Socio-Temporal Trends in Urban Cultural Subpopulations through Social Media

Lander Basterra, Tyler Worthington, James Rogol and Donald Brown
University of Virginia, llb2eu, tjw4ry, rogol, and deb@virginia.edu

***Abstract* – Understanding when, where, and how increasingly diverse and dynamic subpopulations interact in urban environments is critical to the integrity of the city as a whole. This knowledge can facilitate the development of communication strategies, urban planning methodologies, and resource allocation to best serve citizens. Previous research focused either upon large temporal trends for the city as a whole, or mapped citizens in social or geographic space using broad categories (such as shared interests on social media or one's racial designation). Whereas these studies focused on a single geographic area, this paper includes Twitter data from three culturally distinct metropolitan areas over the same 92-day period: Los Angeles, CA, United States; Chicago, IL, United States; and Istanbul, Turkey. The global embrace of the Twitter social media platform provides the primary data source for a methodology applicable to any urban area. For Tweets emanating from within each city, a bag-of-words approach to the messages' textual content creates topical clusters within the most frequently occurring languages. These classifications transcend traditional racial designations. Time series analysis of each Tweet's timestamp reveal that the volume of tweets across topics is significantly correlated with major regional events. Furthermore, certain subpopulations' postings rise and fall in sync with others. Examining the trends among strongly correlated topic groups provides an indication of how these groups might interact.**

Index Terms - Latent Dirichlet Allocation, Time Series Analysis, Twitter.

INTRODUCTION

Globalization and a rise in imigration are causing urban populations to grow at an astounding rate; the United Nations projects the number of individuals living in urban environments to surpass 6 billion by 2045 [1]. As the population grows, so does the dynamism and diversity of the subpopulations within society. Understanding not only which groups are present within a city, but how, when, and where they interact will aid policy makers at the local, regional, and national levels. Privately, local businesses and industries can better target their respective clienteles with this knowledge, as well. Combining these insights at the public and private level can potentially improve the day-to-day lives of ordinary citizens.

This is by no means a new idea – the newly founded United States undertook its first census in 1790. Though the survey evolved to include economic, ethnic, and educational categories, the Census continues to lump individuals into broad groupings. In an effort to expedite data collection, the Census Bureau started conducting an annual American Communities Survey (ACS) in 2002. While this yearly survey suppliments the decentenial Census, profiling the citizens of the United States remains an inexact, time-consuming, and expensive endeavor [2].

Worse still, the tabular format of Census and ACS data can obscure boundaries between even the most broad demographic categories, such as race. Projects like the "Racial Dot Map" made this content more interpretable by placing one dot per person, color-coded by race, on a geographic map [3]. While this clearly locates racial groups in space, this process ignores the not inconsequential differences between ethnic subgroups. For example, despite social and geographic differences, individuals of Japanese, Taiwanese and Malasian descent all fall under the broad heading of "Asian." In the modern age, there must be a better way to represent uniqueness of various cultures.

In particular, cities currently produce vast amounts of data daily. Citizens directly contribute to this flow of information, posting to social media networks like Foursquare, Flickr, and Twitter throughout the day. The rise of social media transcends international boundaries, uniting users from Los Angeles to Dubai and Shanghai to London [4]-[7].

Tapping into the constant flow of messages in urban areas provides a unique opportunity to gather information about a city in near-real time, straight from the source. The textual content of each post offers insight into the thoughts of the populous, while the timestamp can reveal how these ideas fluctuate throughout months and weeks.

This paper provides a more granular analysis of temportal trends in urban areas. Latent Dirichlet Allocation (LDA) clusters Twitter users based upon shared terms, producing subpopulations within language groups. A time series analysis examines the trends underlying each group's posting habits, how they react to external events, as well as the correlations which exist between groups. To best exhibit the flexibility of this approach and incorporate a spatial element, it was developed using the English speaking market of Los Angeles, CA, but was tested in Chicago, IL (a domestic analogue) and Istanbul (as an international market without an English-speaking majority).

PREVIOUS WORK

Whether incorporating venue-based location data like Foursquare and Instagram, or geolocations based on a user's GPS coordinates, social media outlets offer a near instantaneous glimpse at the status of a city. Though the user bases are not explicitly representative of the population, the geographic scope and heavy usage (6.8 million location tagged tweets in the greater Los Angeles area alone between December 2013 and January 2014) have been cross-validated against Census and population survey data in both the United States and United Kingdom [6][7].

Given the ubiquity of so-called smart phones in the pockets of individuals around the globe, these devices provide a reliable source of location throughout the day [4]. The timestamp and location of social media posts revealed broader trends in the ebb and flow of public transport and roadway usage. When visualized, these patterns create a metaphorical "breathing" or "heartbeat" of the large-scale urban environment over time [8][9].

Of existing social media platforms, Twitter has generated considerable interest from the academic community, as well as businesses and policy makers. Linguistic, spatial, and temporal similaries between Tweets have been used to predict the stock market, elections, box office returns, and epidemics [10].

Despite this attention, work seldom explored the spatio-temporal trends of indidivual groups of Twitter users and the interactions among them, focusing instead on either the spatial or time-based properties of social media data. For example, De la Rosa et al. used LDA and K-means clustering to classify Twitter messages based upon frequently co-ocurring hashtags [11]. However, they did not investigate how these key terms changed over time. McKenzie et al. employed LDA to group metropolitan points of interest (universities, concert halls, restaraunts, etc) based upon text culled from venue descriptions on Foursquare. They found thematically similar establishments experienced a similar pattern of user frequency over time [12].

Bag-of-words topic modeling algorithms like LDA can be adapted to create thematic clusters by geographic region, as well. Models trained on place-based documents like travelogues enabled Speriosu et al. to successfully visualize linguistic dissimilarities between spaces [13]. More importantly, this approach worked across multiple languages. Social media provided a similarly ample corpus for regional topic modeling, as shown by Abdelhaq et al [14]. They did not, however, investigate how these key terms changed over time.

In contrast, Becker et al. undertook an episodic approach to social media content. The text and timestamp of Tweets was found to correlate with major real-world events [15]. Similarly, O'Connor et al. explored the link between the sentiment of Twitter messages and public surveys. A time series approach found correlations between Tweets and consumer confidence and political opinion polls as high as 80%, indicating that social media captures overall trends among the public [16]. With regard to forecasting trends, Signorini et al. showed that a focused analysis of Tweets related to influenza symptoms improved predictions of outbreaks by 7 to 14 days [17]. Not only did this study again link social media with public trends, but it worked at both the regional and national levels.

The discoveries of latent groups and trends culled from the geographic, temporal, and texual markers in social media are certainly insightful. Whereas existing work operated at a high level, the opportunity exists to refine these methodologies, and consider the characteristics of the individual users who post online.

DATA

I. Acquisition

Twitter's Streaming API provides access to posts on the social media site in near-real time. Each message in the stream is a self-contained JSON object containing a unique ID, timestamp, textual content, geolocation (determined by a poster's exact coordinates and/or generalized city of origin), and additional metadata for both the Tweet and author [18]. Queries to the API can include a bounding box determined by the latitude and lonitude of the southwest and northeast corners of a rectangular region (in degrees), filtering out messages eminating from beyond the boundary.

Over a 92-day study period (October 28, 2016 to January 27, 2017), Python scripts running on Amazon EC2 instances continuously queried the Streaming API for Tweets originating from three major metropolitan areas (Los Angeles, Chicago and Istanbul). Python extracted the following features from each JSON object and inserted the data in a PostgreSQL database for storage:

- Unique message ID
- Timestamp in Coordinated Universal Time (UTC)
- Source client (i.e. Twitter for Android, Instagram, etc.)
- Message content
- Two-letter code of the most likely language of the text
- Unique user ID and handle
- A user's general home location
- Two-letter code of a user's chosen language
- User description
- Latitude and longitude of geotagged posts

Users can incorporate posts from other authors into their own messages by either "retweeting" or "quoting" them, and should be included with original content when topic modeling. Whereas retweets appear inline with an author's response, quoted tweets are an additional JSON key/value pair. As such, this data was extracted and stored separately in the PostgreSQL database.

Another query to the API obtained a JSON object containing a list of the 36 languages officially supported by Twitter and the corresponding two-letter codes based on Internet Assigned Numbers Authority designations (i.e. "en" for English, "es" for Spanish, etc.) [18].

II. Preprocessing

Messages underwent several cleaning steps in preparation for topic modeling. Exploratory analysis of the Los Angeles area uncovered automated messages pertaining to the weather and job openings. In all three test markets, a closer examination of these Tweets revealed that these "Twitterbots" are not posted by mobile apps, Twitter's web client, nor other consumer platforms. As the sources varied by market (i.e. "altın dükkan twitter robotu," the "gold shop Twitter robot" in Istanbul; or "TweetMyJOBS" in LA), excluding offending sources by hand would prove inefficient and time consuming. Akin to Gao et. al, a PostgreSQL query limited observations to those from a list of universally adopted clients (Instagram, Twitter for iOS, etc.) [14]. In Los Angeles alone, "Twitterbots" accounted for 13.5% of all Tweets. The number of posts before and after filtration appear in Table I.

TABLE I
SUMMARY OF TWEETS, BY CITY

City	Raw Tweets	Filtered Tweets	Top User Languages
Los Angeles	10,476,477	9,063,602	English (98.14%) Spanish (0.77%)
Chicago	5,967,115	4,978,647	English (99.15%) Spanish (0.41%)
Istanbul	6,213,382	5,072,874	Turkish (89.30%) English (9.14%)

Once filtered, individual and quoted messages were stripped of hyperlinks, as the links do not inherently add value. Obtaining additional content by following these links could suppliment the 140-charater Tweet, but doing so was beyond the scope of this paper and will be left to future research.

Twitter employs the "@" character as a prefix for references to other users (i.e. @dsi_culturalmap); and the pound sign prefaces "hashtags," a kind of key word (i.e. #datascience). Removing these special characters rendered user handles and hashtag topics as individual words, enabling topic modeling to consider "#cubs" and "cubs" as the same term.

Each message's timestamp was retreived from the JSON object in UTC. These standardized timestamps were converted to local time in PostgreSQL, so that the study period corresponded with the same calendar dates for each test market. The date, time (rounded to the nearest hour), and day of the week were extracted to aid in time series analysis.

MODELING PROCESS

I. Topic Modeling

Latent Dirichlet Allocation (LDA) clusters text by iterating through a collection of documents (the corpus) and identifying distributions of frequently co-occurring terms (topics). Each document in the corpus is then assigned a topic probabilistically, based upon terms' frequencies within a given document. According to Blei, et al., these probabilities can provide explicit representations of documents with regard to the entire corpus [19]. For the purposes of this research, similar textual content can reveal shared interests.

Twitter places a 140-character limit on individual messages. Given the finite number of terms within a Tweet, LDA's Bayesean approach struggles to find commonalities across such sparse documents. Under the assumption that the entirety of a user's Twitter presence encapsulates his or her interests better than a single message, longer documents were formed by concatenating each author's Tweets, user description, and messages she or he quoted.

Table I shows that a single user language accounts for upwards of 89% of Tweets in a city. To avoid topics biased towards the majority, users were grouped into corpora based on the author's chosen language. This approach created groupings more granular than traditional racial classifications.

Istanbul's Turkish corpus and the English corpora in all three markets accounted for at least 95% of the total Tweets in a region, and were submitted for topic modeling. Each document in a corpus was tokenized on white space using Python's open-source Natural Language Toolkit (NLTK) [20]. Stop words were removed after supplimenting NLTK's tools with stop words from additional languages like Turkish [21]. The remaining tokens of similar denotation were then aggregated via lemmatizing to shrink the feature space.

Python's *gensim* library calculated the TF-IDF scores for every token *t* in every document *d* within a corpus. [22].

$$TFIDF(t,j) = f_{t\in j} \log_2 \frac{D}{f_{t\in d, d\in D}} \quad (1)$$

Equation (1) multiplied the frequency of a token *t* within document *j* by the logarithm of inverse document frequency (the total number of documents *D* divided by the tally of all documents in which *t* appeared).

LDA ran on matrices comprised of terms appearing in at least 75, 65, 50, 30, 25, 15, 7.5 and 3.5% of documents (as calculated from the TF-IDF scores). Lowering the threshold further reduced the dimensionality of the data, producing subjectively more interpretable topics while improving topic assignment distribution. 20 topics were generated a the 3.5% threshold for each city, with 17 Turkish topics and 3 English ones in Instanbul. Table II contains sample topics, along with subjective labels provided by the authors of this paper. Production languages which did account for at least 5% of a city's Tweets were treated as purely language-based topics.

TABLE II
SAMPLE TOPIC MODELING RESULTS: TOP TERMS AND SUBJECTIVE LABELS

City	Top 10 Words	Topic
LA	namm, songwriter, home, cali, music, producer, life, singer, soccer, night	"Music"
LA	la, hillary, obama, clinton, donald, voted, trump, american, russia, racist	"Politics"
Chicago	flythew, worldseries, gocubsgo, wrigley, united, insta, hamilton, field, canada, parade	"Cubs"
Chicago	indiana, basketball, Instagram, coach, varsity, football, central, county, official, baseball	"Sports"

II. ARIMA Models

Time series processes can be estimated by Autoregressive Integrated Moving Average (ARIMA) models. These linear combinations of previous time points (auto-regression) and prior error terms (moving averages) can also remove polynomial trends by incorporating the difference between the response at time x_t and a measurement h points before, x_{t-h}. Removing these trends coerce the data into a series with a constant mean and a covariance structure which depends solely on the time between points.

Models to describe how the topics' volume of Tweets fluctuated over each of the 92 days in the collection period were systematically produced en masse with R's *forecast* package [23]. The auto.arima function explored all possible combinations of autoregressive (p) and moving average (q) terms, in conjunction with a number of differences (d). The optimal model minimized the bias-adjusted version of Akaike's Information Criterion (2).

$$AICc = -2\log L + 2k + \frac{2k(k+1)}{n-k-1} \qquad (2)$$

AICc includes parameters where L is the maximum likelihood of the model, k is the number of parameters, and n is the number of observations. Given the small sample size of the data (n = 92), penalizing the maximum likelihood by adding a term proportional to the number of parameters k helped mitigate overfitting.

RESULTS

The resulting models indicate that no single structure represents the subpopulations within a city. Of the top 20 groups (by raw volume) in Los Angeles, 4 are Auto Regressive, 0 are Moving Averages, 2 ARMA processes, and 14 are ARIMA models. Chicago has 2 AR, 3 MA, 3 ARMA, and 13 ARIMA models; while Istanbul sees 4, 3, 3, and 10, respectively. Examples from each location are summarized in Table III.

TABLE III
ARIMA MODELS FOR SELECT TOPICS PER CITY

Topic	City	Model	μ (Tweets)	Tweets per User
0 - English	LA	ARIMA(1,1,1)	35,668	19.64
3 - English	LA	ARIMA(0,1,2)	8,257	58.39
4 - English	LA	ARIMA(2,1,3)	4,151	98.78
Spanish	LA	AR(3)	755	8.45
0 - English	Chicago	ARMA(3,2)	28,800	24.94
1 - English	Chicago	ARIMA(5,1,1)	12,742	29.71
Spanish	Chicago	ARIMA(0,1,1)	221	10.02
15 - English	Chicago	AR(1)	60	1.93
0 - Turkish	Istanbul	ARMA(2,2)	23,994	21.18
0 - English	Istanbul	ARIMA(1,1,3)	2,427	10.53
4 - Turkish	Istanbul	MA(4)	2,400	58.41
Arabic	Istanbul	AR(1)	212	11.46

Over half of the models (37 of 60) include a difference term. Two factors could contribute to the lack of stationarity in this majority. The study period includes several events which impacted the volume of tweets (see section III below). These extreme fluctuations could affect the variance of the topical time series, which would render the processes non-stationary. Furthermore, there could be a slight decreasing linear trend in volume, perhaps as a consequence of users spending time with family over the winter holidays. Whether this is a distinct trend or part of a larger seasonal trend would require a longer study period.

The overwhelming majority of models (56 of 60) contain some order of auto-regressive term. No one model dominates a city, though six models in Chicago were estimated as ARIMA(5,1,0), including five of the top six topics by volume. One such topic shares this structure and identical slang terms (e.g. "wit," "tryna," and "bout") as a group in Los Angeles. As other ARIMA(5,1,0) models in Chicago contain words of a distinctly different connotaion ("womensmarch," "god" and "peace"), the similarities are likely the effect of chance, as opposed to some inter-regional trend. However, the fact that an autoregressive term appeared in 90% of the top 20 topics in each city demonstrates the influence of the number of past Tweets on the current day's volume.

III. Cross-Process Correlations

Despite the different processes at work, initial time series plots of topics appear to ebb and flow together through time. Statistically significant correlations reveal that a given pair of topics move together through time. To compensate for the disproportionate number of members in different topics, the raw daily volume of Tweets for a group is divided by the count of it's unique users. Hierarchical clustering recombines the top 20 topics into five correlated groups. A complete linkage function maximizes the distance between clusters using (3) as a measure of distance and dissimilarity.

$$dissimilarity = 1 - correlation \qquad (3)$$

Plotting the daily medians for each correlated cluster produces Figure I. The dashed lines represents a 7-day rolling average of the medians of all topics within a cluster. The general trends in the domestic markets are similar, though Los Angeles' daily tallies vary twice as much as Chicago's (σ^2 = 0.0013 and 0.0005, respectively). Both cities experience distinct peaks at the beginning of the collection period, corresponding with the World Series on November 2 and the day after the U.S. presidential election one week later. The unprecedented victories of both the Chicago Cubs and Donald Trump evoked reactions across all topics in both cities, though the former had an understandably greater effect in Chicago. These reactions to major events corroborate the findings of Szell et al [24].

Istanbul is similarly affected by external events. Turkish officials restricted access to social media three times over the study period: November 3 following the arrest of opposition leaders, December 19 after the assassination of the Russian Ambassador in Ankara, and December 23 in reaction to an ISIS propaganda post [25]-[27]. In addition, Turkey issued more than two times as many requests (3,076)

for content to be removed from Twitter than any other nation [28]. Such official edicts against free speech will limit the efficacy of the methodologies presented here in the future.

FIGURE I
CORRELATED CLUSTERS IN LA (TOP), CHICAGO, AND ISTANBUL (BOTTOM)

Heirarchical clustering places Arabic in a cluster by itself in both Los Angeles and Istanbul (cluster 1 in green). The lack of correlation between Arabic and the other topics in these cities implies that the language occupies a social space removed from that of other Twitter users in a city. Comparing the cross-correlations (Figure II) between Los Angeles and Istanbul shows that per-user volume in the latter area leads the former by 3 days.

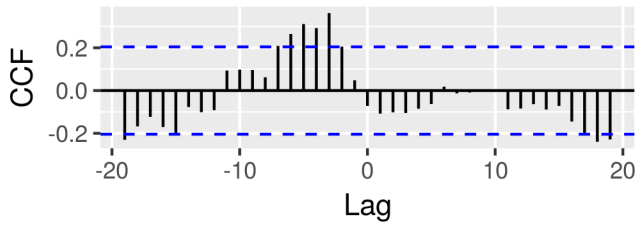

FIGURE II
CROSS-CORRELATION OF ARABIC TOPICS IN LOS ANGELES AND ISTANBUL

Though Arabic is not represented in the top 20 topics in Chicago, a bilingual population is similarly in a group by itself (the purple line). Users within this topic selected English as their user language, yet the top words in the topic are all Spanish stop words ("de," "la," "que," "mi," etc.). Thus, this group also exists in a Twitter space removed from other topics.

Spanish falls into Chicago's cluster 1 (green), alongside the "Cubs" topic from Table II. Intriguingly, this group does not correspond with the largest increase in post-World Series Tweets per user. That distinction belongs to cluster 2 (blue), which contains several sports topics ("wrigley" and "field," "soccer"), a topic paralleling the lyrics of hip-hop and rap ("n*ggas"), and a political one ("womensmarch," "obama"). Appropriately, this cluster also sees a pronounced spike following the U.S. presidential election, as well.

Los Angeles features a similarly bilingual topic, however it enters into a cluster of topics (in red) referencing specific locations in the city ("museum," "pier," "beverly," "venice") or local sports ("usc," "rose," "bowl"). The inclusion of terms "airport" and "california" could indicate this cluster consists of tourists, and merits further research. Conversely, the other English cluster in LA (in purple), consists of topics related to politics ("trump", "clinton"), health ("yoga"), and the arts ("music," "photographer"). Subjectively, these topics appear be more personal in nature.

CONCLUSIONS, LIMITATIONS AND FUTURE WORK

The above work serves as a proof of concept that one can discern cultural subpopulations within an urban area based off of social media. Topic modeling objectively identifies subpopulations more granular than the overarching linguistic or ethnic labels, and ones rooted in the choices and textual content of citizens. A more comprehensive approach would include translators, as subjective analysis of Turkish topics was limited due to lanuage barriers.

Furthermore, LDA struggles with sparse documents like a single 140-character Tweet. Eschewing this bag-of-words approach in favor of vectorized word representations used by Word2Vec and GloVe could improve the topics generated, and the existing class imbalances [22][29].

Time series analysis shows the daily heartbeat of a city, and how regional and national events impact the flow of data. As Istanbul produced a reduced amount of Tweets during government embargoes, this approach will suffer in regions prone to such authoritative actions.

Expanding this approach to include elements of seasonality could reveal additional trends, ones obscured by the brief study period. Conversely, an examination of hourly trends and how they vary from day-to-day or between weekends and weekdays was beyond the scope of this paper, but merits its own analysis.

Locating groups in space would provide an additional layer of understanding. Although geotagged Tweets represent a minority of Twitter data (7.8%, 4.4% and 27.9% of Tweets in Los Angeles, Chicago, and Istanbul, respectively), they could still provide a proxy for where

populations exist physically within the city. Examining how these spaces change over time could prove particularly interesting.

Twitter contains references to additional material not considered for analysis here. Following links would provide additional material for topic modeling, for instance. Additionally, 9.7% (877,111) of all Tweets in Los Angeles alone were posted from Instagram, a photo-sharing site. Analyzing these images would certainly suppliment the text data.

ACKNOWLEDGMENT

The authors would like to thank the Mitre Corporation for sponsoring this project, Mohammad al Boni for his assistance with Twitter listeners, and the professors at the University of Virginia for guiding them along the way. They also extend a special thanks to the cohort at the Data Science Institute for their ever-present support and comeraderie.

REFERENCES

[1] United Nations Department of Economic and Social Affairs. July 2014. "2014 Revision of World Urbanization Prospects." eas.un.org/unpd/wup/. Accessed September 6, 2016.

[2] United States Census Bureau. 2000. "Factfinder for the Nation."

[3] Cable, Dustin A., Martin-Anderson, Brandon and Fisher, Eric. "The Racial Dot Map: One Dot per Person." July 2013. demographics.coopercenter.org/Racial-Dot-Map/. Accessed September 7, 2016.

[4] Gao, Song, Yang, Jiue-An, Yan, Bo, et al. September 2014. "Detecting Origin-Destination Mobility Flows from Geotagged Tweets in the Greater Los Angeles Area." *Proceedings* of the 8th International Conference on Geographic Information Science, Vienna, Austria.

[5] Hu, Yingjie, Gao, Song, Janowicz, Krysztof, et al. 2015. "Extracting and Understanding Urban Areas of Interest Using Geotagged Photos." *Computers, Environment and Urban Systems* 54, pp. 240-254.

[6] Stiger, Enrico; Westerholt, René, Resch, Bernd, and Zipf, Alexander. September 2015. "Twitter as an Indicator for the Whereabouts of People? Correlating Twitter with UK Census Data." *Computers, Environment and Urban Systems* 54, pp. 255-256

[7] Gong, Li, Gao, Song, and McKenzie, Grant. March 2015 "POI Type Matching Based on Culturally Different Datasets." *Proceedings* of the International Conference on Location-based Social Media, Athens, GA.

[8] O'Brien, Oliver. August 2016. "Tube Heartbeat." tubeheartbeat.com/london. Accessed September 7, 2016.

[9] Steiger, Enrico, Ellersiek, Timothy, and Zipf, Alexander. November 2014. "Explorative Public Transport Flow Analysis from Uncertain Social Media Data." *Proceedings* of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Dallas, TX, pp. 1-7.

[10] Ellen, Jeffrey. "All about Microtext: A Working Definition and a Survey of Current Microtext Research within Artificial Intelligence and Natural Language Processing." January 2011. *Proceedings* of the 3rd International Conference on Agents and Artificial Intelligence, Rome, Italy.

[11] De la Rosa, Kelvin; Shah, Rushin; Lin, Bo; et al. "Topical Clustering of Tweets." July 2011. *Proceedings* of the ACM Special Interest Group on Information Retrieval 3rd Workshop on Social Web Search and Mining, Beijing, China.

[12] McKenzie, Grant; Janowicz, Krysztof; Gao, Song; et al. 2015. "POI Pulse: A Multi-Granular, Semantic Signature-Based Approach for the Interactive Visualization of Data." *Cartographica: The International Journal for Geographic Information and Geovisualization* 50(2), pp. 71-85.

[13] Speriosu, Michael; Brown, Travis; Moon, Taesun; et al. January 2010. "Connecting Language and Geography with Region-Topic Models." Proceedings of the Workshop on Computational Models of Spatial Language Interpretation, Portland, OR, pp. 33-40.

[14] Abdelhaq, Hamed; Gertz Michael; and Sengstock Christian. November 2013. "Spatio-temporal Characteristics of Bursty Words in Twitter Streams." *Proceedings* of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, FL, pp. 194-203.

[15] Becker, Hila; Naaman, Mor; and Gravano, Luis. July 2011. "Beyond Trending Topics: Real-World Event Identification on Twitter." International Conference on Weblogs and Social Media, Barcelona, Spain.

[16] O'Connor, Brendan; Balasubramanyan, Ramnath; Routledge, Bryan; et al. May 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." *Proceedings* of the 4th International Conference on Weblogs and Social Media, Washington, DC.

[17] Signorini, Alessio; Segre, Alberto M.; and Polgreen, Philip M. May 2011. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic." *PLoS ONE* 6(5).

[18] Twitter, Inc. "Streaming Overview." https://dev.twitter.com/streaming/overview. Accessed: September 15, 2016.

[19] Blei, David M.; Ng, Andrew Y.; and Jordan, Michael I. March 1, 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3, pp 993-1022.

[20] NLTK Project. 2015. "Natural Language Toolkit." nltk.org. Accessed: September 15, 2016.

[21] Bougé, Kevin. December 2011. "Stop Words." sites.google.com/site/kevinbouge/stopwords-lists. Accessed: October 1, 2016.

[22] Rehurek, Radim. 2009. "genism – Topic Modeling for Humans." radimrehurek.com/gensim/. Accessed: October 1, 2016.

[23] Hyndman, Rob; O'Hara-Wild, Mitchell; Bergmeir, Christoph; Razbash, Slava; Wang, Earo. February 23, 2017. "Package 'forecast'" [Software Manual].

[24] Szell, Michael; Grauwin, Sebastian; Ratti, Carlo. February 2014. "Contraction of Online Response to Major Events." *PLoS ONE* 9(2).

[25] Bulman, May. November 4, 2016. "Facebook, Twitter and Whatsapp Blocked in Turkey after Arrest of Opposition Leaders." independent.co.uk/news/world/asia/facebook-twitter-whatsapp-turkey-erdogan-blocked-opposition-leaders-arrested-a7396831.html. Accessed: March 8, 2017.

[26] McGoogan, Cara. December 20, 2016. "Turkey Blocks Access to Facebook, Twitter and WhatsApp Following Ambassador's Assassination." telegraph.co.uk/technology/2016/12/20/turkey-blocks-access-facebook-twitter-whatsapp-following-ambassadors/. Accessed: March 8, 2017.

[27] "ISIL Video Shows 'Turkish Soldiers Burned Alive.'" December 23, 2016. aljazeera.com/news/2016/12/isil-burns-turkish-soldiers-alive-shocking-video-161223035619947.html. Accessed: March 8, 2017.

[28] Twitter, Inc. March 21, 2017. "Tenth Transparency Report." transparency.twitter.com. Accessed: March 23, 2017.

[29] Pennington, Jeffrey; Socher, Richard; and Manning, Christopher D. August 2014. "GloVe: Global Vectors for Word Representation." nlp.stanford.edu/projects/glove/. Accessed: March 1, 2017.

AUTHOR INFORMATION

Lander Basterra, M.S., Data Science Institute, University of Virginia.

Tyler Worthington, M.S., Data Science Institute, University of Virginia.

James Rogol M.S., Data Science Institute, University of Virginia.

Dr. Donald Brown, Director, Data Science Institute, University of Virginia.