Predicting the Topical Stance and Political Leaning of Media using Tweets

Peter Stefanov[1], **Kareem Darwish**[2], **Atanas Atanasov**[3], **Preslav Nakov**[2]
[1]SiteGround Hosting EOOD, Bulgaria
[2]Qatar Computing Research Institute, HBKU, Doha, Qatar
[3]Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
{stefanov.peter.ps,atanas.atanasov.sf}@gmail.com,
{kdarwish,pnakov}@hbku.edu.qa

Abstract

Discovering the stances of media outlets and influential people on current, debatable topics is important for social statisticians and policy makers. Many supervised solutions exist for determining viewpoints, but manually annotating training data is costly. In this paper, we propose a cascaded method that uses unsupervised learning to ascertain the stance of Twitter users with respect to a polarizing topic by leveraging their retweet behavior; then, it uses supervised learning based on user labels to characterize both the general political leaning of online media and of popular Twitter users, as well as their stance with respect to the target polarizing topic. We evaluate the model by comparing its predictions to gold labels from the Media Bias/Fact Check website, achieving 82.6% accuracy.

1 Introduction

Online media and popular Twitter users, which we will collectively refer to as *influencers*, often express overt political leanings, which can be gleaned from their positions on a variety of political and cultural issues. Determining their leaning can be done through the analysis of their writing, which includes the identification of terms that are indicative of stance (Groseclose and Milyo, 2005; Gentzkow and Shapiro, 2011). Performing such analysis automatically can be done using supervised classification, which in turn would require manually labeled data (Groseclose and Milyo, 2005; Gentzkow and Shapiro, 2011; Mohammad et al., 2016). Alternatively, leanings can be inferred based on which people share the content (blogs, tweets, posts, etc.) on social media, as social media users are more likely to share content that originates from sources that generally agree with their positions (An et al., 2012; Morgan et al., 2013; Ribeiro et al., 2018; Wong et al., 2013). Here, we make use of this observation to characterize influencers, based on the stances of the Twitter users that share their content. Ascertaining the stances of users, also known as stance detection, involves identifying the position of a user with respect to a topic, an entity, or a claim (Mohammad et al., 2016). For example, on the topic of abortion in USA, the stances of left- vs. right-leaning users would typically be "pro-choice" vs. "pro-life", respectively.

In this paper, we propose to apply unsupervised stance detection to automatically tag a large number of Twitter users with their positions on specific topics (Darwish et al., 2020). The tagging identifies clusters of vocal users based on the accounts that they retweet. Although the method we use may yield more than two clusters, we retain the two largest ones, which typically include the overwhelming majority of users, and we ignore the rest. Then, we train a classifier that predicts which cluster a user belongs to, in order to expand our clusters. Once we have increased the number of users in our sets, we determine which sources are most strongly associated with each group based on sharing by each group. We apply this methodology to determine the positions of influencers and of media on eight polarizing topics along with their overall leaning: left, center or right. In doing so, we can also observe the sharing behavior of right- and left-leaning users, and we can correlate their behavior with the credibility of the sources. Further, given the user stances for these eight topics, we train a supervised classifier to predict the overall bias of sources using a variety of features, including the so-called *valence* (Conover et al., 2011a), graph embeddings, and contextual embeddings. Using a combination of these features, our classifier is able to predict the bias of sources with 82.6% accuracy, with valence being the most effective feature. Figure 1 outlines our overall methodology.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 527–537
July 5 - 10, 2020.

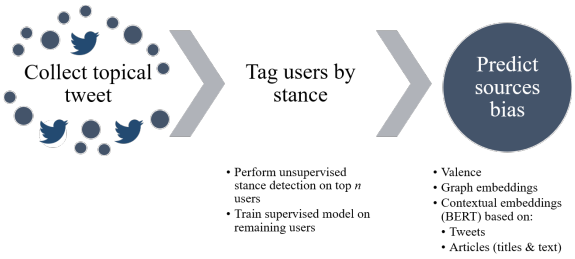

Figure 1: General outline of our methodology.

Our contributions are as follows:

- We use unsupervised stance detection to automatically determine the stance of Twitter users with respect to several polarizing topics.
- We then use distant supervision based on these discovered user stances to accurately characterize the political leaning of media outlets and of popular Twitter accounts. For classification, we use a combination of source valence, graph embeddings, and contextualized text embeddings.
- We evaluate our approach by comparing its bias predictions for a number of news outlets against gold labels from Media Bias/Fact Check. We further evaluate its predictions for popular Twitter users against manual judgments. The experimental results show sizable improvements over using graph embeddings or contextualized text embeddings.

The remainder of this paper is organized as follows: Section 2 discusses related work. Section 3 describes the process of data collection. Section 4 presents our method for user stance detection. Section 5 describes how we characterize the influencers. Section 6 discusses our experiments in media bias prediction. Finally, Section 7 concludes and points to possible directions for future work.

2 Related Work

Recent work that attempted to characterize the stance and the ideological leaning of media and Twitter users relied on the observation that users tend to retweet content that is consistent with their world view. This stems from *selective exposure*, which is a cognitive bias that leads people to avoid the cognitive overload from exposure to opposing views as well as the cognitive dissonance in which people are forced to reconcile between their views and opposing views (Morgan et al., 2013).

Concerning media, Ribeiro et al. (2018) used the Facebook advertising services to infer the ideological leaning of online media based on the political leaning of Facebook users who consumed them. An et al. (2012) relied on follow relationships to online media on Twitter to ascertain ideological leaning of media and users based on the similarity between them. Wong et al. (2013) studied retweet behavior to infer the ideological leanings of online media sources and popular Twitter accounts. Barberá and Sood (2015) proposed a statistical model based on the follower relationships to media sources and Twitter personalities in order to estimate their ideological leaning.

As for individual users, much recent work focused on stance detection to determine a person's position on a topic including the deduction of political preferences (Barberá, 2015; Barber and Rivero, 2015; Borge-Holthoefer et al., 2015; Cohen and Ruths, 2013; Colleoni et al., 2014; Conover et al., 2011b; Fowler et al., 2011; Hasan and Ng, 2014; Himelboim et al., 2013; Magdy et al., 2016a,b; Makazhanov et al., 2014; Trabelsi and Zaïane, 2018; Weber et al., 2013). User stance classification is aided by the tendency of users to form so-called "echo chambers", where they engage with like-minded users (Himelboim et al., 2013; Magdy et al., 2016a), and the tendency of users' beliefs to be persistent over time (Borge-Holthoefer et al., 2015; Magdy et al., 2016a; Pennacchiotti and Popescu, 2011b).

Studies have examined the effectiveness of different features for stance detection, including textual features such as word n-grams and hashtags, network interactions such as retweeted accounts and mentions, and profile information such as user location (Borge-Holthoefer et al., 2015; Hasan and Ng, 2013; Magdy et al., 2016a,b; Weber et al., 2013). Network interaction features were shown to yield better results compared to using textual features (Magdy et al., 2016a; Wong et al., 2013). Sridhar et al. (2015) leveraged both user interactions and textual information when modeling stance and disagreement, using a probabilistic programming system that allows models to be specified using a declarative language.

Trabelsi and Zaïane (2018) described an unsupervised stance detection method that determines the viewpoints of comments and of their authors. It analyzes online forum discussion threads, and therefore assumes a certain structure of the posts.

It also assumes that users tend to reply to each others' comments when they are in disagreement, whereas we assume the opposite in this paper. Their model leverages the posts' contents, whereas we only use the retweet behavior of users.

Many methods involving supervised learning were proposed for stance detection. Such methods require the availability of an initial set of labeled users, and they use some of the aforementioned features for classification (Darwish et al., 2018; Magdy et al., 2016b; Pennacchiotti and Popescu, 2011a). Such classification can label users with precision typically ranging between 70% and 90% (Rao et al., 2010; Pennacchiotti and Popescu, 2011a). Label propagation is a semi-supervised method that starts with a seed list of labeled users and propagates the labels to other users who are similar based on the accounts they follow or retweet (Barberá and Sood, 2015; Borge-Holthoefer et al., 2015; Weber et al., 2013). While label propagation may label users with high precision (often above 95%), it is biased towards users with more extreme views; moreover, careful choice of thresholds is often required, and post-checks are needed to ensure quality.

Abu-Jbara et al. (2013) and more recently Darwish et al. (2020) used unsupervised stance detection, where users are mapped into a lower dimensional space based on user-user similarity, and then clustered to find core sets of users representing different stances. This was shown to be highly effective with nearly perfect clustering accuracy for polarizing topics, and it requires no manual labeling of users. Here, we use the same idea, but we combine it with supervised classification based on retweets in order to increase the number of labeled users (Darwish, 2018). Other methods for user stance detection include collective classification (Duan et al., 2012), where users in a network are jointly labeled and classification in a low-dimensional user-space (Darwish et al., 2017).

As for predicting political leaning or sentiment, this problem was studied previously as a supervised learning problem, where a classifier learns from a set of manually labeled tweets (Pla and Hurtado, 2014; Bakliwal et al., 2013; Bermingham and Smeaton, 2011). Similarly, Volkova et al. (2014) predicted Twitter users' political affiliation (being Republican or Democratic), using their network connections and textual information, relying on user-level annotations.

3 Data Collection

We obtained data on eight topics that are considered polarizing in the USA (Darwish et al., 2020), shown in Table 1.

They include a mix of long-standing issues such as racism and gun control, temporal issues such as the nomination of Judge Brett Kavanaugh to the US Supreme Court and Representative Ilhan Omar's polarizing remarks, as well as non-political issues such as the potential dangers of vaccines. Further, though long-standing issues typically show right–left polarization, stances towards Omar's remarks are not as clear, with divisions on the left as well.

Since we are interested in US users, we filtered some tweets to retain such by users who have stated that their location was USA. We used a gazetteer that included words that indicate USA as a country (e.g., America, US), as well as state names and their abbreviations (e.g., Maryland, MD).

Other data that we used in our experiments is a collection of articles that were cited by users from the tweets collection and that originate from media, whose bias is known, i.e., is discussed on the Media Bias/Fact Check website.

4 User Stance Detection

In order to analyze the stance of influencers on a given topic, we first find the stances of Twitter users, and then we project them to the influencers that the users cite. A central (initial) assumption here is that if a user includes a link to some article in their tweet, they are more likely to agree or endorse the article's message. Similarly, when a user retweets a tweet verbatim without adding any comments, they are more likely to agree with that tweet. We label a large number of users with their stance for each topic using a two-step approach, namely ***projection and clustering*** and ***supervised classification***.

For the projection and clustering step, we identify clusters of core vocal users using the unsupervised method described in (Darwish et al., 2020). In this step, users are mapped to a lower dimensional space based on their similarity, and then they are clustered. After performing this unsupervised learning step, we train a supervised classifier using the two largest identified clusters in order to tag many more users. For that, we use FastText, a deep neural network text classifier, that has been shown to be effective for various text classification tasks (Joulin et al., 2017).

Topic	Keywords	Date Range	No. of Tweets
Climate change	#greendeal, #environment, #climate, #climatechange, #carbonfootprint, #climatehoax, #climategate, #globalwarming, #agw, #renewables	Feb 25–Mar 4, 2019	1,284,902
Gun control/rights	#gun, #guns, #weapon, #2a, #gunviolence, #secondamendment, #shooting, #massshooting, #gunrights, #GunReformNow, #GunControl, #NRA	Feb 25–Mar 3, 2019	1,782,384
Ilhan Omar remarks on Israel lobby	IlhanOmarIsATrojanHorse, #IStandWithIlhan, #ilhan, #Antisemitism, #IlhanOmar, #IlhanMN, #RemoveIlhanOmar, #ByeIlhan, #RashidaTlaib, #AIPAC, #EverydayIslamophobia, #Islamophobia, #ilhan	Mar 1–9, 2019	2,556,871
Illegal immigration	#border, #immigration, #immigrant, #borderwall, #migrant, #migrants, #illegal, #aliens	Feb 25–Mar 4, 2019	2,341,316
Midterm	midterm, election, elections	Oct 25–27, 2018	520,614
Racism & police brutality	#blacklivesmatter, #bluelivesmatter, #KKK, #racism, #racist, #policebrutality, #excessiveforce, #StandYourGround, #ThinBlueLine	Feb 25–Mar 3, 2019	2,564,784
Kavanaugh Nomination	Kavanaugh, Ford, Supreme, judiciary, Blasey, Grassley, Hatch, Graham, Cornyn, Lee, Cruz, Sasse, Flake, Crapo, Tillis, Kennedy, Feinstein, Leahy, Durbin, Whitehouse, Klobuchar, Coons, Blumenthal, Hirono, Booker, Harris	Sept. 28-30, 2018 & Oct. 6-9, 2018	2,322,141
Vaccination benefits & dangers	#antivax, #vaxxing, #BigPharma, #antivaxxers, #measlesoutbreak, #Antivacine, #VaccinesWork, #vaccine, #vaccines, #Antivaccine, #vaccinestudy, #antivaxx, #provaxx, #VaccinesSaveLives, #ProVaccine, #VaxxWoke, #mykidmychoice	Mar 1–9, 2019	301,209

Table 1: Polarizing topics used in study.

Once we have expanded our sets of labeled users, we identify influencers that are most closely associated with each group using a modified version of the so-called *valence score*, which varies in value between -1 and 1. If an influencer is being cited evenly between the groups, then it would be assigned a valence score close to zero. Conversely, if one group disproportionately cites an influencer compared to another group, then it would be assigned a score closer to -1 or 1. We perform these steps for each of the given topics, and finally we summarize the stances across all topics. Below, we explain each of these steps in more detail.

4.1 Projection and Clustering

Given the tweets for each topic, we compute the similarity between the top 1,000 most active users. To compute similarity, we construct a vector for each user containing the number of all the accounts that a user has retweeted, and then we compute the pairwise cosine similarity between them. For example, if user A has only retweeted user B 3 times, user C 5 times and user E 8 times, then user A's vector would be (0, 3, 5, 0, 8, 0, 0, ... 0). Solely using the retweeted accounts as features has been shown to be effective for stance classification (Darwish et al., 2020; Magdy et al., 2016a). Finally, we perform dimensionality reduction and we project the users using Uniform Manifold Approximation and Projection (UMAP). When performing dimensionality reduction, UMAP places users on a two-dimensional plane such that similar users are placed closer together and dissimilar users are pushed further apart. Figure 2 shows the top users for the "midterm" topic projected with UMAP onto the 2D plane. After the projection, we use Mean Shift to cluster the users as shown in Figure 2. This is the best setup described in (Darwish et al., 2020).

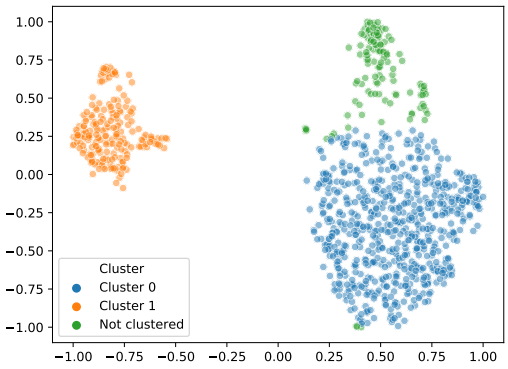

Figure 2: Top active users on the *midterm* topic clustered using UMAP + Mean Shift.

Clustering high-dimensional data often yields suboptimal results, but can be improved by projecting to a low-dimensional space (Darwish et al., 2020).

4.2 Supervised Classification

Since unsupervised stance detection is only able to classify the most vocal users, which only constitute a minority of the users, we wanted to assign stance labels to as many additional users as we can. Given the clusters of users that we obtain for each topic, we retain the two largest clusters for each topic, and we assign cluster labels to the users contained therein. Next, we use all the automatically labeled users for each topic to train a supervised classifier using the accounts that each user retweeted as features (same as the features we used to compute user similarity earlier). For classification, we train a FastText model using the default parameters, and then we classify all other users with five or more retweeted accounts, only accepting the classification if FastText was more than 80% confident (70–90% yielded nearly identical results).

Topic	No. of Users	Clustered Users	Classified Users
climate change	724,470	860	5,851
gun control	973,206	813	11,281
Ilhan Omar	563,706	723	25,484
immigration	940,840	901	22,456
midterm elections	312,954	860	12,765
police brutality & racism	1,175,081	891	18,978
Kavanaugh	809,835	891	10,100
vaccine	194,245	545	556

Table 2: Users per topic: total number of users, umber of clustered users, and number of automatically labeled users.

In order to obtain a rough estimate of the accuracy of the model, we trained FastText using a random 80% subset of the clustered users for each topic and we tested on the remaining 20%. The accuracy was consistently above 95% for all topics. This does not mean that this model can predict the stance for all users that accurately — the clustered users were selected to be the most active ones. Rather, it shows that the classifier can successfully capture what the previous, unsupervised step has already learned. Table 2 lists the total number of users who authored the tweets for each topic, the number of users who were automatically clustered using the aforementioned unsupervised clustering technique, and the number of users who were automatically labeled afterwards using supervised classification. Given that we applied unsupervised stance detection to the most active 1,000 users, the majority of the users appeared in the largest two clusters (shown in Table 2).

4.3 Calculating Valence Scores

Given all the labeled users for each topic, we computed a valence score for each influencer. As mentioned earlier, the valence score ranges between $[-1,1]$, where a value close to 1 implies it is strongly associated with one group of users, -1 shows it is strongly associated with the other group of users, and 0 means that it is being shared or cited by both groups. The original valence score described by Conover et al. (2011a) is calculated as follows:

$$V(u) = 2\frac{\frac{tf(u,C_0)}{total(C_0)}}{\frac{tf(u,C_0)}{total(C_0)} + \frac{tf(u,C_1)}{total(C_1)}} - 1 \qquad (1)$$

where $tf(u,C_0)$ is the number of times (term frequency) item u is cited by group C_0, and $total(C_0)$ is the sum of the term frequencies of all items cited by C_0. $tf(u,C_1)$ and $total(C_1)$ are defined in a similar fashion.

We use the above equation to compute valence scores for the retweeted accounts, but we using a modified version for calculating the score for influencers (I):

$$V(I) = 2\frac{\frac{tf(I,C_0)}{total(C_0)}}{\frac{tf(I,C_0)}{total(C_0)} + \frac{tf(I,C_1)}{total(C_1)}} - 1 \qquad (2)$$

where

$tf(I,C_i) = \sum_{a \in I \cap C_i} [ln(Cnt(a,C_i)) + 1]$

$total(C_i) = \sum_I tf(I,C_i)$

In the latter equation, $Cnt(a,C_i)$ is the number of times article a was cited by users from cluster C_i. In essence, we are replacing term frequencies with the natural log of the term frequencies. We opted to modify the equation in order to tackle the following issue: if users from one of the clusters, say C_1, cite only one single article from some media source a large number of times (e.g., 2,000 times), while users from the other cluster (C_0) cite 10 other articles from the same media 50 times each, then using equation 1 would result in a valence score of -0.6. We would then regard the given media as having an opposing stance to the stance of users in C_0. Alternatively, using the natural log would lead to a valence score close to 0.88. Thus, dampening term frequencies using the natural log has the desired effect of balancing between the number of articles being cited by each group and the total number of citations. We bin the valence scores between -1 and 1 into five equal size bands as follows:

$$Cat(V) = \begin{cases} --, & \text{if } s \in [-1,-0.6) \\ -, & \text{if } s \in [-0.6,-0.2) \\ 0, & \text{if } s \in [-0.2,0.2) \\ +, & \text{if } s \in [0.2,0.6) \\ ++, & \text{if } s \in [0.6,1] \end{cases} \qquad (3)$$

5 Characterizing the Influencers

We use valence to characterize the leaning of all cited influencers for each of the topics. Table 3 shows the valence categories for the top-cited media sources across all topics. It also shows each media's factuality of reporting, i.e., trustworthiness, and bias (ranging from far-left to far-right) as determined by `mediaBiasFactCheck.com`. Since the choice of which cluster should be C_0 and which would be C_1 is arbitrary, we can multiply by -1 the valence scores for any topic and the meaning of the results would stay the same.

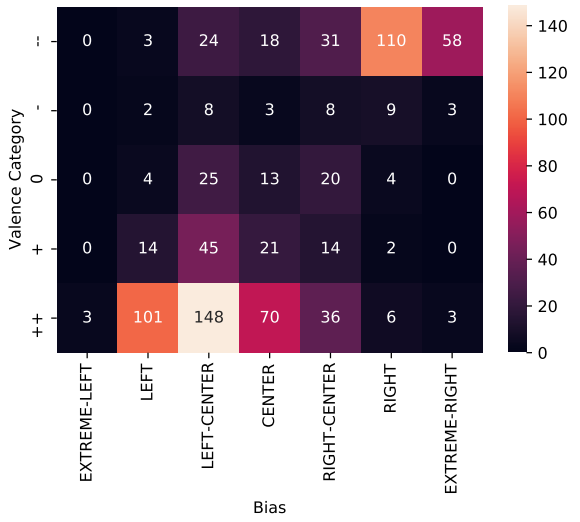

Figure 3: Valence category vs. bias: number of media.

We resorted to doing so for some topics in order to align the extreme valence bands across all topics. Given tweet samples from users in a given cluster for a given topic, labeling that cluster manually was straightforward with almost no ambiguity. Table 4 shows the most frequently cited media source for each topic and for each valence band.

Of the 5,406 unique media sources that have been cited in tweets across all topics, 806 have known political bias from `mediaBiasFactCheck.com`. Figure 3 shows the confusion matrix between our valence categories and the goold labels from `mediaBiasFactCheck.com`.

We notice that many of the media that have a negative valence score (categories $-$ and $--$) are classified on the right side of the political spectrum by `mediaBiasFactCheck.com`, while most media with positive scores (categories $+$ and $++$) are classified as slightly left-leaning. Although there are almost no extreme-left cases, there is a correlation between bias and our valence score. `mediaBiasFactCheck.com` seems to rarely categorize media sources as "extreme-left". This could be a reflection of reality or it might imply that `mediaBiasFactCheck.com` has an inherent bias.

We also computed the valence scores for the top-200 retweeted accounts, and we assigned each account a valence category based on the score. Independently, we asked a person who is well-versed with US politics to label all the accounts as left, center, or right. When labeling accounts, right-leaning include those expressing support for Trump, the Republican party, and gun rights, opposition to abortion, and disdain for Democrats.

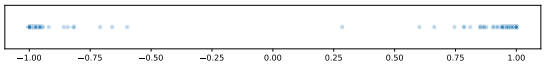

Figure 4: The top-200 retweeted accounts, projected on a number line according to their average valence.

As for left-leaning accounts, they include those attacking Trump and the Republicans, and expressing support for the Democratic party and for Liberal social positions. If the retweeted account happens to be a media source, we used `mediaBiasFactCheck.com`. Table 5 compares the per-topic valence for each retweeted account along with the average category and the true label.

It is noteworthy that all top-200 retweeted accounts have extreme valence categories on average across all topics. Their average valence scores, with one exception, appear between -0.6 and -1.00 for right, and between 0.6 and 1 for left (see Figure 4).

Of those manually and independently tagged accounts, all that were tagged as left-leaning have a strong positive valence score and all that were tagged as right-leaning have a strong negative valence score. Only two accounts were manually labeled as *center*, namely Reuters and CSPAN, which is a US channel that broadcasts Federal Government proceedings, and they had valence scores of 0.55 and 0.28, respectively. Though their absolute values are lower than those of all other sources, they are mapped to the $+$ valence category.

Table 3 summarizes the valence scores for the media across all topics. Table 4 lists the most cited media sources for each topic and for each of the five valence bands. The order of the bands from top to bottom is: $++$, $+$, 0, $-$ and $--$. The table also includes the credibility and the political leaning tags from `mediaBiasFactCheck.com`. The key observations from the table as follows:

1. Most right-leaning media appear overwhelmingly in the $-$ and $--$ valence categories. Conversely, left-leaning media appear in all valence categories, except for the $--$ category. This implies that left-leaning users cite right-leaning media sparingly. We looked at some instances where right-leaning users cited left-leaning media, and we found that in many cases the cited articles reinforced a right-leaning viewpoint. For example, right-leaning users shared a video from `thehill.com`, a left-center site, 2,398 times for the *police racism* topic. The video defended Trump against charges of racism by Lynne Patton, a long-time African-American associate of Trump.

Medium	factuality	bias	Average	climate change	gun control	ilhan	immigration	midterm	police & racism	Kavanaugh	vaccine
thehill.com	H	L-C	+	0	++	+	+	+	+	++	++
theguardian.com	H	L-C	++	++	++	++	++	++	++	++	++
washingtonpost.com	H	L-C	++	++	++	++	++	++	++	++	++
breitbart.com	VL	Far R	−−	−−	−−	−−	−−	−−	−−	−−	−−
foxnews.com	M	R	−−	−−	−−	−−	−−	−−	−−	−−	
nytimes.com	H	L-C	++	+	++	+	+	+	++	++	++
cnn.com	M	L	+	+	++	+	++	+	+	++	+
apple.news			+	0	0	+	0	0	+	+	++
dailycaller.com	M	R	−−	−−	−−	−−	−−	−−	−−	−−	
rawstory.com	M	L	++	++	++	++	++	++	++	++	++
huffingtonpost.com	H	L	++	++	++	++	++	+	++	++	++
truepundit.com	L		−−	−−	−−	−−	−−	−−	−−	−−	−−
nbcnews.com	H	L-C	+	−−	++	+	++	+	+	++	++
westernjournal.com	M	R	−−	−−	−−	−−	−−	−−	−−	−−	
reuters.com	VH	C	+	+	++	++	+	+	+	+	++
washingtonexaminer.com	H	R	−−	−−	−−	−−	−−	0	−−	−−	
thegatewaypundit.com	VL	Far R	−−	−−	−−	−−	−−	−−	−−	−−	
politico.com	H	L-C	+	+	+	+	+	++	+	+	++
npr.org	VH	L-C	+	0	++	++	++	0	++	++	++
townhall.com	M	R	−−	−−	−−	−−	−−	−−	−−	−−	−−
msn.com	H	L-C	+	+	+	+	0	++	0	++	0
nypost.com	M	R-C	−	−−	0	−	−	+	−−	−	
vox.com	H	L	++	++	++	++	++	++	+	++	++
thedailybeast.com	H	L	++	++	++	+	++	++	+	++	++
bbc.com	H	L-C	+	+	+	++	++	0	+	+	++
independent.co.uk	H	L-C	++	++	+	++	++	++	+	++	++
ilovemyfreedom.org	VL	Far R	−−	−−	−−	−−	−−	−−	−−	−−	
thinkprogress.org	M	L	++	++	++	++	++	++	++	++	++
dailywire.com	M	R	−−	−−	−−	−−	−−	−−	−−	−−	++
pscp.tv			−	−−	−−	−−	0	−−	0	−	
dailymail.co.uk	VL	R	−	−	0	−	−	−	−	−−	−−
msnbc.com	M	L	++	++	++	++	++	+	++	++	
dailykos.com	M	L	++	++	++	++	++	+	++	++	
bloomberg.com	H	L-C	+	+	++	0	++	+	0	+	++
usatoday.com	H	L-C	+	+	+	0	+	++	+	0	+

Table 3: Media valence categories for each topic with included average column. Plus (+) and minus (−) signify left or right leaning, respectively. Factuality: Very High (VH), High (H), Mixed (M), Low (L), Very Low (VL). Bias: Left (L), Left-Center (L-C), Center (C), Right-Center (R-C), Right (R), Far Right (Far R). Blank cells mean that we did not have information.

2. Most right-leaning sources in the −− category have mixed, low, or very low factuality. Conversely, most left-leaning sites appearing in the − valence category have high or very high factuality. Similarly for the vaccine topic, where high credibility sources, such as `fda.gov` and `nih.gov`, are frequently cited by anti-vaccine users, mostly to support their beliefs.

3. The placements of sources in different categories are relatively stable across topics. For example, `washingtonPost.com` and `theguardian.com` exclusively appear in the ++ category, while `breitbart.com` and `foxnews.com` consistently appear in the −− category.

6 Predicting Media Bias

Given the stances of users on the aforementioned eight topics, we leverage this information to predict media bias. Specifically, we describe in this section how we make use of the valence scores, as well as other features, namely graph and contextualized text embeddings, to train supervised classifiers for this purpose.

Valence Scores. We use valence scores in two ways. First, we average the corresponding valence across the different polarizing topics to obtain an average valence score for a given target news medium. This is an unsupervised method for computing polarity. Second, we train a Logistic Regression classifier that uses the calculated valence scores as features and annotations from `mediaBiasFactCheck.com` as gold target labels in order to predict the general political leaning of a target news medium. We merged "left" and "extreme left", and similarly we merged "right" and "extreme right". We discarded media labeled as being "left-center" and "right-center". Each news medium was represented by an 8-dimensional vector containing the valence scores for the above topics. In the experiments, we used the lbfgs solver and $C = 0.1$. We used two measures to evaluate its performance, namely accuracy and mean absolute error (MAE). The latter is calculated by considering the different classes as ordered and equally distant from each other, i.e., if the model predicts *right* and the true label is *left*, this amounts to an error equal to 2.

climate change			gun control			Ilhan Omar			immigration		
theguardian.com	H	L-C	thehill.com	H	L-C	washingtonpost.com	H	L-C	theguardian.com	H	L-C
washingtonpost.com	H	L-C	cnn.com	M	L	theguardian.com	H	L-C	washingtonpost.com	H	L-C
independent.co.uk	H	L-C	nytimes.com	H	L-C	mondoweiss.net	H	L	cnn.com	M	L
wef.ch			npr.org	VH	L-C	thinkprogress.org	M	L	huffingtonpost.com	H	L
vox.com	H	L	washingtonpost.com	H	L-C	haaretz.com	H	L-C	npr.org	VH	L-C
nytimes.com	H	L-C	politico.com	H	L-C	nytimes.com	H	L-C	thehill.com	H	L-C
bbc.com	H	L-C	usatoday.com	H	L-C	thehill.com	H	L-C	nytimes.com	H	L-C
cnn.com	M	L	msn.com	H	L-C	politico.com	H	L-C	reuters.com	VH	C
reuters.com	VH	C	bbc.com	H	L-C	cnn.com	M	L	politico.com	H	L-C
bloomberg.com	H	L-C	cnbc.com	H	L-C	apple.news			usatoday.com	H	L-C
thehill.com	H	L-C	apple.news			mediaite.com	H	L	apple.news		
apple.news			sun-sentinel.com	H	R-C	usatoday.com	H	L-C	msn.com	H	L-C
npr.org	VH	L-C	nypost.com	M	R-C	yahoo.com	M	L-C	pscp.tv		
seattletimes.com	H	L-C	dailymail.co.uk	VL	R	timesofisrael.com	H	L-C	whitehouse.gov	M	R
newsweek.com	M	L	mailchi.mp			theatlantic.com	H	L-C	texastribune.org	H	C
change.org	H	L	washingtontimes.com	H	R-C	nypost.com	M	R-C	dailymail.co.uk	VL	R
latimes.com	H	L-C	breaking911.com	VL		jpost.com	H	C	nypost.com	M	R-C
dailymail.co.uk	VL	R	chicagotribune.com	H	R-C	dailymail.co.uk	VL	R	zerohedge.com	M	
climatechangedispatch.com			rt.com	M	R-C	algemeiner.com	H	R-C	ir.shareaholic.com		
cnbc.com	H	L-C	forbes.com	M	R-C	startribune.com	H	L-C	breaking911.com	VL	
forbes.com	M	R-C	breitbart.com	VL	Far R	foxnews.com	M	R	breitbart.com	VL	Far R
breitbart.com	VL	Far R	foxnews.com	M	R	breitbart.com	VL	Far R	illegalaliencrimereport.com		
dailycaller.com	M	R	ammoland.com	H	R	townhall.com	M	R	washingtonexaminer.com	H	R
tambonthongchai.com			dailycaller.com	M	R	change.org	H	L	foxnews.com	M	R
wattsupwiththat.com	L		bearingarms.com	M	R	hannity.com			westernjournal.com	M	R

midterm			police & racism			Kavanaugh			vaccine		
washingtonpost.com	H	L-C	washingtonpost.com	H	L-C	thehill.com	H	L-C	thehill.com	H	L-C
theguardian.com	H	L-C	rawstory.com	M	L	washingtonpost.com	H	L-C	theguardian.com	H	L-C
rawstory.com	M	L	huffingtonpost.com	H	L	cnn.com	M	L	washingtonpost.com	H	L-C
tacticalinvestor.com			theguardian.com	H	L-C	nytimes.com	H	L-C	vaxopedia.org		
vox.com	H	L	nytimes.com	H	L-C	huffingtonpost.com	H	L	nytimes.com	H	L-C
thehill.com	H	L-C	thehill.com	H	L-C	politico.com	H	L-C	cnn.com	M	L
reuters.com	VH	C	apple.news			apple.news			statnews.com	H	C
nytimes.com	H	L-C	cnn.com	M	L	yahoo.com	M	L-C	latimes.com	H	L-C
cnn.com	M	L	nbcnews.com	H	L-C	apnews.com	VH	C	cbc.ca	H	L-C
dailykos.com	M	L	thedailybeast.com	H	L	latimes.com	H	L-C	usatoday.com	H	L-C
apple.news			msn.com	H	L-C	usatoday.com	H	L-C	cdc.gov	VH	
sagagist.com.ng			pscp.tv			mediaite.com	H	L	medium.com	M	L-C
bbc.com	H	L-C	bloomberg.com	H	L-C	theweek.com	H	L-C	newsroom.fb.com		
alzwaaj.com			politics.theonion.com			lawandcrime.com			help.senate.gov		
washingtonexaminer.com	H	R	rollcall.com	VH	C	cnbc.com	H	L-C	msn.com	H	L-C
dailymail.co.uk	VL	R	mediaite.com	H	L	pscp.tv			change.org	H	L
pbs.org	H	L-C	dailymail.co.uk	VL	R	nypost.com	M	R-C	fda.gov		
zerohedge.com	M		news.sky.com	H	L-C	ir.shareaholic.com			variety.com		
ajc.com	H	L-C	newsone.com	H	L-C	rollcall.com	VH	C			
veritablenouvelordre.forumcanada.org			aol.com	H	L-C	c-span.org	VH	C			
breitbart.com	VL	Far R	breitbart.com	VL	Far R	foxnews.com	M	R	ncbi.nlm.nih.gov	VH	
foxnews.com	M	R	defensemaven.io			truepundit.com	L		vaccineimpact.com		
dailycaller.com	M	R	foxnews.com	M	R	dailycaller.com	M	R	naturalnews.com	M	
ilovemyfreedom.org	VL	Far R	thegatewaypundit.com	VL	Far R	breitbart.com	VL	Far R	vaccines.me		
westernjournal.com	M	R	nypost.com	M	R-C	thegatewaypundit.com	VL	Far R	thevaccinereaction.org		

Table 4: Top 5 websites per valence category for each topic.

Account	Truth	Average	climate change	gun control	ilhan	immigration	midterm	police & racism	Kavanaugh	vaccine
realdonaldtrump	R	−−		0	−−	−−	−−	−−	−−	−−
charliekirk11	R	−−			−−	−−	−−	−−	−−	
kylegriffin1	L	++	++	++		++	++	++	++	++
dbongino	R	−−	−−	−−	−−	−−	−−	−−	−−	
kamalaharris	L	++	++	++		++	++	++	++	
mitchellvii	R	−−	−−	−−	−−	−−	−−	−−	−−	
realsaavedra	R	−−	−−	−−	−−	−−		−−	−−	
krassenstein	L	++		++	++	++	++	++	++	++
realjack	R	−−	−−	−−	−−	−−	−−	−−	−−	−−
nbcnews	L	++	++	++	+	++	++	++	++	++
education4libs	R	−−	−−	−−	−−	−−	−−	−−	−−	
nra	R	−−		−−		−−	−−		−−	
donaldjtrumpjr	R	−−		−−		−−	−−	−−	−−	
shannonrwatts	L	++		++	++		++	++	++	
thehill	L	++	++	++	+	++	+	+	++	++
realjameswoods	R	−−		−−	−−	−−	−−	−−	−−	
gopchairwoman	R	−−			−−	−−	−−		−−	
jackposobiec	R	−−	−−	−−	−−	−−	−−	−−	−−	
funder	L	++	++	++	++	++	++	++	++	
cnn	L	++	++	++	++	++	0	++	++	++
ajplus	L	++	++	++	++	++	++	++	0	++
rashidatlaib	L	++		++	++	++		++	+	
stevescalise	R	−−		−−		−−			−−	
jordan_sather_	?	−−	−−			−−	−−		−−	−−
aoc	L	++	++		++	++		++		

Table 5: User valence categories for each topic, preceded by an average column, and a ground truth label. When a cell is blank, there is insufficient data for that particular topic.

	No Valence		With Valence	
	Acc	MAE	Acc	MAE
Baseline 1 (majority class)	43.3	.856	43.3	.856
Baseline 2 (average valence)	–	–	68.0	.330
Valence scores	–	–	75.2	.278
BERT (article title)	60.6	.539	78.3	.264
BERT (article content)	61.1	.526	79.2	.255
BERT (title+content)	62.2	.510	80.8	.228
BERT(Tweet)	64.0	.485	73.6	.302
GraphEmbM	63.5	.468	69.1	.380
GraphEmbH	66.9	.425	71.8	.347
GraphEmbM+H	68.0	.400	79.0	.251
GraphEmbM+H+BERT (tweet)	72.5	.358	80.5	.230
GraphEmbM+H+BERT (tweet, content)	76.1	.311	81.2	.221
GraphM+H+BERT (tweet, title, content)	**78.1**	**.284**	**82.6**	**.206**

Table 6: Predicting media bias.

The results are shown in Table 6, where we can see that using the average valence score yields 68.0% accuracy (0.330 MAE) compared to 75.2% accuracy (0.278 MAE) when using the eight individual valence scores as features.

Graph embeddings. We further use graph embeddings, generated by building a User-to-Hashtag graph (U2H) and a User-to-Mention (U2M) graph and then running node2vec on both (Atanasov et al., 2019), producing two types of graph embeddings. When using graph embeddings, we got worse results compared to our previous setup with valence scores (see Table 6). However, when we combine them with the valence scores, we observe a sizable boost in performance, up to 11% absolute.

Tweets. We also experimented with BERT-base. We used the text of the tweets that cite the media we are classifying. For classification, we fed BERT representations of tweets to a dense layer with softmax output to fine-tune it with the textual contents of the tweets. We trained at the tweet level, and we averaged the scores (from softmax) for all tweets from the same news medium to obtain an overall label for that news medium. The accuracy is much lower than for the valence scores: 64.0% accuracy vs. 75.2% for supervised and 68.0% for unsupervised.

Article titles and text. Using the BERT setup for **Tweets**, we used the titles and the full text of up to 100 articles from each of the target media. When using the full text of articles, we balanced the number of articles per news medium. We trained two separate BERT models, one on the titles and another one on the full text (content). Both models did worse than using valence alone, but the combination improved over valence only.

System Combination. We combined different setups including using all the aforementioned models in combination. Using graph embeddings (GraphH + GraphM) with BERT embeddings (Tweet+Title+Content) and valence yielded the best results with accuracy of 82.6% and MAE of .206. If we remove valence from the combination, the accuracy drops by 4.5% while MAE jumps by .078, absolute. This suggests that valence is a very effective feature that captures important information, complementary to what can be modeled using graph and contextualized text embeddings.

7 Conclusion and Future Work

We have presented a method for predicting the general political leaning of media sources and popular Twitter users, as well as their stances on specific polarizing topics. Our method uses retweeted accounts, and a combination of dimensionality reduction and clustering algorithms, namely UMAP and Mean Shift, in order to produce sets of users that have opposing opinions on specific topics. Next, we expand the discovered sets using supervised learning that is trained on the automatically discovered user clusters. We are able to automatically tag large sets of users according to their stance of preset topics. Users' stances are then projected to the influencers that are being cited in the tweets for each of the topics using the so-called *valence score*. The projection allows us to tag a large number of influencers with their stances on specific issues and with their political leaning in general (i.e., *left* vs. *right*) with high accuracy and with minimal human effort. The main advantage of our method is that it does not require manual labeling of entity stances, which requires both topical expertise and time. We also investigated the quality of the valence features, and we found that valence scores help to predict media bias with high accuracy.

In future work, we plan to increase the number of topics that we use to characterize media. Ideally, we would like to automatically identify such polarizing topics. Doing so would enable us to easily retarget this work to new countries and languages.

Acknowledgments

This research is part of the Tanbih project[1], which aims to limit the effect of "fake news," propaganda and media bias by making users aware of what they are reading.

[1] http://tanbih.qcri.org/

References

Amjad Abu-Jbara, Ben King, Mona Diab, and Dragomir Radev. 2013. Identifying opinion subgroups in Arabic online discussions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 829–835, Sofia, Bulgaria.

Jisun An, Meeyoung Cha, Krishna Gummadi, Jon Crowcroft, and Daniele Quercia. 2012. Visualizing media bias through Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, Dublin, Ireland, pages 2–5.

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.

Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. 2013. Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 49–58, Atlanta, GA, USA.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91.

Pablo Barberá and Gaurav Sood. 2015. Follow your ideology: Measuring media ideology on social networks. In *Proceedings of the Annual Meeting of the European Political Science Association*, Vienna, Austria.

Pablo Barber and Gonzalo Rivero. 2015. Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 33(6):712–729.

Adam Bermingham and Alan Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology*, SAAIP '11, pages 2–10, Chiang Mai, Thailand.

Javier Borge-Holthoefer, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. Content and network dynamics behind Egyptian political polarization on Twitter. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 700–711, Vancouver, BC, Canada.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy! In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM '13, pages 91–99, Cambridge, MA, USA.

Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, 64(2):317–332.

Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011a. Political polarization on Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 89–96, Barcelona, Spain.

Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011b. Predicting the political alignment of Twitter users. In *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 192–199, Boston, MA, USA.

Kareem Darwish. 2018. To Kavanaugh or not to Kavanaugh: That is the polarizing question. *arXiv preprint arXiv:1810.06687*.

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.

Kareem Darwish, Walid Magdy, Afshin Rahimi, Timothy Baldwin, and Norah Abokhodair. 2018. Predicting online islamophobic behavior after #ParisAttacks. *The Journal of Web Science*, 4(3):34–52.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Improved stance prediction in a user similarity feature space. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, pages 145–148, Sydney, Australia.

Yajuan Duan, Furu Wei, Ming Zhou, and Heung-Yeung Shum. 2012. Graph-based collective classification for tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2323–2326, Maui, HI, USA.

James H Fowler, Michael T Heaney, David W Nickerson, John F Padgett, and Betsy Sinclair. 2011. Causality in political networks. *American Politics Research*, 39(2):437–480.

Matthew Gentzkow and Jesse M Shapiro. 2011. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.

Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, IJCNLP '13, pages 1348–1356, Nagoya, Japan.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 751–762, Doha, Qatar.

Itai Himelboim, Stephen McCreery, and Marc Smith. 2013. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, 18(2):40–60.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '17, pages 427–431, Valencia, Spain.

Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. 2016a. #isisisnotislam or #deportallmuslims?: Predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 95–106, Hannover, Germany.

Walid Magdy, Kareem Darwish, and Ingmar Weber. 2016b. #FailedRevolutions: Using Twitter to study the antecedents of ISIS support. *First Monday*, 21(2).

Aibek Makazhanov, Davood Rafiei, and Muhammad Waqar. 2014. Predicting political preference of Twitter users. *Social Network Analysis and Mining*, 4(1):1–15.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, pages 31–41, San Diego, CA, USA.

Jonathan Scott Morgan, Cliff Lampe, and Muhammad Zubair Shafiq. 2013. Is news sharing on Twitter ideologically biased? In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW 13, pages 887–896, San Antonio, TX, USA.

Marco Pennacchiotti and Ana-Maria Popescu. 2011a. Democrats, Republicans and Starbucks afficionados: user classification in Twitter. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 11, pages 430–438, San Diego, CA, USA.

Marco Pennacchiotti and Ana-Maria Popescu. 2011b. A machine learning approach to Twitter user classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 281–288, Barcelona, Spain.

Ferran Pla and Lluís-F. Hurtado. 2014. Political tendency identification in Twitter using sentiment analysis techniques. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pages 183–192, Dublin, Ireland.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, SMUC '10, pages 37–44, Toronto, ON, Canada.

Filipe N Ribeiro, Lucas Henrique, Fabricio Benevenuto, 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.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, AXLL-IJCNLP '15, pages 116–125, Beijing, China.

Amine Trabelsi and Osmar R Zaïane. 2018. Unsupervised model for topic viewpoint discovery in online debates leveraging author interactions. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, ICWSM '18, pages 425–433, Stanford, CA, USA.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 186–196, Baltimore, MD, USA.

Ingmar Weber, Venkata R. Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. Islamist polarization in Egypt on Twitter. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 290–297, Niagara, ON, Canada.

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.