INTERNATIONAL JOURNAL · ENGINEERING RESEARCH ·

International Journal for Innovative
Engineering and Management Research
*A Peer Revieved Open Access International Journal*
www.ijiemr.org

## COPY RIGHT

IJIEMR Transactions, online available on 16th Febraury 2018. Link :
http://www.ijiemr.org/downloads.php?vol=Volume-7&issue=ISSUE-02

Title: Mining Tweets To Generates The Tweets and Retweet and Retweeters.

Volume 07, Issue 02, Page No: 404 – 411.

Paper Authors

***D. KRISHNAVENI, P.CHANDRA MOHANA RAI.**

* Dept of CSE, Krishnaveni Engineering College For Women.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

# Mining tweets to generates the tweets and retweet and retweeters

***D. KRISHNAVENI, **P.CHANDRA MOHANA RAI**

***PG Scholar, Dept of CSE,** Krishnaveni Engineering College for women,keasanupalli, narasaraopet

**Asst.Prof, **Dept of CSE,** Krishnaveni Engineering College for women,keasanupalli, narasaraopet

## ABSTRACT:

We study the implications of the commonplace assumption that most social media studies make with respect to the nature of message shares predominantly positive interaction. In sentiment analysis, Naïve Bayes classifiers with different distributions obtain accuracy 75% and the results reveal positive tweets. We approach the problem as regression and apply linear as well as nonlinear learning methods to predict a user impact score, estimated by combining the numbers of the user's followers and listings. It is important to distinguish between the bias arises from the data serves as the input to the ranking system and the bias that arises from the ranking system itself. The propose framework to quantify these distinct biases and apply this framework to politics-related queries on Twitter. We found the input data and ranking system contribute significantly to produce varying total bias in the search results in different ways. This paper approaches such discussions as a multi faceted data space and applies data mining to identify interesting patterns and factors of influence. We found the time users take to tweet a message is originally posted and useful signal to infer antagonism in social platforms and those surges of out-of-context tweets correlate with sentiment drifts triggered by real-world events. We also discuss how such evidences can be embedded in sentiment analysis models.

**Index Terms**: : President debate, twitter, sentiment analysis, event study, Naïve Bayes, political party classification, Data Analytics, Inference, Signal Processing

## .1. INTRODUCTION

One of the most challenging problems in the intersection of politics and online social media is to use Twitter to predict election outcomes [1]. Although some success has been claimed it has also been argued that the election prediction problem is difficult because of sampling bias among the voter population. In order to correct for bias, it would be helpful to have some prior understanding of the population of study [2].Given that on general purpose social platforms such as Face book and Twitter there are no explicit positive and negative signs encoded in polarized online communities induced by topics such as Politics and public policies do not conduct any explicit analysis of antagonism at the edge granularity and the degree of separation between communities as well as the controversial nature of the topic is accepted as sufficient evidence of polarization [3].We present a method is nonlinear regression using Gaussian Processes a Bayesian non-parametric class of methods proven more effective in capturing the multimodal user features further new aspects of a user's behavior relate to impact by examining the parameters of the inferred model [4]. While the ranking system mitigated the opposing bias in the search results for the most popular democratic candidate, it enhanced it for the most popular republican candidate [5]. Simply the most popular republican candidate is more tweets from the opposing political party than if she searched for the most popular democratic candidate [6]. This may be less than desirable for a popular republican candidate if the users with the opposing polarity primarily post negative tweets about the candidate that result in negatively biased search results for her or him [7]. We propose a general framework, based on latent topic models and user features over a multi-faceted data space [8]. The facets of interest are the topics of tweets their factuality versus sentimentality the inclination of users with regard to the two involved stances and the roles of users with regards to how they affect activity within the discussions [9]. Our technical models is to frame political leaning inference as a convex optimization problem in jointly maximizes tweet and retweet agreement with an error term and user similarity agreement with a regularization term which is constructed to also account for heterogeneity in data [10].

## 2. RELATED WORK

In political science the ideal point find issue intends to assess the political slanting of authorities from move call data and bill message through quantifiable enlistment of their positions in a normal dormant space [11]. The openness of a great deal of political taking legitimate talks, charge substance and social affair decrees, through electronic means has enabled the scope of modernized substance examination for political slanting estimation [12].Americans for Democratic Action (ADA) scores of Congress members 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 members of the Congress [13]. In contrast direct methods quantify media bias find the news content for approval of political parties and new analyzed newspaper editorials on Supreme Court cases to infer the political positions of main newspapers used 60 years of editorial election endorsements to identify a gradual shift in newspapers political results within time [14]. Several studies isexplored political bias in Web search results and search queries. While Weber inferred political leanings of find queries by linking the queries with political blogs [15]. They asked people insufficient the political candidates in election to find the candidates and form opinion based on the results [16].Sentiment analysis is implemented by dictionary-based model to the baseline in our sentiment analysis work or machine learning model [17].. To solve the problems in dictionary based approach, machine learning classifiers is developed supervised and unsupervised machine learning techniques have been studied for many years and achieve good results [18].

## 3. SYSTEM MODEL

On social networks total edge signs are labeled, antagonistic relationships among communities naturally reflected by the number of positive and negative edges flowing from the source community to a target community, and the communities themselves is found by algorithms especially designed to deal with negative edges [19]. The linked media outlets to congress members of think tanks, and then assigned political bias scores to media outlets based on the Americans for Democratic Action (ADA) scores of Congress members and Shapiro results is automated analysis of text content in newspaper articles, and quantified media slant as the tendency of a newspaper to use phrases more commonly used new members of the Congress [20]. The amount of data available for analysis is limited is fast the media sources publish researchers may need to aggregate data created over long periods of time, often years, to perform reliable analysis. Analyzing media sources through their OSN outlets offers many unprecedented opportunities with high volume data from interaction with their audience [21].

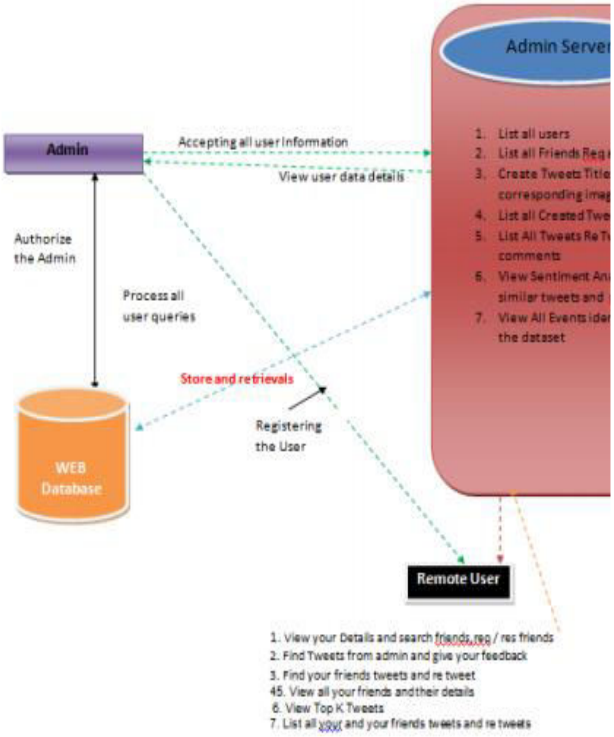

Fig1. System Architecture

## 4. PROPOSED SYSTEMS

Our technical contribution is to frame political leaning inference to convex optimization problem that jointly maximizes tweet-re tweet agreement with an error term and user similarity agreement with heterogeneity in data [22]. 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 a score is the average number of positive/negative tweets expressed tweeting the target user. The liberal-conservative split is balanced. Partisanship also increases with localness of the population. Hash tag 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 [23].

### A. Cross-Ideological Interactions on Social Media

With the rising popularity of social media sites like Twitter and Face book users are increasingly relying on them to obtain news real-time information about ongoing events and public opinion on celebrities [24]. Some others argued that social media usage can result in selective exposure by providing a platform that reinforces users' existing biases [25]. By examining cross-ideological exposure through content and network analysis showed that political talk on Twitter is highly users are unlikely to be exposed to cross-ideological content through their friendship network. Other studies have also confirmed these results by demonstrating users higher willingness to communicate with other like-minded social media users and their inability to engage in meaningful discussions with different minded users [26].

### B. Auditing Algorithms

Today algorithms that curate and present information in online platforms is affect users experiences significantly creating discriminatory ads based on gender different prices for the same products/services to new users and mistakenly labeling a black man as an ape by an image tagging algorithm [27]. These model is lead researchers organizations and even governments towards a new avenue of research called auditing algorithms which endeavors to understand system is biases, particularly when they are misleading or discriminatory to users

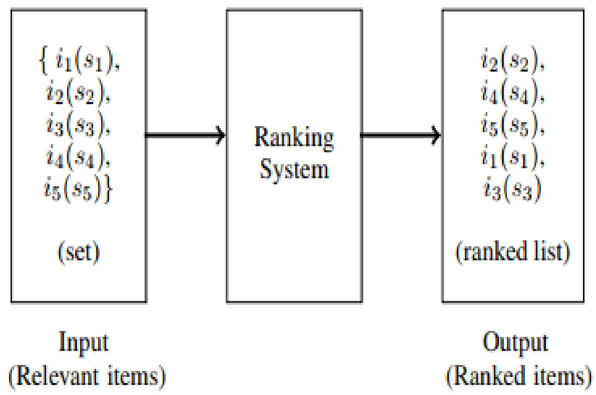
$\{i_1(s_1), i_2(s_2), i_3(s_3), i_4(s_4), i_5(s_5)\}$
(set)
Ranking System
$i_2(s_2), i_4(s_4), i_5(s_5), i_1(s_1), i_3(s_3)$
(ranked list)
Input
(Relevant items)
Output
(Ranked items)

Figure 2: Overview of our search bias quantification framework

**STEP1: Bias of an individual data item:**

As mentioned earlier the search scenario that we are considering is one of the US politics. Each data item is positively biased or negatively biased neutral towards each of these two parties, and the bias score of each item captures the degree to which the item is biased with respect to the two parties. We describe a methodology for measuring the bias score of items in the context of US political searches on Twitter social media [28].

**STEP2: Input Bias:**

This input data captures the bias introduced by the query by filtering the relevant items from the whole corpus of data. Put differently, input bias gives a measure of what bias a user would have observed, had she been shown random items relevant to the query, instead of a list ranked by the ranking system [29].

**STEP3: Output Bias:**

The output bias is the effective bias presented to the user the final ranked list from the search engine. The higher ranked items should be given more importance, since not only are the users more likely to browse through the top search results but they also tend to have more trust in them. We propose a metric for output search bias that is inspired by the well-known metric

Average Precision from the Information Retrieval literature [30].

### C. Data Collection and Preparation

We used Twitter's Streaming API1 to monitor two topics that motivate intense debate on offline and online media and thus are suitable for analysis of formation of antagonistic communities: Politics and Sports provides details on the datasets [31].

Different graphs can be built based on the datasets described in traditionally, a social network G(V, E) represents a set of users V and a set of edges E that connect two users if they exceed a threshold of interaction activity [17].

We performed a validation of the K communities we found using a sampling strategy on the correlation between communities and profiles that make explicit their side. Twitter users that append to their profile names the soccer team or political party they support; and, the content they publish will favor the respective mentioned side, as we observed through manual inspection of a sample [8].

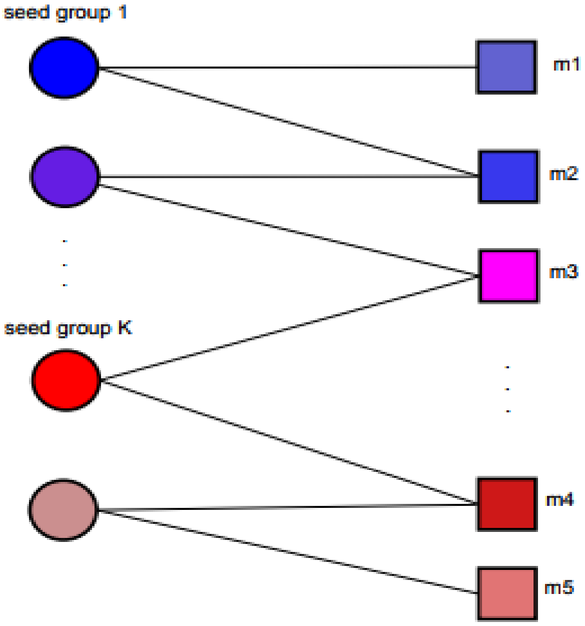

Figure 3: A bipartite user-message graph

### D. Sentiment Analysis

Our sentiment analysis with Naïve Bayes classifier focus on these two main aspects

- Performance under different term distribution Naïve Bayes classifiers. Baseline of Sentiment Analysis: Classify text sentiment based on score calculated from lexicon created Gaussian and Bernoulli Naïve Bayes classifier Tweets are text documents limited in 140 characters so most words will only appear once in a tweet [23]. Multinomial Naïve Bayes to verify our guess of different term distributions under different sentiment circumstances

• Performance under different Laplace smoothing parameter settings. As described in previous section, our idea is very intuitive: we tune Laplace smoothing permanent from close 0 to 2 with smaller gap 0.25 each time and observe the trend of performance of Gaussian Naïve Bayes classifier in sentiment analysis [8].

## 5. RESULTS

Mitt Romney and compare the results with those from a number of algorithms:

**PCA:** We run Principal Components Analysis on A with each column being the feature vector of a source, with or without the columns being standardized, and take the first component.

**Eigenvector:** We compute the second smallest Eigen vector of L, with L becoming computed from S being either the cosine or Jacquard matrix. This is a technique commonly seen in spectral graph partitioning [15] and is the standard approach when only the information is available.

**Sentiment analysis:** We take xi as the average sentiment of the tweets published by source i, using the same methodology in computing y [15]. This is the baseline when only tweets are used.

**SVM on hash tags:** Source we compute its feature vector as the term frequencies of the 23,794 hash tags used by the top 1,000 sources. We then train an SVM classifier using the 900 of the top 1,000 sources that are not labeled by 12 human judges as training data.

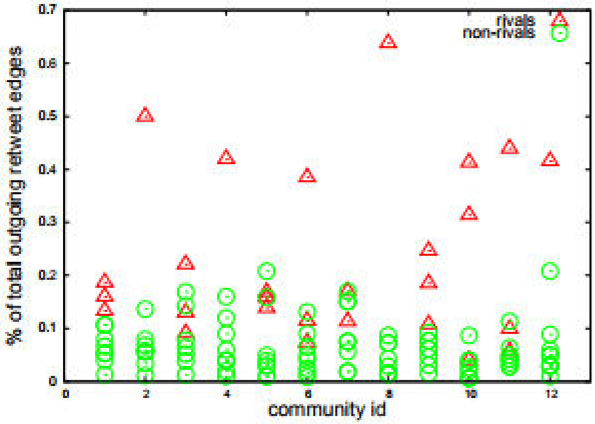

Figure 4: Twitter communities each other polarizing communities

## 6. CONCLUSION AND FUTURE WORK

To our knowledge the present study developed the first framework to quantify bias of ranked results in a search process while being able to distinguish between different sources of bias**.** We explore the observation that, in the vast majority of social media studies, especially those based on Facebook and Twitter data, there is no explicit positive and negative signs encoded in the edges. A well selected Laplace smoothing parameter can help improving accuracy. Sentiment label might help improving user political party classification. We end by calling for mechanisms to make users more aware of the potential biases in search results. We believe this is the first systematic step in this type of approaches in quantifying Twitter users behavior. The Re tweet matrix and re tweet average scores can be used to develop new models and algorithms to analyze more complex tweet-and-re tweet features. Our optimization framework can readily be adapted to incorporate other types of information. In future work we plan to improve various modeling components and gain a deeper understanding of the derived outcomes in collaboration with domain experts. For more general conclusions the consideration of different cultures and media sources is essential.

## 7. **REFERENCES**

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

[2]. References [Adamic and Glance 2005] Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 u.s. election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery, LinkKDD '05, 36–43. New York, NY, USA: ACM.

[3]. [Bakshy, Messing, and Adamic 2015] Bakshy, E.; Messing, S.; and Adamic, L. 2015. Exposure to ideologically diverse news and opinion on facebook.Science.

[4]. Lada A. Adamic and Natalie Glance. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In Proc. LinkKDD.

[5]. Solon Barocas and Andrew D Selbst. 2014. Big data's disparate impact. Available at SSRN 2477899 (2014).

[6]. Parantapa Bhattacharya, Muhammad Bilal Zafar, NiloyGanguly, SaptarshiGhosh, and Krishna P. Gummadi. 2014. Inferring User Interests in the Twitter Social Network. In Proc. ACM RecSys.

[7] Ahmed, S.; Jaidka, K.; Skoric, M.: Tweets and votes: A four-country

comparison of volumetric and sentiment analysis approaches. ICWSM 2016.

[8] Bakshy, E.; Hofman, J. M.; Mason, W. A.; Watts, D. J.: Everyone's an influencer: quantifying influence on twitter. WSDM 2011.

[9] Blei, D. M.; Ng, A. Y.; Jordan, M. I.: Latent dirichlet allocation. Journal of Machine Learning Research 3:993–1022, 2003.

[10] Chang, C.-C., Lin, C.-J.: Libsvm: a library for support vector machines. ACM TIST 2(3):27, 2011.

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

[12]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.

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

[14] .Le Chen, Alan Mislove, and Christo Wilson. 2015. Peeking beneath the hood of uber. In In Proc. of the 2015 ACM Conference on Internet Measurement Conference. ACM, 495–508.

[15] . M. Conover, B. Gonc¸alves, J. Ratkiewicz, A. Flammini, and F. Menczer. 2011a. Predicting the Political Alignment of Twitter Users. In Proc. IEEE SocialCom.

[16] . M. Conover, J. Ratkiewicz, Matthew Francisco, B Gonc¸alves, F. Menczer, and A. Flammini. 2011b. Political Polarization on Twitter.In Proc. AAAI ICWSM.

[17]. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.

[18] .Pennacchiotti, M., &Popescu, A. M. 2011. Democrats, republicans and starbucksafficionados: user classification in twitter. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 430-438.

[19] .Wong, F. M. F., Tan, C. W., Sen, S., & Chiang, M. 2016.Quantifying Political Leaning from Tweets, Retweets, and Retweeters. IEEE Transactions on Knowledge and Data Engineering, 28(8), 2158-2172

[20]P. Barbera,´ "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data," Po-litical Analysis, 2014.

[21]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.

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

[23]. USA Executive Office of the President. 2016. Big Data: A Report on Algorithmic Systems, Opportunity,and Civil Rights. http://tinyurl.com/Big-Data-White-House. (2016).

[24]. S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. 2006. Topical interests and the mitigation of search engine bias. Proc. of the National Academy of Sciences (PNAS) 103, 34 (2006), 12684–12689.

[25]. SaptarshiGhosh, Naveen Sharma, FabricioBenevenuto, NiloyGanguly, and Krishna Gummadi. 2012. Cognos: Crowdsourcing Search for Topic Experts in Microblogs. In Proc. ACM SIGIR.

[26]. Jennifer Golbeck and Derek Hansen. 2011. Computing Political Preference Among Twitter Followers. In ACM SIGCHI.

[27]. AnikoHannak, PiotrSapiezynski, ArashMolaviKakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In Proc. WWW.

[28] Weng, J.; Lim, E.; Jiang, J.; He, Q.: Twitterrank: finding topic-sensitive influential twitterers. WSDM 2010.

[29] Wong, F. M. F.; Tan, C.; Sen, S.; Chiang, M.: Quantifying political leaning from tweets, retweets, and retweeters. IEEE TKDE 28(8):2158– 2172, 2016.

[30] Yuan, Q.; Cong, G.; Ma, Z.; Sun, A.; Magnenat-Thalmann, N.: Who, where, when

and what: discover spatio-temporal topics for twitter users. KDD 2013.

[31] Zhao, W. X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; Li, X.: Comparing Twitter and Traditional Media Using Topic Models. ECIR 2011.

Name of the Guide:

**P.CHANDRA MOHANA RAI**

Designation :Asst.Prof

Mail Id:cmrai.pottasiri@gmail.com

**College Name and Address**: Krishnaveni Engineering College for women,keasanupalli, narasaraopet

**1.**Name of the Student:

**D. KRISHNAVENI**

Reg. No:**15KC1D5804**

krishnaveni.hasini@gmail.com

**(CSE)**