Quantifying Political Leaning from Tweets, Retweets, and Retweeters Systems

N.BasavaRaju[1], Dr.G.V. Ramesh Babu[2]
[1] *Student, Dept. of Computer Science, Sri Venkateshwara University, Tirupati*
[2] *Assistant professor, Dept. of Computer Science, Sri Venkateshwara University, Tirupati.*

***Abstract*- The widespread use of online social networks (OSNs) to disseminate information and exchange opinions, by the general public, news media and political actors alike, has enabled new avenues of research in computational political science. In this paper, we study the problem of quantifying and inferring the political leaning of Twitter users. We formulate political leaning inference as a convex optimization problem that incorporates two ideas: (a) users are consistent in their actions of tweeting and retweeting about political issues, and (b) similar users tend to be retweeted by similar audience. We then apply our inference technique to 119 million election-related tweets collected in seven months during the 2012 U.S. presidential election campaign. On a set of frequently retweeted sources, our technique achieves 94% accuracy and high rank correlation as compared with manually created labels. By studying the political leaning of 1,000 frequently retweeted sources, 232,000 ordinary users who retweeted them, and the hashtags used by these sources, our quantitative study sheds light on the political demographics of the Twitter population, and the temporal dynamics of political polarization as events unfold.**

INTRODUCTION

IN recent years, big online social media data have found many applications in the intersection of political and computer science. Examples include answering questions in political and social science (e.g., proving/disproving the existence of media bias [3, 30] and the "echo chamber" effect [1, 5]), using online social media to predict election outcomes [46, 31], and personalizing social media feeds so as to provide a fair and balanced view of people's opinions on controversial issues [36]. A prerequisite for answering the above research questions is the ability to accurately estimate the political leaning of the population involved. If it is not met, either the conclusion will be invalid, the prediction will perform poorly [35, 37] due to a skew towards highly vocal individuals [33], or user experience will suffer. In the context of Twitter, accurate political leaning estimation poses two key challenges: (a) Is it possible to assign meaningful numerical scores to tweeters of their position in the political spectrum? (b) How can we devise a method that leverages the scale of Twitter data while respecting the rate limits imposed by the Twitter API? Focusing on "popular" Twitter users who have been retweeted many times, we propose a new approach that _ Felix M.F. Wong was with the Department of Electrical Engineering, Princeton University. He is now with Yelp, Inc. Email: mwthree@princeton.edu _ Chee Wei Tan is with the Department of Computer Science, City University of Hong Kong. Email: cheewtan@cityu.edu.hk _ Soumya Sen is with the Department of Information & Decision Sciences, Carlson School of Management, University of Minnesota. Email: ssen@umn.edu _ Mung Chiang is with the Department of Electrical Engineering, Princeton University. Email: chiangm@princeton.edu Preliminary version in [51]. This version has substantial improvements in algorithm, evaluation and quantitative studies. incorporates the following two sets of information to infer their political leaning. Tweets and retweets: the target users' temporal patterns of being retweeted, and the tweets published by their retweeters. The insight is that a user's tweet contents should be consistent with who they retweet, e.g., if a user tweets a lot during a political event, she is expected to also retweet a lot at the same time. This is the "time series" aspect of the data. Retweeters: the identities of the users who retweeted the target users. The insight is similar users get followed and retweeted by similar audience due to the homophily principle. This is the "network" aspect of the data. Our technical contribution is to frame political leaning inference as a convex optimization

problem that jointly maximizes tweet-retweet agreement with an error term, and user similarity agreement with a regularization term which is constructed to also account for heterogeneity in data. Ourtechnique requires only a steady stream of tweets but not the Twitter social network, and the computed scores have a simple interpretation of "averaging," i.e., a score is the average number of positive/negative tweets expressed when nretweeting the target user. See Figure 1 for an illustration. mUsing a set of 119 million tweets on the U.S. presidential election of 2012 collected over seven months, we extensively evaluate our method to show that it outperforms several standard algorithms and is robust with respect to variations to the algorithm. The second part of this paper presents a quantitative study on our collected tweets from the 2012 election, by first (a) quantifying the political leaning of 1,000 frequentlyretweeted Twitter users, and then (b) using their political leaning, infer the leaning of 232,000 ordinary Twitter users. We make a number of findings: Fig. 1. Incorporating tweets and retweets to quantify political leaning: to estimate the leaning of the "sources," we observe how ordinary users retweet them and match it with what they tweet. The identities of the retweeting users are also used to induce a source similarity measure to be used in the algorithm. be liberal as compared to other account types. They also tend to be temporally less stable. _ Liberals dominate the population of less vocal Twitter users with less retweet activity, but for highly vocal populations, the liberal-convservative split is balanced. Partisanship also increases with vocalness of the population. _ Hashtag usage patterns change significantly as political events unfold. _ As an event is happening, the influx of Twitter users participating in the discussion makes the active population more liberal and less polarized. The organization of the rest of this paper is as follows. Section 2 reviews related work in studies of Twitter and quantifying political orientation in traditional and online social media. Section 3 details our inference technique by formulating political leaning inference as an optimization problem. Section 4 describes our dataset collected during the U.S. presidential election of 2012, which we use to derive ground truth for evaluation in Section 5. Then in Section 6 we perform a quantitative study on the same dataset, studying the political leaning of Twitter users and hash tags, and how it changes with time.

EXISTING SYSTEM

- A variety of methods have been proposed to quantify the extent of bias in traditional news media. Indirect methods involve linking media outlets to reference points with known political positions. For example, Lott and Hassett linked the sentiment of newspaper headlines to economic indicators.
- Groseclose and Milyo linked media outlets to Congress members by co-citation of think tanks, and then assigned political bias scores to media outlets based on the Americans for Democratic Action (ADA) scores of Congress members.
- Gentzkow and Shapiro performed an automated analysis of text content in newspaper articles, and quantified media slant as the tendency of a newspaper to use phrases more commonly used by Republican or Democrat members of the Congress.

PROPOSED SYSTEM

- Our technical contribution is to frame political leaning inference as a convex optimization problem that jointly maximizes tweet-retweet agreement with an error term, and user similarity agreement with a regularization term which is constructed to also account for heterogeneity in data.
- Our technique requires only a steady stream of tweets but not the Twitter social network, and the computed scores have a simple interpretation of "averaging," i.e., a score is the average number of positive/negative tweets expressed when retweeting the target user.
- Liberals dominate the population of less vocal Twitter users with less retweet activity, but for highly vocal populations, the liberal-convservative split is balanced. Partisanship also increases with vocalness of the population.
- Hashtag usage patterns change significantly as political events unfold.
- As an event is happening, the influx of Twitter users participating in the discussion makes the active population more liberal and less polarized.

PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- Request Clarification
- Feasibility Study
- Request Approval

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- Operational Feasibility
- Economic Feasibility
- Technical Feasibility

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

4.3.3 REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, it cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

CONCLUSIONS

Scoring individuals by their political leaning is a fundamental research question in computational political science. From roll calls to newspapers, and then to blogs and microblogs, researchers have been exploring ways to use bigger and bigger data for political leaning inference. But new challenges arise in how one can exploit the structure of the data, because bigger often means noisier and sparser. In

this paper, we assume: (a) Twitter users tend to tweet and retweet consistently, and (b) similar Twitter users tend to be retweeted by similar sets of audience, to develop a convex optimization-based political leaning inference technique that is simple, efficient and intuitive. Our method is evaluated on a large dataset of 119 million U.S. election-related tweets collected over seven months, and using manually constructed ground truth labels, we found it to outperform many baseline algorithms. With its reliability validated, we applied it to quantify a set of prominent retweet sources, and then propagated their political leaning to a larger set of ordinary Twitter users and hash tags. The temporal dynamics of political leaning and polarization were also studied. We believe this is the first systematic step in this type of approaches in quantifying Twitter users' behavior. The Retweet matrix and retweet average scores can be used to develop new models and algorithms to analyze more complex tweet-and-retweet features. Our optimization framework can readily be adapted to incorporate other types of information. The y vector does not need to be computed from sentiment analysis of tweets, but can be built from exogenous information (e.g., poll results) to match the opinions of the retweet population. Similarly, the A matrix, currently built with each row corresponding to one event, can be made to correspond to other groupings of tweets, such as by economic or diplomatic issues. The W matrix can be constructed from other types of network data or similarity measures. Our methodology is also applicable to other OSNs with retweet-like endorsement mechanisms, such as Facebook and YouTube with "like" functionality.

REFERENCES

[1] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in Proc. Link KDD, 2005.

[2] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors," in Proc. ICWSM, 2012.

[3] J. An, M. Cha, K. P. Gummadi, J. Crowcroft, and D. Quercia, "Visualizing media bias through Twitter," in Proc. ICWSM SocMedNews Workshop, 2012.

[4] S. Ansolabehere, R. Lessem, and J. M. Snyder, "The orientation of newspaper endorsements in U.S. elections," Quarterly Journal of Political Science, vol. 1, no. 4, pp. 393– 404, 2006.

[5] P. Barber´a, "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data," Political Analysis, 2014.

[6] A. Boutet, H. Kim, and E. Yoneki, "What's in your tweets? I know who you supported in the UK 2010 general election," in Proc. ICWSM, 2012.

[7] d. boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in Proc. HICSS, 2010.

[8] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in Proc. ICWSM, 2010.

[10] J. Clinton, S. Jackman, and D. Rivers, "The statistical analysis of roll call data," American Political Science Review, vol. 98, no. 2, pp. 355–370, 2004.

[11] R. Cohen and D. Ruths, "Classifying political orientation on Twitter: It's not easy!" in Proc. ICWSM, 2013.

[12] M. D. Conover, B. Gonc¸alves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of Twitter users," in Proc. IEEE SocialCom, 2011.

[13] M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonc¸alves, A. Flammini, and F. Menczer, "Political polarization on Twitter," in Proc. ICWSM, 2011.

[14] CVX Research, Inc., "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, Sep. 2012.

[15] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory," Czechoslovak Mathematical Journal, vol. 25, no. 4, pp. 619–633, 1975.

[16] S. Finn, E. Mustafaraj, and P. T. Metaxas, "The coretweeted network and its applications for measuring the perceived political polarization," in Proc. WEBIST, 2014.

[17] J. L. Fleiss, "Measuring nominal scale agreement among many raters," Psychological Bulletin, vol. 76, no. 5, pp. 378– 382, 1971.
[18] M. Gabielkov, A. Rao, and A. Legout, "Studying social networks at scale: Macroscopic anatomy of the Twitter social graph," in Proc. SIGMETRICS, 2014.
[19] D. Gayo-Avello, "All liaisons are dangerous when all your friends are known to us," in Proc. HT, 2011.
[20] M. Gentzkow and J. M. Shapiro, "What drives media slant? Evidence from U.S. daily newspapers," Econometrica, vol. 78, no. 1, pp. 35–71, January 2010.
[21] S. Gerrish and D. Blei, "How the vote: Issue-adjusted models of legislative behavior," in Proc. NIPS, 2012.
[22] "Predicting legislative roll calls from text," in Proc. ICML, 2011.
[23] J. Golbeck and D. Hansen, "A method for computing political preference among Twitter followers," Social Networks, vol. 36, pp. 177–184, 2014.
[24] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," Political Analysis, 2013.
[25] T. Groseclose and J. Milyo, "A measure of media bias," The Quarterly Journal of Economics, vol. 120, no. 4, pp. 1191– 1237, November 2005.