Social Media-based Profiling of Business Locations

Francine Chen   Dhiraj Joshi   Yasuhide Miura   Tomoko Ohkuma
FX Palo Alto Laboratory, Inc.   Research & Technology Group
Palo Alto, CA   Fuji Xerox Co., Ltd., Japan
{chen, joshi}@fxpal.com   {yasuhide.miura, ohkuma.tomoko}@fujixerox.co.jp

ABSTRACT

We present a method for profiling businesses at specific locations that is based on mining information from social media. The method matches geo-tagged tweets from Twitter against venues from Foursquare to identify the specific business mentioned in a tweet. By linking geo-coordinates to places, the tweets associated with a business, such as a store, can then be used to profile that business. We used a sentiment estimator developed for tweets to create sentiment profiles of the stores in a chain, computing the average sentiment of tweets associated with each store. We present the results as heatmaps which show how sentiment differs across stores in the same chain and how some chains have more positive sentiment than other chains. We also created profiles of social group size for businesses and show sample heatmaps illustrating how the size of a social group can vary.

Categories and Subject Descriptors

H3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

location, profiling, social media, data mining, clustering

1. INTRODUCTION

The use of social media for sharing thoughts, opinions and updates about oneself with friends and the general public has been growing rapidly. In turn, these expressions are stored in public social media platforms and can serve as a rich source of information. The applications of mining this information are wide-ranging and include epidemiology, public opinion on political issues, event detection, and public opinion of businesses and their products. In addition to conventional methods for assessing customer satisfaction, such as questionnaires and comment forms, social media is rapidly becoming a widely-used method for expressing judgments about places. As a result, companies employ workers
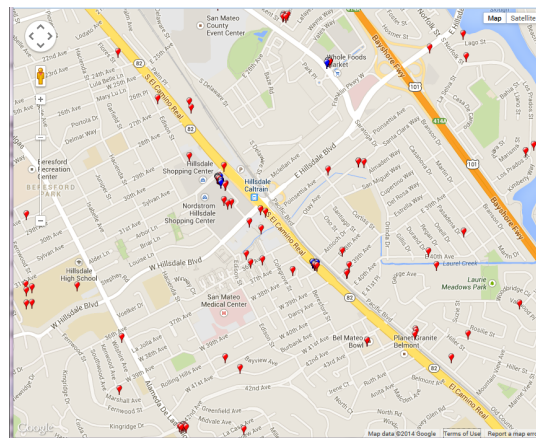
Figure 1: Tweets mentioning Starbucks (red) are often distant from a Foursquare Starbucks venue (blue) (San Mateo, CA).

specifically to track comments and to address issues about their products on public forums and microblogs.

Traditional assessment of customer opinion using questionnaires and comment forms allows a merchant to understand opinion only about the stores in question. With social media, information about all stores is available to anyone. Thus a business can easily collect data, such as tweets, about competitors as well as about themselves, and then mine the data to perform an assessment against their competitors.

While forums such as TripAdvisor and Yelp allow users to post opinions about their experiences with businesses, using these forums requires more effort than sending a quick microblog on Twitter. With Twitter, the casual opinions of many people are expressed.

Tweets may be tagged with geo-coordinates and the percentage of tagged tweets is growing. As of August 2013, about 6% of Twitter users opt-in to broadcast their location[1]. In some locations, an even larger proportion of people tag their tweets with geo-coordinates; [14] noted that out of 26 million tweets in New York City and Los Angeles, 7.57 million tweets, or about 29%, were GPS-tagged.

Geo-tagged tweets provide the longitude and latitude of the tweet; however, the actual place that a user is tweeting from is not provided. Although the geo-coordinates of places are available from cities for businesses and from dictionaries

---

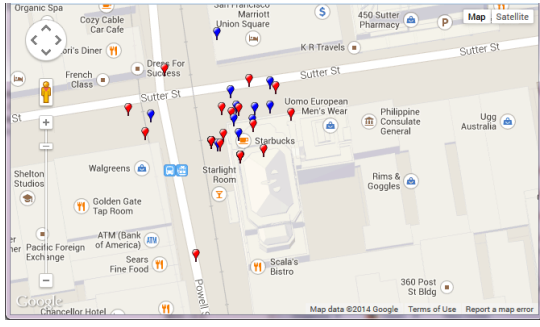[1]http://dornsife.usc.edu/news/stories/1479/mapping-the-twitterverse/

Figure 2: Some tweets (red) and Foursquare venues (blue) associated with a Starbucks store are closer to other stores than the Starbucks store (San Francisco, CA).

of geographic locations, the information is scattered, partially complete, and needs to be reconciled. We chose instead to use Foursquare *venues* for identifying places. Foursquare venues are crowd-sourced places where users check-in. Examples of venue types include stores, stadiums, or points of interest. Each venue is associated with a latitude and longitude. Knowing the actual venue that is being tweeted about can provide much richer information about each of the venues in a collection of geo-tagged tweets. Figure 1 shows the location of Starbucks venues (blue) for three stores, and the location of all tweets where Starbucks is mentioned (red). Note that many of the tweets are not near a Starbucks venue. Figure 2 shows multiple Starbucks venues (blue) associated with one Starbucks store and tweets associated with the Starbucks venues (red). Note that some of the venues and tweets are closer to other businesses and venues than they are to the Starbucks store.

In this paper we present a method for matching geo-tagged tweets that mention a place with the actual venue, and also simultaneously filtering tweets where it is unclear which venue is being referred to. We then give examples of profiling a venue based on the matched tweets. We perform two types of profiling: the sentiment at a given venue, and the social group size of users at a given venue. To estimate social group size, we make use of the fact that Twitter posts can contain photos and analyze them for social groups.

Our contributions include:

- a method for matching geo-tagged tweets with geolocated venues
- visualizations for profiling venues based on information mined from venue-tagged tweets

2. RELATED WORK

A common approach to geo-based investigations is to use locations from the self-reported home location of Twitter users, rather than the geolocation of each tweet. For example, [13] used home locations, which were primarily cities, and [15] mapped home locations to counties. [8] tagged Points of Interest (POI) in tweets, where the set of POI names are extracted from tweets "associated with Foursquare check-ins". However, POI names that correspond to multiple locations, such as chain stores, were not disambiguated. In contrast to these works, [11] visualized the happiness of individual geo-tagged tweets in New York City and the con-

tinental US. Similarly to this work, we also focus on geo-tagged tweets. But in contrast, we map the tweets to specific businesses or venues. To the best of our knowledge, our investigation is the first work to map the geo-tagged tweets to businesses or places at *specific geolocations*.

There have been a number of works on identifying the location of a social media post when the post does not contain geolocation information. For example, from only tweet text [3] was able to place 51% of Twitter users within 100 miles of their actual home location. [10] used an ensemble of classifiers for city, state, and time-zone estimation of a user's home location. [7] created language models for Twitter to predict country, state, town, and zip code locations. [14] used the GPS position of a user's friends to identify the user's location within 100 meters of their actual location with an accuracy of 84.3% when the locations of nine friends are used. The current accuracy of these methods is still too coarse for use in associating the locations with venues; furthermore, none of these works associates locations with places or venues, such as stores, stadiums, or points of interest.

Photos have also been used for geolocation. For example, [12] used gender-based models of Flickr tags to predict location, with a best accuracy of 21.5%. [17] used the information in photos together with compass direction to perform localization. [4] used SVMs to predict the location of photos of landmarks based on visual, textual, and temporal features. And [9] employed visual nearest neighbors ranking to geo-locate a photo. However, even if geolocation performance is high, only a minority of tweets contain at least one photo. In our geo-tagged Twitter corpus, less than 4% of tweets contained an Instagram photo. In addition, not all photos are indicative of a user's location. We also looked at the EXIF information associated with photos, and found that the geo-position information had been stripped. Thus, while geolocation based on photos can be helpful for some tweets, using photo-based methods alone is not sufficient.

In this work, we focus on matching geo-tagged tweets with businesses at specific locations, and then mining the information at each location. For mining, we estimate the sentiment of tweet text using a sentiment analyzer that we implemented and we also estimate social group size from photos.

3. DATA SETS

To profile venues, we need to associate the venues with public posts expressing opinions. For this we collected two sets of data: 1) geo-tagged Twitter tweets and 2) Foursquare venues. A Foursquare venue is tagged with the name of a place and a geo-coordinate. Although Foursquare users may make comments when they check-in to a venue, they are not public on the Foursquare site. To gather public postings, we collected Twitter tweets. We defined a collection area to be inside latitude [37.10, 38.15] and longitude between [-122.6, -121.6], which covers most of the San Francisco Bay Area, including San Francisco and San Jose.

3.1 Geo-tagged Tweets

We collected tweets using the Twitter Streaming API[2]. We specified a geo-query for tweets inside our collection coordinates and collected 16,040,427 geo-tagged tweets during 10-month period from June 4, 2013 to April 7, 2014.

---

[2]https://dev.twitter.com/docs/api/streaming

Tweets are public and provide a sample of user opinions from a wide variety of sources and social media platforms. In addition to posting tweets directly from a Twitter App, e.g., Twitter for iPhone or Twitter for Android, other social media platforms, such as Foursquare, often allow users to publicly post through Twitter as well as on the source itself. We noted over 1100 different sources in our geo-tagged tweets and that the most popular sources, other than Twitter apps, include Instagram and Foursquare.

Some tweets contain one or more links to photos. From the metadata associated with a tweet, we identified links to Instagram photos mentioned in the tweets and downloaded those photos, a total of 601,164, for use in location profiling.

3.2 Foursquare

Foursquare venues are crowd-sourced locations that users identify when they check-in to a place. Foursquare recommends checking into places that you're at, rather than what you're walking by. It also discourages fake check-ins, but we noted that some users are creative in naming locations, especially their homes. For example, in our dataset there are six homes that include "The Chamber of Secrets" in the name. We queried Foursquare using their venue search API[3] for venues near the geo-coordinates of our San Francisco Bay Area tweets. We kept the query rate below Foursquare's rate limit and cached the results to reduce the number of queries. When the maximum number of results was returned, the query was refined to a smaller area to try retrieving all of the closest locations. The meta-data that we extract for each venue includes:

- latitude, longitude
- venue name
- number of check-ins
- number of unique visitors

Algorithm 1 Grouping venue and tweet locations

Input: u: user-specified venue, D: specified maximum geo-distance between a venue and tweet, V: a set of geo-tagged venue locations containing u, T: a set of geo-tagged tweets
Output: *venueTweetGroups*: clusters of venues and tweets associated with each store at a specific location
1: $result \leftarrow \{\}$
2: $venueTweets \leftarrow \{\}$
3: $candTweets \leftarrow \{\}$
4: **for** each tweet t in T **do**
5:      **if** $u \in t$ **then**
6:          $venueTweets \leftarrow t$
7:      **end if**
8: **end for**
9: **for** each venue v in V **do**
10:      **for** each tweet t in $venueTweets$ **do**
11:          **if** $||\text{geo}(v) - \text{geo}(t)|| < D$ **then**
12:              $candTweets \leftarrow t$
13:          **end if**
14:      **end for**
15: **end for**
16: $clusters, outliers \leftarrow \text{DBScan}(candTweets \cup V, \text{minNeighborSize=5})$
17: $venueTweetGroups \leftarrow clusters - outliers$

---

4. METHOD

In this section, we describe our method for matching geo-tagged tweets to Foursquare venues. We also briefly describe

---

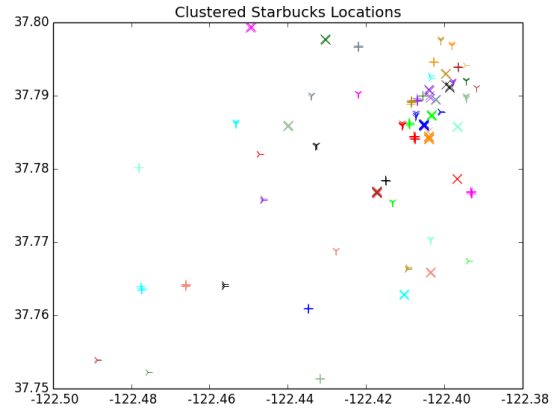[3] https://developer.foursquare.com/docs/venues/search

Figure 3: Results of clustering Starbucks venues and tweets. The plot shows locations in the city of San Francisco; downtown is at the top right.

the sentiment estimator and social group size estimator used for profiling.

4.1 Matching Tweet Geo-coordinates to Stores

To match geo-tagged tweets to Foursquare venues, several factors need to be considered. Although the geo-coordinates of a tweet when Foursquare is the source can be directly mapped to a venue (Foursquare was the source of 492,529 tweets), tweets from other sources may instead reflect the geo-coordinates of the user's current location. Furthermore, a user may refer to a place in their tweet text without actually being there, as shown by the many red markers in Figure 1 that are not near a blue marker. If there are multiple venues with the same name, as in Figure 1, it can be difficult to determine the actual location, if any, to which the user was referring.

We first identify the tweets and venues that correspond to a Foursquare venue with a specified business name. We do this in a multi-step process as shown in lines 1-15 of Algorithm 1. For a specified Foursquare venue name, tweets that mention that venue, and optionally, venue nicknames, are identified. These tweets are then filtered to keep those that are within distance D (we used .0008 degrees, or about 290 ft) from a Foursquare venue with the specified name.

A store at a given location, e.g., a specific Starbucks store, may have multiple check-in locations because Foursquare venues are crowd-sourced. People may create a new venue for different reasons. For example, the store may cover a large area or a user may check in when they are near, but not in, the store. They may also make fake Foursquare venues.

To combine multiple Foursquare check-in venues that correspond to a single store and to reduce the number of fake Foursquare venues, we used clustering to group geolocations. A minimum number of check-ins and unique visitors in each cluster was required, based on the assumption that there will be few check-ins and unique users at a fake venue. Specifically, for a given location name, we used DBSCAN (from http://scikit-learn.org) to cluster over all venues tagged with the location name and all tweets containing the location name. Tweets were included to take advantage of the fact that tweets are not constrained to venue locations, as shown in Figure 2. Thus, results from applying DBSCAN, which performs density-based clustering, may be more robust due
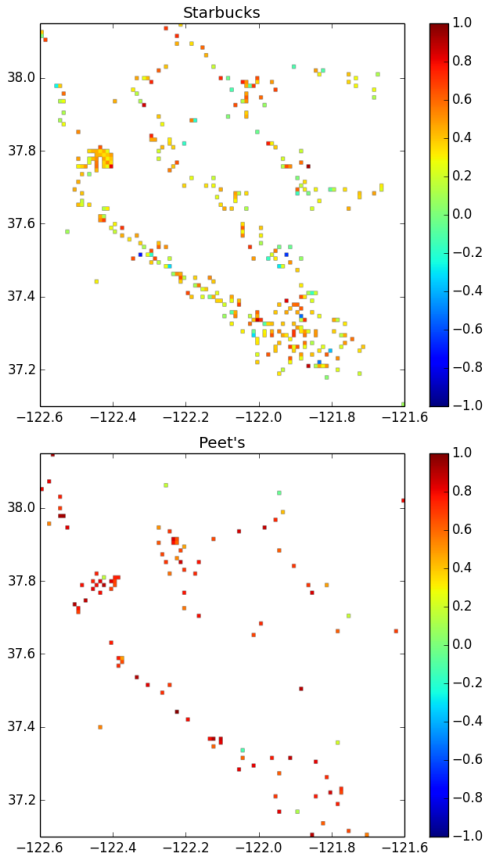
Figure 4: Average sentiment of subjective tweets at Starbucks (top) and Peet's Coffee (bottom) in SFBA. The color bar shows the sentiment mapping from very positive (red) to very negative (blue).
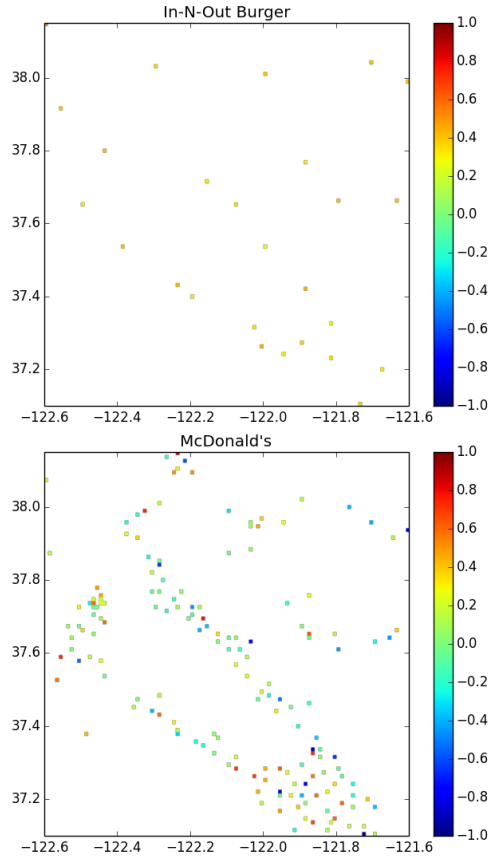

Figure 5: Average sentiment of subjective tweets at In-N-Out Burger (top) and McDonald's (bottom) in SFBA. The color bar shows the sentiment mapping from very positive (red) to very negative (blue).

to the set of unique locations being denser. The maximum allowable distance between two samples was set to be .0008 degrees, or about 290 ft. A minimum of five neighbors per point was required, or else the point was regarded as an outlier. The outlier samples may be due to fake Foursquare venues, as well as non-popular locations or users mentioning a venue when they are somewhere else.

Figure 3 shows clustering results for Starbucks locations in the city of San Francisco. Each mark represents a tweet or Foursquare venue, and a unique color and shape combination is used for each group of venues. Thicker or "fuzzier" marks indicate multiple nearby tweets or venues.

The tweets associated with a cluster are tagged with the "core" venue and its location, where the core venue is defined to be the venue in the cluster with the most check-ins. Outlier samples are not tagged and therefore are not used in profiling.

We next characterize a location with two types of attributes to illustrate the profiling of store locations: average sentiment expressed by customers and the size of the social groups as estimated by the photos people take at a location.

4.2 Estimating Sentiment

There have been many works on general sentiment estimation, and a smaller number focused on estimating the sentiment of tweets. Tweet sentiment estimation methods based on machine learning have been observed to perform slightly better than lexicon-based methods [16]. To estimate the sentiment of tweets at a location, we implemented a logistic-regression based sentiment analyzer trained on Twitter tweets.

Accurate identification of non-opinionated tweets is important because many tweets do not express sentiment. For example, the default for checking in on Foursquare is "I'm at <placename> (<place location>) <URL>". Another common use of Twitter is for people to announce their status: "using Starbucks wifi cause I can", or "Starbucks with _chriiisssss". Subjectivity classification of each tweet was first performed by determining whether the tweet text contained subjective terms from the MPQA subjectivity lexicon [18].

It was observed in [1] that topic-dependent Twitter sentiment models improve performance for only some topics. Since the tweets may cover a variety of topics, we created a topic-independent model.

The polarity of the tweets that were deemed subjective (as opposed to objective) was computed using the distant learning approach described in [5]. The training data from the Sentiment140 tweet corpus[4] was used for distant learning. By pre-filtering objective, or neutral, tweets, with the MPQA-based subjectivity classifier mentioned above, our

---

[4]http://help.sentiment140.com/for-students

sentiment analyzer can be considered as a simple extension of the two class (positive and negative) classifier of [5] to a three class (positive, negative, and neutral) classifier.

The sentiment analyzer outputs two values: 1) whether the tweet is subjective or objective and 2) a score ranging from -1.0 to 1.0 corresponding to very negative to very positive sentiment. From these values, the sentiment score is computed as:

$$score(tweet) = \begin{cases} Pr(positive) & \text{if output = positive} \\ 0.0 & \text{if output = objective} \\ -Pr(negative) & \text{if output = negative} \end{cases}$$

where $Pr(class)$ denotes the probability of "class" estimated by the logistic regression model.

4.3 Estimating Social Group Size

The classification of people in photos into social groups has been used for travel recommendation. Loosely following [2], which classified travel groups into solo, couple, family, and friends, we defined social group size based on the number of faces in a photo. Face detection was performed using the OpenCV face detector, which detected faces in a total of 165,844 photos. The number of faces in a photo was then quantized into one of four classes: single (1 face), pair (2 faces), small group (3-6 faces) and larger group (at least 7 faces), and mapped to a **group size code** of 1, 2, 3, or 4, respectively. These codes were used when computing average group size for heatmaps.

5. RESULTS

To visualize the profiling results, we created heat maps of a profile attribute at different locations of the same venue, e.g., Starbucks at different locations. The collection area described in Section 3 was used in generating the heat maps in Figures 4–6. This area covers most of the San Francisco Bay Area (SFBA), including San Francisco (middle left) and San Jose (bottom right). The longitude and latitude values were each quantized into 100 bins, for a total of 10,000 cells. White areas in a heatmap indicate that a store was not present.

5.1 Sentiment Profiling

To create a sentiment heatmap, for each set of tweets that were clustered to the same "core" venue, the tweets were filtered to keep only those where a sentiment was expressed, i.e., $score(tweet)$ was nonzero. The sentiment scale ranged from -1.0 (very negative) to 1.0 (very positive) which was mapped over the color spectrum from blue to red, respectively. The average sentiment score for the tweets associated with all core venues in a cell was computed and used as the value of the heat map.

In Figure 4, it can be observed that the different Starbucks locations exhibit a variety of average sentiment values. While most of the locations are slightly positive (yellow), some are highly positive (red) and a smaller number are highly negative (dark blue). Peet's Coffee & Tea is a smaller competitor to Starbucks. Comparing the average sentiment for Starbucks locations and Peet's locations, we note that Peet's locations tend to have primarily positive sentiment, noticeably higher than Starbuck's on average. The more positive perception of Peet's is in agreement with the average Yelp scores for the first 20 results returned from queries for Starbucks and Peet's in San Francisco (on July
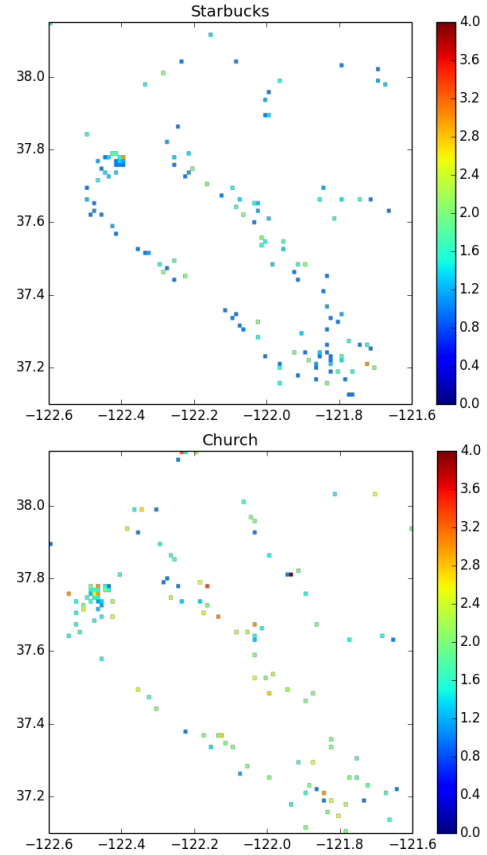

Figure 6: Average social group size code at Starbucks (top) and churches (bottom) in SFBA. The color bar shows the average social group class size code from large (red) to single (blue).

10, 2014), with values of 3.6 and 4.0 (out of a best score of 5.0), respectively.

Comparing two fast food burger chains, In-N-Out Burger, which advertises its ingredients as being freshly made each day, with McDonald's, we see in Figure 5 that while In-N-Out Burger has mildly positive sentiment overall, the sentiment about McDonald's locations varies but is overall more negative. Also, there are several McDonald's locations that exhibit quite negative sentiment. Again, the more positive perception of In-N-Out is in agreement with average Yelp scores of 4.25 and 2.55 for the two In-N-Out stores in or near San Francisco and first 20 results from a query for McDonald's stores in San Francisco, respectively.

This type of store location-based information can be used by management to identify stores with happy customers that are more likely to have good practices and to perhaps use this information to improve more poorly-rated stores.

5.2 Social Group Size Profiling

Knowing the size of social groups who visit a venue or shop (singles, pairs, small, or large groups) can be helpful to commercial businesses for targeting their products and advertisements appropriately. To create a heat map, the average of the group size codes defined in Section 4.3 was computed over all tweet photos with at least one face that clustered to a "core" venue in each cell.

In the heat maps in Fig 6, we visualize the detected social group sizes at Starbucks locations and at churches in the San Francisco Bay Area. We note that the Starbucks heat map is skewed towards single faces. In contrast, the heat map for churches exhibits somewhat larger social groups on average, with some red and orange areas. This observation is intuitive as churches are gathering places that host social events such as weddings, whereas people visit coffee shops more frequently alone than with friends or family.

6. SUMMARY AND FUTURE WORK

We have presented a method for matching geo-tagged Twitter tweets with Foursquare venues at specific locations. Our method also consolidates the multiple crowd-sourced venues that are check-ins for a single store or place. We profiled example venues in terms of sentiment and social group size and presented results from our profiling. For both sentiment-based and social-group-based location profiling, our top-level results are in line with general perceptions about the venues. In addition, we showed that there may be variation by location for a particular venue; this can be used by store owners to identify stores with happy customers and stores that need improvement. In addition, store owners can compare profiles of their stores against competitor stores.

In this work, we conservatively used only tweets where a venue of interest was mentioned to increase the likelihood that the tweet is relevant to the particular venue. However, this results in missing some relevant tweets, such as when a user comments about the food at a restaurant without mentioning the name of the restaurant. Even so, from the reduced set of tweets, we were able to glean differences between different locations of the same store brand as well as differences between store brands in the San Francisco Bay Area. Although we used the Twitter Streaming API to collect tweets, quicker collection methods could be used, such as the Twitter Firehose, or for specific venues, the Twitter Search API.

In the future, we would like to explore methods for identifying tweets related to a venue when the venue has not been explicitly mentioned. One approach is to integrate image-based and textual-based approaches, as was done in [6] for video placing. Additional location-based profiling could also be performed. For example, it may be of interest to determine the gender and age of customers from tweeted images. While our focus has been on stores, our methods can be applied to other venue types, such as Points of Interest (e.g., aquarium, zoo, scenic lookout, stadiums) and public transportation stations (e.g., BART, Caltrain).

7. REFERENCES

[1] F. Chen and S. H. Mirisaee. Do topic-dependent models improve microblog sentiment estimation? In *Proceedings of ICWSM*. AAAI, 2014.

[2] Y.-Y. Chen, A.-J. Cheng, and W. H. Hsu. Travel recommendation by mining people attributes and travel group types from community-contributed photos. *IEEE Transactions on Multimedia*, 15(6):1283–1295, 2013.

[3] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of CIKM*, pages 759–768. ACM, 2010.

[4] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of WWW*, pages 761–770. ACM, 2009.

[5] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12, 2009.

[6] P. Kelm, S. Schmiedeke, J. Choi, G. Friedland, V. N. Ekambaram, K. Ramchandran, and T. Sikora. A novel fusion method for integrating multiple modalities and knowledge for multimodal location estimation. In *Proceedings of the 2nd ACM International Workshop on Geotagging and Its Applications in Multimedia*, pages 7–12. ACM, 2013.

[7] S. Kinsella, V. Murdock, and N. O'Hare. I'm eating a sandwich in Glasgow: modeling locations with tweets. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, pages 61–68. ACM, 2011.

[8] C. Li and A. Sun. Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of SIGIR*, pages 43–52. ACM, 2014.

[9] X. Li, M. Larson, and A. Hanjalic. Geo-visual ranking for location prediction of social images. In *Proceedings of ICMR*, pages 81–88. ACM, 2013.

[10] J. Mahmud, J. Nichols, and C. Drews. Where is this tweet from? Inferring home locations of twitter users. In *Proceedings of ICWSM*. AAAI, 2012.

[11] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE*, 8(5), 2013.

[12] N. O'Hare and V. Murdock. Gender-based models of location from flickr. In *Proceedings of the ACM Multimedia 2012 Workshop on Geotagging and Its Applications in Multimedia*, pages 33–38. ACM, 2012.

[13] D. Quercia, L. Capra, and J. Crowcroft. The social world of twitter: Topics, geography, and emotions. In *ICWSM*. AAAI, 2012.

[14] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. In *Proceedings of WSDM*, pages 723–732. ACM, 2012.

[15] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E. Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E. Seligman, et al. Characterizing geographic variation in well-being using tweets. In *Proceedings of ICWSM*. AAAI, 2013.

[16] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[17] B. Thomee. Localization of points of interest from georeferenced and oriented photographs. In *Proceedings of the 2nd ACM International Workshop on Geotagging and Its Applications in Multimedia*, pages 19–24. ACM, 2013.

[18] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT and EMNLP*, pages 347–354. ACL, 2005.