Quantifying Political Legitimacy from Twitter*

Haibin Liu and Dongwon Lee

College of Information Sciences and Technology
The Pennsylvania State University, University Park, PA
{haibin,dongwon}@psu.edu

Abstract. We present a method to quantify the *political legitimacy* of a populace using public Twitter data. First, we represent the notion of legitimacy with respect to k-dimensional probabilistic topics, automatically culled from the politically oriented corpus. The short tweets are then converted to a feature vector in k-dimensional topic space. Leveraging sentiment analysis, we also consider the polarity of each tweet. Finally, we aggregate a large number of tweets into a final legitimacy score (i.e., L-score) for a populace. To validate our proposal, we conduct an empirical analysis on eight sample countries using related public tweets, and find that some of our proposed methods yield L-scores strongly correlated with those reported by political scientists.

1 Introduction and Related Work

The term *political legitimacy* in political science refers to the acceptance of authority by a law, government, or civil system, and has been the subject of extensive study in the discipline. The concept is often viewed as "central to virtually all of political science because it pertains to how power may be used in ways that citizens consciously accept"[1]. As such, in political science, many proposals have been made to quantify the legitimacy of a populace. Some recent works such as [1, 2] have been well received in the community. While useful, however, such existing works are largely based on hand-picked small-size data from governments or UN based on an ad hoc formula. Therefore, it is still challenging to renew or expand the results from [1, 2] to other regions if there exist no reliable base data. To address this limitation, in this research, we ask a research question "*if it is possible to quantify political legitimacy of a populace from social media data*", especially using Twitter data. As a wealth of large-scale public tweets are available for virtually all populaces, if such a quantification is plausible, the application can be limitless. For instance, in the stochastic simulation environment such as NOEM [3], a quantified legitimacy score forms one of important input parameters. While there is currently no good way to synthetically generate a

* Part of the work was done while Dongwon Lee visited the Air Force Research Lab (AFRL) at Rome, NY, in 2013, as a summer faculty fellow. Authors thank John Salerno at AFRL for the thoughful feedback on the idea and draft. This research was also in part supported by NSF awards of DUE-0817376, DUE-0937891, and SBIR-1214331.

W.G. Kennedy, N. Agarwal, and S.J. Yang (Eds.): SBP 2014, LNCS 8393, pp. 108–115, 2014.

legitimacy score of a populace, one may be able to estimate it from the tweets generated from or closely related to the populace.

In recent years the exploitation of social media such as Twitter and Facebook to predict latent patterns, trends, or parameters has been extensively investigated. For instance, [4] computationally tried to classify tweets into a set of generic classes such as news, events, or private messages. In addition, [5–7] attempted to track and analyze the status of public health via social media data. Some even tried to predict stock market from public mood states collected from Twitter [8]. Studies have also been carried out about the correlation between tweets' political sentiment and parties and politicians' political positions [9, 10]. The case study about 2009 German federal election [9] reported a valid correspondence between tweets' sentiment and voters' political preference. Such studies also verify that the content of tweets plausibly reflects the political landscape of a state or region. Another paper [11] also aggregates text sentiment from tweets to measure public opinions.

While closely related, our method focuses on quantifying the political legitimacy, that is related to not only politics and elections, but also other concepts such as governments, laws, human rights, democracy, civil rights, justice systems, etc. To our best knowledge, this is the first attempt to computationally quantify the political legitimacy of a populace from a large amount of big social media data and conduct a correlation analysis against the results in political science.

2 The Proposed Method

Our goal is to build and validate a model to accurately quantify the political legitimacy score of a populace using tweet messages. The underlying assumption is that some fraction of populace would occasionally express their opinions on the status of political legitimacy. Two such examples are shown in Figure 1.

Fig. 1. Tweets related to legitimacy

Let us use the term **L-score** to refer to the political legitimacy score of a populace, scaled to a range of [0,10]. Then, our overall method consists of three steps: (1) identify and convert relevant tweets into computable feature space, (2) compute L-score of each tweet, and (3) aggregate L-scores to form a time series and compute final L-score of a populace. This overall workflow is illustrated in Figure 2.

2.1 Step 1: Vectorizing Tweets

Each tweet can be up to 140 characters but often very terse. The challenge of this step is to be able to accurately capture and extract critical features from

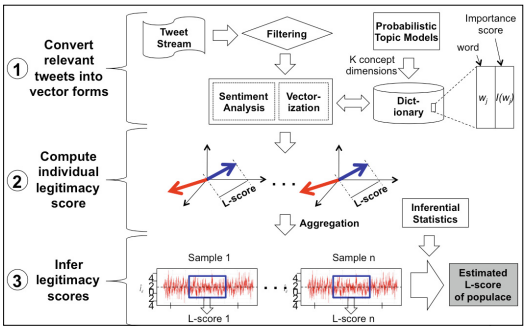


Fig. 2. Overview of the proposed method

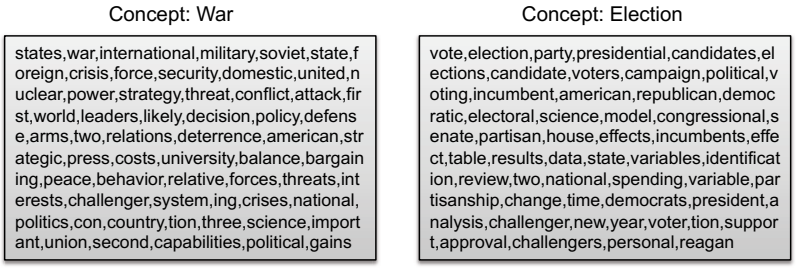


Fig. 3. Two prominent topics found from political science journal articles

short tweets that can indicate the opinion of a writer toward the status of legitimacy. Since there is no widely-accepted "computable" definition of legitimacy, we assume that the notion of political legitimacy is related to k-dimensional topics such as justice system, human rights, democracy, government, etc. While treating k as a tunable parameter in experiments, then, we simply attempt to represent each tweet as a k-dimensional vector, where the score in each dimension indicates the relevance of the tweet to the corresponding topic. Further, we use a dictionary of k dimension where each dimension (i.e., topic) contains a set of keywords belonging to the topic. Finally, we run a probabilistic topic modeling technique such as Latent Dirichlet Allocation (LDA) [12] over politically oriented corpus[1] and build such a k-dimensional dictionary.

Figure 3 illustrates two example topics found by LDA and prominent keywords within each topic (the labels such as "war" and "election" are manually assigned). Note that, although found automatically, such topics represent the main themes of the corpus reasonably well and can be viewed as related to the legitimacy. In addition, prominent keywords within each topic also make sense. Therefore, if a tweet mentions many keywords found in either topic, then the tweet is used to quantify the legitimacy. Suppose k topics are first manually selected and corresponding keywords in each topic are found using LDA. Imagine a k-dimensional dictionary such that a membership of a keyword can be

[1] `http://topics.cs.princeton.edu/polisci-review/`

quickly checked. For instance, one can check if the keyword "military" exists in the "war" dimension of the dictionary. Furthermore, suppose each keyword, w, in the dictionary is assigned an importance score, $I(w)$. In practice, a frequency-based score or LDA-computed probability score can be used to measure the importance of keywords. For instance, an importance of a word can be computed using the following frequency-based formula: $I(w) = \frac{freq(w)}{\sqrt{1+(freq(w))^2}}$. Using this data structure of the k-dimensional dictionary, we can convert tweets into vectors and then compute the L-score.

With such a topic dictionary, we can convert each tweet into a k-dimensional vector by checking membership of words in each dimension. Assume that a tweet, t, is pre-processed using conventional natural language processing (NLP) techniques such as stemming and represented as a bag-of-words, w, with n words: $t \Rightarrow w = \{w_1, w_2, \cdots, w_n\}$. Then, the k-dimensional vector representation of a tweet, v_t, is:

$$v_t \in R^k = [\alpha_1 \sum\nolimits_{\forall m_1 \in |w \cap D_1|} I(m_1), \cdots, \alpha_k \sum\nolimits_{\forall m_k \in |w \cap D_k|} I(m_k)]$$

such that $\sum_1^k \alpha_i = 1$, D_i refers to the i-th dimension of the dictionary, and α_i is the weighting parameter for the relative importance of the i-th dimension.

2.2 Step 2: Computing L-Scores of Tweets

The intuition to compute L-score of a tweet is that when a tweet either positively or negatively mentions keywords related to k-dimensions of the legitimacy, their "strength" can be interpreted as the legitimacy score. The L-score of the tweet, $L-score(t)$, is then defined as the magnitude (i.e., L2-norm) of v_t, with the sign guided by the sentiment of the tweet t–Δ_{sent}. Suppose $v_t = (x_1, ..., x_k)$. Then,

$$L - score(v_t) = \Delta_{sent} \|v_t\| = \Delta_{sent} \sqrt{x_1^2 + \cdots + x_k^2}$$

where Δ_{sent} indicates a $[-1, 1]$ range of sentiment polarity score of the tweet. Note that an alternative to this single Δ_{sent} per tweet is to allow for different sentiment polarity per dimension, Δ_i , in each tweet. However, in our preliminary study, as typical tweets are rather short and there are usually simply not enough information to determine different polarity score per dimension, we maintain a single sentiment score per tweet.

2.3 Step 3: Aggregating L-Scores of Tweets

Once the L-score has been computed for all tweets, we next need to aggregate all the L-scores per some "group" and determine the representative L-score of the group. One example grouping constraint can be a region (e.g., country such as Egypt or city such as Detroit). Suppose we want to aggregate all L-scores of the day d. Assuming the distribution of the daily L-scores follow the Gaussian Distribution, then, we compute the mean L-score of the day and apply the interval-based Z-score normalization, similar to [13], to the L-score.

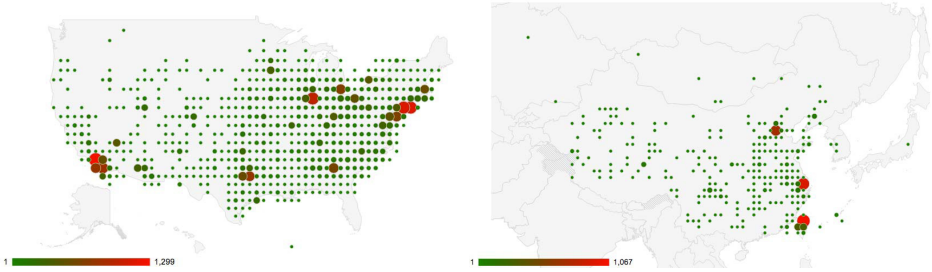
Table 1. L-scores published in [2]

Country	Score	Country	Score	Country	Score	Country	Score
Norway	7.97	Japan	6.13	India	5.21	Bulgaria	3.21
Canada	7.26	Thailand	5.89	France	5.03	Peru	3.44
Vietnam	7.07	United States	5.83	Brazil	4.68	Iran	2.04
New Zealand	6.78	South Africa	5.45	Slovenia	4.33		
Spain	6.64	China	5.36	Turkey	3.96		

Collecting such normalized L-scores over a time interval, finally, we derive a time series and employ standard time series analysis techniques to either compute the overall representative score of the entire time series, or predict future L-scores. For instance, in the current implementation, we used both moving average (MA) and auto-regressive MA (ARMA) models.

3 Empirical Validation

Since there is no ground truth to L-scores of populaces, as an alternative, we aim to see "if our method yields L-scores of populaces similar to those reported in [2]." For instance, Table 1 shows example L-scores reported in [2]. This, computed from UN and WHO data, is widely accepted in political science community. We chose eight countries with varying L-scores in [2]–i.e., Brazil, Iran, China, Japan, Norway, Spain, Turkey, and USA. We prepared two sets of data: (1) *Geo* dataset contains tweets generated within the bounding box of the geo-coordinates of each country of interest, and (2) *Keyword* dataset contains tweets that mention terms related to each country (e.g., a hash tag of "#USA"), regardless of their geo-coordiates. From 9/28/2013 to 11/6/2013, we collected a total of 300,450 tweets using Twitter streaming API that are written in English, and relatively meaningful (e.g., terse tweets with less than 4 words or location-based tweets having the form of "I'm at location" are removed). Figure 4, for instance, shows the geo-coordinates of tweets in the Geo dataset for USA and China. Table 2 summarizes tweets that we used in the experiments. We first present the aggregated mean L-score of crawled tweets during the monitored period.

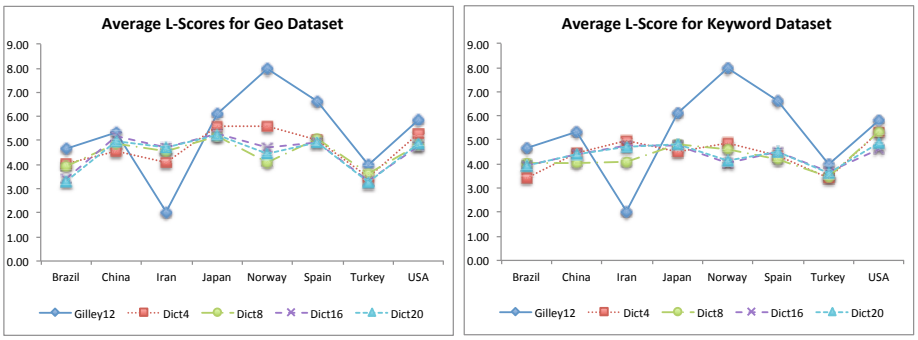


(a) USA (b) China

Fig. 4. Geo-cordinates of tweets in Geo datasets

Table 2. Summary of crawled tweets

	# Keyword tweets	# Geo tweets	# Filtered Geo Tweets
Brazil	10,924	18,788	14,715
China	17,848	8,060	7,569
Iran	51,743	9,600	6,594
Japan	13,112	9,948	9,427
Norway	6,561	5,633	5,554
Spain	15,845	13,094	12,477
Turkey	13,281	38,187	14,634
USA	28,801	39,025	38,662

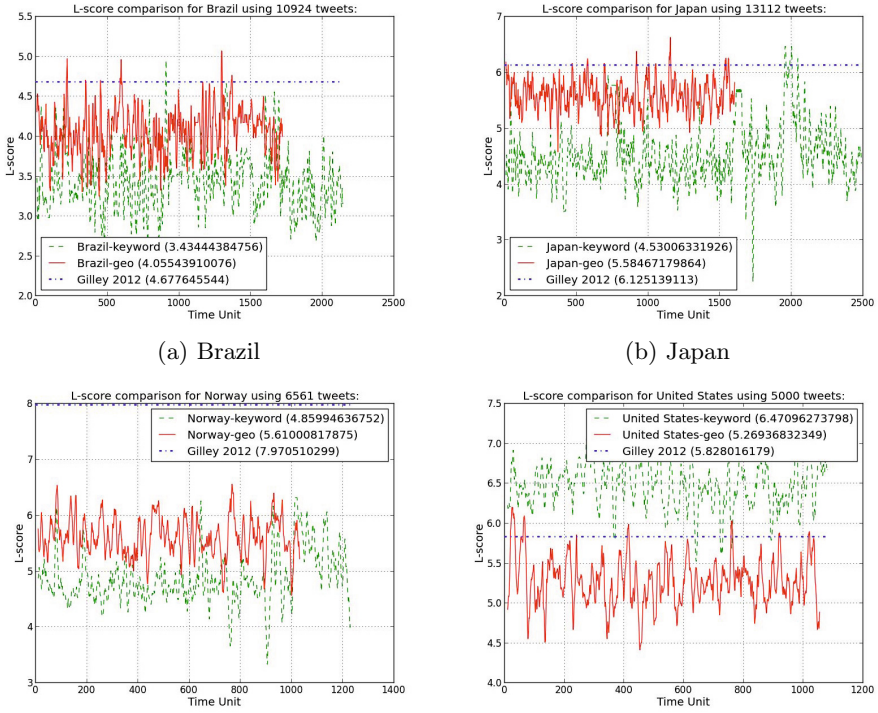


(a) L-scores of Geo dataset (b) L-scores of Keyword dataset

Fig. 5. Aggregated mean L-scores

Several factors are studied that may affect the final L-score. First, the number of topics obtained from LDA may play an important role in quantifying tweets' score. We tried different number of topics from 4 to 20, and the results are shown in Figure 5, where *Dict*4 means result from dictionary with 4 topics. Note that the range of the L-scores are rescaled to $[0, 10]$ to be compliant with the results of [2]. We can see that different number of topics lead to slightly different L-scores on both Geo and Keyword datasets. Studies are also carried out to see the impact of granularity of sentiment analysis in calculating L-scores. While previous results are calculated using sentiment polarity scaled in range $[-1, 1]$, we also tested with only extreme sentiment values of $\{-1, 1\}$. However, the L-scores using this extreme sentiment values show little difference. Figure 6 shows time-series of 4 countries using 4 LDA topics on Geo and Keyword datasets. In most cases, L-scores estimated from Geo tweets match better than those estimated from keyword tweets. Note that compared to L-score of [2], our estimation of L-score matches well for some countries but poor for others (e.g., Norway).

To see the overall correlation with [2], we computed the *Pearson correlation coefficient* (PCC) [14] between the L-scores of all of our methods (using different number of topics or sentiment values) and [2]. As shown in Table 3, the best performer is the *Dict*4 over Geo dataset. With the coefficient value of 0.7997887 (P-value = 0.01717), we can claim a significant correlation between L-score computed



(a) Brazil (b) Japan

(c) Norway (d) USA

Fig. 6. L-score time series of 4 countries with 4 LDA topics on Geo dataset

Table 3. PCC values between L-scores of our proposed methods and [2]

	Keyword	Geo	Keyword-Extreme	Geo-Extreme
Dict4	0.203461755	**0.799788652**	0.214170058	-0.452748888
Dict8	0.472864444	0.233350916	0.401502605	-0.538886828
Dict16	-0.063411723	0.375603634	0.27090464	-0.594296533
Dict20	0.070540651	0.307136136	0.188031801	-0.631019398

using *Dict*4 and Geo dataset and that reported in [2]. This discovery also indicates that tweets directly generated from the territory of a region (i.e., Geo dataset) is a better source to quantify L-score than those conceptually related to a region (i.e., Keyword dataset).

4 Conclusion

We study the problem of quantifying political legitimacy of a populace based on public Twitter data. We propose a solution that converts short tweet text messages into a number of topic dimensions using probabilistic topic modeling. We leverage sentiment analysis to evaluate polarity of each tweet, and aggregate a

large number of tweets into the final legitimacy score of a populace. Our experiments over real tweets collected about eight countries reveal that some configuration of our proposal shows a strong correlation to results reported in political science community. Despite the promising result, there are a set of *limitations* to our study: (1) To derive a more definite conclusion on the validity of our proposed method in quantifying the legitimacy, a more comprehensive experiment is needed–e.g., more number of countries, larger tweet datasets, or topics derived from different corpus; (2) While [2] is a reasonable "beta" ground truth for our study, there is no formal analysis why or how accurate it is. As such, more correlation analysis of our proposal using different methods to compute the legitimacy is needed; (3) In addition to social media such as Twitter, other large-scale data can be used as a source of legitimacy. For instance, a dataset such as GDELT[2] contains a large-scale rich data on world-wide conflicts and can be used to infer the legitimacy status expressed by a populace.

References

1. Gilley, B.: The meaning and measure of state legitimacy: Results for 72 countries. European J. of Political Research 45(3), 499–525 (2006)
2. Gilley, B.: State legitimacy: An updated dataset for 52 countries. European J. of Political Research 51(5), 693–699 (2012)
3. Salerno, J.J., Romano, B., Geiler, W.: The national operational environment model (noem). In: SPIE Modeling and Simulation for Defense Systems & App. (2011)
4. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: ACM SIGIR, pp. 841–842 (2010)
5. Paul, M.J., Dredze, M.: You are what you tweet: Analyzing twitter for public health. In: ICWSM (2011)
6. Lamb, A., Paul, M.J., Dredze, M.: Investigating twitter as a source for studying behavioral responses to epidemics. In: AAAI Fall Symp. (2012)
7. Evans, J., Fast, S., Markuzon, N.: Modeling the social response to a disease outbreak. In: Greenberg, A.M., Kennedy, W.G., Bos, N.D. (eds.) SBP 2013. LNCS, vol. 7812, pp. 154–163. Springer, Heidelberg (2013)
8. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. of Computational Science 2(1), 1–8 (2011)
9. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: ICWSM, pp. 178–185 (2010)
10. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Election forecasts with twitter how 140 characters reflect the political landscape. Social Science Computer Review 29(4), 402–418 (2011)
11. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: ICWSM (2010)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
13. Bollen, J., Mao, H., Pepe, A.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: ICWSM (2011)
14. Rodgers, J., Nicewander, A.: Thirteen ways to look at the correlation coefficient. The American Statistician 42(1), 59–66 (1988)

[2] `http://gdelt.utdallas.edu/`