“Hey #311, come clean my street!”
A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints

Ryan M. Eshleman and Hui Yang
reshlema@mail.sfsu.edu, huiyang@sfsu.edu
Department of Computer Science
San Francisco State University
1600 Holloway Avenue, San Francisco, CA, USA

Abstract— Twitter data has been applied to address a wide range of applications (e.g., political election prediction and disease tracking); however, no studies have been conducted to explore the interactions and potential relationships between twitter data and social events available from government entities. In this paper, we introduce a novel approach to investigate the spatio-temporal relationships between the sentiment aspects of tweets and 311 civil complaints recorded in the 311 Case Database, which is freely available from the City of San Francisco. We also present results from two supporting tasks: (1) We apply sentiment analysis techniques to model the emotional characteristics of five metropolitan areas around the globe, allowing one to gain insight into the relative happiness across cities and neighborhoods within a city; and (2) we quantify the performance of several open-source machine learning algorithms for sentiment analysis by applying them to large volume of twitter data, thereby providing empirical guidelines for practitioners. Major contributions and findings include (1) We have developed a system for the relative ranking of happiness of a geographical area. Our results show that Sydney, Australia is the happiest of the five cities under study; (2) We have found a counterintuitive positive correlation between 311-report frequency and local sentiment; and (3) When performing sentiment analysis of tweets, the inclusion of emoticons in the training dataset can lead to model overfitting, whereas NLP-based features seem to have a great potential to improve the classification accuracy.

Index Terms—311 civil complaints, happiness index, online social networks, sentiment analysis, spatio-temporal analysis, twitter data

I. INTRODUCTION
The proliferation of social data that is freely available from sources such as Twitter and government entities has posed the problem and opportunity of developing new ways to explore the interactions and, as of yet undiscovered, relationships present in the diverse data sets. For example, the San Francisco 311 service has opened its case database for public use [22]. San Francisco 311 serves as the customer service department for the city and its case database represents the city’s immediate and direct responses to the needs of San Francisco residents. Twitter [23] is a micro-blogging platform where users can (almost) freely express their opinions and observations in real time. Both Tweets and 311 records contain geographical coordinates and timestamps indicating where and when a tweet or civil complaint originates. This makes it possible to identify instances that co-occur within the same spatio-temporal range in these two data sources. For example, one can identify the set of Tweets and 311 reports that occurred on Saturday, March 22, 2014, within one mile of San Francisco City Hall. These two sources together provide a unique platform to make potential impact to the society at large, because Twitter data contains the citizenry’s real-time sentiment and 311 contains the government’s responses to citizen’s day-to-day needs. By developing new methods to explore the relationship between these data, for instance, we can gain insights into how the local governments’ responses to civilian needs might influence the citizens’ mood. Furthermore, by establishing a connection between these two data sources, we also forge a path for future research with social datasets beyond 311, such as Police Records, Public Utilities records, even school, traffic and business records. Finally, the lack of existing research in this area and its potential for impact make this integrated data analysis approach even more compelling.

Put into a larger social context, such research is important because government entities often work within limited resources to serve their constituents. The results from the above analysis can facilitate government entities and public service organizations to better understand the people they serve and the effect of their actions, as well as to identify potential issues in a timely manner. Consequently, this could lead to more effective allocation of tax dollars and influence public policy decisions.

We have collected two datasets of tweets. The first dataset, “GlobalTweets” contains tweets originating from five cities worldwide spanning over 19 days. These cities include San Francisco, New York, Los Angeles, London, and Sydney. The second, “SFTweets,” contains only tweets that originate from within San Francisco over a period of 20 days. A third dataset collected from the City of San Francisco contains case data from 311 reports.
Given the large volume of tweets per day (on the order of 400 million, globally), it would be intractable to manually classify the sentiment of each tweet. Our approach is to automate the sentiment classification through the application of machine learning algorithms. The implicit prerequisite task here is to evaluate the performance of various algorithms for sentiment classification. A comparison of six algorithms indicates that the simple Multinominal Naïve Bayes classifier [12] is a top performer in both runtime and accuracy. We therefore choose it as the classifier to analyze the sentiment of twitter data.

After we classify the sentiment of a sufficient number of tweets, we analyze the different spatio-temporal relationships as follows. We first break down the geographic range of each dataset into sub-ranges. GlobalTweets is broken down by city and SFTweets by administrative districts defined by zip code. The sentiment characteristics (or the Happiness Index) of each range can then be evaluated by finding the average sentiment of each district over the entire temporal range. We further refine this analysis by including the temporal information in our datasets. By visualizing the trend lines of this data for each of the geographical ranges, we can identify the daily and weekly patterns in sentiment fluctuation.

We proceed in the analysis of the SFTweets dataset by integrating the 311 case reports data. The frequency of 311 reports is plotted over time for each of the administrative districts and included in the sentiment flow graphs. This leads to a visualization of the relationship between 311 report frequency and local sentiment. During our analysis, we notice that the concept of administrative districts do not always correspond to the “true” spatial regions as manifested by the data. We therefore apply a density-based cluster algorithm [17] to identify the spatial clusters of 311 reports. We then perform the previously described spatio-temporal analysis based on these new geographical districts.

The results of our analysis show that, of the five cities under study, Sydney, Australia exhibits the highest Happiness Index and Los Angeles the lowest. Of the administrative districts of San Francisco, the Embarcadero-Barbary Coast district shows the highest happiness, while Ingleisde-Excelsior showed the lowest.

The results after we include the 311 data in our analysis show a positive correlation between the number of 311 reports and the mean local sentiment. This trend is reinforced when we use the geographic ranges identified by the aforementioned clustering algorithm. We observe that these ranges tend to have a higher average sentiment than the surrounding administrative districts, which is an interesting phenomenon that can be of interest for both government entities and sociologists.

II. RELATED WORK

Twitter data have been widely used to address a host of application domains. For example, researchers have explored the potential of using Twitter data to predict political election outcomes with mixed success [1][24][25]. For instance, Mahmood and colleagues concluded that the tweet generating population in Pakistan is not representative of the general voting population and therefore their tweets don not make for a strong predictor. Anjari et al. have identified conditions where twitter may fail or be successful as a prediction tool. Twitter data has also been used to track and monitor disease epidemics and the aftermath of natural disasters [2][3][4][26][27].

To the best of our knowledge, we are the first to propose to study the potential spatio-temporal interactions between twitter data and social events collected and managed by government entities (e.g., the 311 databases of civil complaints).

Twitter data has been used to model the emotional mood of the U.S. and other countries around the globe, among which are the two notable studies led by Mislove and colleagues [6], and Golder and Macy [7], respectively. These studies provide a national or global map of the emotional characteristics based on twitter data collected over a long period (2-3 years) a few years ago. We strive to provide a relatively microscopic view of the emotional states of the five chosen cities in our present day. As discussed later in Section IV, each city is unique in its own way.

Finally, a few studies have conducted comparative analysis on the performance of different sentiment analysis algorithms when applied to micro blogs such as tweets [8][9]. For example, Gonçalves and colleagues report a comprehensive comparison of 8 commonly used sentiment classification methods ranging from the simple emoticon-based approach to more complex psychometric-based one [8]. In this work, we focus on open-source machine learning algorithms, which include the elegant yet simple Naïve Bayesian classifier [12] and the more recent, state-of-the-art Sentiment Treebank-based deep model [10].

III. METHODS

In this section, we first describe the acquisition of twitter data and San Francisco (SF) 311 case data. We will also brief the pre-processing techniques performed prior to the subsequent data analysis. Next, we introduce the six sentiment classification algorithms examined in this work. We then describe the main idea behind obtaining the sentiment of tweets for across-city and –neighborhood comparison. Finally, we discuss the spatio-temporal analysis using twitter data and SF 311 data.

4. Datasets

We have collected a total of three datasets (See Table 1). We collect tweets using the Twitter Streaming API that allows one to receive live tweets filtered according to user-specified criteria. We use the criterion that allows one to specify the geographical area in the form of a “Bounding Box” and subsequently collect only tweets originated from this area. We collect tweets from the following five cities across the globe: San Francisco (SF), Los Angeles (LA), New York City (NY), London, England, and Sydney, Australia. These metropolitan cities were chosen to allow comparison between similar geographical regions (LA and SF), distant cities in the same country (SF and NY), and cities in different continents. It is often said that different cities have different personalities and we hoped to quantify this difference with our data.
For the GlobalTweets dataset, we collect tweets from the five cities between the dates of July 24, 2014 and August 11, 2014, resulting over 8 million tweets. We use an additional dataset composed of tweets gathered from within San Francisco between the dates of March 11, 2014 and April 1, 2014. This dataset contains over 3 million Tweets. The data included within a tweet contains rich metadata germane to our project, including geographic coordinates of its origin (if enabled by the user), and the specific time at which a tweet was originally sent.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Timespan</th>
<th>Instances</th>
<th>Geographic Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>GlobalTweets</td>
<td>Twitter API</td>
<td>7/24-31/13</td>
<td>8,003,437</td>
<td>SF, LA, NY, London, Sydney</td>
</tr>
<tr>
<td>SFTweets</td>
<td>Twitter API</td>
<td>3/11-4/14</td>
<td>3,710,528</td>
<td>SF</td>
</tr>
<tr>
<td>SF-311</td>
<td>data.sfgov.org/</td>
<td>3/11-4/14</td>
<td>12,866</td>
<td>SF</td>
</tr>
</tbody>
</table>

Table 1. Summary of the three datasets.

The third dataset comes from the City of San Francisco’s 311 Case Database, which contains data about 311 cases that are reported and serviced every month. San Francisco 311 acts as the customer service department for the city, servicing complaints and concerns such as graffiti removal, pothole repair, overturned trashcans, and similar issues. The most common 311 complaints are illegal dumping, followed by graffiti and sewage backup.

The raw format of the two twitter datasets is in JSON (JavaScript Object Notation). We use the MongoDB database [5] to store these datasets directly in JSON. This avoids the hurdle of developing a relational database schema. Additionally, MongoDB supports sophisticated queries on geographical data (in GEOJson format) to facilitate geocentric queries.

To integrate these heterogeneous data sources, we need to normalize many data fields, especially the timestamps and geographical coordinates, to follow the same convention. For the timestamps, we convert the timestamps in the SF-311 dataset to the unix-time representation as used in the Twitter datasets. As for the geographical coordinates, we convert the (latitude, longitude) format adopted in the SF-311 dataset to GEOJson format, which is readily integratable with the twitter datasets.

B. Sentiment Classification

Following data collection, we apply sentiment analysis techniques to a sampling of tweets from each location. To determine which classification algorithm to use in the subsequent steps, we examine six algorithms for accuracy and speed. Five of the six algorithms are implemented in the WEKA Machine Learning suite[11]. These include Naïve Bayes[12], Multinomial Naïve Bayes [13], Logistic Regression [14], Bagging [15], and Random Forests [16]. The sixth classifier is the Stanford NLP Sentiment Classifier.

A pervasive problem in Sentiment Classification is the acquisition of high quality training data. Hand annotation is often costly and time consuming. Accordingly, methods to circumvent this practice with regard to Twitter data have been developed. For our experiments we implement the ‘Distant Supervision’ approach described in [19]. This approach is based on the hypothesis that Tweets containing emoticons can be accurately classified based on the polarity associated with the emoticon. We label tweets containing positive emoticons such as “:)” as holding a positive sentiment. Conversely, a tweet containing “(“ is labeled as holding a negative sentiment.

One detail of this method we have examined is whether to remove the emoticon from the text of the tweet after labeling. The effects of this detail are examined and the results presented in section IV.

We test each of the five WEKA classifiers in four different configurations. We train each classifier in two ways and test each of the trained classifiers with two methods. First, each classifier is trained with the emoticons present in the training dataset. Two accuracy measurements are considered. The first accuracy is acquired from 10-Fold cross validation and the second is by testing a classifier against a hand-annotated data set obtained from an external source [20]. A total of 500 tweets from the hand-annotated set are used. Second, each classifier is trained on our training dataset with the emoticons removed. Their accuracy is measured with the two methods mentioned above.

The Stanford NLP Sentiment Classifier is only measured against the hand-annotated data set in the same way as the other five. We use the language model included with the standard distribution because the required format of the training data is not amenable to the “Distant Supervision” method of training data acquisition. Therefore, this classifier is not trained with Twitter data. [See the results in Section IV.]

C. Spatio-Temporal Sentiment Analysis

We choose to use the Multinomial Naïve Bayes classifier included with WEKA for our subsequent analysis because of its balance of speed and accuracy. Two levels of geographical sentiment analysis are undertaken, intra-city and inter-city. We have developed Spatio-Temporal Sentiment (SpaSe) Charts for visualizing the sentiment patterns that emerge from the twitter data. Two types of SpaSe charts are subsequently generated to represent these two levels.

For our GlobalTweets dataset, each city is defined by the bounding box that encloses it. For the SFTweets dataset, we use zip code boundaries to differentiate administrative districts. Zip code boundaries encoded in GEOJson format are obtained from [22] and used as query parameters to the database.

The two datasets are each too large to apply the classification to all of the tweets, so we use a process of targeted sampling. We divide the total timespan of the data in each database into three-hour segments. For each of these segments, we query 200 tweets from each geographical district. Our classifier determines the polarity of each tweet. A tweet is then annotated according to its polarity, 1 for positive and -1 for negative.

With sentiment data generated across all districts and across all time segments, time series data is then generated by computing the mean sentiment of each time segment for each
district. Pseudocode for this algorithm is presented in Algorithm 1. From this data we construct a SpaSe Chart by plotting the trend lines of sentiment flow for each of the districts discussed. See Section IV for the results.

**LocalSentiments** (Locations[], startTime, endTime, interval)

//Locations[]: list of target cities or districts
//interval: the temporal granularity, e.g., daily, or hourly
foreach (location in Locations)
    sentiments[] = 0
    current = startTime
    while (current < endTime)
        tweets[] = db.query(tweets form location sent between current and (current + interval) limit 200)
        foreach (tweet in tweets)
            sentiments[current] += classifySentiment(tweet)
        sentiments[current] /= length(tweets)
        current += interval
print location, sentiment, and time to a csv file for visualization

**Algorithm 1**: LocalSentiments computes average sentiment scores for each location in Locations and outputs in a time series format for analysis.

**D. Sentiment and 311 Analysis**

Charting the sentiment flow of a district provides a unique reference for further analysis. We are interested in exploring what relationships, if any, exist between 311 reports within a given time and location and the average sentiment. We apply the same spatio-temporal analysis described above to the 311 data, but instead of average sentiment, we use the frequency of reports for each time segment normalized to a range between 0 and 1 by dividing each value by the maximum frequency. The trend lines of this data are plotted to see the daily fluctuation in 311 reports. Plotting this 311 time series data on top of the previously derived SpaSe Charts allows us to visualize the relationship in an integrated 311-SpaSe chart.

To dig deeper into the relationships between regional sentiment and 311 reports, we postulate that the administrative districts based on zip codes may not be representative of the actual local communities because zip code boundaries may be artificially drawn and historical in nature. To discover new data driven geographical boundaries, we search for clusters of 311 reports by applying a clustering algorithm to the geographical coordinates of the 311 reports. We use the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [17] provided by WEKA to identify geographical clusters in the 311 dataset and the twitter dataset. The data is clustered at three different temporal ranges: daily, weekly, and over the entire temporal range of each set. By comparing the clustering solutions at the different ranges we identify several relatively stable clusters in the 311 dataset.

“Bounding Box” coordinates are then derived from each of the stable clusters. These coordinates are now used as the boundaries for the new data-driven districts. SpaSe Charts could now be generated for newly identified geographical regions.

**IV. RESULTS**

**A. Sentiment Classifier Performance Evaluation**

Sentiment Classification is central to the generation of SpaSe charts and related research, so understanding the performance of our Classification tools was important. We test two variations of the “Distant Supervision” method of gathering training data and two methods of evaluation for each of the classifiers included with the WEKA package. See Table 2 for a summary of the comparison.

“Distant Supervision” is extremely cost effective regarding training data generation, however the accuracy of classifiers trained with this method needs to be examined with care. The results show that 10-fold Cross Validation produces significantly inflated accuracy measurements when compared to measurements using an outside test set, in some cases, upwards of 30% higher. We believe the act of querying the database for tweets containing emoticons introduces a strong bias in the resulting dataset. Additionally, classifiers tend to show slightly better performance when the emoticons are removed from the training set before the training set is used to train the classifier. This holds true for four out of the five classifiers that we tested. The exception is the Random Forests classifier.

**B. Inter-City Sentiment**

At this juncture we have sentiment data across all selected cities for the entire temporal range of July 24th 2014 to August 11, 2014. The first level of analysis we perform is to try to characterize the overall sentiment characteristics of each city.

**Happiness Index**: To the extent that twitter users are representative of the population of a region as a whole, the sentiment expressed by these users can be used to make an approximation of the overall happiness of the region. Happiness Index is scored on a scale from -1 to 1. A score of 1 indicates the highest possible sentiment where all tweets are uniformly positive, -1 the lowest.
Comparing the Happiness Index of diverse regions lends to some interesting comparisons. As shown in Figure 1, San Francisco and Los Angeles, two cities in California separated by a little over 350 miles, have similar Happiness Index ratings, notably the lowest two on the list. Los Angeles holds the lowest with -0.062 and San Francisco is close behind with -0.061, a Happiness Differential of 0.001.

San Francisco and New York City are separated by nearly 3,000 miles. NYC’s Happiness Index is rated at -0.013, the Happiness Differential between SF and NYC is 0.049. Of the three cities in the United States that we sampled, none of them crossed the Neutral Sentiment Threshold, so all of them showed a net-negative overall sentiment. This is in contrast to the two cities we sampled that are not located in the US.

The mean sentiment expressed by twitter users in London, England was measured at 0.046. The highest Happiness Index that we measured came from Sydney, Australia, with an index of .141. This is nearly .1 above the second highest, London.

These results allow for the formation of a “Relative Happiness Ranking.” By pointing to this data, one can say that Sydney, Australia is happier than New York City.

One potential limitation of our sentiment classification model that may affect the happiness indexes and resulting rankings is that the adopted model does not consider the potential linguistic differences across different geographic regions. For example, it is well known that specific words and idioms bear distinct semantics in American English, British English and Australian usage. This is certainly an interesting topic warranting further research.

The second level of analysis is performed by plotting the time series sentiment data to reveal the trend lines of the sentiment flow over time. This allows us to visualize how the sentiments shift at each location throughout the day as well as throughout the week.

Sentiment trend anomalies like this may point towards sentiment effecting social occurrences. The cause of local sentiment trend anomalies may be an avenue for further research.

C. Intra-City Sentiment

We next narrow the scope of this kind of analysis by limiting the geographic range of the data to one city. This allows us to drill down and understand how the local
sentiment is distributed on a neighborhood or administrative
district level.

Finding the mean value of the sentiment in each
neighborhood over the full temporal range reveals a stark
difference in the overall sentiment character of the different
neighborhoods in San Francisco (Figure 4).

With this data, we also generated SpaSe charts to compare
the sentiment trends between districts. See Figure 5.

The local context that might explain the difference in
sentiment between the different administrative districts is not
strictly accessible at our level of analysis at this point, since
we do not consider that semantic aspects of each tweet to
derive its sentiment. The only contextual information that we
incorporate in our analysis is the local frequency of 311
reports. It is possible that a deeper integration of the textual
information of tweets may point towards explanations of the
difference. It is also likely that sociological and economic
factors influence local sentiment. All of these are grounds for
further study and exploration.

Additionally, tracking tweet sentiment of individual users as
their location changes throughout the day may prove
insightful. We could examine whether residents of a
neighborhood tend to show higher sentiment, or if people who
happen to be in the neighborhood when they send a twitter
message.

D. Integrated Intra-City SpaSe-311 Charts

To examine relationships between local sentiment flow and
other indicators of the character of the local environment, we
integrated the local 311 reporting frequency with the local
sentiment flow data. The resulting chart (Figure 6) suggests
that there is a relationship between the two datasets.
Specifically, we observe that peaks in 311 frequency often co-
occur with spikes in sentiment and oftentimes 311 frequency
spikes precede sentiment spikes slightly. This observation
anticipates a trend presented in the next section, where areas
of high 311 frequency are targeted for analysis.

Given that the overall mean sentiment of San Francisco
was measured to be below 0, it is no surprise that the majority
of the districts showed a net-negative Happiness Index. In
fact, only one district showed a net positive Happiness Index,
this is the “Embarcadero, Barbary Coast” neighborhood.

The Mean Happiness Index data, as with the GlobalTweets
dataset, allows for a Relative Happiness Ranking of the
different administrative districts.

E. Data Driven Community Identification and Analysis

The contrived geographical boundaries as delineated by zip
codes have little relationship with the communities that grow
organically within a city. As described in the Methods section, in order to identify new communities, we clustered the 311 reports based on location. Examining these clusters over different temporal ranges revealed several stable clusters indicating areas that show a high density of 311 reports. See Figure 7. Because 311 issues are reported by civilian witnesses to the issue, these clusters serve at least to identify communities of highly active 311 observer-reporters. A likely correlation here is also that the geographic area related to the clusters are areas of higher general activity (economic and social) and are likely to have a higher population density.

Figure 7: Comparison of Administrative District mean sentiment and Data Driven district sentiment

We have identified both a positive temporal relationship between 311 report frequency and sentiment as well as a spatial relationship. Etiological research regarding this correlation is beyond the scope of this paper, but several conjectures seem plausible. For instance, sentiment may have a positive correlation with the attention that a community receives by civic entities. If a community feels like the city’s 311 service is available to attend to the local issues, then a higher Happiness Index is observed. An alternative, but related conjecture is that communities who tend to be vocal about their needs, i.e., reporting local issues with a high frequency, show a higher Happiness Index.

Figure 8: Density Based Clustering of San Francisco 311 Reports

V. CONCLUSION AND DISCUSSION

The novel approach that we develop for analyzing the diverse public datasets opens up new opportunities for understanding the interactions hidden therein. With the combination of machine learning supported sentiment analysis, data driven community identification and integration of data from both Twitter and government 311 records, we have identified a positive relationship between local sentiment and government responsiveness.

Additionally, the system and method we employ for the relative ranking of local happiness has the potential to help identify broader civic issues and trends that can be acted upon by local governments.

The performance evaluations that we provide for open-source machine learning algorithms provide a performance baseline for practitioners who wish to pursue this line of research further, or any line of research that relies on off the shelf machine learning techniques.

Our methods do leave room for improvement. The Stanford NLP Sentiment classifier’s accuracy could potentially be improved upon by retraining it on our “Distant Supervision” Twitter data. This would be costly due to the complex structure enforced upon its training data. But the expended effort could yield greatly improved accuracy results for the classifier.

Another area of complimentary research is the study of regional language usage and its effect on sentiment classification algorithms.

There is abundant opportunity for future research in the areas explored in this paper. This paper did not explore the rich contents of the both the tweets and 311 databases. Analysis of the additional data available from twitter such as user tracking information and the textual content of the tweets themselves could provide valuable insight. 311 databases include information on the type of issue being reported (e.g. graffiti, or overturned trashcans). Including this information could provide new perspectives and nuanced understanding of the relationships in the data.

Another avenue of research is the integration of additional public datasets beyond 311 data. Possible datasets include crime statistics, school and business records.
REFERENCES


[18] https://data.sfgov.org/


[22] City and County of San Francisco—311 Case Database: http://www.sf311.org/

[23] https://twitter.com/


[25] Anjaria, M.; Guddeti, R.M.R. Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on


[27] Nan Cao ; Yu-Ru Lin ; Xiaohua Sun ; Lazer, D. ; Shixia Liu ; Huamin Qu Visualization and Computer Graphics, IEEE Transactions on