Geo-Fingerprinting Social Media Content

Hatim Gazaz
George Mason University
hgazaz@gmu.edu

Arie Croitoru
George Mason University
acroitor@gmu.edu

Paul L. Delamater
George Mason University
pdelamat@gmu.edu

Dieter Pfoser
George Mason University
dpfoser@gmu.edu

ABSTRACT

With the percentage of Twitter users approaching 20% of the US population by 2019, tweets provide a good sample of the public's sentiment and opinion. Consequently such data has been excessively used in commercial and research efforts. While works have analyzed the content of tweets in relation to the underlying social network of a discussion, somewhat less attention has been paid to the spatial distribution of messages and topics. This work tries to assess the locality of discussions using the concepts mentioned in tweets. Based on a global distribution of topics across the 48 contiguous states, we try to ascertain spatial topic dissimilarity by recursively subdividing the space into smaller and smaller partitions and using statistical testing to compare the distributions. Experimenting with a large Twitter dataset for the US, we can observe that locality of a discussion occurs at specific thresholds and that only 14 of the 49 most populous urban areas feature a unique discussion. Overall, this work establishes trends as to when locality in a discussion in social media occurs.

CCS Concepts

• **Information systems~Spatial-temporal systems**
• **Information systems~Document representation** • *Information systems~Data mining*

Keywords

Twitter; Entity Extraction; Topics; Chi-Square; Lack-of-Fit.

1. INTRODUCTION

Social media usage has exploded in recent years as evident by the fact that in January 2015 there existed 186 million social media accounts in the USA alone [17]. Users of social media share their observations and opinions instantaneously. Using the ever popular Twitter data, studies show how many aspects of human activity such as mobility [10] or even National Football League (NFL) game events [20] can be observed by mining such social media data. Using smartphones, tweets are increasingly geotagged, i.e., geographic coordinates are attached as metadata to a tweet. Although Twitter limits free access to streaming data to 1% of all tweets, more than 90% of all geotagged tweets are captured when using a bounding box as a filter parameter [15].

The objective of this work is to assess the locality of topics mentioned in tweets. Using a large collection of tweets collected from the public Twitter streaming API for several days, we use entity extraction techniques, specifically the Textrazor services, to identify concepts in those tweets. All concepts mentioned in a set of tweets taken together are considered a topic of discussion. The objective of this work is to see how topics differ at various levels of spatial granularity. In our experimental setup we partition the area of the US (48 contiguous states) into hierarchical bins (bounding boxes), extract named entities from geo-located tweets for those areas, categorize the named entities into 9 high-level topics, calculate the topic distribution for the entire US and for all hierarchical bins, and then compare all bins to the global distribution using a statistical test to see if they differ.

The remainder of this paper is structured as follows. Section 2 briefly discusses related work. Section 3 introduces the data used in the experimentation and the tools used for entity extraction. Section 4 discusses the overall approach and presents the experimental results. Finally, Section 5 gives conclusions and directions for future work.

2. RELATED WORK

A large number of geo-tagged tweets does provide a good sampling of topics being discussed in relation to space. The diverse ways in which such a resource can be utilized can be broadly categorized into the two major areas of (i) inferring users' location and (ii) a analyzing content in relation to space.

2.1 Inferring Users' Location

Since only a small fraction of tweets are geo-tagged, using additional methods to geolocate tweets are needed. A simple approach is to geocode the toponyms mentioned in tweets using NER approaches and gazetteers as well as public APIs such as Google Geocoder [11]. The works that rely on a single characteristic of a tweet, i.e., tweet text, yield lower rates of correctly geolocating tweets than those using a combination of characteristics. A probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets is proposed in [5]. Around 51% of users where assigned coordinates within 100 miles of their correct locations. A method to predict the location of tweets related to dengue fever is proposed in [6]. The prediction only utilized information relating to follower or following relationships in Twitter to identify friends of users that don't have geo-tagged tweets. Then, a voting system was used to figure out the most probable location to assign to the tweet. This method increased the number of geo-tagged tweets by 45%.

GeoRich'16, June 26, 2016, San Francisco, CA, USA. Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-4309-1/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2948649.2948654

As expected, using multiple data aspects can increase the scope and accuracy of geocoding such as in the case of a hierarchical classifier that combines tweet content, tweeting behavior and explicitly mentioned geographic locations [13]. The classification estimates the time zone of Twitter users based on the pattern of tweeting activity. Then, using the place names and tweet content they refine the estimation to state and city level.

2.2 Geographic Analysis of Tweets

Aggregating the Twitter content with respect to a point-of-interest yields a better understanding of what is going on at that location. Various works try to capitalize on this and extract events. GeoScope [3] conducts statistical and geospatial analysis to derive the most frequently used hashtags within cities around the world. The authors conducted their analysis on a large dataset containing 63 million tweets. In our work we will investigate their suggestion to conduct a hierarchical trend detection coupled with named entities. Another study [19] considered also toponyms in tweets to locate events. Observing the duration of trending topics can help marketers or organizers of campaigns to assess impact. An "Attention Automaton" system [16] assesses collective user interests either in relation to geographic extent or a virtual network of followers on Twitter. Several works have investigated the correlation between geo-tagged tweets and population characteristics. Gore et al. [9] have found a correlation between an increased use of words like "espresso", "yoga", and happiness (mood) and low obesity in urban areas. The mentioning of physical activities has also been positively correlated with low obesity. Mitchell et al. [14] investigated correlations between the tweets' content and happiness levels of states and cities. They also observed happiness being positively correlated with low obesity and they propose to predict and monitor the changing levels of obesity and happiness in real-time.

3. DATA and TOOLS

What follows is a brief discussion of the data used in our experiments and the tools used to collect them. Two large sets of geo-tagged tweets were collected using the public Twitter streaming API and a bounding box around the 48 contiguous states. The tweets were stored in a PostgreSQL/PostGIS database and all experiments were implemented in Python.

3.1 Data

We collected two datasets. An initial dataset of around 222,000 tweets was collected on 8/18/2015 between 9am and 11pm ET. This dataset was used to establish the database schema and the skeleton of the processing code. Another dataset was collected on 1/29/2016 for a 24h period ET resulting in 245,000 geo-tagged tweets.

Cleaning up the data required trimming the tweets geographically to only include tweets within the 48 contiguous states. The public Twitter API only allows a rectangular bounding box for collection. The bounding box used in the collection was: MinX,Y = -124.72, 24.56 MaxX,Y = -66.72, 49.56. This bounding box also produced tweets from Mexico, Canada, and other regions.

From the collected tweet content and metadata we only used id, tweet content, coordinates and timestamp information.

This study relies on the content of the tweets, which might be influences by bots and other spam information. In creating a simple filter, we eliminate vocal users that produce more than 30 tweets in the dataset. When analyzing the data we observed that this affects less than 1% of the users. Additionally, there were high numbers of automatically generated users who have common patterns. The two most prevalent words in tweets from such users were (i) tmj (abbreviating "That's My Job") and (ii) job. Tweets of users with user ids that include any of these two words have been deleted. As the numbers in Table 1 indicate, although the percentage of spam users is only 6% and 7%, they generate 57% and 58% of the total tweets in the two respective datasets.

Table 1. Twitter Data Collections

Dataset	08/18/2015	01/29/2016
Total Users	76,077	82,081
Spam Users	4,711	5,797
% Spam Users	*6 %*	*7 %*
Total Tweets	222,364	244,919
Spam Tweets	126,926	142,445
% Spam Tweets	*57 %*	*58 %*

Figure 2 shows the spatial distribution of spam tweets (red dots) for the 8/18/2015 data. Spam tweets are typically within populous areas and do not follow a random spatial distribution.

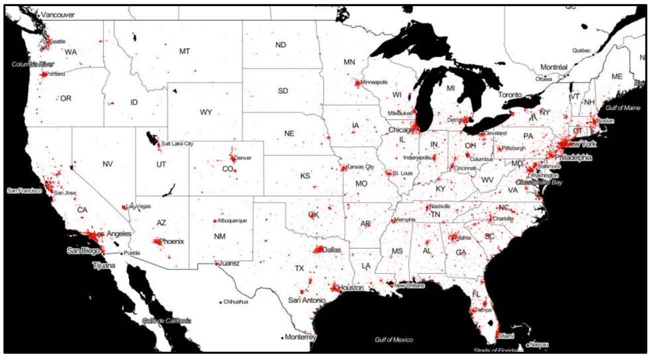

Figure 1. Spam Tweets (red dots) for 8/18/2015 data

3.2 Tools

The collected tweets were stored in a PostgreSQL/PostGIS database to be able to query the tweets within a specific geographic area. The analysis was implemented in Python. Scipy was used for the majority of the statistical functions like minimum, maximum, variance and the Chi-Square Test (see next section). The TextRazor API [18] was used from within Python to extract named entities from tweets. All extracted entities were linked to the corresponding tweets and stored in a separate table to simplify analysis.

4. ANALYSIS & RESULTS

The objective of this work is to detect locality in Twitter discussions with respect to a "global" topic distribution of the US. In the following, we describe our analysis approach and respective results.

The first step uses an initial spatial tweet distribution to create hierarchical bins that will be used in this work. Named entities are extracted from all the tweets and a topic distribution is computed for each of the bins. For each bin, a statistical test is used to assess its similarity to the overall US distribution topics.

4.1 Hierarchical Bins

One of the goals of this paper is to find a spatial threshold at which topics in an area differ from the distribution of the entire US. Essentially, by considering topics for smaller and smaller areas (cf. Figure 4) and comparing them to the entire US, we expect to detect an increasing dissimilarity, i.e., a more "local" discussion.

We generated a hierarchical spatial grid such as shown in Figure 3 that covers initially the entire US and is then subdivided into regular partitions. We refer to each partition as a bin of topics, or, short, a bin.

As shown in Figure 3, the bin generation and numbering starts at the root level with the entire US. We recursively calculated the number of tweets for each bin. If the number of tweets was more than 20, we divide the partition into 4 smaller equal-sized rectangular partitions. This partitioning is continued until either the number of tweets drops below 20, or the size of a partition becomes smaller than 0.00625 degrees, i.e., a size of ~800x800*m*.

A labeling system is introduced that reflects the hierarchical nature of the grid and, thus, the binning of topics. A bin is labeled based on the name of the parent bin and appending one more digit to identify the child bin. The root bin, which is level 1 in the hierarchy, has the number 0. The second level contains ten bins numbered from 00 to 09. Starting with level 3 and until level 13, there can only be 4 children and thus numbers. This facilitates the generation of child-parent aggregation without the need of conducting spatial intersection operations. To find the level of a bin we just need to count the length of the numbering digits. In Figure 3, when examining the top right bin, we find that it's at level 4 and belongs to the following parent bins: root, bin 9, and then bin 4 (within 9).

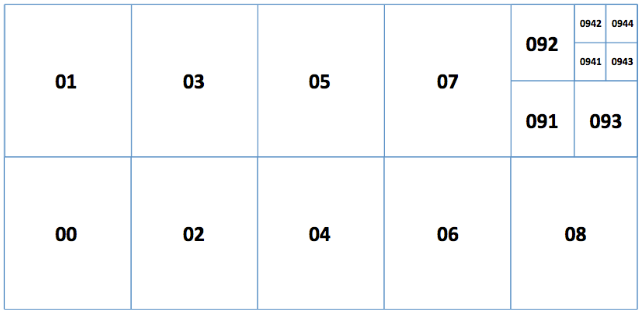

Figure 2. Hierarchical grid numbering system

We generated bins for both collected Twitter datasets, with the spatial distribution of the data differing between the two datasets. However, in all our experimentation, we use the bins that are used in both datasets. The total tweet counts for bins at all level are shown in Table 2. The total number of bins is 5210.

Figure 4 shows all bins up to level 8 (out of 13). A higher concentration of bins is evident around urban areas such as New York City, Los Angeles, etc.

Table 2. Twitter Data Collections

Level	Dimension (Degrees)	18-Aug	29-Jan	Chosen
2	12.8	9	9	9
3	6.4	30	30	30
4	3.2	96	93	93
5	1.6	253	244	236
6	0.8	514	486	445
7	0.4	739	699	591
8	0.2	1010	748	790
9	0.1	1311	891	999
10	0.05	1171	978	777
11	0.025	831	1057	530
12	0.0125	673	1285	409
13	0.00625	529	1317	301
Totals	**-**	**7166**	**7837**	**5210**

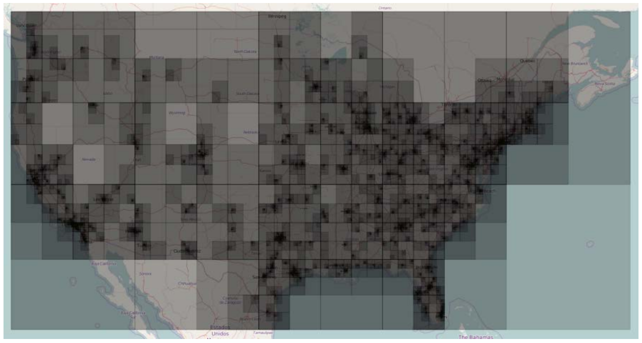

Figure 3. Chosen Bins

4.2 Entity Extraction

To assess the topics of a discussion in Twitter we use entity extraction methods, which effectively discover the concepts of a knowledge base such as DBpedia or Freebase in tweets. In our work we used TextRazor[1]. In order to improve the quality of matched entities, the tweet text was cleaned before being sent to TextRazor. Usernames, hashtags and URLs were removed. TextRazor uses Freebase [8] and/or DBpedia [7] to identify entities. In our work we selected Freebase as a knowledge base. Freebase contains 423 million entities that belong to 62 million topics that fall under 77 domains. Each entity can belong to several topics and domains and this information is returned as part of the entity extraction, along with the listing of entities and their confidence score, i.e., a percentage giving the likelihood that a returned entity matches the given word or phrase.

4.3 Topics

An entity identified by TextRazor can belong to several of the *77 domains (categories).* We group these domains into high-level topics to simplify the definition of a topic distribution. Also, having a small number of categories is a prerequisite to using the Chi-Square Test to assess the similarity of bins in terms of topics. Table 3 lists the chosen domain groupings and the domains assigned to each group. Also, the percentages of topics and

[1] http://www.textrazor.com

entities in each group are listed. As can be seen, “culture” attracts most concepts.

Table 3. Topic categories – 77 domains are mapped to 9 categories

Group	Domains	Topics %	Entities %
Culture	music, film, tv, fictional_universe, cvg, theater, broadcast, fashion, comedy, radio, amusement_parks	61%	71%
Geography	location, geography, transportation, travel, zoos, protected_sites	4%	5%
Literature	visual_art, book, media_common, periodicals, library, comic_books, exhibitions, interests	20%	8%
Sports	sports, soccer, olympics, american_football, baseball, basketball, cricket, ice_hockey, boxing, martial_arts, games	1%	2%
People	people, food, language, celebrities	7%	5%
Science	biology, medicine, astronomy, chemistry, spaceflight, meteorology, engineering, geology, physics	2%	3%
Society	organization, award, education, projects, aviation, boats, law, rail, event	3%	4%
Technology	business, internet, computer, digicams	2%	1%
Politics	military, government, royalty, symbols, religion	0%	0%

TextRazor finds for each tweet its named entities. Each of these entities belongs to one of the 9 categories shown in Table 3.

We can now compute the total counts and percentages for each topic category for both datasets. Table 4 lists the percentages for each group for both datasets. Although collected half a year apart, the respective percentages for both Twitter datasets are almost identical.

For each bin we now record the number of tweet occurrences on a category basis, i.e., how many tweets are recorded in the area of a bin in a specific category.

4.4 Find Different Bins Using Lack-of-Fit

Having topic distributions for all bins, we want to determine the *spatial resolution at which a locality of a discussion* can (if at all) be observed. Here we compare the distribution of all bins to the overall topic distribution of the US by modifying the Chi-Square Test [12] into a test called lack-of-fit X^2 test. The primary purpose of the lack-of-fit test is to test the hypothesis that a sample categorical frequency fits the known population frequency. Equation 1 shows the calculation of the X^2 using the expected frequency (E) and the sample observed frequency (O) and summing up all the values for the 9 categories.

Equation 1. Chi Square (X^2)

$$X^2 = \sum_{i=1}^{9} \frac{(E_i - O_i)^2}{E_i}$$

Observed values in the equation are calculated using the percentage of entities in a category and multiplying it with the number of tweets. The expected values are the product of multiplying the percentage of the category in the root bin by the number of the tweets in the tested bin. So, the number of tweets in a bin is a crucial parameter in the calculation.

In order to *reject the hypothesis that the sample bin distribution fits the root bin distribution* an α value of 0.01 will be used. Since the number of topic groups we are using to calculate X^2 is 9, the degrees of freedom is equal to 8. When using these parameters, the X^2 value needs to exceed 20.09 to reject the hypothesis that the specific topic distribution of a bin fits the root data distribution. The Python package Scipy is used for the calculation of the X^2 and the p-value. Thus, if any bin has a calculated p-value < 0.01 it is considered significantly different from the root bin.

Table 4. Topic Group Percentages for Both Datasets

Dataset	Culture	Geography	Literature	Sports	People
8/18/15	13.3%	15.0%	18.3%	6.3%	6.1%
1/29/16	13.2%	14.7%	18.1%	6.8%	5.9%
	Science	**Society**	**Technology**	**Politics**	
8/18/15	5.2%	15.3%	12.2%	8.4%	
1/29/16	5.1%	15.0%	11.9%	9.2%	

Since using all available tweets in a bin pushes the chi square equation beyond its limitations, smaller samples should be drawn randomly to calculate the lack-of-fit test. We set this max sample size to 250 tweets per bin. Thus, if there are 250 or fewer tweets in a bin, all tweets will be used to calculate the X^2 and p-value. If more than 250 tweets are in a bin, we will randomly sample 250 tweets. This random sampling of 250 tweets is performed 1000 times and each time the p-value will be calculated and stored in a Python array. Then, we store the mean p-values from the 1000 experiments.

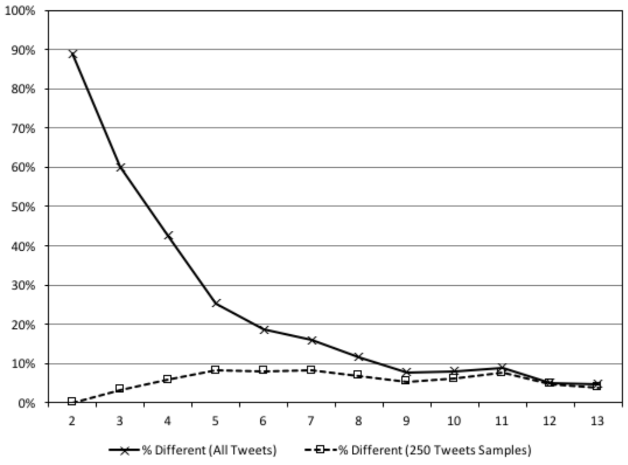

Figure 4. Percentage of different bins - all tweets vs. sampled tweets (Aug 18 Dataset)

Figures 4 and 5 show the results of using the mean p-value of sampled tweets (dashed) in both datasets compared to using all tweets (solid) with no sampling. The solid lines in both figures show decreasing percentage of different bins with increasing spatial granularity due to a large number of tweets. When the number of tweets in a bin is much larger than 250, the value of X^2 rapidly increases causing the bin to *falsely appear different*. Meanwhile, the dashed lines show the expected behavior of an increasing percentage of different bins with an increasing level of spatial granularity. Starting with Level 7, at which the bins have an approximate side length of 51*km*, the percentage of different bins per level remains the same up until Level 11. *This indicates that locality of twitter discussion starts to emerge at this level of spatial granularity.*

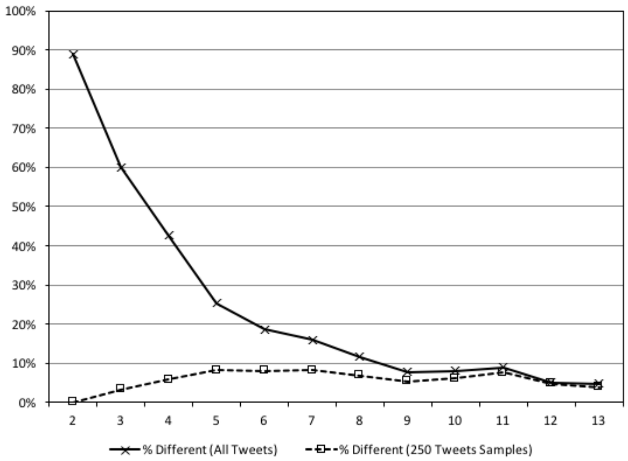

Figure 5. Percentage of different bins - all tweets vs. sampled tweets (Jan 29 Dataset)

Figures 6 and 7 show maps of the 48 contiguous states and the bins that are different from the root level (entire US). Transparency is used to better distinguish overlapping bins. Thus, darker red areas indicate several overlapping bins.

4.5 Different US Urban Areas

In this section we investigate the percentage of urban areas at higher levels (10 – 13) that have a Twitter discussion differing from the global topics. The urban area extents were obtained from the United States Census Bureau [4] for 2014. The population figures are from the 2010 census [1].

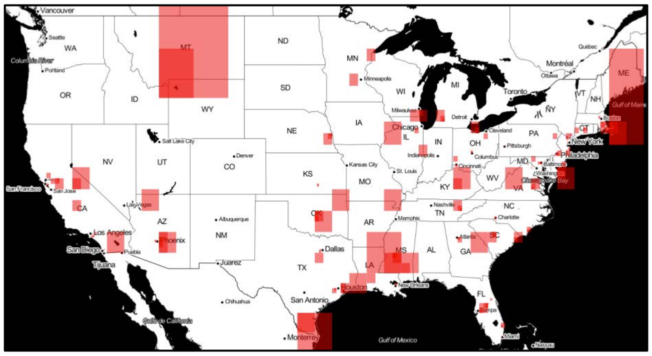

Figure 6. Different bins using sampled tweets at all levels (8/18/15 dataset)

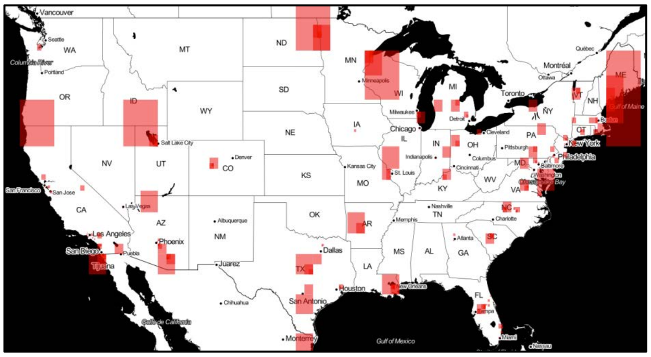

Figure 7. Different bins using sampled tweets at all levels (01/29/16 dataset)

Table 5. Urban areas – thematic difference to global topic distribution

Rank	Name	08/15	01/16
1	New York--Newark, NY--NJ--CT	YES	YES
2	Los Angeles--Long Beach--Anaheim, CA	YES	YES
3	Chicago, IL--IN	YES	YES
4	Miami, FL	YES	YES
5	Philadelphia, PA--NJ--DE--MD	YES	YES
6	Dallas--Fort Worth--Arlington, TX	YES	YES
7	Houston, TX	YES	NO
8	Washington, DC--VA--MD	NO	YES
9	Atlanta, GA	NO	YES
10	Boston, MA--NH--RI	YES	YES
12	Phoenix--Mesa, AZ	NO	YES
13	San Francisco--Oakland, CA	YES	YES
15	San Diego, CA	YES	YES
16	Minneapolis--St. Paul, MN--WI	NO	YES
17	Tampa--St. Petersburg, FL	NO	YES
19	Baltimore, MD	YES	YES
20	St. Louis, MO--IL	NO	YES
23	Las Vegas--Henderson, NV	YES	YES
24	Portland, OR--WA	NO	YES
27	Pittsburgh, PA	YES	NO
30	Cincinnati, OH--KY--IN	YES	NO
32	Orlando, FL	YES	YES
36	Columbus, OH	YES	NO
37	Austin, TX	NO	YES
38	Charlotte, NC--SC	YES	YES
43	Louisville/Jefferson County, KY--IN	YES	NO
49	New Orleans, LA	YES	YES
50	Raleigh, NC	NO	YES
95	Winston-Salem, NC	YES	NO
117	Kissimmee, FL	NO	YES
120	Greensboro, NC	NO	YES
297	Norman, OK	YES	NO
1708	Aspen, CO	NO	YES

We intersect the spatial extent of our bins with the city shapes. Using both collected Twitter datasets, 08/15 and 01/16, the areas that show a difference in topics exhibit a higher level of uniqueness. The results are shown in Table 5 by indicating topic distributions differed. The numbers of different bins are 63 and 103 for the 08/15 and 01/16 datasets, respectively.

Only 33 out of all 3601 urban areas had differing topics in the case of at least one dataset. Only 14 (highlighted in green in Table 5) had different distributions for both datasets.

5. CONCLUSIONS

With millions of users generating content, it is very easy to analytically drown in this wealth of tweets. This effect is compounded by the high percentage of unimportant chatter [2]. Viewing tweets from a high-level perspective, this work tries to identify locality in tweets. We extracted named entities from geo-tagged tweets for the 48 contiguous states. The entities belong to 77 domains and were grouped into 9 broader topic categories. We used a hierarchical subdivision of space to aggregate the tweets and respective categories in bins of varying spatial size. With the various topic distributions of bins of different sizes, a lack-of-fit test was used to compare all bins to the topic distribution of the global level (entire US) and to identify the bins that differ topic wise. An important observation was that the percentage of differing bins remains constant for Levels 7 (corresponding to 51*km* side length) and above (even smaller bins). This finding suggests that a locality in discussion can be observed starting at such spatial extents. I.e., in larger cells the distribution of topics resembles that of the entire US. since they generate most of the traffic on Twitter, when examining urban areas, we observed that out of 3601 areas only 33 differed in terms of discussed topics. These findings should increase the understanding and utilization of geo-tagged tweets to encourage novel applications. Content uniqueness and differentiation can inform better event detection and facilitate the detection of interesting local patterns.

The directions for future work are as follows. Many of the findings of the paper should be applicable to tweets with no geo-tags. It would be interesting to see if the topic distributions are similar during other time periods and/or for non-US regions. Specific regions can be studied to identify what differentiates some urban areas. The topic classification in this work can be improved and varying mappings of topics to categories should be used. Also, we are using Freebase in our work and other knowledge bases should be used to validate and confirm these results.

ACKNOWLEDGMENTS

We would like to thank George Lamprianidis for helping with the collection of Twitter data.

REFERENCES

[1] 2010 Census Urban Lists Record Layouts: *https://www.census.gov/geo/reference/ua/ualists_layout.html*. Accessed: 2016-03-29.

[2] Balasubramanyan, R., Kolcz, A., 2013. "W00T! Feeling Great Today!": Chatter in Twitter: Identification and Prevalence. *Proc of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 306–310.

[3] Budak, C., Georgiou, T., Agrawal, D., El Abbadi, A., 2013. GeoScope: Online Detection of Geo-correlated Information Trends in Social Networks. *Proc. VLDB conf.*, 7(4):229–240.

[4] Cartographic Boundary Shapefiles - Urban Areas: *https://www.census.gov/geo/maps-data/data/cbf/cbf_ua.html*. Accessed: 2016-03-29.

[5] Cheng, Z., Caverlee, J., Lee, K., 2010. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. *Proc. CIKM conf,*, pp. 759–768.

[6] Davis, C.A., Pappa, G., Rocha de Oliveira, D.R., de L. Arcanjo, F., 2011. Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS*. 15(6):735–751.

[7] DBpedia: *http://wiki.dbpedia.org/*. Accessed: 2016-03-28.

[8] Freebase: *https://www.freebase.com/*. Accessed: 2016-03-28.

[9] Gore, R.J., Diallo, S., Jose Padilla, J., 2015. You Are What You Tweet: Connecting the Geographic Variation in America's Obesity Rate to Twitter Content. *PLoS ONE*, 10(9).

[10] Hawelka, B, Sitkoab, I., Beinata, E., Sobolevskyb, S., Kazakopoulosa, P., Ratti, C. 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271.

[11] Lee, S. et al. 2015. Read Between the Lines: A Machine Learning Approach for Disambiguating the Geo-location of Tweets. *Proc. 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 273–274.

[12] Lynch, S.M. 2013. Statistical Approaches for Nominal Data: Chi-Square Tests. *Using Statistics in Social Research: A Concise Approach*, pp. 107–116.

[13] Mahmud, J., Nichols, J., Drews, C., 2014. Home Location Identification of Twitter Users. ACM Transactions on Intelligent Systems and Technology (TIST), 5(3):1–21.

[14] Mitchell, L. et al. 2013. The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. *PLoS ONE*, 8(5).

[15] Morstatter, F., Pfeffer, J., Liu, H., Carley 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *Int'l conf. on Weblogs and Social Media*, pp. 400–408.

[16] Roy, S.D., Lotan, G., Zeng, W., 2015. The Attention Automaton: Sensing Collective User Interests in Social Network Communities. *IEEE Transactions on Network Science and Engineering,* 2(1):40–52.

[17] statistics: *http://wearesocial.net/tag/statistics/*. Accessed: 2015-12-12.

[18] TextRazor - The Natural Language Processing API: *https://www.textrazor.com/*. Accessed: 2015-12-12.

[19] Unankard, S., Li, M., Sharaf, M., 2015. Emerging event detection in social networks with location sensitivity. *World Wide Web Journal*, 18(5):1393–1417.

[20] Zhao, S., Zhong, L., Wickramasuriya, J., Vasudevan, V. 2011. Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games. *Technical Report TR0620-2011, Rice University and Motorola Labs, June 2011.*