Do the Communities We Choose Shape our Political Beliefs? A Study of the Politicization of Topics in Online Social Groups

Benjamin Kane[1] and Jiebo Luo[2]

***Abstract*— Social media has become a ubiquitous part of the lives of many people, and provides a channel for ordinary people to voice and to hear political opinions. However, many believe that the rise of social media has lead to an increasing polarization of political views, with political bias becoming intertwined with seemingly non-political interests and behavior. In this paper, we aim to use natural language processing techniques to analyze the political bias of online social groups and the degree to which this bias correlates with non-political topics. Whereas this phenomenon has been studied extensively on networks such as Twitter, the popular social media domain Reddit has been relatively unexplored despite the structure of the platform allowing users to easily create their own interest-based subcommunities, thus providing an important and unique data source relevant to the topic at hand. This paper analyzes comment data from approximately 3,300 message boards on Reddit, with the goal of providing novel empirical knowledge about how group topic informs political bias in Reddit subcommunities. Topics of Reddit subcommunities are determined using a Latent Dirichlet Allocation model, and political bias of subcommunities is measured through direct comparison with external corpora of politically biased vocabulary; results are discussed within. Furthermore, to test the politicization of topics, we train a classifier on topic features to obtain a high accuracy of 85.2% (compared to a random-guessing baseline of 64.8%), suggesting fairly strong correlations between group topic and politically biased language in communities.**

I. INTRODUCTION

With the rising prevalence of the media in modern culture, many have voiced concern about the formation of online "echo chambers", where users are selectively exposed to information which aligns with their political opinions (Barber et al. 2015; Mutz and Martin 2001). Indeed, the tendency of people to divide themselves along political beliefs on internet blogs, news channels, and social media sites such as Twitter has been extensively documented (Adamic and Glance 2005; Iyengar and Hahn 2009; Yardi and Boyd 2010) although it has also been observed that rather than being systematically adverse to confrontational opinions, the phenomenon of selective exposure is primarily driven by a desire for opinion-reinforcement (Garrett 2009a; Garrett 2009b).

This latter tendency is related to *homophily*, the tendency for individuals to associate with similar others. Homophily has been shown to be a defining factor in the separation of individuals online into different social groups, based on both political factors (Colleoni, Rozza, and Arvidsson 2014) as well as non-political factors (McPherson, Smith-Lovin, and Cook 2001). Furthermore, an experimental study based on a social news aggregation site indicated a significant "herding effect" in how positive/negative comment votes were received, in which an arbitrary positive initial vote would lead to inflated subsequent scores (Muchnik, Aral, and Taylor 2013).

Another similar phenomenon linked with social media is the tendency for online social groups and subcommunities to form slight variations in language structure and vocabulary. For instance, a study of social cliques on IRC channels found a relationship between strong social network ties of a group and the use of in-group vernacular language variants (Paolillo 2001). Social bonds online are often communicated through "ambient affiliation", where users share similar interests and values through variations in language, in particular, "the emerging searchable talk of microblogging" (Zappavigna 2012). Of particular focus to us are the categories of "social media memes", which are certain words or phrasal templates particular to a group, as well as the use of political terminology to express group affiliation (Zappavigna 2012). Studies on the nature of linguistic change in online communities have shown that users tend to enter a distinct stage in which they learn and adopt the language of that community, followed by a conservative stage in which the user stops adapting and possibly loses track of community norms (Danescu-Niculescu-Mizil et al. 2013).

Previous research on congressional speeches over time has also indicated significant differences in the terminology and phrases used by Democrat and Republican candidates. A small sample of these partisan phrasal differences are shown in Table 1 (Gentzkow, Shapiro, and Taddy 2017).

TABLE I

A SAMPLE OF PARTISAN PHRASES USED BY CANDIDATES

Democrat	Republican
depart homeland	illegal immigrants
gun violence	mental health
african american	radical islam
climate change	american energy
affordable care	taxpayer dollar
voting rights	religious freedom

In this paper, we study the role that non-political interests play in political homophily with the aim of contributing new empirical knowledge about political bias across non-political social groups in the Reddit domain, as well as analyzing the extent to which these topics become associated with political bias. While this topic has been studied extensively using

[1]B. Kane is with the Department of Computer Science, University of Rochester, Rochester, NY 14627, USA.

[2]J. Luo is with Faculty of Computer Science, University of Rochester, Rochester, NY 14627, USA. http://www.cs.rochester.edu/u/jluo/

Twitter social networks, networks such as Reddit have gone relatively unexplored. This is unfortunate, as Reddit provides a high-potential dataset due to the fact that any user in Reddit is able to create their own *subreddit*, resulting a large set of groups, representing a diverse set of people and interests, with which one can study group dynamics. In particular, Reddit offers an important source of new political bias data with unique measurement challenges to overcome.

In contrast with previous Twitter research, in which "ground truth" classifications for a set of users can be obtained from domain knowledge (for instance, by determining which politicians a user subscribes to), determining political bias on Reddit is considerably more difficult as individual users may interact with a number of groups without necessarily agreeing with the predominant biases of that group. Hence, determining the latent political bias of a group, if one exists at all, is no simple task. Our work employs a different strategy by using Latent Dirichlet Allocation (LDA) to create a topic model to represent groups of similar interests, and inferring political bias of subcommunities through direct comparison of bigrams and trigrams with external corpora of politically slanted vocabulary. For each topic, we report the average bias of all subcommunities falling into that topic. Finally, to further analyze the correlation between group topics and political bias, we use the topic model is used to train a classifier, with the task of predicting the political bias of a group based on the topic of the group. The accuracy of the classifier is evaluated over a number of variations in the parameters of the topic model and classifier choice.

II. Related Work

Earlier work attempting to study political polarization on Twitter has been numerous. One such study uses SVM, LSTM, and VGG classifiers in an attempt to predict whether a user was a Trump follower or a Clinton follower from tweet features and profile pictures, and is able to achieve an accuracy of 69% (Wang et al. 2017). Another Twitter study uses sentiment analysis and tweet volume to attempt to predict election results, and determines that prediction results were significantly better with right-wing users than with left-wing users (Chen, Wang, and Sheth 2012). A paper attempting to predict a user's supported UK political party from Twitter content was able to achieve 86% accuracy using a Naive Bayesian classifier based on hashtag volume (Boutet, Kim, and Yoneki 2012). Our approach differs from much of the available Twitter classification studies because we focus on the topical content of overall social groups, as opposed to the bias of individual users. One study on Twitter attempted to propose a new domain-invariant method of measuring controversy by building a "conversation graph" and partitioning it into potential sides of the controversy (Garimella et al. 2017). While we apply a simpler domain-specific method of measuring bias on Reddit, the use of this alternative method to studying polarization on Reddit presents an opportunity for future exploration.

Other studies have attempted to link non-political statistics to language and behavior on social media, particularly Twitter. These approaches typically involve a transformation of textual features to topic models, and then employ SVM or Gaussian Process (GP) classifiers, using RBF kernels (Preoiuc-Pietro et al. 2015a; Preoiuc-Pietro, Lampos, and Aletras 2015b). Furthermore, previous studies have taken differing approaches to identifying "social media memes" and group-specific vocabulary. One study identifies "policy memes" from UK House of Commons debates by extracting the most frequent bigrams and trigrams from the speech texts (Gurciullo et al. 2015). Another study analyzing the relationship between Reddit post contents and scores also employs the method of extracting the most frequent bigrams and trigrams, as well as proposing a more advanced method termed "meme clustering" (Jin, Mai, and Setter 2015). However, for the purposes of this study, we will rely on the former method of using only vocabulary from frequent bigrams and trigrams.

III. Data and Preprocessing

In this study, we primarily use the popular social media and news aggregation site Reddit. This data source is particularly useful due to the fact that any user in Reddit is able to create their own *subreddit*, resulting a large set of groups representing a diverse set of people and interests. As of 2015, there were over 50,000 active subreddits[1], each one dedicated to a particular topic. However, for the purposes of this study, we use a subset of Reddit comment data over the course of May 2015, and look at only the 3,385 most active subreddits – subreddits with more than 500 comments over the course of the month.

In order to best characterize the overall social structure of a group, based on the findings of (Muchnik, Aral, and Taylor 2013), we select all comments from every subreddit which pass a particular score-threshold, s, as it is assumed that comments which do not align with the values of a particular group would not be given positive votes. We choose the threshold $s > 5$ to ensure that we still have a sufficiently large dataset. Finally, we remove comments less than 10 characters in length in order to ensure quality responses, as well as cleaning the data to remove text contents such as URLs and HTML code. In total, the dataset following preprocessing contains approximately 5 million comments. An example of a typical Reddit comments section is shown in Figure 1, with relevant fields labeled.

In building a political vocabulary for inferring the political bias of a group, we make use of two separate corpora: first, we use an external corpus with explicitly political content. This corpus is generated from the 2016 US presidential debates, with approximately 400 lines for each political party. However, this has the possible disadvantage of missing out on a consistent base of political "memes" and phrases which may be specific to the Reddit domain. In order to account for this possibility, we also introduce a corpus generated directly from various political subreddits from the same May 2015 dataset (which are held out from the subsequent

[1] https://expandedramblings.com/index.php/reddit-stats/

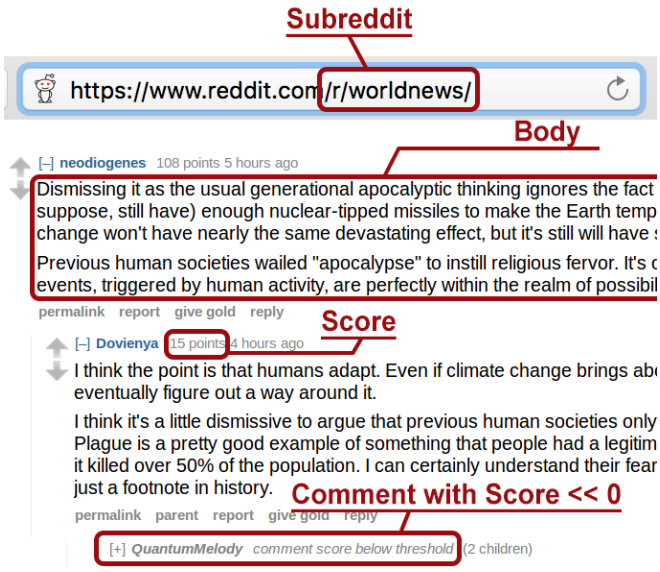

Fig. 1. The anatomy of a typical Reddit comment section.

analysis). We use the subreddits 'Liberal', 'Progressive', and 'socialism' to constitute Liberal ideology, whereas we use the subreddits 'Conservative' and 'Libertarian' to constitute Conservative ideology. The combined Liberal set contains about 600 comments, while the combined Conservative set contains about 1,000 comments – however, these are trimmed to the same size during preprocessing.

IV. Methodology

In order to understand the relationship between the bias of a group and the topic of that group, we first must extract from the documents a finite number of topics and use these topics to group each document. Two methods that are frequently employed in document grouping are document clustering based on a clustering technique such as K-means or concept factorization (Steinbach, Karypis, and Kumar 2000; Xu and Gong 2004), and topic modeling using a probabilistic approach such as LDA (Lu, Mei, and Zhai 2011). As a first approach, we tried to use document clustering to group documents. However, we ultimately found that the clustering approach performs poorly on Reddit and we obtained better results using LDA topic modeling, as discussed in the following section.

We first concatenate the comments of each subreddit into a full document for that subreddit, prior to tokenizing the text. In doing so, we remove common stopwords and apply a stemmer, thus reducing each inflected word to its root form. To do this, we use tools from Python's *nltk* library. We then convert each document to a bag-of-words vector using a *CountVectorizer* from Python's *sklearn* library.

A. Topic Modeling

As a first approach to grouping similar topics, we used document clustering techniques using the word frequencies as features, and manually labeled each cluster based on the ten most common words in each cluster center. An experimental study on document clustering found that higher quality results were achieved using k-means clustering than hierarchal algorithms (Steinbach, Karypis, and Kumar 2000), so used k-means for clustering.

In measuring the quality of our clusters, we had no "ground truth" to make use of. One experimental study determined that, among internal cluster quality measures, Davies-Bouldin Index and Silhouette Index yielded the best results for k-means clustering (Rendn, et al. 2011).

Consequently, we found the ideal amount of clusters to be 45. The resulting Silhouette coefficient of our clustering was $\bar{s_k} = 0.039$. This value indicates significant overlap between clusters, suggesting that topic clustering performs poorly on a domain such as Reddit. On top of the low Silhouette score, the cluster assignments were found to be too sensitive to particular shared words; for instance, the "Medical" cluster was assigned the subreddit "doctorwho", which is not a true medical subreddit, but rather a group for fans of a fantasy television show.

Because of these issues, we ultimately discard the clustering technique and instead take a second approach of creating a topic model using latent Dirichlet allocation (LDA). Previous research has found LDA topic models to be an appealing and successful option in text categorization (Lu, Mei, and Zhai 2011; Ramage et al. 2009). In order to validate our topic model and choose the number of topics to use, we employ the UMass intrinsic topic coherence metric (Mimno, et al. 2011). If $D(v)$ is the document frequency of word v, and $D(v, v')$ is the co-document frequency of words v and v' (i.e. number of documents containing both), then the UMass coherence of that topic is calculated as follows:

$$C\left(t; V^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D\left(v_m^{(t)}, v_l^{(t)}\right) + 1}{D\left(v_l^{(t)}\right)} \quad (1)$$

Where $V^{(t)} = \left(v_1^{(t)}, ..., v_M^{(t)}\right)$ is a list of the M most probable words in topic t. We choose to set $M = 20$. A plot of average topic coherence scores across various numbers of topics is shown in Figure 2, where higher (less negative) coherence scores represent a better model.

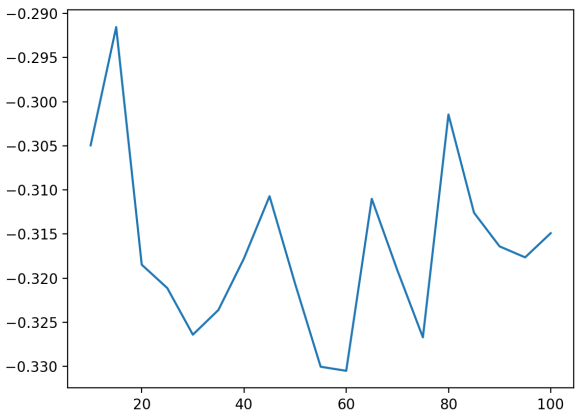

Fig. 2. Average UMass topic coherence at each 5-topic increment in number of topics.

To maximize UMass coherence while ensuring a sufficiently interesting set of topics, we choose the number of topics to be 80 in accordance with Figure 2. This resulted in an average coherence of $\bar{C_{T_k}} = -0.301$ indicating overall high quality topics (Mimno, et al. 2011). To assign a label to each topic, we manually review the top words in each topic, as well as the subreddits to which that topic is assigned. In doing so, 9 topics could not be confidently decided or were found to be incoherent. These topics were left unlabeled, and omitted from future analysis. Word clouds for a sample of topics and their manually assigned labels are shown in Figure 3.

Fig. 3. Word clouds for a sample of topics and their manually assigned labels.

B. Determining Political Bias

In the next stage of our analysis, we determine the political bias of each individual subreddit using the method to be outlined. Although this method calculates scores individually for each subreddit, there may be hundreds of subreddits for each topic in our model, so we report our calculated scores as averages for the subreddits of each topic.

As detailed in the *"Data and Preprocessing"* section, we use two corpora to measure the political bias of a cluster or topic: one corpus created from Liberal-leaning and Conservative-leaning subreddits, and another corpus created from candidate speeches during the 2016 presidential debates.

In accordance with the methods proposed by (Gurciullo et al. 2015; Jin, Mai, and Setter 2015), we extract all bigrams and trigrams (after removing stopwords and stemming vocabulary) from each political document, in order to center on "political memes" and particular phrases. A sample of these bigrams and trigrams for each political vocabulary are shown in Table 2. The Liberal and Conservative biases of each cluster or topic are measured by calculating the Jaccard correlation coefficient between the vector of extracted bigrams and trigrams from that subreddit and the vector of the corresponding political vocabulary, using the following formula:

$$J(D_i, P_j) = \frac{|D_i \cap P_j|}{|D_i \cup P_j|} \tag{2}$$

Where D_i represents the vector of a subreddit, and P_j represents the vector of a given political vocabulary. The numerator is the number of bigrams and trigrams which belong to both vectors, while the denominator is the number of unique bigrams and trigrams belonging to either vector. Additionally, once the full range of coefficients has been obtained for both Liberal and Conservative vocabularies, the coefficients are normalized using *min-max normalization* to lie within the range [0,1].

TABLE II
A SAMPLE OF BIGRAMS AND TRIGRAMS FOUND IN EACH POLITICAL VOCABULARY.

Democrat	Republican
sanction iran	war iraq
undocu immigr	swat team
help haiti	prosecut fullest
afford lower price	suprem law land
social democrat parti	protect free speech
john stuart mill	filibust patriot act

The political bias for a subreddit is then calculated by subtracting the Conservative bias index from the Liberal bias index. Hence, Liberal biases are indicated by positive values, and Conservative biases are indicated by negative values. We group subreddits by topic and plot the average political bias for each topic in Figure 4, as well as error bars showing the calculated standard deviation of our bias measure. These meaning of these results, as well as potential shortcomings, are discussed in the *Qualitative Discussion* section.

C. Classification and Experimental Results

To further analyze the politicization of topics, we attempt to train a classifier with the goal of predicting the political bias of a topic – liberal or conservative – based on a bag-of-words vector of the N most common words in that topic. To transform the political bias measures from the previous section into a classification problem, we assign a bias value of "1" to topics with bias greater than zero, and "0" to topics with bias less than zero. We choose to omit topics with no

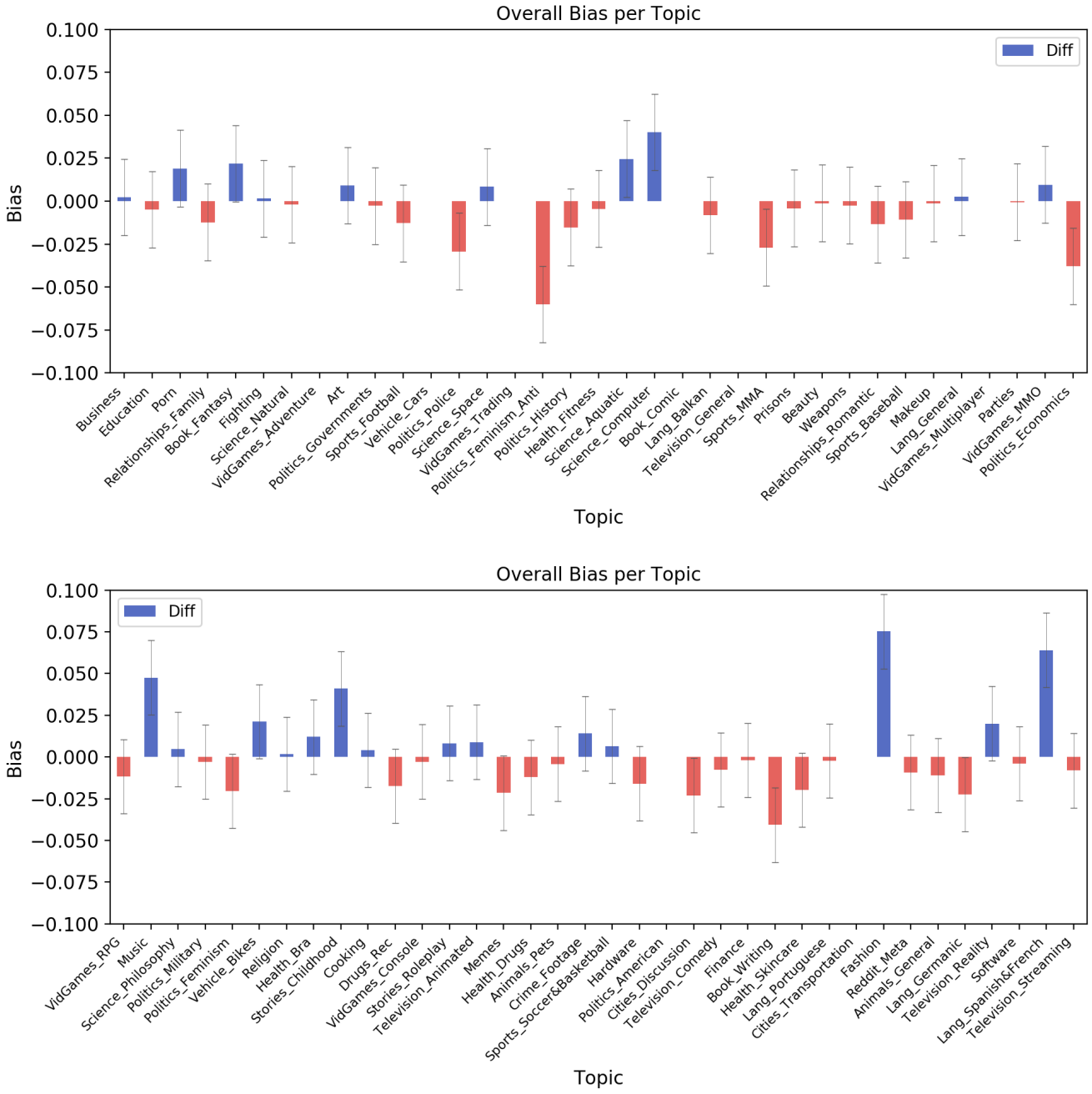

Fig. 4. Overall political bias for each LDA topic.

measured bias, as a bias measure of zero does not necessarily indicate that the political stance of a group is neutral, so this class could possibly be misleading.

While it is not ideal that the topic word distribution is built from the same data as used to calculate average political bias of topics, LDA creates a generative model and hence cannot memorize the individual subreddit vocabularies which are used to determine average bias for each topic. Thus, although this setup is not optimal, it still provides a useful method of testing the relationship between topic vocabulary and bias.

Naive Bayes classifiers are often "used as a baseline in text classification because it is fast and easy to implement", however "its severe assumptions [...] also adversely affect the quality of its results" (Rennie et al. 2003). Nonetheless, some practical studies have still been able to apply a Naive Bayes classifier with reasonably high accuracy (Boutet, Kim, and Yoneki 2012).

SVM classifiers have been found to be useful for text classification with many features, due to their robust behavior toward non-linearities (Joachims 1998). In particular, we use an SVM classifier with an RBF kernel.

Lastly, previous studies involving similar uses of topic

modeling or document clustering employ a Gaussian Process (GP) classifier, with higher accuracy than a basic SVM classifier in general (Preoiuc-Pietro et al. 2015a; Preoiuc-Pietro, Lampos, and Aletras 2015b).

We test the performance of each of these classifiers on our LDA topic model, along with a stratified random guessing classifier (RG) to provide a benchmark. In order to analyze the accuracy of our classifiers, we use 5-fold cross-validation on our set of topics, with the mean accuracy and standard error calculated from the scores for each fold. Python's *sklearn* library is used for each of the classifiers. For our experimental setup, we vary the number of word features, N, used to train the classifiers. We also try increasing the number of topics from the 80 used initially to 40 and 200 in order to test the performance of the classifier as the differences between topics become more or less fine-tuned.

These results are shown in Table 3. The highest accuracy we were able to achieve is 85.2% using an SVM classifier with 40 topics (compared to a random-guessing baseline of 64.8%), suggesting that variations in vocabulary between groups can be a strong predictor of the overall political bias of the group. Classification results appear to improve with a lower number of topics, and the SVM classifier appeared to be more accurate compared to the NB and GP classifiers.

TABLE III
EXPERIMENTAL RESULTS.

# Topics	# Features	RG	NB	SVM	GP
40	500	64.8±10.0	65.4±7.00	**85.2±4.00**	77.3±6.50
40	1000	67.5±16.5	70.2±10.0	**85.2±4.00**	85.1±9.00
40	2000	72.0±11.0	80.6±10.0	**85.2±4.00**	83.1±9.50
80	500	68.2±13.5	59.8±9.50	**71.1±2.50**	61.4±8.00
80	1000	59.6±8.50	59.9±14.5	**71.1±2.50**	59.6±12.0
80	2000	66.5±11.0	52.9±19.5	**71.1±2.50**	58.1±6.50
200	500	56.2±16.5	59.4±6.50	**67.9±2.50**	64.5±7.50
200	1000	42.4±14.5	59.4±6.50	**67.9±2.50**	64.5±7.50
200	2000	57.6±5.00	61.1±6.00	**67.9±2.50**	66.2±9.00

V. QUALITATIVE DISCUSSION

The efficacy of our topic model and the applicability of it to the real world can be tested by comparing our results to those of other studies analyzing political bias in particular topics.

First, we note that the results from the LDA topic model, shown in Figure 4, appear to indicate a slight overall conservative bias among all topics, with an approximately equal proportion of topics with significant bias between both sides. While there are no large-scale studies on the political bias of Reddit individually, previous empirical studies on political bias of USENET groups found that "liberal or left-wing political groups are less active and more poorly organized" (Hill and Hughes, 2010), and that in ideological groups, conservatives tend to generate more messages per author than liberals (Kelly, Fisher, and Smith, 2005). These findings point to two potential implications for our work: first, as we measured political bias through comment data, it makes sense that we would observe a slight conservative bias if it is indeed true that the phenomenon observed on USENET carries over to Reddit. Second, it may not necessarily be the case that conservative *users* outnumber liberal users on Reddit. Furthermore, our results do not rule out the possibility that the largest subreddits on Reddit may have primarily liberal bias, as we currently do not weight by number of users.

One useful benchmark to make to similar studies involves comparing the biases of sport topics. A recent study on Twitter attempted to predict whether users were Trump supporters or Clinton supporters based on tweet content and profile pictures–this study found that users who fell into the cluster of sports profile pictures were far more likely to be Trump followers (Wang et al. 2017). Additionally, a survey analysis by members of the research firm *National Journal* found that fans of many popular sports, for instance American football and baseball, had moderate conservative leanings, with the exceptions of NBA basketball and soccer which both had significant liberal leanings. Though none of our sports topics have been found to have significant bias, with the exception of "Sports_MMA", our results generally corroborate these previous findings. Each sports topic was found to have a moderate conservative bias, with one exception being "Sports_Soccer&Basketball", which had a slight liberal bias.

Another point which warrants some discussion is the bias of the "Politics_Feminism" and "Politics_Feminism_Anti" topics. Although the latter topic records a strongly conservative bias, as expected, "Politics_Feminism" records a moderate conservative bias as well. While it is puzzling that both the feminist and anti-feminist topics have net conservative bias, upon further investigating samples of comments in some of the feminist subreddits we found that a substantial amount of the political content consisted of debate between feminists and anti-feminists.

In general, one potential drawback of our approach is that it does not account for the fact that people may borrow from the vocabulary of an opposite political bias if both sides visit a particular community to debate . Hence, topics which are explicitly political may show bias which misrepresents the true ideological leaning of the members of that group. Intuition suggests, therefore, that our results should hold more reliably for boards of a non-political topic, such as sports or music communities, than for groups which are dedicated to discussing political content.

This issue presents a possible focus for future research; in particular, a survey-based approach could instead be taken to measure and compare the political bias of each subreddit. Additionally, future work could attempt to employ a more sophisticated method of determining political vocabulary than extracting bigrams and trigrams–for instance, one study suggested using a "meme clustering" algorithm to improve upon the results from bigram/trigram generation (Jin, Mai, and Setter 2015). Finally, some studies have successfully avoided the need for external bias measures altogether by utilizing a conversation graph and creating a partition to represent sides of a debate (Garimella et al. 2017; Agrawal et al. 2003).

VI. Conclusion

While much work has been done studying the process of political polarization of individuals on online social media, the method by which political homophily becomes established in groups of people and how topics become politicized has remained somewhat enigmatic. In this paper, we have presented new empirical data on the relationships between political bias and non-political topics in Reddit communities, and discussed the implications of our findings. Furthermore, by training classification models using the text features of a topic model, we are able to predict the political bias of particular topics with a high accuracy of 85.2% (compared to a random-guessing baseline of 64.8%), suggesting a strong consistency in the type of vocabulary indicative of political biases in different communities.

References

[1] Adamic, L. A.; Glance, N. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *Proc. of the 3rd ACM International Workshop on Link Discovery* 36-43.

[2] Agrawal, R.; Rajagopalan, S.; Srikant, R.; Xu, Y. Mining Newsgroups Using Networks Arising From Social Behavior. *In Proceedings of the 12th International Conference on World Wide Web*, 529-535. Budapest, Hungary.

[3] Barber, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; Bonneau, R. 2015. Tweeting From Left to RightIs Online Political Communication More Than an Echo Chamber? *Psychological Science* 26(10): 1531-1542.

[4] Boutet, A.; Kim, H.; Yoneki, E. 2012. What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election. *ICWSM* 12: 411-414.

[5] Chen, L.; Wang, W.; Sheth, A. P. 2012. Are Twitter Users Equal in Predicting Elections? A Study of User Groups in Predicting 2012 U.S. Republican Presidential Primaries. *In Proceedings of the Fourth International Conference on Social Informatics (SocInfo)*, 379-392. Lausanne, Switzerland.

[6] Colleoni, E.; Rozza, A.; Arvidsson A. 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.

[7] Danescu-Niculescu-Mizil C.; West, R.; Jurafsky, D.; Leskovec, J.; Potts, C. 2013. No country for old members: user lifecycle and linguistic change in online communities. *In Proceedings of the 22nd international conference on World Wide Web*, 307-318. Rio de Janeiro, Brazil.

[8] Garimella, K.; De Francisci Morales, G.; Gionis, A.; Mathioudakis, M. 2017. Quantifying Controversy on Social Media. *In Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, 33-42. San Francisco, USA.

[9] Garrett, R. K. 2009a. Echo Chambers Online? Politically Motivated Selective Exposure Among Internet Users. *Computer-Mediated Communication* 14(2): 265-285.

[10] Garrett, R. K. 2009b. Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate. *Journal of Communication* 59(4): 676-699.

[11] Gentzkow, M.; Shapiro, J. M.; Taddy, M. 2017. Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech, Working Paper 22423, National Bureau of Economic Research, Cambridge, Massachusetts.

[12] Gurciullo, S.; Herzog, A.; John, P.; Mikhaylov, S. 2015. How Do Policy Memes Spread? An Analysis of the UK House of Commons Debates. Paper presented at 2015 European Political Science Association (EPSA) Annual Conference, Vienna, Austria, 25-27 June.

[13] Hill, K. A.; Hughes, J. E. Computer-Mediated Political Communication: The USENET and Political Communities. *Political Communication* 14(1): 3-27.

[14] Iyengar, S.; Hahn, K. S. 2009. Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication* 59(1): 19-39.

[15] Jin, A.; Mai, D.; Setter, J. 2015. Learning to Rank Comments Within Subreddit Submissions. Dept. of Computer Science, Stanford Univ.

[16] Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *In Proceedings of the 10th European Conference on Machine Learning*, 137-142. Chemnitz, Germany.

[17] Kelly, J.; Fisher, D.; Smith, M. Debate, Division, and Diversity: Political Discourse Networks in USENET Newsgroups. Paper presented at 2005 Online Deliberation Conference, Stanford University, USA, 24 May.

[18] Lu, Y.; Mei, Q.; Zhai, C. 2011. Investigating Task Performance of Probabilistic Topic Models: an Empirical Study of PLSA and LDA. *Information Retrieval* 14(2): 178-203.

[19] McPherson, M.; Smith-Lovin, L.; Cook, J. M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27: 415-444.

[20] Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; McCallum, A. Optimizing Semantic Coherence in Topic Models. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262-272. Edinburgh, United Kingdom.

[21] Muchnik, L.; Aral, S.; Taylor, S. J. 2013. Social Influence Bias: A Randomized Experiment. *Science* 341(6146): 647-651.

[22] Mutz, D. C.; Martin, P. S. 2001. Facilitating Communication across Lines of Political Difference: The Role of Mass Media. *American Political Science Review* 95(1): 97-114.

[23] Paolillo, J. C. 2001. Language Variation on Internet Relay Chat: A Social Network Approach. *Journal of Sociolinguistics* 5(2): 180-213.

[24] Preoiuc-Pietro, D.; Volkova, S.; Lampos, V.; Bachrach, Y.; Aletras, N. 2015a. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE* 10(9).

[25] Preoiuc-Pietro, D.; Lampos, V.; Aletras, N. 2015b. An Analysis of the User Occupational Class through Twitter Content. *In Proceedings of the 53nd Annual Meeting of the Association for Computational Linguistics*, 17541764. Beijing, China.

[26] Ramage, D.; Daniel, H.; Nallapati, R.; Manning, C. D. 2009. A Supervised Topic Model for Credit Attribution in Multi-Labeled Corpora. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 248-256. Singapore.

[27] Rendn, E.; Abundez, I.; Arizmendi, A.; Quiroz, E. M. 2011. Internal Versus External Cluster Validation Indexes. *International Journal of Computers and Communications* 5(1): 27-34.

[28] Rennie, J. D. M.; Shih, L.; Teevan, J.; Karger, D. R. 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *In Proceedings of the Twentieth International Conference on Machine Learning*, 616-623. Washington, DC, USA.

[29] Shannon, M.; Feltus, W. 2012. Play Ball: What Your Favorite Sports Say About Your Politics. *Hotline on Call*, NationalJournal.com, 26 Oct.

[30] Steinbach, M.; Karypis, G.; Kumar, V. 2000. A Comparison of Document Clustering Techniques, Technical Report, #00-034, Dept. of Computer Science and Engineering, University of Minnesota.

[31] Wang, Y.; Feng, Y.; Hong, Z.; Berger, R.; Luo, J. 2017. How Polarized Have We Become? A Multimodal Classification of Trump Followers and Clinton Followers. *In Proceedings of the Ninth International Conference on Social Informatics (SocInfo)*, 440-456. Oxford, UK.

[32] Xu, W.; Gong, Y. 2004. Document Clustering by Concept Factorization. *In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 202-209. Sheffield, United Kingdom.

[33] Yardi, S.; Boyd, D. 2010. Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. *Bulletin of Science, Technology & Society* 30(5): 316-327.

[34] Zappavigna, M. 2012. *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. London: Bloomsbury Academic.