BiasWatch: A Lightweight System for Discovering and Tracking Topic-Sensitive Opinion Bias in Social Media

Haokai Lu
Texas A&M University
hlu@cse.tamu.edu

James Caverlee
Texas A&M University
caverlee@cse.tamu.edu

Wei Niu
Texas A&M University
weiniu.2010@gmail.com

ABSTRACT

We propose a lightweight system for (i) semi-automatically discovering and tracking bias themes associated with opposing sides of a topic; (ii) identifying strong partisans who drive the online discussion; and (iii) inferring the opinion bias of "regular" participants. By taking just two hand-picked seeds to characterize the topic-space (e.g., "pro-choice" and "pro-life") as *weak labels*, we develop an efficient optimization-based opinion bias propagation method over the social/information network. We show how this approach leads to a 20% accuracy improvement versus a next-best alternative for bias estimation, as well as uncovering the opinion leaders and evolving themes associated with these topics. We also demonstrate how the inferred opinion bias can be integrated into user recommendation, leading to a 26% improvement in precision.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

Keywords

opinion analysis; bias; social media

1. INTRODUCTION

Social media has increasingly become a popular and important platform for "regular" people to express their opinions, without the need to rely on expensive and fundamentally limited conduits like newspapers and broadcast television. These opinions can be expressed on a variety of themes including politically-charged topics like abortion and gun control as well as fun (but heated) rivalries like android vs. iOS and Cowboys vs. 49ers. Our interest in this paper is in creating a flexible tool for discovering and tracking the themes of opinion bias around these topics, the strong partisans who drive the online discussion, and the degree of opinion bias of "regular" social media participants, to determine to what degree particular participants support or oppose a topic of interest.

However, assessing topic-sensitive opinion bias is challenging. First, the opinion bias of "regular" users may not be as pronounced as prominent figures, so discerning this bias will require special care. Second, how opinion bias manifests will inevitably vary by topic, so a system should be adaptable to each topic. Third, the themes by which people express their opinions may change over time depending on the circumstances (e.g., gun control debates may take different forms based on the ebb and flow of elections, recent shooting incidents, and so forth). As a result, assessing bias should be adaptive to these temporal changes.

Hence in this paper, we develop a lightweight system – BiasWatch – for discovering and tracking opinion bias in social media. BiasWatch begins by taking just two hand-picked seeds to characterize the topic-space (e.g., "pro-choice" and "pro-life" for abortion) as *weak labels* to bootstrap the opinion bias framework. Concretely, we leverage these hand-picked seeds to identify other emerging (and often unknown) themes in social media, reflecting changes in discourse as new arguments and issues arise and fade from public view (e.g., an upcoming election, a contentious news story). We propose and evaluate two approaches for expanding the hand-picked seeds in the context of Twitter to identify supporting and opposing hashtags – one based on co-occurrence and one on signed information gain. We use these discovered hashtags to identify strong topic-based partisans (what we dub *anchors*). Based on the social and information networks around these anchors, we propose an efficient opinion-bias propagation method to determine user's opinion bias – based on both content and retweeting similarity – and embed this method in an optimization framework for estimating the topic-sensitive bias of social media participants. In summary, this paper makes the following contributions:

- First, we build a systematic framework – BiasWatch – to discover biased themes and estimate user-based opinion bias quantitatively under the context of controversial topics in social media. We propose an efficient optimization scheme – called User-guided Opinion Propagation [UOP] – to propagate opinion bias. By feeding just two opposing hashtags, the system can discover bias-related hashtags, find bias anchors, and assess the degree of bias for "regular" users who tweet about controversial topics.
- Second, we evaluate the estimation of users' opinion bias by comparing the quality of the proposed opinion bias approach versus several alternative approaches over multiple Twitter datasets. Overall, we see a significant improvement of 20.0% in accuracy and 28.6% in AUC on average over the next-best method.
- Third, we study the effect of different approaches for biased theme discovery to measure the impact of newly discovered biased hashtags as additional supervision. We observe that the newly discovered hashtags are often associated with the underlying community of similar opinion bias, and that they temporally fluctuate due to the impact of new controversial events.
- Finally, we demonstrate how these inferred opinion bias scores can be integrated into user recommendation by giving similar-minded users a higher ranking. We show that the integration

CIKM'15, October 19–23, 2015, Melbourne, VIC, Australia.
© 2015 ACM. ISBN 978-1-4503-3794-6/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806416.2806573.

can improve the recommendation performance by 26.3% in precision@20 and 13.8% in MAP@20. This result implicitly confirms the principle of homophily in the context of opinion bias, and demonstrates how topic-sensitive opinion bias can enrich user modeling in social media.

2. RELATED WORK

There has been considerable research effort devoted to exploring political polarization, assessing media bias of major news outlets, and assessing user sentiment towards particular topics.

Political polarization. Political polarization has been a topic of great interest in the past decade and studied in news articles [33], online forums [3] and social media [1, 5, 6, 7, 11, 16, 23]. Adamic and Glance [1] demonstrated the divided community structure in the citation network of political blogs. Conover et al. [7] and Livne et al. [16] showed that there exists a highly segregated network structure using modularity. Guerra et al. [11] compared polarized and non-polarized networks and proposed a new measure to determine whether a network is polarized given that the network is also modular. Since knowing users' political orientation can be of great importance for understanding the overall political landscape, many approaches have been proposed to classify a user's political identity. Conover et al. [6] and Pennacchiotti and Popescu [23] exploit text and network features for classification. Akoglu [3] proposed to use signed bipartite opinion networks for the classification and ranking of user's political polarity on forum data. Zhou et al. [33] applied semi-supervised learning methods to classify news articles and users' political standing. Cohen et al. [5] employ supervised methods to classify political users into groups with different political activities, and conclude that it is hard to infer "ordinary" users' political orientation. In our work, instead of simply focusing on the classification of user's political orientation, we are interested in developing a flexible tool to explore controversial themes and discover their underlying users' degree of opinion bias on a topic basis. We show that user's opinion bias can be leveraged to improve other applications such as user recommendation.

Media bias. Apart from user-oriented political orientation, some works have explored media bias. Groseclose et al. [10] proposed a new measure to quantify media bias by comparing the number of citations to think tanks and policy groups to those of Congress members. Gentzkow et al. [9] also proposed a media bias measure which considers the frequency of phrases quoted by Congressional members of Republican and Democratic parties in newspapers. Lin et al. [15] focused on the measure of coverage quantity to compare the extent of bias between blogs and news media. Wong et al. [29] quantified the political leanings of media outlets on Twitter using aggregated retweeting statistics. Our work differs from these in that we target the opinion bias of "regular" users instead of prominent media, and with respect to different controversial topics instead of only political leanings.

User sentiment. There are also prior works which infer user's sentiment toward a topic in social media or online forums. Tan et al. [26] proposed a semi-supervised approach to inferring users' sentiment using social network information. Kim et al. [12] and Gao et al. [8] proposed to use a collaborative filtering like approach to estimate user-level sentiment. Lu et al. [17] proposed to use content and social interactions to discover opinion networks in forum discussions. However, our work has two differences from these and other sentiment-oriented approaches. The first is that many of these works require a significant amount of manually labelled tweets or users as ground truth. In our work, we develop automatic approaches using crowdsourced hashtags as seeds to substantially reduce manual labor. The second is that we focus on intrinsic opinion bias instead of sentiment. Sentiment [22] centers around users' attitude or emotional state, usually reflected by the use of emotional words. However, opinion bias can also be reflected by the news or factual information she chooses to post, which may lack any prominent emotional words.

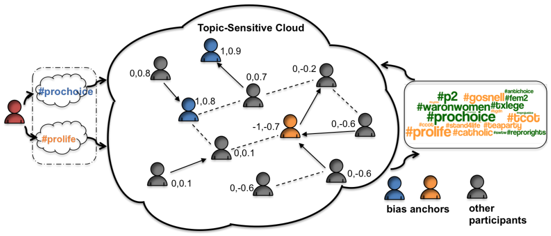

Figure 1: Overall BiasWatch Framework

3. LIGHTWEIGHT BIAS DISCOVERY

Problem Statement. We assume there exists a set of users $U = \{u_1, u_2, ..., u_n\}$ sampled from Twitter. Each user has their corresponding tweets $D = \{d_1, d_2, ..., d_n\}$ related to a controversial topic T, where d_i is a collection of tweets by u_i. Since a person's opinion bias represents the intrinsic tendency that she chooses to support or oppose a concept under a controversial context, we choose to quantize the degree of her opinion bias by a numeric score ranging from -1 to 1. Specifically, we assign $B = \{b_1, b_2, ..., b_n\}$ for each user in U, respectively, where $b_i \in [-1, 1]$. When b_i is close to 1, it denotes that user u_i has a strong positive standing toward the topic; when b_i is close to -1, it represents the opposite. Thus, given a controversial topic T, a sampled set of users U and their on-topic tweets D, we identify the following tasks of the system framework: (i) Discovering biased themes that are discussed by opposing sides of users. We denote P as the set of positive themes and N as the set of negative themes; (ii) Finding bias anchors who show strong degree of opinion bias, which we denote as U_{anchor}; (iii) Determining "regular" participants' opinion bias B.

Overall Approach. In order to tackle these tasks, we propose a lightweight framework that propagates opinion bias scores based only on a few hand-picked seeds that characterize the topic-space (e.g., "pro-choice" and "pro-life" for abortion). The BiasWatch framework, illustrated in Figure 1, takes as input these hand-picked seeds and then proceeds through the following three key steps:

- **Finding Bias Anchors.** This first step identifies topic-based partisans whose opinion bias is strongly revealed through their choice of hashtags. We develop two automatic approaches to: (i) identify biased themes in the form of hashtags through initial seeds; (ii) expand the pools of bias anchors with these identified biased themes.
- **Propagating Bias.** This second step builds a user similarity network around these expanded anchors and other "regular" participants, and propagates bias along this network. The edges here measure the similarities of two users through content and link features from tweets.
- **Noise-Aware Optimization.** Lastly, we propose to embed the previous two steps into a noise-aware optimization framework where anchors' opinion bias can be effectively propagated to each "regular" participant throughout the network. A key facet is that this optimization is tolerant of noisy labels on the initial bias anchors, so that initial errors made in identifying bias anchors need not lead to cascading errors in "regular" participants.

3.1 Finding Bias Anchors

Our first challenge is to identify strong topic-based partisans (what we dub *anchors*). These anchors serve as the basis for propa-

Table 1: Top ten themes at different times for "fracking" discovered by seed expansion; red for pro-fracking; blue for anti-fracking.

Dec 2012	Mar 2013	June 2013	Sept 2013
#shale	#dontfrackny	#shale	#balcombe
#natgas	#shale	#energy	#shale
#oil	#energy	#natgas	#frackoff
#energy	#oil	#oil	#energy
#gas	#gas	#gas	#greatgasgala
#tcot	#natgas	#banfracking	#natgas
#jobs	#frack	#fracked	#banfracking
#frack	#banfracking	#frack	#gas
#marcellus	#nokxl	#tcot	#oil
#naturalgas	#tcot	#dontfrackny	#frack

gating opinion bias throughout the social and information network. One reasonable method for identifying anchors is to manually label a number of users, among whom we hope that there exist a portion of users whose opinion bias is clearly shown. However, there are two disadvantages of this approach: (i) it is potentially expensive and time-consuming; and (ii) because of the random nature of labeling, typically, we have to label many users whose opinion bias is not clear or neutral in order to obtain anchors.

To overcome these difficulties, we propose to exploit crowd-generated hashtags in Twitter. Hashtags are often used to tag tweets with a specific topic for better discoverability or to indicate the community to which the tweets are posted [30]. Some hashtags may be viewed as a rich source of expressing opinions [28], potentially indicating user's opinion bias. For example, some most popular hashtags for "gun control" include #guncontrolnow and #2ndamendment, which reveal strong user bias, with the former often used by supporters of gun control and the latter by opponents. Hashtags of this nature can be essentially considered as *weak labels* of the polarity of tweets with respect to a controversial topic. Hence, to find bias anchors we first seek to identify a candidate set of hashtags that provide support for the topic and a candidate set of hashtags that express opposition. Since the hashtags used to express an opinion may change over time as new issues arise and fade from public view (e.g., an upcoming election, a contentious news story) and as new arguments are reflected in online discourse, we leverage these seeds to identify emerging themes in social media via seed expansion. To show a concrete example, Table 1 lists top ranking opposing and supporting themes at different times for the topic "fracking" after seed expansion is performed. We can see that new biased themes emerge as controversial events occur. In the following, we consider two approaches for expanding the hand-picked seeds to identify supporting and opposing hashtags:

Seed Expansion via Co-Occurrence. The first approach relies on hashtag co-occurrence statistics to find other hashtags used by users with similar opinion bias. The intuition is that if two hashtags are often used together by different users (whether in the same tweet or different tweets with respect to a topic), these two hashtags are likely to indicate the same opinion bias. Here, hashtag co-occurrence is based on users instead of tweets, considering that a user's opinion bias is not likely to change. Thus, for hashtags which occur in different tweets, as long as they are used by the same user, they are still considered to occur together.

Let f_i and f_j represent the frequency of the hashtag h_i and h_j related to a topic, respectively. Let f_+ and f_- represent the frequency of the pro-seed h_+ and the anti-seed h_-, respectively. The similarity between two hashtags, denoted as $\sigma(h_i, h_j)$, is computed by *Jaccard coefficient (JC)* as:

$$\sigma(h_i, h_j) = \frac{f_{h_i \cap h_j}}{f_{h_i \cup h_j}}$$

where $f_{h_i \cap h_j}$ represents the co-occurrence frequency of h_i and h_j, and $f_{h_i \cup h_j}$ represents the total occurrence frequency of either h_i or h_j. Thus, $\sigma(h_+, h_i)$ and $\sigma(h_-, h_i)$ represents the similarity of h_i to the pro-seed and anti-seed, respectively. We select top hashtags with the largest similarity to the pro-seed and anti-seed as the candidate set C_+ and C_-. However, since some hashtags, for example, "#guncontrol", are generic and co-occur with both the pro-seed and anti-seed, these hashtags do not indicate any opinion bias. To filter out common hashtags like these, we impose the following constraint on hashtag from C_+ for pro-seed:

$$\sigma(h_+, h_i)/\sigma(h_-, h_i) > \epsilon$$

where a large ϵ reflects more correlation with the pro-seed; similar constraint can also be imposed to filter hashtags in C_-. We then use the resulting m top hashtags for pro-seed as positive theme set P, and m top hashtags for anti-seed as negative theme set N.

Seed Expansion via Signed Information Gain. Although the approach above does find other biased hashtags, there are two disadvantages: (i) it often gives niche hashtags which are only used by a small number of participants; (ii) it often misses event-related short-lasting biased hashtags. In light of these issues, we propose the second approach which relies on weak supervision to select the most distinguishing hashtags for each side. Specifically, we perform the following procedure:

1. *Training with pro-seed and anti-seed.* First, we aggregate a user's tweets and use a bag-of-words model to compute TFIDF for each user. Users who have tweeted with at least one hashtag are then selected and used. From these users, we treat users with only pro-seed as positive class c_+, users with only anti-seed as negative class c_-, and the rest for prediction. Finally, an SVM classifier is learned on the training data and used to predict the polarity of users which are left. We now have an expanded set of users who are positive, and an expanded set of users who are negative.

2. *Selecting hashtags.* From the expanded sets of users, we use *signed information gain (SIG)* proposed by Zheng et al. [31] as the measure to select hashtags for pro-seed and anti-seed, respectively.

$$SIG(h_i, c) = sign(AD - BC) \cdot \sum_{c \in \{c_+, c_-\}} \sum_{h \in \{h_i, \bar{h}_i\}} p(h, c) \cdot log \frac{p(h, c)}{p(h) \cdot p(c)}$$

where A is the number of users with h_i and in class c_+, B is the number of users with h_i and in class c_-, C is the number of users without h_i and in class c_+ and D is the number of users without h_i and in class c_-. Also, $\bar{h}_i$ represents hashtags other than h_i. Here, the probability p is obtained by maximum likelihood estimation. We select m hashtags with the largest SIG as the finalized hashtag set P for pro-seed, and m hashtags with the smallest SIG as the finalized hashtag set N for anti-seed.

As a result, this approach can not only filter out common hashtags used by both sides of users without manually specifying any extra parameters, but also can discover popular yet distinguished biased hashtags.

Bias Anchors. Given the expanded set of hashtags (both supporting and opposing a particular topic), we identify as our strong partisans users who consistently adopt hashtags from only one opinion standpoint, which we denote as U_{anchor}. We assign an initial bias score $\tilde{b_i}$ to these anchors as follows:

$$\tilde{b_i} = \begin{cases} +1, & \text{if } u_i \in U_P \\ -1, & \text{if } u_i \in U_N \end{cases} \quad (1)$$

where U_P and U_N is the set of anchors adopting hashtags from P and N, respectively, and $U_{anchor} = U_P \cup U_N$. These opinion bias anchors serve as the basis for propagating bias throughout the social and information network, which we tackle next.

3.2 Bias Propagation Network

After the discovery of bias anchors, how do we determine the opinion bias of those remaining participants? We propose to build a propagation network where two users are only connected if their similarity passes a threshold. In the following, we adopt both content and link features to determine user similarity.

Content-Based Propagation. The assumption of content induced propagation is that if two users have a high textual similarity in their posts, it is likely that they may share similar opinion bias. To compute the content similarity of two users, we aggregate each user's topic-related tweets and treat each user as a document. Thus, content similarity of two users can be computed with document similarity. Here, we adopt cosine similarity of the TFIDF of the two documents with a standard bi-gram model. Tokenization of tweets is done through the tool provided by Owoputi et al. [21] for its robustness. Hashtags and mentions are also included in the model as features. To reduce the size of feature dimensions, we performed stop-word removal and kept only unigrams and bi-grams with occurrence frequency greater than two. The content similarity between u_i and u_j can be written as:

$$w_{ij}^{content} = \begin{cases} C_{ij}, & \text{if } u_j \in \mathcal{N}^c(u_i) \\ 0, & \text{if } u_j \notin \mathcal{N}^c(u_i) \end{cases} \tag{2}$$

where C_{ij} is the cosine similarity of u_i and u_j. To reduce the propagation complexity, we construct a sparse network by only considering k-nearest neighbors $\mathcal{N}^c(u_i)$ for each user u_i. We choose k to be 10 for its efficiency without compromising much accuracy in experiments.

Link-Based Propagation. Retweeting can be considered a form of endorsement for users in Twitter [4, 7]. Based on this observation, it is expected that if a user retweets another user on a topic, both users tend to share similar opinion bias. Thus, we define the link similarity between u_i and u_j as follows:

$$w_{ij}^{link} = \begin{cases} 1, & \text{if } u_j \in \mathcal{N}^l(u_i) \\ 0, & \text{if } u_j \notin \mathcal{N}^l(u_i) \end{cases} \tag{3}$$

where u_j is in the neighbors of u_i if u_j retweeted u_i or u_i retweeted u_j. Besides retweeting links, we can also take advantage of the follower/following network information in Twitter. Here, we choose not to use it since following link is not as strong a signal as retweeting. One reason is that a user whose opinion is on one side may choose to follow someone on the opposite side for the purpose of receiving any topic related statuses or refuting their arguments. Furthermore, we notice that the resource of retweeting activities is usually sparse so that the propagation network constituted only from retweeting links is separated into many isolated networks. Thus, a retweeting-based propagation is not enough to be treated by itself but needs to be combined with content-based propagation, leading to the following fusion of similarity between users:

$$w_{ij} = w_{ij}^{content} + \lambda w_{ij}^{link} \tag{4}$$

where λ is a weighting parameter for content and link similarity.

3.3 Optimization Framework

Finally, we embed the discovered anchors and bias propagation network into an optimization setting to propagate the opinion-bias score of all users more effectively. Since the initial input to the approach is a set of weak labels (two user-specified opposite hashtags), we call this optimization *User-guided Opinion Propagation [UOP]*. Specifically, by allowing each user's true opinion bias b_i to change as an optimization variable, we force the following conditions: (i) for bias anchors, b_i should be as close to the bias indicated by adopting biased hashtags (Eqn. 1); (ii) for other participants, b_i and b_j should be close to the degree indicated by their content and link similarity (Eqn. 4). Their opinion bias is initialized randomly between $[-1, 1]$ and can now be iteratively propagated through optimization. Thus, we have the following objective function:

$$\begin{aligned} \min_{b_i \in B} \quad & f = \sum_{u_i \in U_{anchor}} (b_i - \tilde{b_i})^2 + \mu_1 \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij}(b_i - b_j)^2 \\ \text{subject to} \quad & -1 \leq b_i \leq 1 \qquad \forall i \in \{1, ..., n\} \end{aligned} \tag{5}$$

where μ_1 is the tradeoff weight for different components. Thus, by solving this optimization, each user's opinion bias is propagated through the network in an optimized fashion. Since the objective function is convex, we can use the standard L-BFGS method with constraints to solve it efficiently. Another advantage of this framework is that other similarity signals such as location, profile demographics, and so on can be easily incorporated into Equation 5.

Handling Noisy Bias Anchors. The essence of the above optimization framework is that we propagate users' opinion bias which we know with confidence to other users who we have seldom knowledge of. We can see that anchor's opinion bias $\tilde{b_i}$ is treated as the golden truth and stays unchanged. However, a prominent issue is that users who consistently adopt hashtags from either P or N are not guaranteed to have the corresponding opinion bias. To give an example, one user's tweet reads, "#Gosnell certainly a tragedy, also cautionary tale, but not an argument against abortion rights in the US. Want more Gosnells? Ban abortion." This sarcastic pro-choice user adopted the hashtag #Gosnell, as we can observe from many tweets, is a primary hashtag pro-life users would use. Hence, according to Equation 1, this user is falsely identified as a pro-life anchor. Without manual inspection of user's profile and their related tweets, it is very hard to judge whether a bias anchor determined from Equation 1 is correctly identified. To relieve this problem, we introduce another variable y_i for $u_i \in U_{anchor}$ as the ideal opinion bias. Intuitively, it should satisfy: (i) y_i should be close to the opinion bias inferred from neighbors; (ii) most y_i should be consistent with $\tilde{b_i}$, with a few of them being noisy. Inspired by the annotation of noisy web images in [27], we propose a modified minimization function as follows:

$$\begin{aligned} \min_{b_i \in B} \quad & f = \sum_{u_i \in U_{anchor}} (b_i - y_i)^2 + \mu_1 \sum_{i=1}^{n} \sum_{j=i+1}^{n} w_{ij}(b_i - b_j)^2 \\ & \quad + \mu_2 \sum_{u_i \in U_{anchor}} |y_i - \tilde{b_i}| \\ \text{subject to} \quad & -1 \leq b_i \leq 1 \quad -1 \leq y_i \leq 1 \quad \forall i \in \{1, ..., n\} \end{aligned} \tag{6}$$

where μ_1 and μ_2 are the weighting parameters. We use l-1 norm to constrain the ideal variable y_i to $\tilde{b_i}$ since normally, only a small portion of bias anchors are noisy and l-1 norm could force most anchors to stay as biased. We solve the above minimization through the following steps:

(i) Initialize y_i to $\tilde{b_i}$ and solve Equation 5 to obtain b_i.

(ii) Solve the following sub-minimization problem with b_i to obtain y_i:

$$\begin{aligned} \min_{b_i \in B} \quad & f = \sum_{u_i \in U_{anchor}} (b_i - y_i)^2 + \mu_2 |y_i - \tilde{b_i}| \\ \text{subject to} \quad & -1 \leq y_i \leq 1 \qquad \forall i \in \{1, ..., n\} \end{aligned} \tag{7}$$

We employ the package L1 General [24] to solve the problem.

(iii) Replace $\tilde{b_i}$ with y_i in Equation 5 to get final b_i. We could repeat step (i) and (ii) for several times for further optimization but

Table 2: Datasets

Topic	Users	Tweets	Retweets
gun control	70,387	117,679	60,293
abortion	119,664	173,236	93,690
obamacare	67,937	123,320	70,008
vaccine	27,362	36,822	13,108
fracking	22,231	34,485	14,524

usually two to three iterations are enough shown by experiments. In this way, errors made in identifying bias anchors can be mitigated, leading to more accurate opinion bias estimation.

4. EXPERIMENTAL EVALUATION

In this section, we perform several sets of experiments to evaluate the BiasWatch framework for topic-sensitive opinion bias discovery. We investigate the impact of seed expansion, the quality of bias propagation via both content and retweeting links, and compare the performance versus alternative opinion bias approaches. We couple this study with an application of the system on two more controversial datasets.

4.1 Data

The datasets that we use are collected with Twitter's streaming API from October 2011 to September 2013. To create topic-related datasets for opinion discovery, we selected three controversial topics: “gun control", “abortion" and “obamacare". We select these topics because they are popular controversial topics discussed by a large number of Twitter users with both opposing sides of opinion expressed in the time period. For each topic, we extracted a base set of tweets (and their corresponding users) containing at least one topic-related keyword: for “gun control": *gun control, gun right, pro gun, anti gun, gun free, gun law, gun safety, gun violence*; for “abortion": *abortion, prolife, prochoice, anti-abortion, pro-abortion, planned parenthood*; and for “obamacare": *obamacare, #aca*. Additionally, we created another two datasets on the topics “vaccine” and “fracking” for demonstration. We select these two topics for further evaluation because they are relatively recent controversial topics compared to the previous ones, and also their opposing sides may not be fully entrenched in traditional left/right party politics. To extract “vaccine” related tweets, we use the following keywords: *vaccine, vaccination, vaccinate, #vaxfax*; for “fracking”, we use: *fracking, #frack, hydraulic fracturing, shale, horizontal drilling*. We summarize the datasets in Table 2.

4.2 Gathering Ground Truth

In order to evaluate the framework, we need to know the true opinion of a randomly sampled user set against which we can compare the optimization results. Without direct access to user's bias and considering the inherent difficulty of knowing a user's bias degree with respect to a controversial topic, we rely on an external labeling scheme using Amazon Mechanical Turk. Since the bias score obtained from Equation 5 is continuous, we discretize the opinion bias of a Twitter user into the following five categories: strong support [+2], some support [+1], neutral or no evidence [0], some opposition [-1], strong opposition [-2].

Thus, we can map the continuous range into the above categories for evaluation. For each topic, we randomly selected 504 Twitter users from the total users in Table 2, and assigned eight users in each human intelligence task (HIT), then ask the turkers (human labeler) to select the most appropriate category for these users. For each user, we show her twitter user ID and her topic related tweets for each turker to examine. We also highlight the hyperlinks embedded in the tweets and make them clickable. To ensure good quality of assessment, we follow the suggestions by Marshall and Shipman [18]. For each human intelligence task, we put two additional users in random positions, making a total of ten users in one HIT. Those users' bias are already known through experts, which we refer to as the golden users. If the label given by a turker for any of these golden users is very different from that by experts, we discard the entire answer by this turker for the HIT. Moreover, we ask five turkers to label one user and take the majority vote as the final label for the user. The results are shown in Table 3.

Table 3: Turker labeling results of HITs

topic	Number of users for each category				
	+2	+1	0	-1	-2
gun control	116	40	60	54	234
abortion	115	54	55	26	254
obamacare	82	26	33	26	337

Agreement of Opinion Bias Labels. To measure the reliability of the above human labeling tasks, we investigate the inter-rater agreement of the obtained assessment with Fleiss' κ statistic. Specifically, we obtained the 5-category κ statistic of 0.264, 0.393 and 0.418 for “gun control”, “abortion” and “obamacare”, respectively. These values lie in the interpretation of fair agreement by Landis and Koch [13]. In addition, we also adopt the accuracy of agreement provided by Nowak [20] and adapt it into the following formula since each user is assessed by more than two turkers:

$$\mathit{accuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{\# of votes of the majority category for user i}}{\text{\# of votes for user i}}$$

where N is the total number of users to be assessed by turkers for each topic. The accuracy ranges from 0.304 when the majority is obtained by chance to 1 when every user's bias category is agreed by all turkers. The lower bound 0.304 is obtained by calculating the average number of people in a majority out of five when they select one category from five by chance. Hence, an accuracy of 0.6, for example, means that on average, 3 out of 5 turkers agree on a category. The accuracy for “gun control”, “abortion” and “obamacare” is 0.646, 0.727 and 0.788, respectively, which means, on average, at least 3 turkers agree on the majority category.

Furthermore, we aggregated the same polarity into one category, namely, category [+1] and [+2] are combined to one category and vice versa. The 3-category κ statistic increases to 0.461, 0.588 and 0.649, correspondingly, while the accuracy increases to 0.811, 0.857 and 0.906. The κ values can now be interpreted as moderate agreement. The accuracy now means at least four out of five people agree on the majority bias polarity on average. This indicates that humans are more capable of discerning the polarity of users' opinion bias than determining the extent of users' bias.

In the following, we choose to use the 3 bias categories as ground truth since it has the most consistent and reliable performance by human labelers. We additionally consider only the support and opposition categories (ignoring the minority of users who are neutral or do not show evidence, namely, category [0]) so we can cast the evaluation as a binary class problem. We adopt the standard classification measures of accuracy and area under the curve (AUC).

4.3 Alternative Opinion Bias Estimators

To evaluate our approach of determining users' opinion bias, we consider the following alternative opinion bias estimators:

- **SentiWordNet [SWN]**. This is a simple sentiment detection approach, where we assign a sentiment score to each user's tweets according to SentiWordNet and classify each user's opinion bias with the relative portion of positive and negative tweets. Specifically, for each tweet d_i of user u_i, we classify it positive if

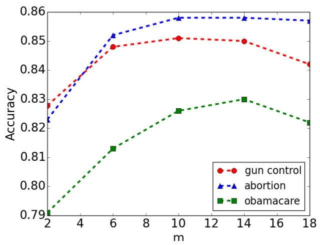

Figure 2: Effect of m for seed expansion via SIG.

$\sum_{w_j \in d_i} pos(w_j) > \sum_{w_j \in d_i} neg(w_j)$, and vice versa. We then classify user u_i as positive if the number of positive tweets is greater than the number of negative tweets, and vice versa.

- **User Clustering with Content [uCC]**. In this baseline, we construct a user graph and perform user clustering with tweets. The nodes of the graph are users and the edges are constructed based on k nearest neighbors with the largest content similarities of tweets. We choose to use cosine similarity of the TFIDF of bigrams as the similarity measure. To perform graph clustering, we apply normalized cuts for graph partitioning by Shi and Malik [25] due to its simplicity and good performance. This is essentially a max-cut problem explored in [2, 19] to partition newsgroup and online debates into opposite positions, respectively. The purpose of using this unsupervised baseline is to examine whether the selected biased hashtags as a form of weak supervision can provide much improvement.
- **User Clustering with Content and Links [uCCL]**. This baseline is the modified version of uCC in which the edge weight combines both content and link similarity. Specifically, if there exists a retweeting link between two users, we add a constant to its content similarity, i.e., $w = w^{content} + \theta * w^{link}$, where θ is used to balance the weights. The purpose of this baseline is to examine if retweeting links can help distinguish opposing sides of users' opinion bias compared to uCC.
- **Weakly-supervised SVM [wSVM]**. We train an SVM with a bi-gram model of bias anchors' tweets, which is then used to classify the test dataset. The parameters of SVM are determined through 5-fold cross-validation. We denote the trained classifier with bias anchors found through initial seeds (IS) as *wSVM+IS*, and the other two classifiers trained on seed expansions as *wSVM+JC* and *wSVM+SIG*. Here, wSVM+IS is treated as the baseline, and the other two as our improved versions.
- **Local Consistency Global Consistency [LCGC].** This is a semi-supervised method proposed by Zhou et al. [32] and applied in [33] for the classification of the political learning of news articles. This method optimizes the tradeoff between local consistency and global consistency among node labels. Here, we use Equation 4 as the affinity between nodes for the method and adapt LCGC into our own version by incorporating seed expansion from SIG, denoted as *LCGC+SIG*.
- **UOP***. This is the framework in which we only consider content based bias propagation without handling noisy bias anchors.
- **UOP**†. This is the framework in which we consider both content and link based bias propagation without handling noisy bias anchors, indicated by Equation 5.
- **UOP**. This is the full blown-approach indicated by Equation 6.

4.4 Biased Theme Discovery

Before experiments, we first select the following pro-seed and anti-seed manually as the input to the system: #guncontrolnow and #2ndamendment for "gun control"; #prochoice and #prolife for "abortion"; #ilikeobamacare and #defundobamacare for "obamacare". We later show in the experiments the effect of different seed selections.

We then highlight the seed expansion methods – both hashtag co-occurrence (JC) and signed information gain (SIG) – used to identify biased hashtags adopted by users with similar opinion bias. For seed expansion via SIG, we need to determine the value of parameter m. To study the influence of this parameter, we adopt the method UOP* to evaluate performance changes with values from $\{2, 6, 10, 14, 18\}$. Figure 2 shows that the accuracy is highest when m is approximately at 10 for "gun control" and "abortion", and is slightly larger for "obamacare". After those values, accuracy levels or even decreases, possibly because the additional discovered hashtags are noisy or do not imply much opinion bias. Thus, in the following experiments, m is fixed at 10 for all topics. For seed expansion via co-occurrence, we choose m to be 10 using the similar approach, and empirically set ϵ to 3. The expanded hashtags, as we observed from the output, can be approximately categorized as:

(i) *Sentiment-oriented.* These hashtags can be easily discerned and used directly by participants to show opinion bias, such as #nowaynra for "gun control" and #dontfundit for "obamacare".

(ii) *Community identification.* These hashtags indicate personal or political identities and are often used in a community, such as #p2, #tcot, #teaparty and #fem2 (for feminists);

(iii) *Thematic.* These hashtags often indicate arguments used by participants to express their opinion. For example, gun control antagonists say #gunrights as a constitutional right protected by #2ndamendment (also #2a); abortion protagonists may emphasize #reprorights or #reprojustice in their arguments;

(iv) *Action-oriented.* Examples include #momsdemandaction, #stand4life, #standwithcruz (stand with Ted Cruz), #swtw (stand with Texas women) and #demandaplan (as in "#demandaplan to end gun violence");

Overall, we can conclude that seed expansion through our proposed approaches is able to find other biased hashtags which are used by people with similar opinion bias. The above categorization also serves as guidance for users to pick initial opposing seeds as input to the system. Now that we have discovered the biased themes related to each side of polarity for different controversial topics, can we leverage those to determine "regular" participants' opinion bias? Specifically, we ask the following questions:

(i) Can these newly discovered biased hashtags help to identify user's opinion bias? If so, how much better can they do?

(ii) Do different pro-seed and anti-seed selections affect performance? If so, can seed expansion help us with the selection?

(iii) Can social ties, in the form of retweeting links, help us determine user's opinion bias?

Effect of Seed Expansion. To evaluate the performance of these expanded hashtags, we use wSVM and UOP* as the base methods since both of them only rely on the information provided by bias anchors and content. For UOP*, the weight parameter μ_1 in Equation 5, is empirically determined to be 0.1.

We now compare the performance between the version when opinion bias is propagated only through initial seeds and the version when the seeds are expanded. Since the result of accuracy is similar with AUC, we only show AUC in Figure 3. We can see that both seed expansion approaches outperform initial seeds, with seed expansion via SIG giving the best performance. Specifically, AUC from seed expansion via SIG gives a 19.5% and 6.8% improvement over IS and JC for wSVM, respectively, while the improvement is 12.5% and 5.6% for UOP*, respectively. This shows that (i) the newly discovered hashtag set through seed expansion provides additional amount of bias information for users; and (ii) the quality of

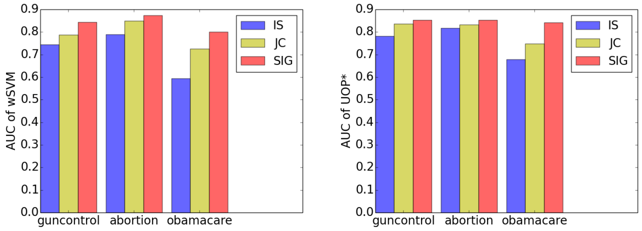

Figure 3: Effect of seed expansion.

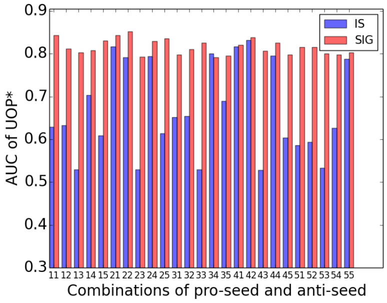

Figure 4: Performance for 25 pro-seed and anti-seed combinations.

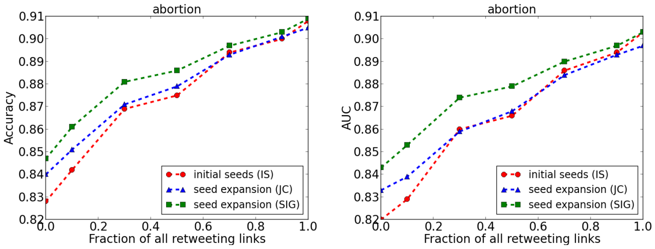

Figure 5: Performance for different seed expansion approaches with respect to different fraction of retweeting links for abortion.

expanded hashtag set via SIG is generally the best, i.e., it is able to discover higher quality of biased themes.

Effect of Different Seeds. Here, we are interested in studying the influence of different choices of initial seeds to the system. To that end, we first rank hashtags by occurrence frequency and then manually select top five pro-seeds and top five anti-seeds for "gun control" by observing tweets with those hashtags. The top five pro-seeds are: 1. #nowaynra; 2. #guncontrolnow; 3. #demandaplan; 4. #newtown; 5. #whatwillittake. The top five anti-seeds are: 1. #tcot; 2. #2ndamendment; 3. #nj2as; 4. #2a; 5. #gunrights. We represent a pro-seed and anti-seed combination with their corresponding number. For example, "12" represents the combination "#nowaynra #2ndamendment". We then use the method UOP* to evaluate the performance of each combination for two cases: one with initial seeds; the other with seed expansion via SIG. The results are shown in Figure 4. Overall, we can see that for different choices of input seeds, the performance without seed expansion is very sensitive to the choices, while it gives consistent high accuracy for seed expansion with SIG. We can also observe that #nj2as is not a good anti-seed choice since all combinations with #nj2as performs bad. Also, all combinations with #gunrights give unsatisfactory results except "55". These results indicate that it is difficult to choose the most effective seed combinations, since very often a single pair of seeds can be noisy and do not cover enough bias related themes. However, seed expansion can mitigate this effect by introducing other and often more complete biased hashtags for bias propagation. Hence, even though the initial seeds are not well selected, the final performance does not suffer much due to the benefits of seed expansion. For other topics, similar results are observed, and thus are not reported due to the space limit.

Effect of Social Ties. Here, we evaluate the effect of social ties in the form of retweeting links for different seed expansion approaches. The retweeting links are randomly sampled for ten times for each fraction since a different sample of retweeting links gives a different propagation network. Here, we adopt UOP† for the evaluation, with the weight given to the link-based propagation λ in Equation 4 empirically chosen to be 1. The final results are averaged and plotted in Figure 5. We only show the results for abortion since it is similar with other topics. We can see that as more retweeting links are added for propagation, the performance increases for all three approaches, which confirms the assumption that users linked via retweeting tend to share similar opinion bias.

Furthermore, among these three approaches, the performance obtained via SIG is always the best for different fractions of retweeting links at all topics. As the fraction gets larger, the improvement generally gets smaller. This indicates that the effect of seed expansion gets reduced when the fraction increases. For seed expansion via JC, the performance is generally better than that of initial seeds when the fraction is small. However, this initial improvement gets reduced or even disappears when retweeting link is abundant, hence, making the extra hashtags obtained via co-occurrence less effective. This is probably because of the noisiness of those extra hashtags. Though beneficial at a small number of retweeting links, they prevent the correct bias from being propagated when more of them are added for optimization. This shows again that the key of seed expansion is that not only more biased hashtags should be discovered, but also they should be of high quality.

4.5 Comparison with Baselines

In this section, we compare our proposed BiasWatch framework with the alternative opinion bias estimators. For uCCL, the weight-balancing parameter θ is selected from $\{0.05, 0.1, 0.15, 0.2, 0.25\}$ to be 0.1 for best performance. For LCGC+SIG, we use the same parameter setting as in [33]. Other parameter settings in UOP*, UOP† and UOP are the same as in previous experiments. For μ_2 in Equation 6, we empirically set it to 0.2 for handling noisy bias anchors. We choose the best seed expansion approach (via SIG) for all methods. The final results are shown in Table 4.

Overall, user-guided approaches give much better performance than unsupervised methods, indicating that by just a small amount of human guidance — two opposite seed hashtags, the performance can be boosted significantly. Moreover, UOP gives the best performance, reaching an average accuracy and AUC of 0.923 and 0.913, respectively (an improvement of 20.0% and 28.6% over supervised baseline wSVM+IS). Note that the sentiment based approach SWN gives unsatisfactory results, probably because user's opinion bias is multifaceted and can be reflected by the topical arguments or factual information published by the user. For example, one of the anti-obamacare tweets reads, "Double Down: Obamacare Will Increase Avg. Individual-Market Insurance Premiums By 99% For Men, 62% For Women. #Forbes". Also, we can see that UOP† gives better results than LCGC+SIG, indicating that our framework works better in capturing user's opinion bias. UOP, however, gives better performance than UOP†, confirming that initial bias anchors determined through biased hashtags are noisy, and that UOP is able to correct some wrongly determined bias anchors due to the l-1 norm regularization on ideal bias scores.

Table 4: Comparison of performance with alternative opinion bias estimators. Boldface: the best result for each topic among all methods. '$*$' marks statistically significant difference against the best of alternative opinion bias estimators (with two sample t-test for $p \leq 0.05$).

Method	Accuracy				AUC			
	gun control	abortion	obamacare	average	gun control	abortion	obamacare	average
SWN	0.560	0.527	0.465	0.517	0.570	0.531	0.541	0.547
uCC	0.534	0.537	0.516	0.529	0.533	0.527	0.522	0.527
uCCL	0.586	0.530	0.520	0.545	0.584	0.531	0.546	0.554
wSVM+IS	0.696	0.825	0.786	0.769	0.745	0.790	0.594	0.710
wSVM+SIG	0.860	0.884	0.727	0.824	0.844	0.874	0.800	0.839
UOP*	0.851	0.847	0.826	0.841	0.853	0.843	0.842	0.846
LCGC+SIG	0.858	0.900	0.811	0.856	0.857	0.900	0.864	0.874
UOP†	0.881	0.906	0.894	0.894	0.861	0.903	0.915	0.893
UOP	**0.908****	**0.915**	**0.945***	**0.923***	**0.883***	**0.910**	**0.945***	**0.913***

Furthermore, we see from the table that UOP† has an improvement of 0.053 and 0.047 for accuracy and AUC over UOP*, respectively. Compared to the corresponding improvement of 0.016 and 0.027 by uCCL over uCC, it is considerably higher. This indicates that retweeting links are more effective in contributing to bias propagation with the help of bias anchors. When some users are correctly "labeled" by discovered biased hashtags, opinion bias can be propagated more effectively through retweeting links.

4.6 Multi-Category Classification

In previous experiments, we mainly evaluate our framework as a binary class problem. However, in this way, we lose the finer granularity of the inferred opinion bias score by ignoring users who are neutral or do not show evidence. This category of users can not only be beneficial to our understanding of the general landscape of controversial topics, but also can be targeted by polarized activists to influence their bias. Thus, in this section, we are interested in a finer level of evaluation by casting it as a three-class problem. Here, instead of combining the category +1 and +2, and the category -1 and -2, we aggregate users in the category of +1, 0 and -1 into one neutral category 0 for a relatively larger pool of users in the middle. We then partition the datasets into 50% for training and 50% for testing, and compare the performance of wSWM+SIG, LCGC+SIG, UOP† and UOP. We are interested to see how these methods perform in the recall defined by $\frac{1}{3}\sum_i \frac{tp_i}{tp_i+fn_i}$. This measure indicates the ability of finding out the correct user category in a three-category setting for different methods.

For wSVM+SIG, we adopt the one-vs-rest multi-class implementation. For LCGC, we modified it to consider three labels. For UOP based methods, we select two optimal thresholds as boundaries: θ_p, used to distinguish category +2 and 0 and θ_n, used to distinguish category -2 and 0. These thresholds are selected by searching the range (-1,1) with a step of 0.05 with the training data. For UOP, θ_n and θ_p are determined as -0.25 and 0.2 for "gun control", -0.35 and 0.35 for "abortion", -0.2 and 0.2 for "obamacare". The final results are shown in Table 5. As we can see, UOP gives the best performance of all, indicating that it is the most capable in finding out users who are polarized and neutral. wSVM and LCGC perform worse, probably because of the small amount of training samples for category 0. UOP based methods are able to mitigate this effect by propagating constrained bias score without explicit consideration of neutral users and only with bias anchors. Thus, we can conclude that UOP is effective in representing the degree of user's opinion bias under the controversial topics.

4.7 Case Study: Fracking and Vaccines

Now that we have demonstrated the effectiveness of the system in finding general users' opinion bias, we would like to test the framework further in another two datasets: "fracking" and "vaccine". Since these two topics are relatively new compared to the above politically-driven well-known topics, it would be interesting to discover what is being posted and debated by opinionated users of both sides. For "fracking", we pick the seeds #dontfrackny and #jobs as the input to the system. We select #jobs as the pro-seed for "fracking" because protagonists tend to emphasize the benefit of creating jobs as one of the arguments for "fracking". For "vaccine", we pick the seeds #health and #autism as the input to the system. Again, we select #autism as the anti-seed because antagonists tend to focus the negative effect of vaccination. Overall, selection of input seeds is intuitive due to the recognizability of some of the crowd-generated tags and thus does not require much human effort.

We first demonstrate the biased themes discovered via SIG. Table 1 shows the top-ranking biased hashtags for different time periods. We can see that some themes such as #natgas and #frack remain constantly used by users supporting and opposing fracking, respectively; some other themes, such as #dontfrackny and #balcombe, arise and fade as related controversial events occur. For example, the occurrence of #dontfrackny corresponds with a rally of New Yorkers in the mid February 2013 to urge Governor Cuomo to resist fracking in the state of New York; the occurrence of #balcombe corresponds with a protest against a license to drill near Balcombe in England granted by Environment Agency. These changing themes indicate a strong degree of opinion bias and emerge as a group of users start to use them together, making them a very use-

Table 5: Multi-category classification performance.

Method	gun control	abortion	obamacare
wSVM+SIG	0.611	0.592	0.573
LCGC+SIG	0.632	0.657	0.646
UOP†	0.690	0.699	0.673
UOP	0.705	0.701	0.716

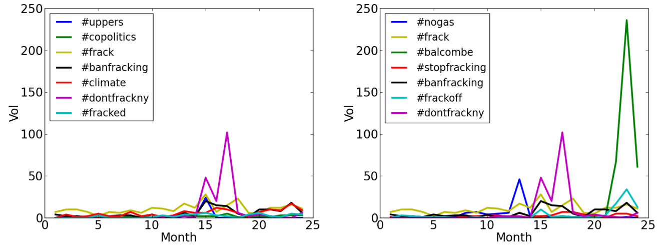

Figure 6: Temporal volumes of top anti-fracking themes for seed expansions via co-occurrence (Left) and via SIG (Right).

Table 6: Top ten themes at different times for “vaccine”; red for pro-vaccine; blue for anti-vaccine.

Feb 2012	June 2012	Nov 2012	Apr 2013
#vaxfax	#health	#health	#vaccineswork
#health	#vaxfax	#vaxfax	#health
#flu	#polio	#flu	#measles
#polio	#hpv	#autism	#mmr
#hpv	#vaccineswork	#news	#hiv
#autism	#pakistan	#polio	#autism
#thrillers	#autism	#hiv	#vaxfax
#measles	#news	#suspense	#polio
#action	#suspense	#thrillers	#flu
#flushot	#action	#hpv	#news

ful signal to determine and propagate user’s opinion bias. Furthermore, the transient property of these changing themes also makes the approach of seed expansion via co-occurrence less effective. Figure 6 illustrates the temporal characteristics of discovered anti-fracking themes for different seed expansion approaches. We can see that seed expansion via co-occurrence failed discovering #balcombe and #nogas, as both of these hashtags do not co-occur with #dontfrackny. In contrast, SIG tackles the problem by tapping into the power of content to discover more related biased themes. Table 6 shows the top-ranking biased hashtags at different time periods for the dataset “vaccine”. To illustrate how the system can uncover strong partisans, Table 7 shows two uncovered bias anchors – @Energy21, a pro-fracking account from the Institute for 21st Century Energy, and @Duffernutter, an anti-fracking account associated with TheEnvironmentTV.

5. INTEGRATING OPINION BIAS INTO USER RECOMMENDATION

In this section, we demonstrate the application of integrating opinion bias for user recommendation in social media. The principle of homophily, observed both in political blogosphere [1] and social media [7], states that people tend to associate with others who are like-minded. Thus, when social media services recommend users to follow, it is natural to consider recommending users who have similar opinion bias on shared topic interests. User recommendation can also be considered as a task of link prediction in graphs, and there has already exist many works [14] which specifically address this task, mostly taking advantage of graph structure. Here, our goal is to demonstrate how opinion bias can be utilized for user recommendation as a different dimension.

The task can be formally described as follows: Given a controversial topic T, a sampled set of users U and their corresponding on-topic tweets D, recommend k friends to follow for the target user. Note that we have only tweets to rely on to determine the recommendations. We provide two approaches for the task.

Content-Based approach. In this collaborative filtering approach, users with the highest content similarities to the target user are recommended. Specifically, we aggregated the corresponding tweets for each user and applied vector space model (VSM) with unigrams and bi-grams, with the similarity computed by cosine measure.

Opinion-Weighted (OW) approach. Here, user similarity is considered as a weighted sum of content similarity and opinion similarity. The content similarity is computed in the same way as VSM. The opinion similarity is obtained as follows. First, user’s opinion bias score is obtained through UOP. Then, we characterize the opinion similarity of two users as:

$$sim_{opi}(u_i, u_j) = 1 - |f(b_i) - f(b_j)|$$

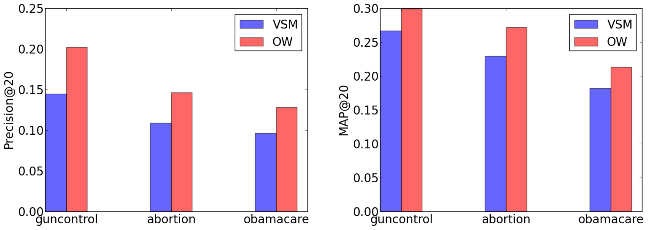

Figure 7: Performance comparison for VSM and OW.

where $f(x)$ is a normalizing function with the form as $\frac{1}{1+e^{-c \cdot x}}$. The value of parameter c should be chosen to make b’s transition to opposite sign steep enough, so that two bias scores with small difference but opposite signs can still result in large difference after transformation. Also, when b_i and b_j have the same sign, they can have a relatively large opinion similarity. To this end, c is chosen to be 5. The final similarity is computed as the weighted sum (α is the weighting parameter), which we use for ranking users:

$$sim(u_i, u_j) = \alpha sim_{vsm}(u_i, u_j) + (1 - \alpha) sim_{opi}(u_i, u_j)$$

5.1 Evaluating User Recommendation

For evaluation, we additionally crawled following links for the dataset “gun control”, “abortion” and “obamacare”. These following links are used as the ground truth in our experiments. We randomly sampled 500 users who have at least 20 followees for each topic and used the following metrics to evaluate: (i) precision@K: which measures the percent of the correct followees out of the top K recommended users; and (ii) mean average precision@K: which is the average of the precision at the position of each correct followee out of the top K recommended users. This measure considers the positions of the recommended users. K is chosen to be 20.

Figure 7 shows the performance comparisons between the two approaches. Here, α is set to 0.5. For both metrics, the OW approach has a better performance than the vanilla VSM for each topic. On average, it gives an improvement of 26.3% in precision@20 and 13.8% in MAP@20. These results indicate that user’s opinion similarity boosted the rank of some of the true followees who have similar opinion bias as the target user, while lowering the similarity with users who hold different opinion. Hence, it implicitly confirms the principle of homophily that people tend to make friends who share similar opinions.

In Figure 8, we also show the performance at different values of α. As we can see, the best performance is achieved neither at $\alpha = 0$ or 1 for all topics, but at a mixed weight between content and opinion similarity. Even when the weight given to opinion similarity is small, i.e., when $\alpha = 0.9$, the improvement can reach 20.2% for precision@20 and 9.2% for MAP@20 on average. The figure shows that the performance is not very sensitive to the value of α when α is approximately in the range of (0.2, 0.8), indicating it does not require fine tuning to reach a better performance.

6. CONCLUSION AND NEXT STEPS

We have seen how the BiasWatch system can lead to an improvement of 20% in accuracy over the next-best alternative for bias estimation, as well as uncover opinion leaders and bias themes that may evolve over time. We also demonstrated how the inferred opinion bias can be integrated into user recommendation and showed that it gives a performance improvement of 26% in precision. While our investigation has focused on textual and relational features, it does not limit us from integrating new signals such as location and profile demographics for better performance in future work. We are also interested in incorporating opinion bias into a

Table 7: Sample opinionated users and their corresponding tweets for "fracking"; positive bias score represents pro-fracking.

users	bias anchor	bias score	tweets
@Energy21	yes	0.91	Four real-life examples of how #shale #energy is creating #jobs and improving the economy: via @FreeEnterprise.
@Duffernutter	yes	-0.92	RT @ABFalecbaldwin: Together we can help @NYGovCuomo see there is no place for #fracking in NY. RT to let him know #DontFrackNY.
@IntellisysUK	no	0.15	RT @JimWoodsUK: Provocative article from @James_BG on New Environmentalism. Worth remembering fracking has reduced US CO2 emissions.
@janet_ewan	no	-0.39	RT @IshtarsGate: A 5.7-magnitude earthquake linked to fracking in Oklahoma in 2011 knocked down 14 homes and injured two people.

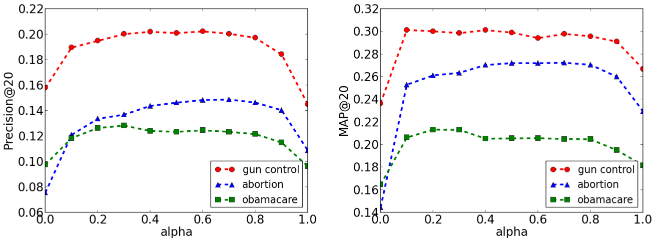

Figure 8: Performance at different values of parameter α.

social media dashboard, so that participants can be aware of their own opinion dynamics as well as those of others.

7. ACKNOWLEDGMENTS

This work was supported in part by AFOSR grant FA9550-12-1-0363 and NSF grant IIS-1149383. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] L. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: Divided they blog. In *LinkKDD*, 2005.

[2] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *WWW*, 2003.

[3] L. Akoglu. Quantifying political polarity based on bipartite opinion networks. In *ICWSM*, 2014.

[4] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS*, 2010.

[5] R. Cohen and D. Ruths. Classifying political orientation on twitter: It's not easy! In *ICWSM*, 2013.

[6] M. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of twitter users. In *SocialCom*, 2011.

[7] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on twitter. In *ICWSM*, 2011.

[8] H. Gao, J. Mahmud, J. Chen, J. Nichols, and M. Zhou. Modeling user attitude toward controversial topics in online social media. In *ICWSM*, 2014.

[9] M. Gentzkow and J. Shapiro. *What Drives Media Slant? Evidence from US Daily Newspapers*. National Bureau of Economic Research Cambridge, Mass., USA, 2006.

[10] T. Groseclose and J. Milyo. A measure of media bias. *The Quarterly Journal of Economics*, 2005.

[11] P. Calais Guerra, W. Meira Jr., C. Cardie, and R. Kleinberg. A measure of polarization on social media networks based on community boundaries. In *ICWSM*, 2013.

[12] J. Kim, J. Yoo, H. Lim, H. Qiu, Z. Kozareva, and A. Galstyan. Sentiment prediction using collaborative filtering. In *ICWSM*, 2013.

[13] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 1977.

[14] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 2007.

[15] Y. Lin, J. Bagrow, and D. Lazer. Quantifying bias in social and mainstream media. *SIGWEB Newsl.*, 2012.

[16] A. Livne, M. Simmons, E. Adar, and L. Adamic. The party is over here: Structure and content in the 2010 election. In *ICWSM*, 2011.

[17] Y. Lu, H. Wang, C. Zhai, and D. Roth. Unsupervised discovery of opposing opinion networks from forum discussions. In *CIKM*, 2012.

[18] C. Marshall and F. Shipman. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *WebSci*, 2013.

[19] A. Murakami and R. Raymond. Support or oppose?: Classifying positions in online debates from reply activities and opinion expressions. In *COLING*, 2010.

[20] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *MIR*, 2010.

[21] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *NAACL*, 2013.

[22] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2008.

[23] M. Pennacchiotti and A. Popescu. Democrats, republicans and starbucks afficionados: User classification in twitter. In *SIGKDD*, 2011.

[24] Mark Schmidt. L1 general, May 2015.

[25] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2000.

[26] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *SIGKDD*, 2011.

[27] J. Tang, R. Hong, S. Yan, T. Chua, G. Qi, and R. Jain. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM TIST*, 2011.

[28] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *CIKM*, 2011.

[29] F. Wong, C. Tan, S. Sen, and M. Chiang. Quantifying political leaning from tweets and retweets. In *ICWSM*, 2013.

[30] L. Yang, T. Sun, M. Zhang, and Q. Mei. We know what@ you# tag: does the dual role affect hashtag adoption? In *WWW*, 2012.

[31] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. *SIGKDD Explor. Newsl.*, 2004.

[32] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2004.

[33] X. Zhou, P. Resnick, and Q. Mei. Classifying the political leaning of news articles and users from user votes. In *ICWSM*, 2011.