Geovisual text analytics for exploring public discourse on Twitter: A case study of immigration tweets before and after the January 27, 2017 Travel Ban

Caglar Koylu, Bryce J. Dietrich, Ryan L. Larson

ABSTRACT: Social media provides a unique opportunity to study geographic variation and evolution of content and sentiment of publicly shared opinions. However, it is challenging to identify the general patterns and trends in public discourse due to the complexity of the tone, topic, geographic and temporal variation in the way people express their opinions. This paper examines how public discourse varies across political geographies and over time by introducing a geovisual analytics environment that integrates topic modeling and sentiment analysis with spatiotemporal visualization. Our proposed study makes two major contributions. First, we contribute to the analysis of public discourse on immigration by introducing a natural experiment by collecting and analyzing tweets related to immigration a month before and a month after the 2017 Travel Ban. Second, we introduce a geovisual analytics framework that allows simultaneous and linked views of topical themes and sentiment patterns of public discourse on immigration as well as the variation in patterns across states and the eight-week time period before and after the ban. Preliminary results of this study revealed that the intensity of tweets about immigration substantially increased right after the immigration ban, and the geographic intensity of online discourse was correlated with the protest events held at a number of airports.

KEYWORDS: Geovisual analytics, topic modeling, sentiment analysis, Twitter, immigration, public discourse

1. Introduction

Diffusion of policies are often influenced by citizens who express their opinion through public discourse on Twitter (Vasi et al., 2015). Public discourse often varies by states, political and administrative areas, and evolves in reaction to policies and real world events. Social media provides a unique opportunity to study geographic variation and evolution of content and sentiment of publicly shared opinions, and identify political geographies of public discourse. For example, on January 27, 2017, President Donald Trump suspended the entry of people into the U.S. from a number of predominantly Muslim countries. In response, thousands of people flooded airports across the country to protest what the travel ban meant for democracy in the United States and elsewhere. Similar protests were observed on Twitter. This paper explores the regional, national and temporal progression of tenor of these online protests, both in terms of their tone and topic.

Our proposed study makes two major contributions. First, we contribute to the analysis of public discourse on immigration by introducing a natural experiment by collecting and analyzing tweets related to immigration a month *before* and a month *after* President Trump signed Executive Order 13769. The vast majority of previous studies analyzed Twitter protests, including those that focused on the Arab Spring, looked at tweets after the even occurred. For example, Bruns, Highfield and Burgess (2013) studied the Arab

Spring using tweets from January to November 2011 which is a full month *after* the Arab Spring began on December 17, 2010. Our data collection and analysis enable the natural experiment in which we can determine how and in what context the discussion of immigration changed *after* the travel ban was announced. Second, we introduce a geovisual text analytics framework that allows simultaneous and linked views of topical themes and sentiment patterns of public discourse on immigration as well as the variation in patterns across states and the eight-week time period before and after the ban.

2. Data and Methods

Data

Using the Twitter Streaming API, we collected tweets that contain the keywords related to immigration, and specifically Muslim refugees and immigrants (i.e., "immigration", "immigrant", "muslim", "Islam", "refugee") in a number of languages including English, Arabic, French, Spanish, Turkish, and Persian a month before and a month after the Immigration Ban put in place on the January 27th , 2017. There were 84,159,489 immigration tweets world-wide from December 27, 2016 to February 27, 2017. The locations of 99 % of these tweets can be identified at the state or country level, and 21% (~17 million tweets) of the geo-located immigration tweets were generated in the Continental U.S by 759,171 users. Figure 1 illustrates the temporal distribution of immigration tweets which highlights a large increase following the day the first travel ban was announced. This provides an initial evidence of the extensive online protest we reference in the introduction and demonstrates the "natural experiment" we aim to leverage in our study.

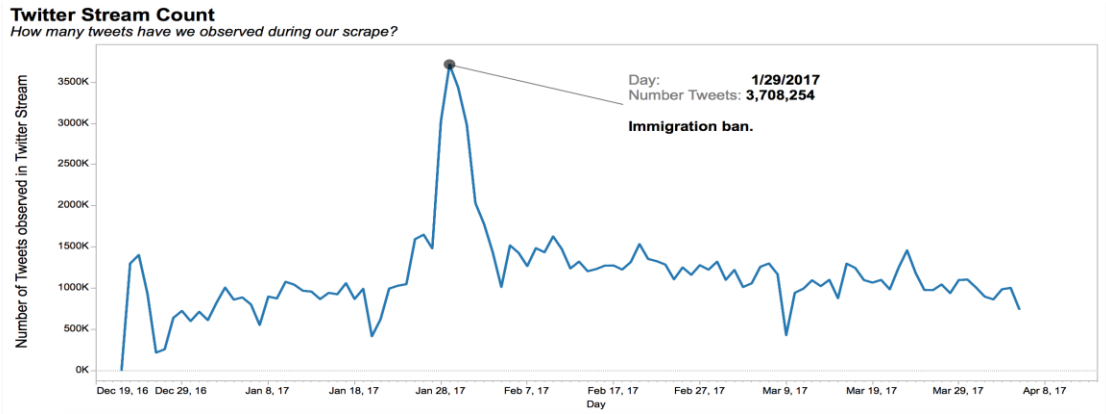

Figure 1: Immigration tweets per 1,000 people

Out of the 17 million tweets generated within the U.S., 49% of the tweets were re-tweets. Retweets were excluded in topic modeling and sentiment analysis, while they were used in identifying opinion leaders. Figure 2 shows the multiplicity the diversity of Twitter participants including celebrities, journalists, academics, politicians and activists whose tweets were retweeted extensively. This finding reveals that the online protest was not focused on one specific group, but ultimately reached a diverse audience.

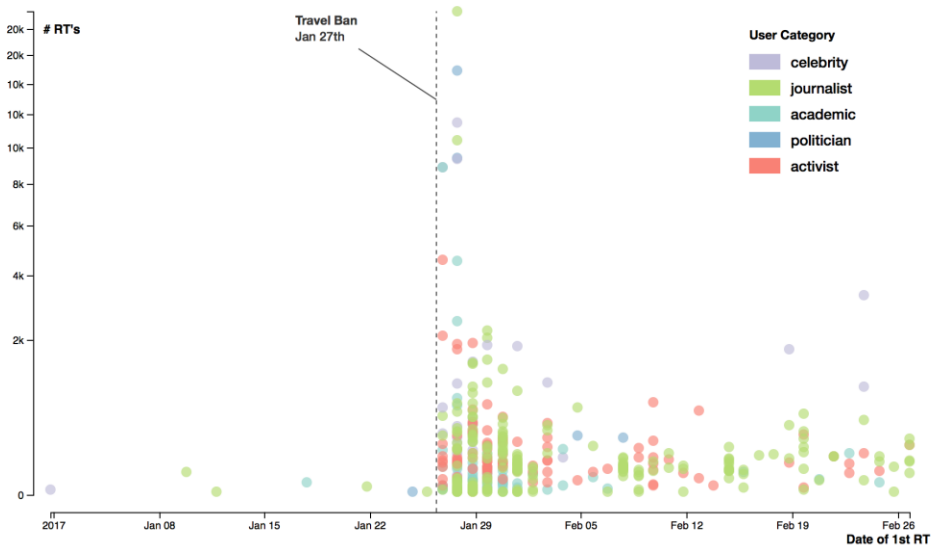

Figure 2: Opinion leaders: users with the most number of retweets before and after the ban

Extracting content and sentiment from tweets

First, pre-processing phase is used to clean and partition the data into eight weekly time periods a month before and after the travel ban. At this phase, tweets within each state are combined into documents into the eight week-period before and after the ban to obtain temporally organized document collections. Aggregation by states address the short-text problem in topic modeling (Yan et al., 2013) caused by the 140 character limit, and inability of words to occur within a single tweet. We employ a separate Latent Dirichlet Allocation (LDA) on the document collection for each of the eight-week time period before and after the ban. LDA is a generative process that relies on term frequency-inverse document frequency (tf-idf), which reflects how important each word is to a document in a collection of documents or corpus, and its value increases proportionally to the number of times a word appears in a document (Goldstone & Underwood, 2012). LDA has been used in a variety of studies to identify event locations, geographical variation of linguistics, and topical themes, and provide recommendations based on location from Twitter data (Chae et al., 2012; Lai, Cheng, & Lansley, 2017; Lansley & Longley, 2016; Liu, Ester, Hu, & Cheung, 2015; Pozdnoukhov & Kaiser, 2011).We employ LDA to identify representative topics, ultimately ranging from specific issues (i.e., xenophobia or border security) to different types of phrasing (i.e., formal versus informal language).

In addition to extracting topical themes from immigration tweets, we employ the Linguistic Inquiry Word Count (LIWC) in order to identify positive and negative sentiment, and the patterns across states over the eight-week period. Previous studies (Alpers et al., 2005; Bantum & Owen, 2009; Kahn et al., 2007) found that the "negative emotion" category outperformed all other LIWC categories, leading these authors to conclude that this category could be used to track changes in expression of negative emotions in on-line groups. Kahn et al. (2007) argued that the LIWC provides a meaningful indicator of emotion that may be used as an alternative or complement to

self-reports of emotion. Similar to these previous studies, we utilize the count of the number of "positive" and "negative" words in identifying overall sentiment of public discourse grouped by states and the eight-week time period.

3. Preliminary Results

Figure 3 illustrates the number of immigration tweets within the eight-week period normalized by state population. Figure 3 suggests that the highest density of tweets centered on protest events in important airports like those found in New York, Chicago, and Boston. Since airports were the focal point of many protest events, we can use the proximity to airports to explore whether those demonstrations actually influenced the way people discussed immigration. If they did have an effect, then you would expect the greatest change in both the topic and tone of Twitter discussions to occur in close proximity to airports. Figure 3 demonstrates the initial layout, and the potential of the geovisual analytics framework for spatiotemporal comparisons of public discourse. This layout will be combined with the coordinated views of (1) the temporal progression of the public discourse both in terms of its content and sentiment; (2) topic word clouds illustrating the extracted themes from topic modeling, and (3) the sentiment visualizations that highlight the changing patterns of the tone in public discourse on immigration.

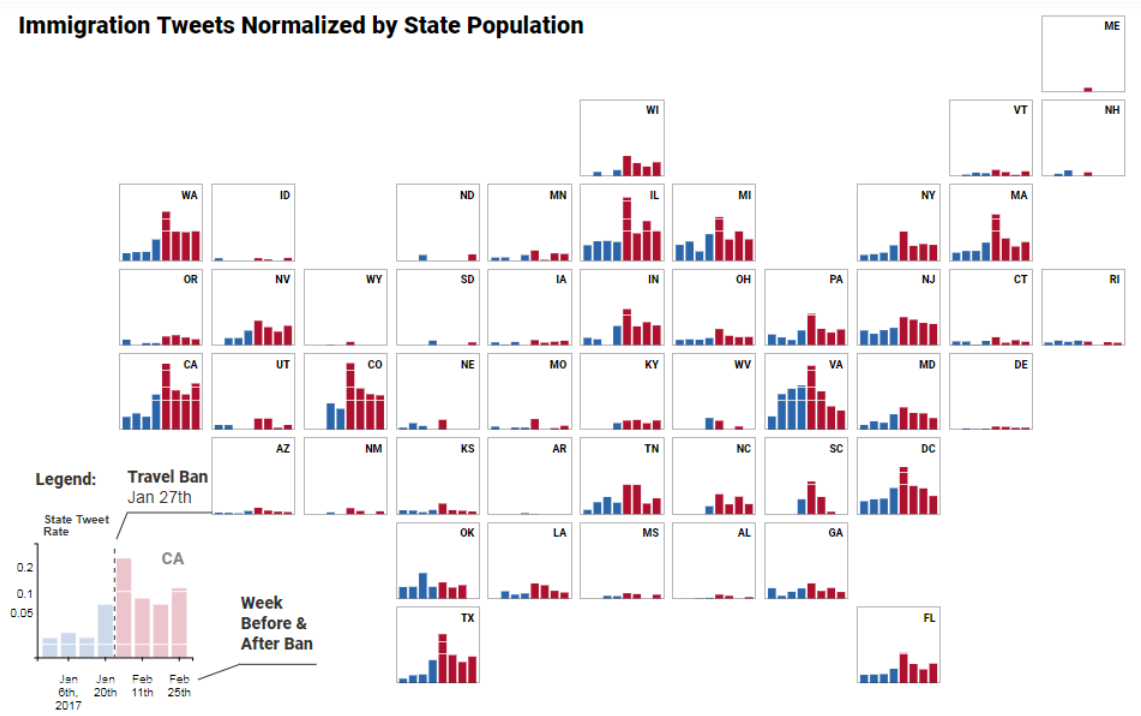

Figure 3: Spatiotemporal cartogram that illustrates the normalized frequency of immigration tweets per state.

Conclusion and Future Work

With this paper, we not only provide a methodology for gaining a greater understanding of a very important recent moment in American political history, but we also contribute to the broader discussion of how online protests emerge and ultimately influence broader political discussions. The next step is to integrate the results of topic modeling and

sentiment analysis in the geovisual analytics environment to explore both regional and national variation, and the changing patterns of content and sentiment of public discourse on immigration over the course of the eight-week period including the month before and after the first travel ban.

References

Alpers, G. W., Winzelberg, A. J., Classen, C., Roberts, H., Dev, P., Koopman, C., & Taylor, C. B. (2005). Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior, 21*(2), 361-376.

Bantum, E. O. C., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological assessment, 21*(1), 79.

Bruns, A., Highfield, T., & Burgess, J. (2013). The Arab Spring and social media audiences: English and Arabic Twitter users and their networks. *American Behavioral Scientist, 57*(7), 871-898.

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., & Ertl, T. (2012). *Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition.* Paper presented at the Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on.

Kahn, J. H., Tobin, R. M., Massey, A. E., & Anderson, J. A. (2007). Measuring emotional expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, 263-286.

Lai, J., Cheng, T., & Lansley, G. (2017). Improved targeted outdoor advertising based on geotagged social media data. *Annals of GIS, 23*(4), 237-250.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems, 58*, 85-96.

Liu, Y., Ester, M., Hu, B., & Cheung, D. W. (2015). *Spatio-Temporal Topic Models for Check-in Data.* Paper presented at the Data Mining (ICDM), 2015 IEEE International Conference on.

Pozdnoukhov, A., & Kaiser, C. (2011). *Space-time dynamics of topics in streaming text.* Paper presented at the Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks.

Vasi, I. B., Walker, E. T., Johnson, J. S., & Tan, H. F. (2015). "No fracking way!" Documentary film, discursive opportunity, and local opposition against hydraulic fracturing in the United States, 2010 to 2013. *American Sociological Review, 80*(5), 934-959.

Caglar Koylu, Assistant Professor, Department of Geographical and Sustainability Sciences, University of Iowa, Iowa City, IA, 52242

Bryce Dietrich, Assistant Professor, Department of Political Science, University of Iowa, Iowa City, IA, 52242

Ryan L. Larson, Computer Science, University of Iowa, Iowa City, IA, 52242