Visualizing 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 livelihood 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. Applying a bag-of-words approach to the textual content of Tweets emanating from within a city created topical clusters within the most frequently occurring languages. Such classifications transcend traditional racial designations. Time series analysis of each Tweet’s time stamp revealed that the volume of tweets across topics is significantly correlated with major regional events. Furthermore, certain subpopulations’ postings rose and fell in sync with others. Examining the long-term and weekly trends between strongly correlated topic groups provided 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 too 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. While the survey evolved to include economic, ethnic and educational categories, the Census continues to lump individuals into broad categories. In an effort to expedite data collection, the Census Bureau started conducting the American Communities Survey (ACS) in 2002. Though this annual survey suppliments the decentenial Censuses, profiling the citizens of the United States remains an inexact and 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. Despite social and geographic differences, individuals of Japanese, Taiwanese and Malasian descent all fall under the broad heading of “Asian,” for example. In the modern age, there should 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 London and Shanghai to Dubai [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 can offer insight into the thoughts of the populous, while the timestamp can reveal how these ideas fluctuate throughout months and weeks.

This paper provided a more granular analysis of temportal trends in urban areas regardless of geographic location. Latent Dirichlet Allocation produced subpopulations within language groups, clustering Twitter users based upon shared terms. A time series analysis examined the trends underlying each group’s posting habits, how they react to external events, as well as the correlations which existed 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, but was tested in Chicago (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 real-time glimpse at the status of a city. Though the user bases are not explicitly indicative 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 [cite, 7-9?].

Of all the social media platforms, Twitter in particular, has generated the most interest from both the academic community and businesses and policy makers [L1].

<<WHY>>

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. [L2] used Latent Dirichlet Allocation (LDA) and K-means clustering to classify Twitter messages based upon from frequently co-ocurring hashtags. However, they did not investigate how these key terms changed over time XXX and XXX [POI Pulse] employed LDA to group metropolitan points of interest (universities, concert halls, restaraunts, etc) based upon text culled from venue descriptions on Foursquare. Thematically similar establishments experienced a similar pattern of user frequency over time.

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, Brown, Moon, et al. to successfully visualize linguistic dissimilarities between spaces [ http://ceur-ws.org/Vol-620/paper5.pdf]. More importantly, this approach worked across multiple languages. Social media provided a similarly ample corpus for regional topic modeling, as show by Abdelhaq, Gertz, and Sengtock [L4]. They did not, however, investigate how these key terms changed over time.

In contrast, Becker, Naaman, and Gravano [L3] undertook an episodic approach to social media content. The text and timestamp of Tweets was found to correlate with major real-world events. Similarly, O’Connor et al. [L5] 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. With regard to forecasting trends, Signorini, Segre and Polgreen [L6] showed that a focused analysis of Tweets related to influenza symptoms improved predictions of outbreaks by 7 to 14 days. 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), language most likely used in the text, hashtags, mentions of other users, and metadata pertaining to the account posting the Tweet [cite]. Queries to the API can include a bounding box determined by the latitude and lonitude of the southwest and northeast corners of a bounding box (in degrees), filtering out messages eminating outside of these boundaries.

Over a 92-day study period (October 28, 2016 to January 27, 2017), Python scripts running on Amazon EC2 instances 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 insert 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)
* Message content
* Two-letter code of the most likely language of the text
* Unique user ID and hhandle
* General user\_location
* Two-letter code of a user’s chosen language
* User description
* Latitude and longitude of geotagged posts

Users can incorporate posts other authors into their own posts 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 also extracted and stored in the PostgreSQL database.

A separate 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.) [ CITE ].

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 or 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, or other consumer platforms. Excluding offending sources by hand would prove inefficient and time consuming, as the source varied by market (i.e. “altın dükkan twitter robotu,” the “gold shop Twitter robot” in Istanbul; or “TweetMyJOBS” in LA). Akin to Gao, Yang, Yan et. al [DETECTING O-D], a PostgreSQL query limited observations to those from a list of universally adopted clients (Instagram, Twitter for iOS, etc.). 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 researchers.

Twitter uses the “@” character as a prefix for references to other users (i.e. @dsi\_culturalmap), and the pound sign to 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 modelling to consider “#cubs” and “cubs” as the same term.

Each message’s time stamp was stored in the JSON object in Coordinated Universal Time (UTC). These standardized time stamps converted to local time in PostgreSQL, so that the study period corresponded with the same calendar dates for each test market. Once converted, the date, time (rounded to the nearest hour), and day of the week were extracted to aid in time series analysis.

Process & Results

I. Topic Modeling

Latent Dirichlet Allocation (LDA) provides a clustering framework for grouping documents 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 proabilistically, based upon terms’ frequencies within a given document. According to Blei, XXX, and XXX, these probabilities can provide explicit representations of documents with regard to the entire corpus [cite]. 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 Baysean 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 s/he quoted.

A single user language accounts for upwards of 89% of Tweets in a city, as shown in Table I. 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 market accounted for at least 5% 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). NLTK provides tools in XXX languages, including English and Turkish, which removed language-specific stop words and aggregated tokens of similar denotation via lemmatizing [cite].

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

(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. Lowering the threshold 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. Production languages which did not meet the 5% volume threshold were treated as purely language-based topics.

TABLE II

Topic Modeling Results: Top Terms and Subjective Labels

|  |  |  |
| --- | --- | --- |
| City | Top 10 Words | Topic |
| LA | namm, songwriter, home, cali, music, producer, life, inger, 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” |
| Istanbul |  | “We don’t Speak Turkish” |

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 *xt* and a measurement *h* points before, *xt-h*. Differencing removes polynomial trends in a process so as to coerce the data into a series with both a constant mean and covariance which depends solely on the time between points.

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

(2)

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.

The resulting models indicated that no single structure represented the subpopulations within a city. Of the top 20 groups (by raw volume) in Los Angeles, 4 were Auto Regressive, 0 were Moving Averages, 2 ARMA processes, and 14 were ARIMA models. Chicago had 2 AR, 3 MA, 3 ARMA, and 13 ARIMA models; while Istanbul saw 4, 3, 3, and 10, respectively. Examples from each location were 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 |

The majority of models contain some order of auto regressive term, indicating that the current day’s volume depends to varying degrees on the previous days’ counts. While differenced models describe at least half of thegroups with each city, no one model dominates in Istanbul and Los Angeles.

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 the other 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 an autoregressive term appeared in 90% of the top 20 topics in each city demonstrated the influence of the number of past tweets on the current day’s volume.

III. Cross-Process Correlations

Despite the different processes at work, topics appeared to ebb and flow together through time <<figure XX?>>. Statistically significant correlations revealed 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 was divided by the count of it’s unique users. Hierarchical clustering recombined the top 20 topics in to five correlated groups. A complete linkage function maximized the distance between clusters using (3) as a measure of distance.

(3)

Plotting the daily medians for each correlated cluster prodced Figure XXX. The general trends in the domestic markets were similar, though Los Angeles’ daily tallies varied twice as much as Chicago’s ( = 0.0013 and 0.0005, respectively). Both cities experienced 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 corroboratted the findings of Szell, Grauwin, and Ratti [Contraction of Online Events].

Istanbul Twitter usage across groups was 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 [cite x3]. In addition, Turkey issued more than two times as many requests (3,076) for content to be removed from Twitter than any other nation [cite]. Such official edicts against free speech will limit the efficacy of the methodologies presented here moving forward.

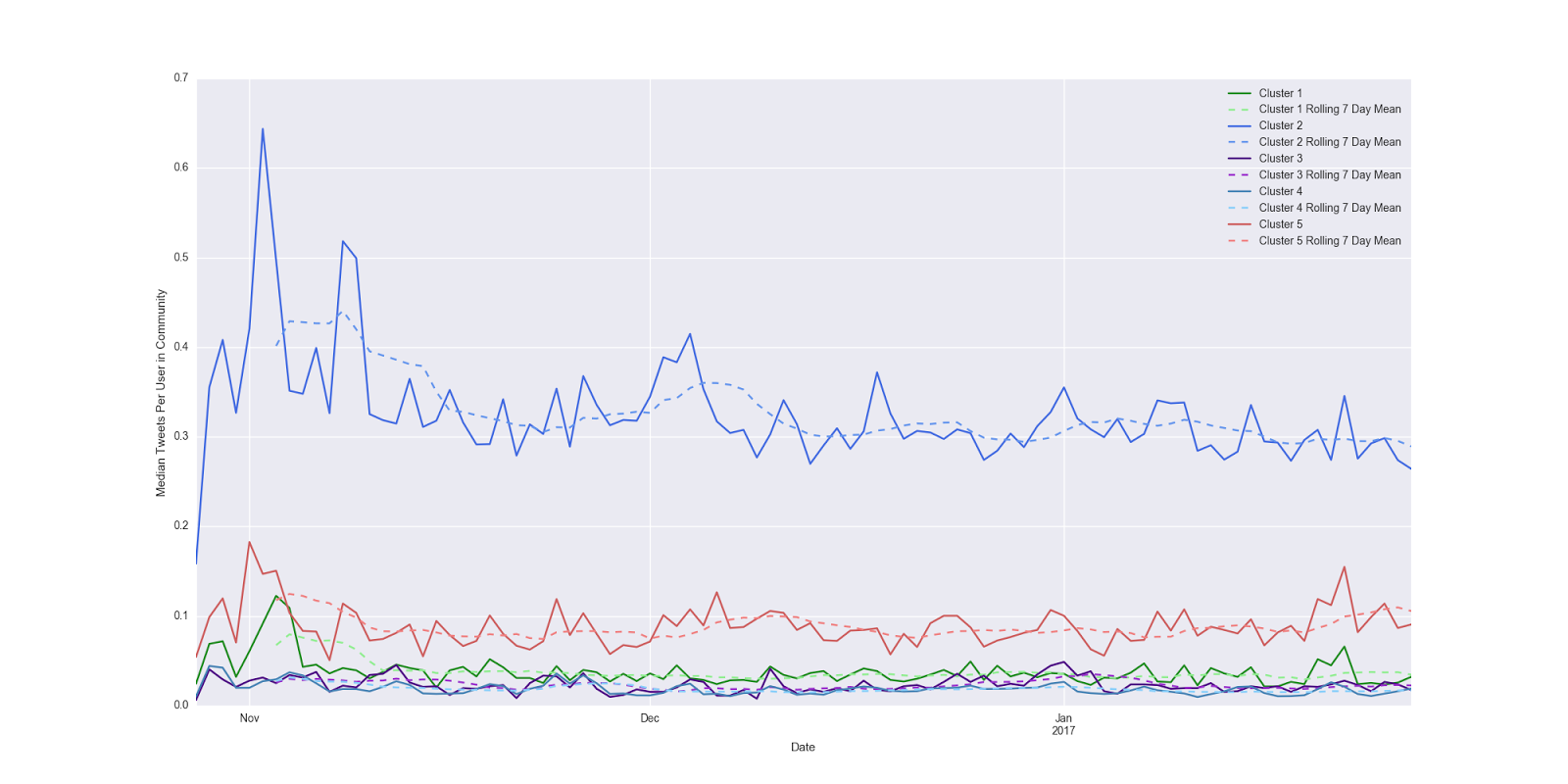
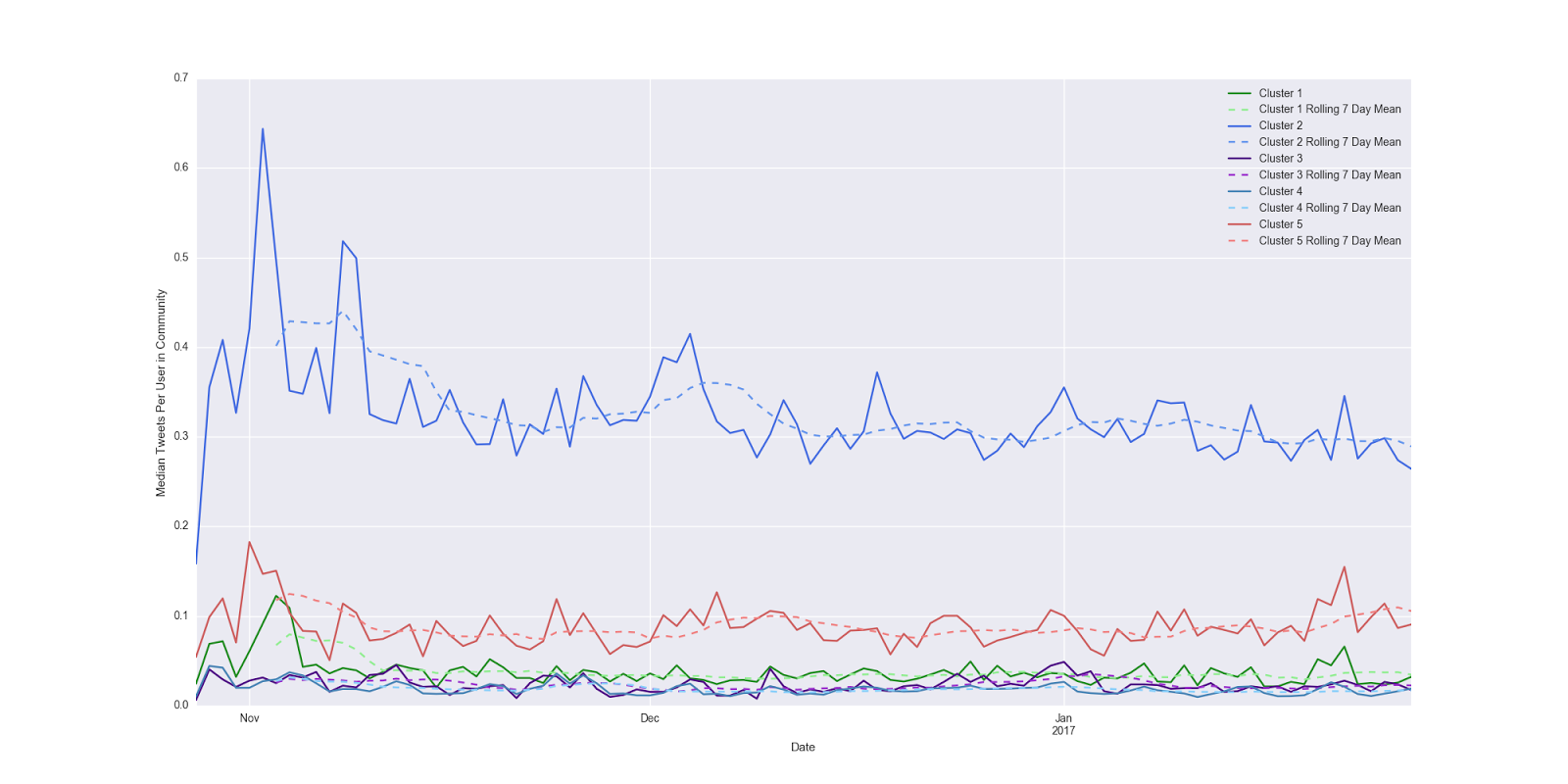


Figure I

Correlated Clusters in Chicago (Top) and Istanbul (Bottom)

IV. Within-Week Variations

Weekly trends were considered in addition to longitudinal ones, utilizing the same the median Tweets per user metric established after heirarchical clustering. Despite the cyclic nature of weeks, polar coordinates obscured within-week variations. Hence, traditional linear plots were employed.

Weekends in all three locales experiences a slight increase in the median number of Tweets per user. LA cluster 1 peak on Saturday <EXPLAIN>. CHI cluster 2 is almost sinusoidal, strong dip on Friday from Thursday <EXPLAIN>. Istanbul was the most consistent.

Removing the strong presence of the World Series and Election (coincidentally, both peaked on Wednesdays), did lower the median Tweets per user, as seen by the dotted lines in Figure II. CHI cluster 2 drops the most <<are these cubs fans??>> , whereas LA’s Cluster 4 experiences a slight increase with the outliers removed <<investigate!>>

Conclusions and Limitations

Using social media successfully enabled research in cities both domestic and international. Topic Modelling produced subpopulations that were more granular than the overarching linguistic or ethnic labels. Subjective analysis of Turkish topics was limited due to lanuage barriers (ironically). Linguistic specialists would help.

Time Series analysis showed a daily heartbeat of a city. Users across topics were generally more active on weekends. Major events evoked strong reactions, whereas governmental interference could prohibit the utility of the approaches. However, Istanbul did produce a small number of tweets despite the embargoes.

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, and can be extended in a variety of ways. Collecting tweets over a longer study period could reveal the presence of seasonal trends. Conversely, one could also examing the variation across the hours of the day.

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 [cite][cite]

Furthermore, one could create a streaming protocol which processes and classifies Tweets as they are posted, creating a near real-time “heartbeat” of an urban area.

In Los Angeles alone, XX% of tweets were posted from Instagram, a photo-sharing site. Image analysis was well beyond the scope of this paper, but could certainly suppliment the text data.

Following links and incorporating that data into the topic model.

Lastly, as some Tweets contain geolocation, future research could investigate the correlation between topical groups in space in addtion to time.

Acknowledgment

The authors would like to thank the MITRE corporation for sponsoring the project, Mohammad al Boni for his assistance with Twitter listeners, as well as their cohortmates for putting up with them.

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, D.A.;Martin-Anderson, B. and Fisher, E. 2013. “The Racial Dot Map: One Dot per Person.” demographics.coopercenter.org/DotMap/index.html. Accessed September 7, 2016.
4. Gao, S.;Yan, B.; et al. “Detecting Origin-Destination Mobility Flows from Geotagged Tweets in the Greater Los Angeles Area.”
5. Hu, Y.; Gao, S.; et al. 2015. “Extracting and Understanding Urban Areas of Interest Using Geotagged Photos.” *Computers, Environment and Urban Systems* 54, pp. 240-254.
6. [6] Stiger, E.; Westerholt, R.; et al. 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, L.; Gao, S. and McKenzie, G. “POI Type Matching Based on Culturally Different Datasets.”
8. O’Brien, O. 2016. “Tube Heartbeat.” tubeheartbeat.com/london. Accessed September 7, 2016.
9. Steiger, E.l Ellersiek,T.; and Zipf, A. 2014. “Explorative Public TransportFlow Analysis form Uncertain Social Media Data.” *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information.*
10. Troy, Dave. 2014. “Social Maps that Reveal a City’s Intersections – and Separations.” TEDGlobal. www.ted.com/talks/dave\_troy\_social\_maps\_that\_reveal\_a\_city\_s\_intersections\_and\_separations. Accessed September 4, 2016.
11. Calabresee, D.; Dahlem, D.; et al. 2011 “The Connected States of America” senseable.mit.edu/csa. Accessed September 6, 2016.
12. Hyndman, Rob; O’Hara-Wild, Mitchell; Bergmeir, Christoph; Razbash, Slava; Wang, Earo. February 23, 2017. “Package ‘forecast’” [Software Manual].
13. Twitter, Inc. “Streaming Overview.” https://dev.twitter.com/streaming/overview. Accessed September 15, 2016.
14. GloVe - https://nlp.stanford.edu/projects/glove/
15. Word2Vec – tensorflow.org https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjQg9-in\_\_SAhWIVyYKHVitCGQQFggaMAA&url=https%3A%2F%2Fwww.tensorflow.org%2Ftutorials%2Fword2vec&usg=AFQjCNEuWmN12xU45d-u\_b7j9O92a2rR4Q&sig2=8oPvDd7QfRllpDBg3Ehy7w&bvm=bv.151325232,d.eWE
16. Arrest of opposition leaders 3 Nov (http://www.independent.co.uk/news/world/asia/facebook-twitter-whatsapp-turkey-erdogan-blocked-opposition-leaders-arrested-a7396831.html)
17. 12/19 assassination of RUS ambassador (http://www.telegraph.co.uk/technology/2016/12/20/turkey-blocks-access-facebook-twitter-whatsapp-following-ambassadors/)
18. 12/23 ISIS propaganda video (http://www.aljazeera.com/news/2016/12/isil-burns-turkish-soldiers-alive-shocking-video-161223035619947.html)

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.