Spatial-Temporal Characterization of an Urban Environment Using Twitter

Víctor R. Martínez

Department of Computer Science
Instituto Tecnológico Autónomo de México (ITAM)
Río Hondo #1, Mexico City, 01080, Mexico
{victor.martinez, victor.gonzalez}@itam.mx

Víctor M. González

Abstract— This paper presents a study of the usage of Twitter within the context of urban activity. We retrieved a set of tweets submitted by users located in Mexico City. Tweets were labeled as either positive or negative mood using a sentiment analyzer implementation. By calculating the average mood, we were able to run a Mann-Whitney’s U test to evaluate differences in the calculated mood per day of week. We found that all days of the week had significantly different medians with Sunday being the most positive day and Thursday the most negative. Additionally, we study the location for the tweets as an indicator important events and landmarks around the city.

Keywords— Visualization of Collaborative Processes & Applications

I. INTRODUCTION

Twitter, a popular micro-blogging platform, allows users to broadcast brief text updates to a public or selected group of contacts, using their computers or mobile phones [1,2]. Users in Twitter answer the question “What are you doing?” by using fewer than 140 characters. A status update message, called a tweet, is often used as a message to friends and colleagues [2]. A user can follow other users and read those tweets from users they are following. The link between users is not reciprocal: a user can follow another user without being followed back [1, 2]. After its launch on July 2006, Twitter users have increased rapidly. As of 2013, the number of monthly active users is estimated around 241 million worldwide and 500 million tweets are sent per day [3]. Just in Mexico, Twitter has become the 2nd most important social network having 10.7 million active users [4] and 55 million tweets are sent per day [5]. Almost a third of Mexican tweets come from a smartphone, which roughly translates to 18.3 million tweets a day from those type of devices [5].

Several works have focused on understanding Twitter from a multitude of perspectives. In 2007, Java et al. [6] found out the main uses for Twitter: daily chatter, posts dedicated to daily routine; conversations, replies to another user using the @ symbol; sharing information and URLs, posts containing some URL in them, and reporting news, which include weather reports and news stories. B. Huberman et al. [7] analyzed the interactions between users and their friends (defined as a message whom a user has sent using the “@” symbol) finding that the friendship relationship is far better at describing interactions than the follow/followee model. Other recent works have shown that it is possible to analyze the public available broadcasts to infer population attitudes in a similar manner as those obtained by pollsters [8, 9]. By understanding the sentiment portrayed in each message, Bollen et al. [8] were able to predict outcomes in the stock market. O’Connor et al. [9] connected measures of public opinion with sentiment measured from text by analyzing surveys on consumer confidence and political opinion over the 2008 and 2009 period.

Additionally, other works focus on the location of tweets (the coordinates where the message was submitted) as a spatial indicator for event detection and information inference. Sakaki et al. [2] proposed a particle filtering algorithm to detect earthquake centers in Japan by using tweets collected during the event. Pontes et al. [10] studied simple methods to infer the user home location using a collection of tweets. Hecht et al. [11] performed an in-depth study of user behavior with regard of the user location field in Twitter where they found that although a huge percentage of users did not specify their location beyond a city level scale, it was possible to deduce their state and city using machine learning techniques.

A preliminary version of this work was presented at URBAI 2013 [12], were we retrieved a set of tweets submitted in Mexico City, labeled each tweet as either positive or negative mood using our own sentiment analyzer implementation, and calculated an average mood for the tweets submitted in a certain day and time, yielding one time series for each day of the week. In this work, we focus our attention to finding important locations where Mexicans are more inclined to tweet. In doing so, we would like to discover if there are any relations between the collective mood and the location of the user at the time of submission.

The paper is organized as follows: in Section II, we describe the dataset, the sentiment classification tool, and a sketch of the visualization designed. Section III covers a quick summary of the results presented in the preliminary work with
the addition of the spatial analysis. Finally, we summarize our findings and conclude with Section IV.

II. DATA AND METHODOLOGY

A. Sentiment classifier

We trained a sentiment labeler using an automatically labeled dataset [13] of 118,092 positive and 77,265 negative examples. The model used was the Naive Bayes Classifier implementation from Scikit-learn [14] over unigrams and bigrams. We normalized each tweet to lowercase and replaced URLs and user mentions (@USER) with a unique token for each. We also removed additional whitespaces and substituted hash-tags (#WORD) with the word after the hash. All punctuation was removed, which also included common smilies (";)", “:"(”, “:P”, etc...). We did not delete stop-words nor used any stemming. The final model learnt 145,520 features with over 85% precision.

B. Captured dataset

We obtained a collection of public tweets recorded from June 15th to July 19th, 2013 (1,226,981 tweets from 1,805 distinct locations around Mexico City). For each tweet, we recorded a unique identifier, date-time of submission (GMT – 6), coordinates of submission and place identifier, alongside with the text of the tweet.

C. Average Mood

Using the previously mentioned sentiment analyzer, we assigned each tweet to either a positive or negative mood. We calculated an average mood score \( m_{d,t} \) for tweets submitted in day \( d \) and 5-minute interval \( t \), as described by the following equation

\[
m_{d,t} = \frac{\sum_{w \in T_{d,t}} \mathbf{1}(w)}{|T_{d,t}|}
\]

(1)

where \( T_{d,t} \) denotes the set of all tweets captured in day \( d \) and in time \([t - 5, t]\). The numerator is the sum of characteristic functions for positive tweets; the denominator is the cardinality of \( T_{d,t} \) (i.e., the number of tweets captured at that day and time). Therefore, each \( m_{d,t} \) represents a percentage of positive tweets with respect of the total tweets captured at the time.

Grouping by day of week yields seven 24-hour time-series with values ranging from 0 (no positive tweets at the time) to 1 (no negative tweets at the time).

D. Data visualization

We developed an application to visualize the moods of captured tweets using Processing 1.5. First, it displays a map of Mexico City using the Open Street Map API along with the positions of the captured tweets (see Figure 1). The application collects tweets submitted in a certain timeframe, and evaluates the average mood for each square in a grid. A color is assigned according to the mood of each zone with blue being the most positive and red being the least positive. Zones without tweets at the time appear empty. A text label at the bottom of the window lets the user know the day and time portrayed at the moment. Buttons at the top of the screen allows the user to speed-up / slow-down or pause/resume the simulation or increase the minute interval (which for default is set to 20 minutes).

III. RESULTS

We now present our results in two sections: mood analysis and spatial analysis.

A. Mood analysis

A typical day in Mexico City starts around 6 or 7 am with the beginning of labors. At that time, the average mood for weekdays is around the 60% positive mark. From around 7 am to 11 am the mood quickly declines to below 50%. By lunch hour (1 pm – 2pm) reaches its peak with 90% positive mood.

The authors would like to thank Consejo Nacional de Ciencia y Tecnologia (CONACYT) – Proyecto Pulso de la Ciudad - and Asociación Mexicana de Cultura A.C. for all their support.
During the post-meal hours and throughout the night, the mood stabilizes in a steady 70%.

Figure 2 shows the aggregated mood score. X-axis shows hours and Y-axis the average mood calculated by equation (1).

Figure 2 shows the aggregated mood for each day of the week. In a similar manner as that of the heat-maps, we can color our mood progression for each day of the week in order to better visualize the changes in the collective feeling. In Figure 3 we show the results using columns to represent days and a row for each 5-minute interval starting from [00:00, 00:05).

It can be noticed that even when all days tend to follow a similar pattern, weekdays (with the exception of Friday) are more likely to end in a neutral mood whereas Fridays, Saturdays and Sundays end in a more positive tone.

We ran a non-parametric Mann-Whitney test to find if any day of the week tends to have larger values (i.e., a more positive mood) than others. We found out that Friday’s mood tends to be above Wednesday ($W = 44816.5$, $p < 0.05$) and above Thursday ($W = 45986$, $p < 0.05$). Saturday’s mood was above Monday’s ($W = 47358$, $p < 0.01$), Tuesday’s ($W = 48796$, $p < 0.001$), Wednesday ($W = 49544$, $p < 0.001$), Thursday ($W = 50938.5$, $p < 0.001$), and Friday’s ($W = 46058$, $p < 0.05$). Finally Sunday’s mood was above Monday’s ($W = 49586.5$, $p < 0.001$), Tuesday’s ($W = 51391.5$, $p < 0.001$), Wednesday’s ($W = 51790$, $p < 0.001$), Thursday’s ($W = 53289.5$, $p < 0.001$), Friday’s ($W = 48950$, $p < 0.001$) and Saturday’s ($W = 45000$, $p < 0.05$). Hence the worst day for inhabitants in Mexico City is Thursday ($\mu = 0.7050$, $\sigma = 0.06120$) while the best is Sunday ($\mu = 0.7279$, $\sigma = 0.04816$).

B. Spatial Analysis

From the 16 available boroughs we were only able to register 13 (81.25%) in our data. The three missing compose the east and southwest corners of Mexico City: Cuajimalpa de Morelos and La Magdalena Contreras to the east and Milpa Alta to the southwest. We theorize this data is missing due to an error in the calculation of the coordinates submitted to the Streaming API.

Most of the tweets in a given day were originated from Cuauhtemoc borough, home to the Mexican Stock Market, the tourist attractions of the historic center and Zona Rosa, skyscrapers such as Torre Mayor (the tallest building in the city) and Torre Latinoamericana (one of the city’s most important landmarks). It also contains numerous museums, libraries, government offices, markets and other commercial centers which bring as many as 5 million people each day to work, shop or visit [15].

The second place were most tweets were submitted from is the Coyoacan borough, a suburban area south of Cuauhtemoc. This borough shows a slight increase in activity on Sundays due to the touristic nature of the borough. Finally the third place where most tweets originated from was the Benito Juarez borough, a densely populated commercial area south of the historic center.

Our final approach to the data was to create a square grid for Mexico City. We used squares of 1.002 Km by 1.002 Km resulting in 1548 separations ranging from (19.2° N, 98.98° W) to (19.58° N, 99.3° W). We collected the number
of tweets submitted in each square of the grid regardless of their mood.

This resulted in a bubble map shown in figure 4 where each bubble size and color is in relation with the number of tweets submitted.

We identified each of the bubbles according to their location and time of our study. For example, the point around (19.3 N°, 99.15 W°) in figure 4 corresponds to the Estadio Azteca, Mexico City’s biggest sport stadium and where, during our time frame of study, the National Soccer Team disputed most of their World Cup’s classification games. Just north of the Estadio Azteca is the Coyoacan borough center, which holds colonial-era estates hidden, several interesting churches, museums and artisans’ markets. We were also able to identify Bosque de Chapultepec—one of the largest city parks in among other attractions the Chapultepec Castle, the National Museum of Anthropology, and the Rufino Tamayo Museum— as the point in the middle of the city around coordinates 19.4 N°, 99.15 W°. The second to northern concentration point corresponds to Arena Ciudad de Mexico, an indoor arena used to host events such as concerts, sports events and more. Finally we could not find any meaningful association for the northmost point, which lies just outside the boundaries of Mexico City but we think this point corresponds to all those citizens tweeting just before leaving or arriving to the city.

We did not find any meaningful differences between the moods and the locations of the tweets. That is, we could not detect any place where the submitted tweets were considerably more negative or positive. This could be due the size of our grid, inaccuracies in our sentiment analyzer or simply because such relation does not exists at all.

IV. DISCUSSION

Our study and work-in-progress aim to use public available social-media broadcasts, which are commonly found in Twitter, as a spatial indicator for the population mood in an urban environment. Previous works have shown that it is possible to use Twitter to get similar data as that obtained by public opinion polls. By analyzing collective mood (measured as a relation between positive and negative tweets in a certain time) we were able to study the progression in which mood changes in a certain day. We found meaningful differences between the weekday’s average mood and the weekends’. With our current data, we were able to identify locations were Mexicans are more inclined to tweet (such as, stadiums and touristic attractions, etc.). In a future work, we would like to develop a real-time city sensor which would allow for decision-takers to visualize the state of the city in any given time in order for them to take a more informed decision towards a public policy without the expenses associated with polling.

ACKNOWLEDGMENT

We would like to thank everyone involved during the course of this study, Consejo Nacional de Ciencia y Tecnología (CONACYT), and to the Asociación Mexicana de Cultura A.C. for all their support.
REFERENCES


