We Know Where You Are Tweeting from: Assigning a Type of Place to Tweets using Natural Language Processing and Random Forests

Abdulkareem Alsudais
Center of Information Systems & Technology
Claremont Graduate University
Claremont, USA
E-mail:
abdulkareem.alsudais@cgu.edu

Gondy Leroy
Center of Information Systems & Technology
Claremont Graduate University
Claremont, USA
Eller College of Management
University of Arizona
Tucson, USA
E-mail:
gondyleroy@email.arizona.edu

Anthony Corso
Center of Information Systems & Technology
Claremont Graduate University
Claremont, USA
E-mail:
anthony.corso @cgu.edu

Abstract— **Identifying the type of the place a user is tweeting from is important for many business and social applications, e.g., user profiles can help local businesses identify current and potential clients and their interests. We used Random Forest to identify six location categories. They are active life, eating out, hotels, nightlife, shopping, and shows. We evaluated 16 features for use in classification. The features are generated from the textual contents in the tweet, the metadata associated with the tweet, and the geographical area the user is tweeting from. We trained our classifier by analyzing 43,149 reviews from Yelp and by examining two twitter datasets. The first is an original dataset consisting of 6,359 tweets and the second is a stratified one containing 2,400 tweets uniformly distributed between the six categories. We evaluated our approach by creating a gold standard. Using 60% of our tweets for training and 40% for testing, our approach classified 74% of tweets in the original dataset, and 77% of tweets in the stratified dataset, correctly with the right location category. The results could be beneficial for research and business.**

Keywords-Natural Language Processing; Random Forests; location analytics.

I. INTRODUCTION

Twitter is a platform that allows users to send short messages known as tweets. In the US, Twitter attracts 49 million monthly active users [1]. According to a survey conducted by Pew Research Center, 16% of American adults use Twitter [2], and due to its omnipresence, Twitter users create more than half a billion tweets per day [3]. This allows family, friends, and colleagues to publicize all aspects of their life. Not only can people tweet about every aspect of their life, but users can also follow each other. Twitter provides instantaneous information with respect to content and even the location of user if their tweet is geocoded.

Geocoded tweets are tweets that have location variables associated with them. The location variables are beneficial when locating the tweet's origin and putting it on a map. Moreover, they help in performing various spatial queries, thus making geocoded tweets significantly valuable for researchers and Big Data application developers. Conversely, lack of geolocation means less precision when identifying a network participant's exact location. With the increasing popularity of mobile devices giving Twitter users the ability to tweet from various locations, e.g., a restaurant, bar, or gym, identifying one's location or suspected location would seem like a valuable task.

Successfully identifying the type of place a tweet originated from is interesting for both research and business and provides human-friendly insight into the domain of Big Data research. First, the generated information could potentially be used in building a profile of users and their interests. Second, local business owners could use the collected data and information to learn more about current and potential clients; advertisers would better be able to target these people. For instance, users tweeting from a gym could receive ads about supplements, and the gym owner could instantaneously send coupons and special offers to the users. Finally, it attempts to close the gap between traditional research methods as applied to social media corpora and generates an innovative solution to a problem historically vexing the Twitter research development community. Much of a tweet's value is hard to unlock without its exact origin and only about 1% of tweets contain location variables, our classifier not only addresses but overcomes this issue.

To our knowledge, currently, there is not a solution that successfully generates the type of place a tweet originates from. It is possible for users to associate the name of the place they are in with the tweet itself. From there, identifying the type of place occurs via named entity recognition by looking at the tweet's content; our classifier surpasses this approach. Hence, it is a significantly novel idea to identify the type of place a tweet originates from even if a user does not reveal a certain business entity within the content of the tweet.

Specifically, the goal of this study is to identify the type of location users are tweeting from. For example, if a user is tweeting while at a restaurant, a successful label

978-1-4799-5057-7/14 $31.00 © 2014 IEEE
DOI 10.1109/BigData.Congress.2014.91

would associate that tweet with "eating out." We are concerned in maximizing the number of tweets being correctly labeled with an appropriate location category and with minimizing the number of tweets being wrongfully labeled with an inaccurate location category.

II. RELATED WORK

Predicting locations based on tweets has gained much interest in recent years. Most of the work has been on predicting the city and country of users based on the textual content of the tweets. Chang et al. built a model to predict the city-level location of users based on tweets content [4]. Their approach mainly relied on identifying local words that could be associated with certain cities. They tested their model on 5,113 Twitter users and were able to accurately locate 49% of them within 100 miles of their home location. Cheng et al. [5] developed a similar locations words list that they used to predict users location from their tweets content. Their location estimator accurately placed 51% of tested Twitter users within 100 miles of their actual physical location. In their evaluation, they relied on what users reported in their Twitter location field and compared that to the location their model suggested. Hecht et al. [6] examined the location field of Twitter users and built a classifier that predicts their home state and country. The classifier successfully predicted the home state for 30% of approximately 100,000 users. Kinsella et al. [7] predicted the location of tweets and users by creating language models for areas. Thom et al. [8] attempted at predicting location by utilizing knowledge about users as well as the textual content of the tweets. They were able to accurately locate 74% of tweets in a specific city. Sadilek et al. [9] looked at tweets from two cities and predicted the exact location of a user at a particular time using dynamic Bayesian networks. Rout et al. [10] predicted city-level locations of users. Their method differs from the others in that they did not use a tweet's content, instead, they constructed a social tie framework. McGee et al. [11] tested a different approach by using tie-strength and one's friends list to predict user location.

Other researchers developed interesting applications that take advantage of geolocated social media data. For example, in one study, Takehara et al. [12] used a restaurant reviews website to generate area-related and restaurants-related keywords. They used these keywords to identify areas and restaurants in tweets. In addition, their model suggested new restaurants to users based on their restaurants-related tweets. Ghosh and Guha [13] used topic modeling to discover health-related themes in Twitter. Subsequently, they underlined the spatial patterns of the themes to compare between, for example, northern and southern states and coasts and inland states. Finally, Okazaki and Makoto [14] created an earthquake prediction system. It tracks tweets and raise flags when there is a spark in earthquake-related tweets.

III. LOCATION-AWARE TWEET CLASSIFICATION

In this study, we identify six categories of places representing the type of location a user is tweeting from. The categories are *active life, eating out, hotels, nightlife, shopping,* and *shows*. We aim to assign a tweet to one of the six location categories.

Active life refers to places where people are performing physical activities. Examples of places in this category are gyms, dance studios, and hiking trails. Eating out includes places such as restaurants and coffee places. Nightlife places are ones where people go out and socialize. This includes bars and night clubs. The choice of these six categories was not arbitrary. We followed a two step process: examining the tweet corpus and examining Yelp's search filtering category list.

First, we looked at tweets that had a named entity referenced in them. For example, the tweet "*Brunch on The Farm with stevejmcbride jeremyrudy & @heatherhazzan @ Morning Glory Cafe http://t.co/iUaPRNu0LR"* has the place "Morning Glory Cafe" referenced in it. As such, we reviewed the list of tweets that contained named entities; upon reviewing 34,521 tweets, we were able to identify certain categorical patterns. For instance, a high number of tweets had restaurants and coffee places referenced in them. As a result, restaurants and coffee places were considered in our list of location categories. Conversely, there were very few tweets that had pet stores or religious places referenced in them. Therefore, in this processing step, we did not consider including them in the final list.

Based on these findings, we created initial location categories consisting of gyms, coffee places, bars, restaurants, and shopping. At this point, the list was still problematic since there was an overlap between some of the categories. Furthermore, other relevant types of places, overlooked by manual observation, were not on the list. A broader list, one that includes a variety of social places, was needed.

The second step of generating location categories was examining Yelp's review categories [15]. Yelp is a website users mainly visit to seek information about local businesses [16]. According to the official Yelp blog, as of March 2014, the websites averages 120 million unique visitors every month [17]. Yelp uses 22 super categories with a total of 512 subcategories linked to them

Given this large number, we created category groups that generically matched all 512 subcategories to the six popular categories we discovered in step one. To further mitigate falsely assigning tweets to one these many categories, a final list was constructed further combining analogous categories. For instance, restaurants and coffee places are combined in the "eating out" category. Additionally, music venues and movie theaters are both in the "shows" category. As a result of this process we

identified 6,359 tweets that belong to one of the six categories.

I. FEATURES

To predict a category of a tweet, we test 16 tweet features that belong to one of three groups: natural language processing (NLP) features, metadata features, and establishments density (ED) features.

A. Natural Language Processing Features

We are generating The NLP features by examining the textual content of the tweets. The purpose of these features is to identify how relevant each tweet is to each of the location categories. These features are generated by scoring tweets on six categories. The scoring is accomplished by comparing tweet content with Yelp categories.

First, we generate a representative vocabulary for each of the six categories by collecting 43,149 reviews from Yelp. Since Yelp links all every single review to a business and associates each business with a category, we were able to identify at least 944 reviews for each category; Table 1 represents this. We are using these reviews to discover the relevancy of each word to each one of the categories

From these reviews, we created a word frequency list for each category. We use term frequency–inverse document frequency (tf-idf) to score how representative and unique each word is for a category. For example, using this method, words such as "sushi" and "food" receive a high score in the category "eating out" but a low score in the "shows'' category. After generating the tf-idf scores for each category, we normalized all scores to ensure a fair comparison between the categories.

To generate the NLP features for each tweet, we split the tweets into words. For each word, and according to the result of the previous step, we calculated the six corresponding scores that demonstrated how relevant the word is to each one of the six categories. This process resulted in generating six NLP features that indicate how significant the tweet is to each category. Table 2 shows an example of the scores for a sample tweet.

B. Metadata Features.

We are generating the metadata features by examining information associated with the tweets. The features are *day of week, time of day, tweet length,* and *number of users mentioned*.

TABLE 1. TOTAL NUMBER OF REVIEWS ANALYZED FOR EACH CATEGORY

Category	**Number of reviews**
Active life	1325
Eating out	34383
Hotels	1685
Nightlife	5618
Shopping	4671
Shows	944

For time of day, we specified five time periods: morning, afternoon, evening, night, and late night. We chose these based on the assumption that tweets content will differ in these general time slots. For example, one assumption is that the majority of tweets in the "nightlife" category will be happening in either the night or late night slots. On the other hand, tweets in the category "active life" will be less likely to happen in these two slots. Morning period is from 04:00 to 11:59. Afternoon period is from 12:00 to 17:59. Evening period is from 18:00 to 20:59. Night period is from 21:00 to 22:59 and late night period is from 23:00 to 03:59.

Tweet length refers to the total number of characters in the tweet. Since the text we are capturing from the twitter streaming API has an html length tag associated with each tweet, we are also transforming all the tags back to their original format, for instance, "&amps" in the tweets have been changed back to "&."

Number of users mentioned represents the total number of people a user has tagged in his or her tweet. In Twitter, users can mention other users in their tweets, e.g., in the tweet "Brunch on The Farm with stevejmcbride jeremyrudy & @heatherhazzan @ Morning Glory Cafe http://t.co/iUaPRNu0LR," the user "heatherhazzan" is being mentioned and thus tagged in the tweet. This feature is represented by a numeric value and represents the total number of users we are able to identify in the tweet

C. *Establishments Density Features.*

We are generating the establishment density features by examining the geographical area of the tweet. Each feature corresponds to one of the six categories we are trying to predict. The features are *active life ed score, eating out ed score, hotels ed score, shopping ed score, nightlife ed score,* and *shows ed score*. Each feature shows how many businesses of that particular category are within three miles of the location of the tweet.

TABLE 2. EXAMPLE OF THE NLP FEATURES OF A TWEET.

Tweet: Brewlee Burger. #yummy #food	**Active life**	**Eating out**	**Hotels**	**Shopping**	**Nightlife**	**Shows**
Brewlee	0	0	0	0	0	0
Burger	0.002	0.1	0.01	0.01	0.2	0.01
Yummy	0.002	0.06	0.01	0.02	0.03	0.006
Food	0.06	0.8	0.3	0.3	0.6	0.4
Total	0.064	0.96	0.32	0.33	0.83	0.416

TABLE 3. AN EXAMPLE OF METADATA FEATURES

Tweet	Good dinner joeyhousey @McCaffinated amir_garrett @jacko71422 @ Yard House http://t.co/laif32NJjD
Day of week	Tue
Time of day	Morning
Tweet length	97
Number of users mentioned	2

TABLE 4: AN EXAMPLE OF THE ESTABLISHMENTS DENSITY FEATURES.

Tweet	One last beer! #bigbeer
Active_life_ed_score	21
Eating_out_ed_score	142
Hotels_ed_score	11
Shopping_ed_score	71
Nightlife_ed_score	11
Shows_ed_score	3

An example of the features is *eating out ed score*. This feature finds how many businesses categorized as "eating out" places are within three miles of the tweet. The five other features are similarly calculated.

II. DATASETS

For our study, we use Yelp and Twitter data. The Yelp dataset comprised the existing dataset created from the Yelp Dataset Challenge [18]. This dataset contains 229,901 reviews of 11,537 businesses. We used the information associated with businesses to identify the type of business entity. Table 5 shows a sample of data available for businesses in Yelp.

The Twitter dataset was collected using Twitter's streaming API, which includes filtering techniques that help in returning relevant tweets. One such technique allows applications to filter tweets based on a certain location. We used latitude and longitude and filtered using the boundaries of the Phoenix metropolitan area. The reason we only used tweets from that area is that it represents the same area as the Yelp dataset. Collecting the tweets took place during the period from April 14, 2013 to June 12, 2013. The total number of tweets collected in that period was 718,117 tweets. However, most of those tweets did not have a place attach to them. Therefore, we cannot use them to train or test the classifier. After eliminating tweets that do not meet the requirements, the remaining number of tweets was 6,359.

III. GOLD STANDARD

To train and test the classifier, we created a gold standard. Our gold standard consists of tweets that we can definitely link to one of the six location types. To build our gold standard, we used two services Twitter users commonly link to their Twitter account: Foursquare and Instagram. Foursquare enables users to "check in" to places such as restaurants and gyms. In Instagram, users can upload and share photos and attach them to a physical location.

The steps to creating the gold standard are as follow.

a) Collect tweets from the Phoenix area.

b) Check if the tweet originated from either Foursquare or Instagram (because users of these services can attach places in the tweet)

c) For the ones that did, check if the tweet contains a name of a place. For example, for the tweet "What was your favorite place to visit today? (@AlmightyTrainingStudio) [pic]: http://t.co/lL1f3iRK", there a name of a place and it is "Almighty Training Studio"

d) If the tweet contains a name of place, check if the name of the place exists on Yelp's list of businesses in the Phoenix area.

e) If the place exists on Yelp, the next step is to check if the category of the place matches one of the categories in our list of categories.

f) Add the tweet to the gold standard.

As a result of this process, we created a gold standard consisting of 6,359 tweets. Table 6 shows a sample of tweets in the gold standard.

The total number of the tweets we are using is 6,359 tweets. By examining the data, we observed that the number of tweets for each category varies. For instance, the "eating out" category has significantly more tweets than categories such as "shows" and "active life". To evaluate the impact of this uneven representation of each category, we train and test our classifier using two datasets. The first, the original dataset contains all tweets. Then, we created a stratified dataset, which has an equal number of tweets, randomly selected, for each category with a total of 400 tweets per category.

TABLE 5: A SAMPLE OF THE ATTRIBUTES FOR YELP'S BUSINESSES

Field	**Description**	**Example**
Business_ID	A unique ID to identify each business.	--5jkZ3-nUPZxUvtcbr8Uw
Full_address	The full address of the place.	4252 S 48th St Phoenix, AZ 85040
Open	Whether the place is still active	0
Categories	The type of place	['Nightlife', 'Dance Clubs']
City	The name of the city	Phoenix
Name	The name of the place	Milagro Grill
Longitude	The longitude of the place	-111.9891177
Latitude	The latitude of the place	33.495074
Stars	The rating the place received	4.0
Review count	The total number of reviews	15

TABLE 6: SAMPLE OF TWEETS FROM THE GOLD STANDARD.

Original Tweet	Place Name	Category
Getting my grape kool-aid on! @ Lo-Lo's Chicken & Waffles http://t.co/Kj4TPVlI7B	Lo-Lo's Chicken & Waffles	Eating out
I missed this gorgeous girl a lot this weekend. Date time! @mrssavdog @ Sushi Station http://t.co/1h792GXIop	Sushi Station	Eating out
Morning cocktail to round out an amazing weekend of Bocce play. Art (@ Coach House) http://t.co/1mn0r96vXU	Coach House	Nightlife
What a coincidence #leopard eddy223 @misskushner @ Nordstrom http://t.co/YpMfWEmFuE	Nordstrom	Shopping
Squat like your toosh depends on it. (@ Youfit) http://t.co/mqEqao6FE4	Youfit	Active Life

IV. STUDY RESULTS

We tested the 16 features described above using random forests classifier. Random forests are a method of classification that relies on building a myriad of decision trees for training and on generating random vectors to enhance the construction of each tree [19]. To train the classifier, we used 60% of the tweets. The remaining 40% served to test the model. We selected precision, recall, f-measure, and accuracy to evaluate the classifier.

A. Features Analysis

Using all features, for the original dataset, our classifier was able to accurately predict the correct category for 74% of the tweets. Similarly, the classifier's accuracy for the stratified dataset was 77%. The overall accuracy is similar in both cases.

To evaluate the relevance and usefulness of each group of features, we systematically tested the three types, first by themselves and then combined (see table 7.) Furthermore, we compared the accuracy of the same type in the two datasets. In all cases, the accuracy was higher than the baseline of 20% (random location type).

TABLE 7 SUMMARY OF THE RESULTS FOR EACH GROUP OF FEATURES

Group of features	Accuracy for original dataset	Accuracy for stratified dataset
Metadata	46%	29%
NLP	51%	36%
ED	88%	82%
Meta and NLP	52%	38%
ED and meta	82%	79%
ED and NLP	77%	78%
All	74%	77%

The accuracy for the **metadata** features was 46% for the original dataset and 29% for the stratified dataset. The accuracy was the lowest among the group of features in both cases. A possible reason for this is that we are currently using only four metadata features. Unexpectedly, the overall accuracy was higher for the original dataset. The reason may be that many of the tweets in the original dataset belong to the "eating out" category. For this category, metadata features such as time of day and day of week are more relevant.

The results for the **NLP features** were similar to the metadata group. The accuracy was 51% for the original dataset and 36% for the stratified dataset. Similarly, having many of the tweets in the original dataset belonging to the "eating out" category likely resulted in the relatively higher accuracy. Since NLP features are related to the textual content of the tweet, we expected the results of using them to be the highest. Combining the two groups resulted in generating a higher accuracy.

Using **only ED features** resulted in getting the highest percentage of accuracy. Combining them with any other group decreased the overall accuracy. It is important to note that for the ED features, the random forests algorithm does not simply assign the category based on the most common type of places in the area. Instead, it is trained to compare the relative popularity of each type of locations in that area. When comparing the two datasets, the performance of the ED observed the least decrease in the accuracy.

B. Category Analysis

For precision, recall, and f-measure, the results varied for each one of the categories. Table 8 shows a summary of the results for each one of the six categories. As the table shows, there are clear differences between the results for the original and the stratified datasets.

For the original dataset, since 46% of the tweets belong to the category "eating out", the classifier is classifying most tweets either accurately into the correct category or inaccurately into the category "eating out". For "eating out", as expected, the recall is very high. That is because the classifier is not failing to classify many "eating out" tweets. However, the precision for the category is the lowest because the classifier is classifying many non-"eating out" tweets as such. The precision for all the categories other than "eating out" was high. Unexpectedly, the category "shows" performed the best among all the other categories.

For the stratified dataset, the classifier performed differently. As anticipated, the precision, recall, and f-measure for the category "eating out" decreased. Moreover, the recall for all the other categories increased and the precision for most decreased. The categories "hotels", "nightlife", and "shows" performed better than the other categories.

TABLE 8 SUMMARY OF THE RESULTS FOR EACH OF THE CATEGORIES

Category	Original dataset			Stratified dataset		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Active life	90%	40%	55%	70%	73%	71%
Eating out	67%	96%	79%	54%	64%	58%
Hotels	91%	46%	61%	83%	89%	85%
Nightlife	86%	73%	79%	90%	89%	89%
Shopping	77%	27%	40%	69%	54%	60%
Shows	98%	86%	92%	97%	93%	94%

V. CONCLUSION AND FUTURE WORK

In this study, we built a classifier to predict the type of location a user is tweeting from. We identified six possible types of places and applied the location-aware tweet classification classifier to 6,359 tweets. Experimental results showed 77% accuracy when using a stratified dataset and 74% accuracy when using the original dataset. In both cases, the accuracy is higher than the baseline of 20%. Based on our numbers, we concluded that assigning tweets to a location solely based on their content result in low accuracy. If there is some indication of the surrounding businesses, the ED features, then the results improve significantly.

There are a number of limitations in our study. First, there are certain types of places that the classifier does not cover such as health-care related facilities and religious organizations. Second, the establishments density features in the classifier trigger execution only if the tweet is geocoded.

Future work includes a number of possible improvements to the current classifier. First, a possible enhancement is adding new categories to the classifier. By doing so, the coverage in terms of geocoded tweets and physical locations would be enhanced. Another possibility is improving the method we are currently using to generate the NLP features implemented; this will increase the overall accuracy of outcomes. Last, adding a new group of features to increase the performance of the classifier, e.g., features related to the sentiment of the tweet.

In addition, this work could be extended in various ways. For instance, the NLP method for generating scores could be applied to other areas of research. Moreover, this work could be extended by applying the classifier on a set of tweets from users instead of only looking at single tweets. Finally, the results of the classifier could be the foundation for apps and services.

VI. REFRENCES

[1] S. Fiegerman,, "Will Twitter Ever Be Mainstream?", Internet: http://mashable.com/2013/10/04/twitter-ipo-mainstream/, Oct. 4, 2013 [Dec. 10, 2013]

[2] G. Gao. "Looking for news on Twitter's IPO? many young adults may find it … on Twitter.". Internet: http://www.pewresearch.org/fact-tank/2013/11/07/looking-for-news-on-twitters-ipo-many-young-adults-may-find-it-on-twitter/, Nov. 7, 2013 [Dec. 10, 2013]

[3] Twitter. About. Internet: https://about.twitter.com/company

[4] H. Chang, D. Lee, M. Eltaher, and J. Lee, "@Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage" Proc. of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM '12), Aug 2013, pp. 111-118, doi=10.1109/ASONAM.2012.29

[5] Z. Cheng, J. Caverlee, and K. Lee. "You Are Where You Tweet: a Content-based Approach to Geo-locating Twitter Users" Proc. the 19th ACM international conference on Information and knowledge management (CIKM '10), Oct 2013, pp. 759-768, doi=10.1145/1871437.1871535

[6] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from Justin Bieber's Heart: the Dynamics of the Location Field in User Profiles" Proc. the SIGCHI Conference on Human Factors in Computing Systems (CHI '11), May 2011, pp. 237-246, doi=10.1145/1978942.1978976

[7] S. Kinsella, V. Murdock, and N. O'Hare, "I'm Eating a Sandwich in Glasgow: Modeling Locations With Tweets" Proc. the 3rd International Workshop on Search and Mining User-generated Contents (SMUC '11), Oct 2011, pp. 61-68, doi=10.1145/2065023.2065039.

[8] D. Thom, H. Bosch, R. Krueger, and Thomas. E, "Using Large Scale Aggregated Knowledge for Social Media Location Discovery," 2014 47th Hawaii International Conference on System Sciences (HICSS), 6-9 Jan. 2014, pp.1464,1473, doi: 10.1109/HICSS.2014.189.

[9] A. Sadilek, H. Kautz, and J. P. Bigham, "Finding Your Friends and Following Them to Where You Are" Proc. the fifth ACM international conference on Web search and data mining (WSDM '12), Feb 2012, pp. 723-732, doi=10.1145/2124295.2124380.

[10] D. Rout, K. Bontcheva, D. Preoțiuc-Pietro, and T. Cohn, "Where's @wally?: a Classification Approach to Geolocating Users Based on Their Social Ties" Proc. the 24th ACM Conference on Hypertext and Social Media (HT '13), May 2013, pp. 11-20, doi=10.1145/2481492.2481494.

[11] J. McGee, J. Caverlee, and Z. Cheng, "Location Prediction in Social Media Based on Tie Strength " Proc. the 22nd ACM international conference on Conference on information & knowledge management (CIKM '13), Oct 2013, pp. 459-468, doi: 10.1145/2505515.2505544.

[12] T. Takehara, S. Miki, N. Nitta, and N. Babaguchi, "Extracting Context Information from Microblog Based on Analysis of Online Reviews" Proc. IEEE International Conference on Multimedia and Expo Workshops (ICMEW), July 2012, pp. 248-253, doi: 10.1109/ICMEW.2012.49.

[13] D. Ghosh and R. Guha (2013) "What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System". Cartography and Geographic Information Science 40 2 90-102

[14] M. Okazaki, and Y. Makoto. "Semantic Twitter: Analyzing Tweets for Real-Time Event Notification." *Recent Trends and Developments in Social Software. International Conference on Social Software. BlogTalk 2008 and BlogTalk 2009. Revised Selected Papers*, p. 63.

[15] Yelp. Yelp for Developers. "Category List". Internet: http://www.yelp.com/developers/documentation/category_list/

[16] A. Hicks, S. Comp, J. Horovitz, M. Hovarter, M. Miki, and J. Bevan. "Why people use Yelp.com: An exploration of uses and gratifications ", Computers in Human Behavior, Volume 28, Issue 6, November 2012, Pages 2274-2279, http://dx.doi.org/10.1016/j.chb.2012.06.034.

[17] Yelp. Want More Customers? Try Yelp Ads. Internet: http://officialblog.yelp.com/2014/03/want-more-customers-try-yelp-ads.html

[18] Yelp. "Yelp Dataset Challenge". Internet: http://www.yelp.com/dataset_challenge/

[19] L. Breiman (2001). "Random Forests" Machine Learning, vol 45 (1), Oct 2001, pp. 5–32, doi:10.1023/A:1010933404324.