-checked claims (Mukherjee and Weikum, 2015; Popat et al., 2017, 2018). Therefore, stance detection became an essential component in fact-checking systems (Baly et al., 2018b).

In political science, media profiling is essential for understanding media choice (Iyengar and Hahn, 2009), voting behavior (DellaVigna and Kaplan, 2007), and polarization (Graber and Dunaway, 2017). The outlet-level bias is measured as a similarity of the language used in news media to political speeches of congressional Republicans or Democrats, also used to measure media slant (Gentzkow and Shapiro, 2006). Article-level bias was also measured via crowd-sourcing (Budak et al., 2016). Nevertheless, public awareness of media bias is limited (Elejalde et al., 2018).

Political bias was traditionally used as a feature for fact verification (Horne et al., 2018b). In terms of modeling, Horne et al. (2018a) focused on predicting whether an article is biased or not. Political bias prediction was explored by Potthast et al. (2018) and Saleh et al. (2019), where news articles were modeled as left vs. right, or as hyperpartisan vs. mainstream. Similarly, Kulkarni et al. (2018) explored the left vs. right bias at the article level, modeling both textual and URL contents of articles.

In our earlier research (Baly et al., 2018a), we analyzed both the political bias and the factuality of news media. We extracted features from several sources of information, including articles published by each medium, what is said about it on Wikipedia, metadata from its Twitter profile, in addition to some web features (URL structure and traffic information). The experiments on the Media Bias/Fact Check (MBFC) dataset showed that combining features from these different sources of information was beneficial for the final classification. Here, we expand this work by extracting new features from the existing sources of information, as well as by introducing new sources, mostly related to the social media context, thus achieving sizable improvements on the same dataset.

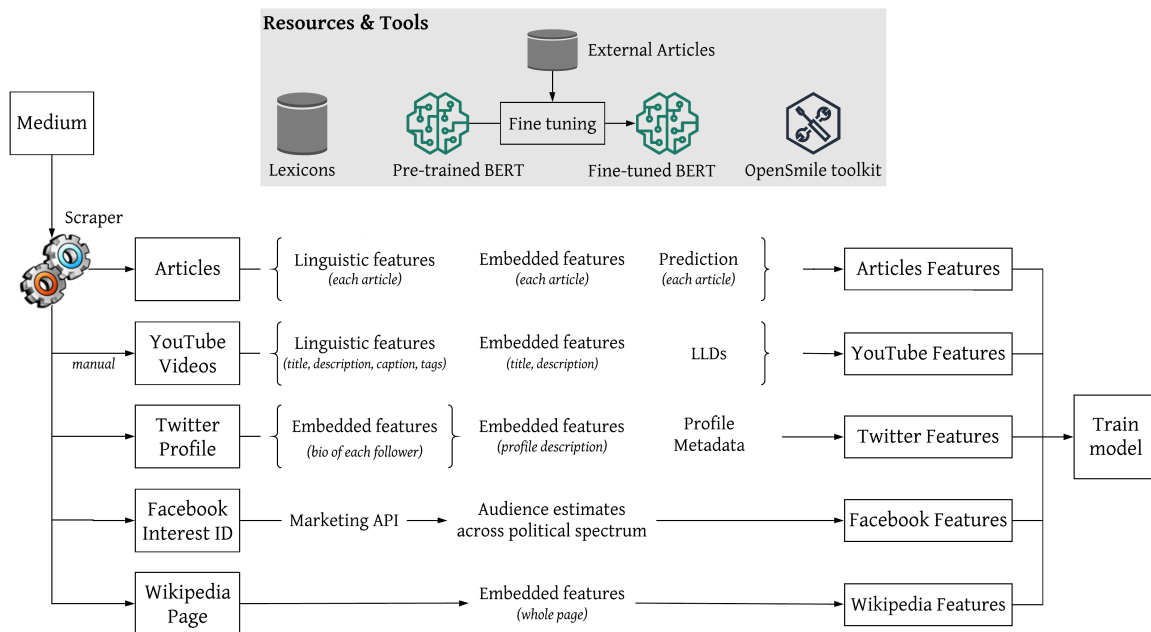


Figure 1: The architecture of our system for predicting the political bias and the factuality of reporting of news media. The features inside {curly brackets} are calculated at a finer level of granularity and are then aggregated at the medium level. The upper gray box shows the resources used to generate features, e.g., the OpenSmile toolkit is used to extract low-level descriptors (LLD) from YouTube videos; see Section 3 for further details.

In follow-up work (Baly et al., 2019), we showed that jointly predicting the political bias and the factuality is beneficial, compared to predicting each of them independently. We used the same sources of information as in (Baly et al., 2018a), but the results were slightly lower. While here we focus on analyzing political bias and factuality separately, future work may analyze how the newly proposed features and sources affect the joint prediction.

3 System and Features

In this section, we present our system. For each target medium, it extracts a variety of features to model (i) what was written by the medium, (ii) the audience of the medium on social media, and (iii) what was written about the medium in Wikipedia. This results in multi-modal (text, speech, and metadata) feature set, which we use to train a classifier to predict the political bias and the factuality of reporting of news media. Figure 1 illustrates the system architecture.

3.1 What Was Written

We describe the features that we used to model the content generated by the news media, analyzing both the articles they publish on their website as well as relevant activity on social media.

3.1.1 Articles on the News Medium Website

Given a target news medium, we first collect a number of articles it has published. Then, we extract various types of features from the text of these articles. Below we describe these features in more detail.

Linguistic Features: These features focus on language use, and they model text structure, topic, sentiment, subjectivity, complexity, bias, and morality. They have proved useful for detecting fake articles, as well as for predicting the political bias and the factuality of reporting of news media (Horne et al., 2018b; Baly et al., 2018a). We extracted such features using the News Landscape (NELA) toolkit (Horne et al., 2018b), and we will refer to them as the NELA features in the rest of this paper. We averaged the NELA features for the individual articles in order to obtain a NELA representation for a news medium. Using arithmetic averaging is a good idea as it captures the general trend of articles in a medium, while limiting the impact of outliers. For instance, if a medium is known to align with left-wing ideology, this should not change if it published a few articles that align with right-wing ideology. We use this method to aggregate all features that we collected at a level of granularity that is finer than the medium-level.

Embedding Features: We encoded each article using BERT (Devlin et al., 2019) by feeding the first 510 WordPieces² from the article³ and then averaging the word representations extracted from the second-to-last layer.⁴ In order to obtain representations that are relevant to our tasks, we fine-tuned BERT by training a softmax layer on top of the [CLS] output vector to predict the label (bias or factuality) of news articles that are scrapped from an external list of media to avoid overfitting. The articles’ labels are assumed to be the same as those of the media in which they are published (a form of distant supervision). This is common practice in tasks such as “fake news” detection, where it is difficult to manually annotate large-scale datasets (Nørregaard et al., 2019). We averaged the BERT representations across the articles in order to aggregate them at the medium level.

Aggregated Probabilities: We represent each article by a C -dimensional vector that corresponds to its posterior probabilities of belonging to each class $c_i, i \in \{1, \dots, C\}$ of the given task, whether it is predicting the political bias or the factuality of the target news medium. These probabilities are produced by training a softmax layer on top of the [CLS] token in the above-mentioned fine-tuned BERT model. We averaged the probability representations across the articles in order to aggregate them at the medium level.

3.1.2 YouTube Video Channels

Some news media post their video content on YouTube. Thus, we use YouTube channels by modeling their textual and acoustic contents to predict the political bias and the factuality of reporting of the target news medium. This source of information is relatively underexplored, but it has demonstrated potential for modeling bias (Dinkov et al., 2019) and factuality (Kopev et al., 2019).

Due to the lack of viable methods for automatic channel retrieval, we manually looked up the YouTube channel for each medium. For each channel marked as English, we crawled 25 videos (on average) with at least 15 seconds of speech content. Then, we processed the speech segments from each video into 15-second episodes by mapping the duration timeline to the subtitle timestamps.

²There is a limit of maximum of 512 input tokens, and we had to leave space for the special tokens [CLS] and [SEP].

³This is recommended in (Adhikari et al., 2019) when encoding full documents using Transformer-based models.

⁴This is common practice, since the last layer may be biased towards the pre-training objectives of BERT.

We used the OpenSMILE toolkit (Eyben et al., 2010) to extract low-level descriptors (LLDs) from these speech episodes, including frame-based features (e.g., energy), fundamental frequency, and Mel-frequency cepstral coefficients (MFCC). This set of features proved to be useful in the Interspeech Computational Paralinguistics challenge of emotion detection (Schuller et al., 2009). To complement the acoustic information, we retrieved additional textual data such as descriptions, titles, tags, and captions. This information is encoded using a pre-trained BERT model. Furthermore, we extracted the NELA features from the titles and from the descriptions. Finally, we averaged the textual and the acoustic features across the videos to aggregate them at the medium level.

3.1.3 Media Profiles in Twitter

We model how news media portray themselves to their audience by extracting features from their Media Twitter profiles.

In our previous work, this has proven useful for political bias prediction (Baly et al., 2018a). Such features include information about whether Twitter verified the account, the year it was created, its geographical location, as well as some other statistics, e.g., the number of followers and of tweets posted.

We encoded the profile’s description using SBERT for the following reasons: (i) unlike the articles, the number of media profiles is too small to fine-tune BERT, and (ii) most Twitter descriptions have sentence-like structure and length. If a medium has no Twitter account, we used a vector of zeros.

3.2 Who Read it

We argue that the audience of a news medium can be indicative of the political orientation of that medium. We thus propose a number of features to model this, which we describe below.

3.2.1 Twitter Followers Bio

Previous research has used the followers’ networks and the retweeting behavior in order to infer the political bias of news media (Wong et al., 2013; Atanasov et al., 2019; Darwish et al., 2020). Here, we analyze the self-description (bio) of Twitter users that follow the target news medium. The assumption is that (i) followers would likely agree with the news medium’s bias, and (ii) they might express their own bias in their self-description.

We retrieved the public profiles of 5,000 followers for each target news medium with a Twitter account, and we excluded those with non-English bios since our dataset is mostly about US media. Then, we encoded each follower’s bio using SBERT (Reimers and Gurevych, 2019). As we had plenty of followers’ bios, this time fine-tuning BERT would have been feasible. However, we were afraid to use distant supervision for labeling as we did with the articles since people sometimes follow media with different political ideologies. Thus, we opted for SBERT, and we averaged the SBERT representations across the bios in order to obtain a medium-level representation.

3.2.2 Facebook Audience

Like many other social media giants, Facebook makes its revenues from advertisements. The extensive user interaction enables Facebook to create detailed profiles of its users, including demographic attributes such as age, gender, income, and political leaning. Advertisers can explore these attributes to figure out the targeting criteria for their ads, and Facebook returns an audience estimate based on these criteria. For example, the estimated number of users who are female, 20-years-old, very liberal, and interested in the NY Times is 160K. These estimates have been used as a proxy to measure the on-line population in various domains (Fatehikia et al., 2018; Araujo et al., 2017; Ribeiro et al., 2018).

In this study, we explore the use of political leaning estimates of users who are interested in particular news media. To obtain the audience estimates for a medium, we identify its Interest ID using the Facebook Marketing API⁵. Given an ID, we retrieve the estimates of the audience (in the United States) who showed interest in the corresponding medium. Then, we extract the audience distribution over the political spectrum, which is categorized into five classes ranging from *very conservative* to *very liberal*.

3.2.3 YouTube Audience Statistics

Finally, we incorporate audience information from YouTube videos. We retrieved the following metadata to model audience interaction: number of views, likes, dislikes, and comments for each video. As before, we averaged these statistics across the videos to obtain a medium-level representation.

⁵<http://developers.facebook.com/docs/marketing-api>

3.3 What Was Written About the Target Medium

Wikipedia contents describing news media were useful for predicting the political bias and the factuality of these media (Baly et al., 2018a). We automatically retrieved the Wikipedia page for each medium, and we encoded its contents using the pre-trained BERT model.⁶ Similarly to encoding the articles, we fed the encoder with the first 510 tokens of the page’s content, and used as an output representation the average of the word representations extracted from the second-to-last layer. If a medium had no page in Wikipedia, we used a vector of zeros.

4 Experiments and Evaluation

4.1 Dataset

We used the Media Bias/Fact Check (MBFC) dataset, which consists of a list of news media along with their labels of both political bias and factuality of reporting. Factuality is modeled on a 3-point scale: *low*, *mixed*, and *high*. Political bias is modeled on a 7-point scale: *extreme-left*, *left*, *center-left*, *center*, *center-right*, *right*, and *extreme-right*. Further details and examples of the dataset can be found in (Baly et al., 2018a).

After manual inspection, we noticed that the *left-center* and *right-center* labels are ill-defined, ambiguous transitional categories. Therefore, we decided to exclude news media with these labels. Also, to reduce the impact of potentially subjective decisions made by the annotators, we merged the *extreme-left* and *extreme-right* media with the left and right categories, respectively. As a result, we model political bias on a 3-point scale (*left*, *center*, and *right*), and the dataset got reduced to 864 news media. Table 1 provides statistics about the dataset.

| | Political Bias | | Factuality | |
|--------|----------------|-------|------------|--|
| Left | 243 | Low | 162 | |
| Center | 272 | Mixed | 249 | |
| Right | 349 | High | 453 | |

Table 1: Label counts in the dataset.

We were able to retrieve Wikipedia pages for 61.2% of the media, Twitter profiles for 72.5% of the media, Facebook pages for 60.8% of the media, and YouTube channel for 49% of the media.

⁶Similarly to Twitter descriptions, the number of news media with Wikipedia pages is too small to fine-tune BERT.

4.2 Experimental Setup

We evaluated the following aspects about news media separately and in combinations: (i) what the target medium wrote, (ii) who read it, and (iii) what was written about that medium. We used the features described in Section 3 to train SVM classifiers for predicting the political bias and the factuality of reporting of news media. We performed an incremental ablation study by combining the best feature(s) from each aspect to obtain a combination that achieves even better results. We used 5-fold cross-validation to train and to evaluate an SVM model using different features and feature combinations. At each iteration of the cross-validation, we performed a grid search to tune the hyper-parameters of our SVM model, namely the values of the cost C and of the γ value for the RBF kernel. In the process of search, we optimized for macro-average F_1 score, i.e., averaging over the classes, since our dataset is not balanced, which is true for both tasks. Finally, we evaluated the model on the remaining unseen fold. Ultimately, we report both macro- F_1 score, and accuracy.

We compared our results to the majority class baseline and to our previous work (Baly et al., 2018a). The latter used (i) NELA features from articles, (ii) embedding representations of Wikipedia pages using averaged GloVe word embeddings, (iii) metadata from the media’s Twitter profiles, and (iv) URL structural features. Since we slightly modified the MBFC dataset, we retrained the old model on the new version of the dataset.⁷

To fine-tune BERT’s weights, we trained a softmax layer on top of the [CLS] token of the pre-trained BERT model to classify articles for the task at hand: either predicting the articles’ political bias as *left*, *center*, or *right*, or predicting their level of factuality as *low* or *high*.⁸ To avoid overfitting, we scrapped articles from news media listed in the Media Bias/Fact Check database, but not included in our dataset: 30K articles from 298 such media.

Finally, we used two strategies to evaluate feature combinations. The first one trains a single classifier using all features. The second one trains a separate classifier for each feature type and then uses an ensemble by taking a weighted average of the posterior probabilities of the individual models.

⁷The data and the corresponding code, both old and new, are available at <https://github.com/ramybaly/News-Media-Reliability>

⁸We ignored *mixed* as it does not apply to articles.

Note that we learn different weights for the different models, which ensures that we pay more attention to the probabilities produced by better models. We used the sklearn library to obtain probabilities from an SVM classifier as a function of the distance between the data point and the learned hyperplane using Platt scaling (for the binary case) or an extension thereof (for the 3-way case).

4.3 Political Bias Prediction

Table 2 shows the evaluation results for political bias prediction, grouped according to different aspects. For each aspect, the upper rows correspond to individual features, while the lower ones show combinations thereof.

The results in rows 3–5 show that averaging embeddings from a fine-tuned BERT to encode articles (row 4) works better than using NELA features (row 3). They also show that using the posterior probabilities obtained from applying a softmax on top of BERT’s [CLS] token (row 5) performs worse than using average embeddings (row 4). This suggests that it is better to incorporate information from the articles’ word representations rather than using [CLS] as a compact representation of the articles. Also, since our BERT was fine-tuned on articles with noisy labels obtained using distant supervision, its predictions for individual articles are also noisy, and so are the vectors of posterior. Yet, this fine-tuning seems to yield improved article-level representations for our task.

The results in rows 7–10 show that captions are the most useful type of feature among those extracted from YouTube. This makes sense since captions contain the most essential information about the contents of a video. We can further see that the BERT-based features outperform the NELA ones. Overall, the YouTube features are underperforming since for half of the media we could not find a corresponding YouTube channel, and we used representations containing only zeroes.

Rows 11-16 show the results for systems that combine article, Twitter, and YouTube features, either directly or in an ensemble. We can see on rows 13–16 that the YouTube and the Twitter profile features yield loss in performance when added to the article features (rows 11–12). Note that the article features already outperform the individual feature types from rows 3–10 by a wide margin, and thus we will use them to represent the *What Was Written* aspect of the model in our later experiments below.

| Group | # | Features | Dim. | Macro F_1 | Accuracy | |
|---|-----------------------|---|---------------------------------|--------------|--------------|-------|
| <i>Baselines</i> | 1 | Majority class | – | 19.18 | 40.39 | |
| | 2 | Best model from (Baly et al., 2018a) | 764 | 72.90 | 73.61 | |
| <i>A. What Was Written</i> | 3 | Articles: NELA | 141 | 64.82 | 68.18 | |
| | 4 | Articles: BERT representations | 768 | 79.34 | 79.75 | |
| | 5 | Articles: BERT probabilities | 3 | 61.21 | 62.27 | |
| | 6 | Twitter Profiles: Sentence BERT | 768 | 59.23 | 60.88 | |
| | 7 | YouTube: NELA (title, description) | 260 | 45.78 | 50.46 | |
| | 8 | YouTube: OpenSmile (LLDs) | 385 | 46.13 | 50.69 | |
| | 9 | YouTube: BERT (title, description, tags) | 768 | 48.36 | 53.94 | |
| | 10 | YouTube: BERT (captions) | 768 | 49.14 | 53.94 | |
| | 11 | Articles: ALL (<i>c</i>) | 912 | 81.00 | 81.48 | |
| | 12 | Articles: ALL (<i>en</i>) | 912 | 81.27 | 81.83 | |
| | 13 | Articles + Twitter Prof. (<i>c</i>) | 1,691 | 76.59 | 77.20 | |
| | 14 | Articles + Twitter Prof. (<i>en</i>) | 1,691 | 80.00 | 80.56 | |
| | 15 | Articles + Twitter Prof. + YouTube cap. (<i>c</i>) | 2,315 | 75.73 | 76.39 | |
| | 16 | Articles + Twitter Prof. + YouTube cap. (<i>en</i>) | 2,315 | 79.70 | 80.32 | |
| | <i>B. Who Read It</i> | 17 | Twitter Follower: Sentence BERT | 768 | 62.85 | 65.39 |
| | | 18 | YouTube: Metadata | 5 | 40.05 | 46.53 |
| 19 | | Facebook: Political Leaning Estimates | 6 | 27.87 | 43.87 | |
| 20 | | Twitter Fol. + YouTube Meta. (<i>c</i>) | 773 | 63.72 | 65.86 | |
| 21 | | Twitter Fol. + YouTube Meta. (<i>en</i>) | 773 | 65.12 | 66.44 | |
| 22 | | Twitter Fol. + YouTube Meta. + Facebook Estimates (<i>c</i>) | 779 | 63.63 | 65.74 | |
| 23 | | Twitter Fol. + YouTube Meta. + Facebook Estimates (<i>en</i>) | 779 | 64.18 | 66.20 | |
| <i>C. What Was Written About the Medium</i> | 24 | Wikipedia: BERT | 768 | 64.36 | 66.09 | |
| <i>Combinations</i> | 25 | All features: rows 3–11; 18–20; 25 (<i>c</i>) | 5,413 | 78.17 | 78.70 | |
| | 26 | All features: rows 3–11; 18–20; 25 (<i>en</i>) | 5,413 | 79.42 | 80.32 | |
| | 27 | A+B: rows 12 & 21 (<i>c</i>) | 1,685 | 84.28 | 84.87 | |
| | 28 | A+B: rows 12 & 21 (<i>en</i>) | 1,685 | 84.15 | 84.64 | |
| | 29 | A+C: rows 12 & 24 (<i>c</i>) | 1,680 | 81.53 | 81.98 | |
| | 30 | A+C: rows 12 & 24 (<i>en</i>) | 1,680 | 82.99 | 83.48 | |
| | 31 | A+B+C: rows 12, 21 & 24 (<i>c</i>) | 1,691 | 83.53 | 84.02 | |
| | 32 | A+B+C: rows 12, 21 & 24 (<i>en</i>) | 1,691 | 84.77 | 85.29 | |

Table 2: **Political bias prediction:** ablation study of the proposed features. *Dim* refers to the number of features, whereas (*c*) and (*en*) indicate whether the features are concatenated or an ensemble was used, respectively.

We can further notice that the ensembles consistently outperform feature concatenation models, which is actually true for all feature combinations in Table 2.

Next, we compare rows 6 and 17, which show results when using Twitter information of different nature: from the target medium profile (row 6) vs. from the profiles of the followers of the target medium (row 17). We can see that the latter is much more useful, which confirms the importance of the *Who Read It* aspect, which we have introduced in this paper. Note that here we encode the descriptions and the self-description bio information using Sentence BERT instead of the pre-trained BERT; this is because, in our preliminary experiments (not shown in the table), we found the former to perform much better than the latter.

Next, the results in rows 20–23 show that the YouTube metadata features improve the performance when combined with the Twitter followers’ features. On the other hand, the Facebook audience features’ performance is deficient and hurts the overall performance, i.e., these estimates seem not to correlate well with the political leanings of news media. Also, as pointed by (Flaxman et al., 2016), social networks can help expose people to different views, and thus the polarization in news readership might not be preserved.

Row 24 shows that the Wikipedia features perform worse than most individual features above, which can be related to coverage as only 61.2% of the media in our dataset have a Wikipedia page. Nevertheless, these features are helpful when combined with features about other aspects; see below.

| Group | # | Features | Dim. | Macro F_1 | Accuracy | |
|---|-----------------------|---|---------------------------------|--------------|--------------|-------|
| <i>Baselines</i> | 1 | Majority class | – | 22.93 | 52.43 | |
| | 2 | Best model from (Baly et al., 2018a) | 764 | 61.08 | 66.45 | |
| <i>A. What Was Written</i> | 3 | Articles: NELA | 141 | 55.54 | 62.62 | |
| | 4 | Articles: BERT representations | 768 | 61.46 | 67.94 | |
| | 5 | Articles: BERT probabilities | 3 | 51.39 | 61.46 | |
| | 6 | Twitter Profiles: Sentence BERT | 768 | 49.96 | 56.71 | |
| | 7 | YouTube: NELA (title, description) | 260 | 32.52 | 51.04 | |
| | 8 | YouTube: OpenSmile (LLDs) | 385 | 37.17 | 52.08 | |
| | 9 | YouTube: BERT (title, description, tags) | 768 | 38.19 | 54.28 | |
| | 10 | YouTube: BERT (captions) | 768 | 38.82 | 55.56 | |
| | 11 | Articles: ALL (<i>c</i>) | 912 | 59.34 | 64.82 | |
| | 12 | Articles: ALL (<i>en</i>) | 912 | 48.27 | 59.95 | |
| | 13 | Articles: BERT + Twitter Prof. (<i>c</i>) | 1,691 | 61.06 | 66.09 | |
| | 14 | Articles: BERT + Twitter Prof. (<i>en</i>) | 1,691 | 61.50 | 68.63 | |
| | 15 | Articles: BERT + Twitter Prof. + YouTube: cap. (<i>c</i>) | 2,315 | 60.23 | 65.51 | |
| | 16 | Articles: BERT + Twitter Prof. + YouTube: cap. (<i>en</i>) | 2,315 | 58.21 | 66.44 | |
| | <i>B. Who Read It</i> | 17 | Twitter Follower: Sentence BERT | 768 | 42.19 | 58.45 |
| | | 18 | YouTube: Metadata | 5 | 31.92 | 52.78 |
| 19 | | Facebook: Political Leaning Estimates | 6 | 27.24 | 53.70 | |
| 20 | | Twitter Fol. + YouTube Meta. (<i>c</i>) | 773 | 42.48 | 58.76 | |
| 21 | | Twitter Fol. + YouTube Meta. (<i>en</i>) | 773 | 39.66 | 57.64 | |
| 22 | | Twitter Fol. + YouTube Meta. + Facebook Estimates (<i>c</i>) | 779 | 42.28 | 57.76 | |
| 23 | | Twitter Fol. + YouTube Meta. + Facebook Estimates (<i>en</i>) | 779 | 39.33 | 57.99 | |
| <i>C. What Was Written About the Medium</i> | 24 | Wikipedia: BERT | 768 | 45.74 | 55.32 | |
| | 25 | All features: rows 3–10; 17–19; 24 (<i>c</i>) | 5,413 | 62.42 | 67.79 | |
| <i>Combinations</i> | 26 | All features: rows 3–10; 17–19; 24 (<i>en</i>) | 5,413 | 45.24 | 60.42 | |
| | 27 | A+B: rows 14 & 24 (<i>c</i>) | 1,680 | 65.45 | 70.40 | |
| | 28 | A+B: rows 14 & 24 (<i>en</i>) | 1,680 | 61.80 | 69.25 | |
| | 29 | A+C: rows 14 & 20 (<i>c</i>) | 1,685 | 67.25 | 71.52 | |
| | 30 | A+C: rows 14 & 20 (<i>en</i>) | 1,685 | 62.53 | 69.90 | |
| | 31 | A+B+C: rows 14, 20 & 24 (<i>c</i>) | 1,691 | 64.14 | 69.36 | |
| | 32 | A+B+C: rows 14, 20 & 24 (<i>en</i>) | 1,691 | 60.35 | 68.90 | |

Table 3: **Factuality of reporting:** ablation study of the proposed features. *Dim* refers to the number of features, whereas (*c*) and (*en*) indicate whether the features are concatenated or an ensemble was used, respectively.

Finally, rows 25–32 in Table 3 show the evaluation results when combining all aspects. We can see that the best results are achieved when using the best features from each of the three aspects, where the combination is performed as an ensemble (row 32). This combination improves over using information from the article only (row 12) by +3.5 macro- F_1 points absolute. It further yields sizeable absolute improvements over the baseline system from (Baly et al., 2018a), by +11.87 macro- F_1 points absolute. While this improvement is due to a large extent to improved techniques for text representation such as using fine-tuned BERT instead of averaged GloVe word embeddings, modeling the newly-introduced media aspects further yielded a lot of additional improvements.

4.4 Factuality Prediction

Table 3 demonstrates the evaluation results when using the proposed sources/features for the task of predicting the factuality of reporting of news media.

Similarly to the results for political bias prediction, rows 3–10 suggest that the features extracted from articles are more important than those coming from YouTube or from Twitter profiles, and that using BERT to encode the articles yields the best results. Note that overall, the results in this table are not as high as those for bias prediction. This reflects the level of difficulty of this task, and the fact that, in order to predict factuality, one needs external information or a knowledge base to be able to verify the published content.

The results in rows 11–16 show that combining the Twitter profile features with the BERT-encoded articles improves the performance over using the article text only.

Comparing rows 6 and 17 in Table 3, we can see that the Twitter follower features perform worse than using Twitter profiles features; this is the opposite of what we observed in Table 2. This makes sense since our main motivation to look at the followers’ profiles was to detect political bias, rather than factuality. Moreover, the metadata collected from media profiles about whether the corresponding account is verified, or its level of activity or connectivity (counts of friends and statuses) are stronger signals for this task.

Finally, rows 25–32 show the results for modeling combinations of the three aspects we are exploring in this paper. The best results are achieved using the best features selected from the *What was written* and the *What was written about the target medium* aspects, concatenated together. This combination achieves sizeable improvements compared to the baseline system from (Baly et al., 2018a): by +6.17 macro- F_1 points absolute. This result indicates that looking at the audience of the medium is not as helpful for predicting factuality as it was for predicting political bias, and that looking at what was written about the medium on Wikipedia is more important for this task.

5 Conclusion and Future Work

We have presented experiments in predicting the political ideology, i.e., left/center/right bias, and the factuality of reporting, i.e., high/mixed/low, of news media. We compared the textual content of what media publish vs. who read it on social media, i.e., on Twitter, Facebook, and YouTube. We further modeled what was written about the target medium in Wikipedia.

We have combined a variety of information sources, many of which were not explored for at least one of the target tasks, e.g., YouTube channels, political bias of the Facebook audience, and information from the profiles of the media followers on Twitter. We further modeled different modalities: text, metadata, and speech signal. The evaluation results have shown that while what was written matters most, the social media context is also important as it is complementary, and putting them all together yields sizable improvements over the state of the art.

In future work, we plan to perform user profiling with respect to polarizing topics such as *gun control* (Darwish et al., 2020), which can then be propagated from users to media (Atanasov et al., 2019; Stefanov et al., 2020). We further want to model the network structure, e.g., using graph embeddings (Darwish et al., 2020). Another research direction is to profile media based on their stance with respect to previously fact-checked claims (Moftarami et al., 2018; Shaar et al., 2020), or by the proportion and type of propaganda techniques they use (Da San Martino et al., 2019, 2020). Finally, we plan to experiment with other languages.

Acknowledgments

This research is part of the Tanbih project⁹, which aims to limit the effect of “fake news,” propaganda and media bias by making users aware of what they are reading. The project is developed in collaboration between the Qatar Computing Research Institute, HBKU and the MIT Computer Science and Artificial Intelligence Laboratory.

References

- Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. DocBERT: BERT for document classification. *arXiv preprint arXiv:1904.08398*.
- Matheus Araujo, Yelena Mejova, Ingmar Weber, and Fabricio Benevenuto. 2017. Using Facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *Proceedings of the 2017 ACM Conference on Web Science, WebSci ’17*, pages 253–257, Troy, NY, USA.
- Atanas Atanasov, Gianmarco De Francisci Morales, and Preslav Nakov. 2019. Predicting the role of political trolls in social media. In *Proceedings of the 2019 SIGNLL Conference on Computational Natural Language Learning, CoNLL ’19*, pages 1023–1034, Hong Kong, China.
- Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018a. 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, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news

⁹<http://tanbih.qcri.org/>

- media. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 2109–2116, Minneapolis, MN, USA.
- Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018b. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 21–27, New Orleans, LA, USA.
- Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.
- Kevin R. Canini, Bongwon Suh, and Peter L. Pirolli. 2011. Finding credible information sources in social networks based on content and social structure. In *Proceedings of the IEEE International Conference on Privacy, Security, Risk, and Trust, and the IEEE International Conference on Social Computing*, SocialCom/PASSAT '11, pages 1–8, Boston, MA, USA.
- Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, Hyderabad, India.
- Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence*, IJCAI-PRICAI '20, Yokohama, Japan.
- Giovanni Da San Martino, Seunghak Yu, Alberto Barron-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, EMNLP '19, pages 5636–5646, Hong Kong, China.
- Kareem Darwish, Michael Aupetit, Peter Stefanov, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '20, Atlanta, GA, USA.
- Stefano DellaVigna and Ethan Kaplan. 2007. The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, MN, USA.
- Yoan Dinkov, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Predicting the leading political ideology of Youtube channels using acoustic, textual and metadata information. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, INTERSPEECH '19, pages 501–505, Graz, Austria.
- Erick Elejalde, Leo Ferres, and Eelco Herder. 2018. On the nature of real and perceived bias in the mainstream media. *PloS one*, 13(3):e0193765.
- Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. openSMILE – the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462, Florence, Italy.
- Masoomali Fatehikia, Ridhi Kashyap, and Ingmar Weber. 2018. Using Facebook ad data to track the global digital gender gap. *World Development*, 107:189–209.
- Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320.
- Matthew Gentzkow and Jesse M Shapiro. 2006. Media bias and reputation. *Journal of political Economy*, 114(2):280–316.
- Doris A Graber and Johanna Dunaway. 2017. *Mass media and American politics*. SAGE Publications.
- Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. 2018a. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Proceedings of the The Web Conference*, WWW '18, pages 235–238, Lyon, France.
- Benjamin D. Horne, Sara Khedr, and Sibel Adali. 2018b. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the Twelfth International Conference on Web and Social Media*, ICWSM '18, pages 518–527, Stanford, CA, USA.
- Shanto Iyengar and Kyu S Hahn. 2009. Red media, blue media: Evidence of ideological selectivity in media use. *Journal of communication*, 59(1):19–39.
- Daniel Kopev, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Detecting deception in political debates using acoustic and textual features. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, ASRU '19, pages 652–659, Singapore.

- Vivek Kulkarni, Junting Ye, Steven Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '18*, pages 3518–3527, Brussels, Belgium.
- Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence, IJCAI '16*, pages 3818–3824, New York, NY, USA.
- Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, pages 1751–1754, Melbourne, Australia.
- Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT '18*, pages 767–776, New Orleans, LA, USA.
- Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, pages 353–362, Melbourne, Australia.
- Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adali. 2019. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the International AAAI Conference on Web and Social Media, ICWSM '19*, pages 630–638, Munich, Germany.
- Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the Web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17*, pages 1003–1012, Perth, Australia.
- Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A credibility lens for analyzing and explaining misinformation. In *Proceedings of The Web Conference 2018, WWW '18*, pages 155–158, Lyon, France.
- Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylistic 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.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China.
- Filipe N Ribeiro, Lucas Henrique, Fabricio Benvenuto, Abhijnan Chakraborty, Juhi Kulshrestha, Mahmoudreza Babaei, and Krishna P Gummadi. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media, ICWSM '18*, pages 290–299, Stanford, CA, USA.
- Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. 2019. Team QCRI-MIT at SemEval-2019 task 4: Propaganda analysis meets hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval '19*, pages 1041–1046, Minneapolis, MN, USA.
- Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The INTERSPEECH 2009 emotion challenge. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH '09*, pages 312–315, Brighton, UK.
- Shaden Shaar, Giovanni Da San Martino, Nikolay Babulov, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the Annual Conference of the Association for Computational Linguistics, ACL '20*, Seattle, WA, USA.
- Petar Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the Annual Conference of the Association for Computational Linguistics, ACL '20*, Seattle, WA, USA.
- Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.
- Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2013. Quantifying political leaning from tweets and retweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, ICWSM '13*, pages 640–649, Boston, MA, USA.
- Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow. 2014. A bayesian approach for predicting the popularity of tweets. *Ann. Appl. Stat.*, 8(3):1583–1611.
- Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11(3):1–29.