**¡Chévere! Forecasting Tweet Location with Big Data**

Laura H. Kahn

School of Informatics and Computing

Indiana University

Bloomington, IN, USA

E-mail:lkahn@indiana.edu

***Abstract***

Social media data from countries with limits to free speech is a reliable, alternative form of journalism. Over 37,000 Spanish Tweets from the city of Caracas, Venezuela were used to observe reactions to the food shortage crisis within each of the city’s five municipalities. The number of Tweets from December 2014 - October 2016 is compared to trending Tweets from the week of July 19, 2017. Machine learning techniques show that certain Tweets are particular to a municipality location.

***Keywords****-* ***Social media mining; location-based estimation; open source indicators; supervised machine learning; crowdsourced journalism; event forecasting; food security***

# INTRODUCTION

Popular microblogging services such as Twitter have yielded bigger volumes of location and time-referenced data. Analysis of microblogs is interesting for a number of applications, including understanding people’s reactions in an event. Social media mining has Twitter users that act as social sensors to form a collective, crowdsourced reporter. This data can be obtained to supplement traditional media.  
  
Food shortages in Venezuela have been prevalent since 2013 [2] and reached a crisis-level nationwide shortly thereafter. In May 2016, the newspaper El Nacional reported scarcity of basic food necessities levels above 80%+. The government-controlled media stopped reporting crisis details in January 2015 [5]. Even though the food shortages began in 2013, the temporal subset for this project begins in December 2014 since this was prior to the end of official news media reporting crisis details.   
  
Social media mining (SMM) can be used to give insight into how citizens express themselves*.* Tweeting patterns were shown with hashtags and other related terms within each of Caracas, Venezuela’s five municipalities and all of Caracas from December 2014 - October 2016 (670 days) and were compared to more recent reactions trending during the week of July 19, 2017 (7 days).

Twitter was chosen because according to EModeration.com,   
Venezuela led Latin America in the number of active Twitter users compared to internet users at 14% (4 million). Using Twitter data is a reasonable approximation to describe the food shortage event.

Tweets were limited to geocoded Tweets from the city of Caracas in Spanish in accordance with Tobler’s first law of geography [17], which states that *near things are more related than distant things.* Although Twitter allows a user to declare their location, such metadata is unstructured and ad hoc.

Supervised machine learning models were used to see whether patterns could be detected in the Tweet corpus and within each Caracas municipality.

# RELATED WORK

Korkman uses Twitter with other heterogeneous data sources from 3,072 Venezuela events and a logistic regression model to predict where civil unrest activities will occur in a city [9]. He uses Twitter and machine learning to predict neighborhood hotspots for protests in St. Louis [8]. Han discusses the likelihood of geotagging Tweets by city [7]. Zhao describes a model where 75-91% of Venezuelan Tweets are trustworthy content [20]. Cranshaw studies the social dynamics of neighborhoods in a city using geo-tagged Tweets [4]. Kounadi uses Twitter to analyze how the public responds to crime events and finds that over half of the Tweet authors live within 10 km of the event [10].  Priedhorsky uses 30,000 Tweets to infer user location with a 67.5% accuracy [15]. Kahn examines the use of different machine learning classifiers for Caracas, Venezuela from December 2014 – October 2016 [9].

Peca defines an event ‘as any physical or abstract discrete object having a particular position in space and time’ [15]. Examples of events include geo-referenced photos, occurrence of earthquakes, disease cases, or in this case, the food shortage crisis in Caracas, Venezuela.

Caracas was selected since prior research shows that more densely populated, urban areas have a higher volume of Tweets [3].

Tweets retrieved without hashtags have a much higher signal-to-noise ratio than tweets with hashtags [3]. Tweets are only an approximation of events since prior research shows that the number of Tweets can be up to four times higher than the number of protestors [3]. There is no indication if the severity of an event affects the total number of Tweets in a city or municipality. However, based on previous research [1, 8, 13], it is appropriate to use only geocoded Tweets and only Tweets originating in Caracas.

# DATA PREPARATION

## *Ethics Statement*

The data analyzed are publically available as they come from a public online social media site (Twitter). There is no private data in the final dataset. No Tweet information was retained that would identify a user’s personal information.

## *Extraction*

The data used in my research comes from a corpus of Spanish Tweets collected from December 2014 to October 2016 and compares it to trending Tweets the week of July 19, 2017. Only geotagged tweets using Twitter’s Streaming API (<https://dev.twitter.com>) were collected to avoid API rate limits. Each tweet is up to 140 characters of text and is associated with a user id, timestamp, latitude and longitude. All other fields obtained from Twitter, including the user’s twitter handle are disregarded. Even though it is not included in this analysis, all of the URLs and pictures were extracted.

## *Geocoding*

Latitude and longitude of each municipality (Baruta, Chaco, El Hatillo, Libertador and Sucre) in Caracas, Venezuela were determined to four decimal places using Place Beam and Mapa and were included in the search criteria.



Only Tweets with geocoded tags of latitude and longitude were mined. The radius was set to 10 km from the center of Caracas, 6 km from Baruta, 4 km from Chacao, 6 km from El Hatillo, 6 km from Libertador municipality and 6 km from Sucre. Although prior research has the smallest radius set to 10 km [10], the radius ranges were chosen because the entire size of Caracas is 4.714 km2. Although other research uses coordinates to six decimal points, four decimal points from the latitude and longitude geocodes were used in the data extraction process [3]. During the first collection period, over 37,000 geotagged tweets were retrieved.

## *Features*

The following features were included in the initial data set: Tweet ID, text, date, retweet count and favorite count. There were 37,216 Tweets. Only Tweets with either 5 or more ‘Retweets’ or 5 or more ‘Favorites’ were included in the filtered data set. Sentiment analysis was presumed to not add any value to the analysis since almost all Tweets about a food crisis event would be negative. There were 2,833 filtered Tweets.

Labels 1 - 6 were manually assigned to each of the municipalities in Caracas (1 = Baruta, 2 = Caracas, all, 3 = Chacao, 4 = El Hatillo, 5 = Libertador, 6 = Sucre).

Table 1 shows the approximate number of Internet users in the entire country of Venezuela, the number of Twitter Users and the population of Caracas. The number of Tweets about the food crisis during the sample period is about 35% of all Tweets from Caracas implying the significance of the event to citizens living in the city.

TABLE 1. Social Media Use in Venezulea

|  |  |
| --- | --- |
| Venezuela Internet Users+ | 17.5 million |
| World-Wide Tweets / Day | 58 million |
| Venzuela Twitter Users | 4 million |
| Caracas Population\* | 5.5 million |
| Average Caracas Tweets  (670 day \* 1969 tweets/day) | 1.32 million |
| Food Crisis Tweets Total | 37,216 |

+ [http://data.worldbank.org](http://data.worldbank.org/indicator/IT.NET.USER.P2) (Accessed 20 November 2016)

\* <http://data.un.org>  (Accessed 24 November 2016)

Trending patterns for the city of Caracas and all of Venezuela were extracted using Python for the week of July 19, 2017. The name of the trend and the total number of users Tweeting about that particular trend was obtained.

For analysis purposes, the user's’ Twitter location is also considered to be either their home location or another location near where the event is occurring. The users’ Twitter location and corresponding Tweets can be made from a desktop or mobile phone, therefore introducing an element of true location uncertainty during the analysis.

## *Phrase Filtering*

Since the project uses text-centered data mining and no subject matter expert was available, search terms were developed using phrase filtering [15]. According to Wikipedia, the most read uncensored traditional media sources in Venezuela are: La Patilla, Últimas Noticias, El Nacional, El Mundo, CNNEspañol and El Tiempo [20]. A manual search of these sources using food scarcity terms yielded 38 initial search terms. Most of these terms were filtered out because they would have had a high signal-to-noise ratio (e.g. ‘pan’ (bread), ‘leche’ (milk), ‘guerra económica’ (economic war), etc.). Terms were also filtered out because they would have duplicated results from the final phrases. For example, searching ‘inseguridad de alimentos’ (food insecurity) would have duplicated results from searching ‘comida’ (food).

Final search terms included three adjectives and nouns [12]. The top three most popular hashtags reported by traditional media outlets were also chosen for comparison.

Final phrases included ‘[#AnaquelesVaciosEnVenezuela](http://www.maduradas.com/lo-ultimo-tuiteros-posicionan-hashtag-sobre-anaqueles-vacios-en-venezuela-a-nivel-mundial/)’ ( #EmptyShelvesinVenezuela), ‘[#NosCayóLaDietadeMaduro](http://periodicoellibertario.blogspot.com/2016/07/el-hashtag-noscayolabrian-representa-al.html)’ (#WeFellonMaduro’sDiet) , ‘[#VzlaTieneHambre](http://primeraemision.com/el-hashtag-venezuela-tiene-hambre-se-posiciono-en-twitter-como-protesta-de-los-usuarios)’ (#VenezuelaisHungry), ‘Escasez’ (scarcity - noun), ‘Hambre’ (hungry - adjective), and ‘Alimentos’ (foods - noun). After running a query for ‘#[NosCayóLaDietadeMaduro](http://periodicoellibertario.blogspot.com/2016/07/el-hashtag-noscayolabrian-representa-al.html)’ in two different municipalities with 0 results, this term was removed from the analysis.

# DATA ASSUMPTIONS

Understanding why people are Tweeting is beyond the scope of this type of big data analysis. Also by using social media data, interviewer effects of bias used for traditional journalism that could alter the data was avoided [6].

Because about 14% of the population in Venezuela uses Twitter, the SMM methodology has limited but powerful analytic power. By its nature, Twitter excludes certain demographics without Internet access or that are not using Twitter [19]. The resulting dataset is a self-selected population which creates confounding variables.

Also, a hashtag (#) is often loaded with assumptions, meanings and cultural or political structure.

It is also important to note that only even though only 0.833% of all Tweets are tagged with geographical coordinates [8], they are an appropriate representative sample [13].

Tweets retrieved without the use of hashtags have a higher signal-to-noise ratio than Tweets archived with hashtags [3].

A user’s reported location in their Twitter profile was assumed to be their actual location. There could be variation in reported versus actual user location.

The Tweet text is an approximation of the local event in the same way that the machine learning classifiers are the best approximation of whether a Tweet is from a particular municipality.

Looking at this subset of data presents a snapshot of Tweeting patterns but does not disclose the entire socio, economic or political story of the food shortage crisis in Caracas or Venezuela.

The analysis reported in this paper was performed from Tweets originating in Caracas, Venezuela. All Twitter messages were sampled to perform the Support Vector Machine, Logistic Regression, Naïve Bayes and k-Nearest Neighbor algorithms with the municipality used as the feature label.

# METHODS

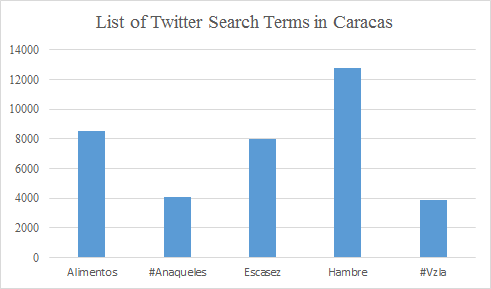
The analyses reported in this paper were performed using Python. Natural language processing (NLP) techniques such as stemming, tokenizing and punctuation were performed on the Tweet text corpus. The Snowball Stemmer Python package was used to remove stop words (“de”/of, “que”/that, “y”/and, etc).

A Spanish corpus CESS-ESP was used to train the data [17]. The corpus was automatically annotated using the SKLEARN Python library.

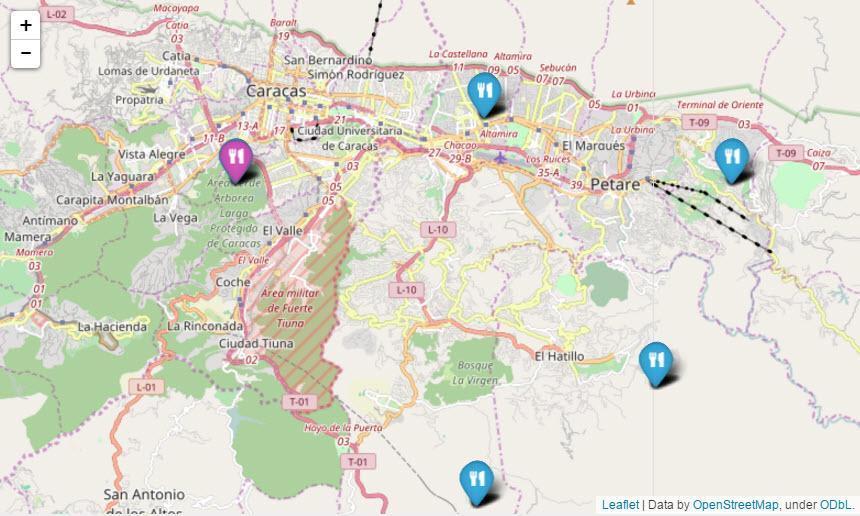
Machine learning models predict the output of new data based on past data and detect any patterns. Selecting relevant features and choosing how to encode them helps to extract a good model. Support vector, Logistic Regression, Naïve Bayes and k-nearest neighbor algorithms were chosen for this analysis to predict whether certain words could identify a Caracas municipality. A cross validation of 10 was used for all three classifiers.

# DISCUSSION

The search terms (‘Alimentos’, ‘Escasez’, and ‘Hambre’) were two to three times more popular than the hashtags cited in the traditional media outlets for each municipality. The word ‘Hambre’ (hunger) was the most popular in all municipalities except the Libertador municipality. In the Libertador municipality, the phrases ‘Alimentos’ (foods), ‘#AnaquelesVaciosEnVenezuela’ (#EmptyShelvesinVenezuela), and ‘#VzlaTieneHambre’ (#VnzlaisHungry) were as popular as the word ‘Hambre’ (hunger).



Each of the most popular terms was represented visually using the Folium Python library. The interactive map allows you to click on the leaflet-style marker to see the most popular term for that municipality. A fork and knife icon was added on the visualization to each marker.

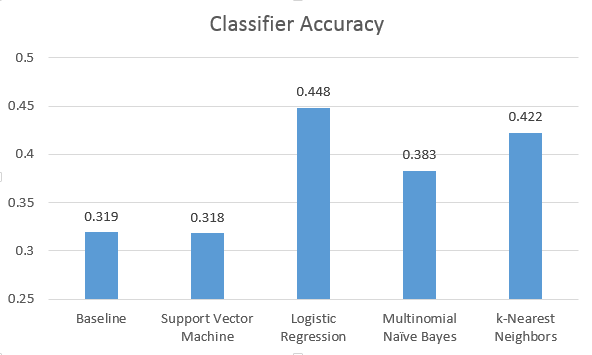




In addition to looking at reactions of the food crisis events from December 2014 – October 2016, the top ‘trending’ searches for the city of Caracas and the entire country of Venezuela were identified for each day during the week of July 19, 2017. An item is defined as ‘trending’ if Twitter users are mentioning the phrase or hashtag most frequently at one point in time.

The top trending search term for Caracas and all of Venezuela during this week was ‘#TrumpVenezuelaSeRespeta’. This hashtag, like many others, is loaded with political implication(s) and may refer to the U.S. President’s statements preceeding Venezuela’s July 30, 2017 Constituent Assembly vote. This hashtag may also have implications about the Venezuelan government, which would include its response to the food crisis. The reason the phrase is trending is unknown and is beyond the scope of this big data analysis.

There are a total of 2,820 labeled data points and the largest category (Chacao) has a count of 900. Therefore a majority baseline classifier would get an accuracy of 900/2820, or 0.319. The Support Vector Machine (SVM) accuracy is 0.318, the logistic regression (LR) accuracy is 0.448, the Multinomial Naïve Bayes (MNB) accuracy is 0.383, and the k-NN accuracy is 0.422. The SVM underperforms the baseline, the MNB accuracy outperforms the baseline by 16.7%, and the k-NN outperforms it by 24.4%. Of the three types of machine learning classifiers used, logistic regression was the most accurate with an accuracy of 0.448.





Since all but the SVM model perform better than the baseline classifier, a reasonable conclusion is that words used in a Tweet in Caracas are signals of their municipality location. This is a significant finding that Tweet texts can be discovered at the micro / municipality level. Machine learning classifiers can also be used with a certain confidence to predict where Tweets might occur.

# CONCLUSION

SMM can be used to help provide insight into how citizens express themselves in space and time*.* Social scientists, researchers, policy makers, governments and international aid relief organizations could use this knowledge to make data-driven decisions and/or analysis of the food shortage crisis at a city-wide or municipality level.

Tweeting patterns were detected using Twitter hashtags and other related Spanish terms within each of Caracas, Venezuela’s five municipalities and the entire city from December 2014 - October 2016 and July 2017 from over 37,000 Spanish language Tweets.

The hypothesis whether certain words are particular to a municipality was tested using support vector, naïve Bayes and k-nearest neighbor machine learning classifiers. Words used in a Tweet in Caracas are signals of their municipality location. Overall reactions to the crisis in changed over time with the most Tweets occurring right after the start of the analysis period.

The research is significant because it is the first known data mining analysis of the Venezuelan food crisis using machine learning techniques. The research presents a new depth of detail about this event by looking at the texts from a micro/municipality level. The research is also significant since it presents a new feature identifier to predict the municipality where a Tweet text originates.

There is not a clear explanation of why certain Tweet texts are more popular in some municipalities than others or why there was a spike in the number of tweets in January 2015.

There are several opportunities to build upon this research in the future, including an analysis of the pictures extracted from this Twitter corpus.

Future research could also be done on the adoption of hashtags during an event versus adoption of key search terms in South America, Venezuela and Caracas to see if the hashtags differ than how they are adopted in the rest of the world.

# ACKNOWLEDGMENTS

My thanks to Vincent Malic of Indiana University for advice and assistance.

# REFERENCES

[1] Andrienko, G., Andrienko, N., Bosch, B., Ertl, T., Fuchs, G., Jankowski, P., Thom, D. (2013): Discovering Thematic Patterns in Geo-Referenced Tweets through Space-Time Visual Analytics. In: Computers in Science and Engineering, v.15(3), pp.72-82.

[2] Barnola, T., R. Corredor, M. López de Blanco, C. Allison and A. Aristiguieta. Análisis de la Crisis de Alimentos en Venezuela. Informe de la Comisión, 24 de augosto 2010.

[3] Bastos, M.T., Recuero, R., Zago, G.Taking Tweets to the Streets: A Spatial Analysis of the Vinegar Protests in Brazil.*First Monday*, 2014.  
  
[4] Cranshaw, J, Schwartz, R, Hong, J and  N. Sadeh. 2012. The  livehoods project: Utilizing social media to understand the dynamics of a city. ICWSM’12.

[5] Freedom on the Net 2015. Venezuela. <https://freedomhouse.org/sites/default/files/resources/FOTN%202015_Venezuela.pdf>

[6] Fricke, M. 2015. Big Data and Its Epistemology. Journal of the Association for Information Science and Technology, 66(4): 651-661, 2015.

[7] Han, B., P. Cook, T. Baldwin. Text-Based Twitter User Geolocation Prediction. 2014. Journal of Artificial Intelligence Research 49 (2014): 451-500.

[8] He, J., L. Hong., V. Frias-Martinez, P. Torrens. Uncovering Social Media Reaction Pattern to Protest Events: A Spatiotemporal Dynamics Perspective of Ferguson Unrest.  2015. Social Informatics: 7th International Conference, SocInfo2015, pp. 67-81.

[9] Kahn, L. Spatiotemporal twitter analysis of the Venezuelan food crisis. Proceedings from the 2017 International Conference on Food Security. <https://www.omicsonline.org/conference-proceedings/2157-7110-C1-062-011.pdf> , p. 51.

[10] Korkmaz, G., J. Cadena, C. Kuhlman, A. Marathe, A. Vullikanti, N. Ramakrishnan. Combining Heterogeneous Data Sources for Civil Unrest Forecasting. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 258-265.

[11] Kounadi, O., T. Lampoltshammer, E. Groff, I. Sitko, M. Leitner. Exploring Twitter to Analyze the Public’s Reaction Patterns to Recently Reported Homicides in a London.  PLoS ONE, 10(3), e0121848. doi: 10.1371/journal.pone.0121848.

[13] Liu, B. Sentiment Analysis and Opinion Mining: on optimal parameters and parameters. 2015. Data Mining and Knowledge Discover, v5 (5): pp.255-263.

[14] Morstatter, F., J. Pfeffer, H. Liu, K. Carley. Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose. International Conference on Weblogs and Social Media, 400-408.

[15] Muthiah, S., B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz and N. Ramakrishnan. Planned Protest Modeling in News and Social Media.AAAI Publications, Twenty-Seventh IAAI Conference. 2015.

[16] Peca, I., G. Fuchs, K. Vrotsou, N. Andrienko and G. Andrienko. Scalable Cluster Analysis of Spatial Events. (2012) International Workshop on Visual Analytics, pp. 19-23.

[17] Priedhorsky, R., A. Culotta, S. Del Valle. 2014. Inferring the origin location of tweets with quantitative confidence. CSCW ’14 Proceedings of the 17th ACM conference on Computer supported cooperative work and social computing, pp. 1523-1536.

[18] Tobler, WR. (1970) A computer movie simulating urban growth in the Detroit region. Economic Geography, 234-240.

[19] Tufekci, Z. Big Data: Pitfalls, Methods and Concepts for an Emergent Field (March 7, 2013).

[20] Wikipedia. Media of Venezuela. <https://en.wikipedia.org/wiki/Media_of_Venezuela> (Accessed 19 July 2017).

[21] Zhao, L., T. Hua, C. Lu and I. Chen. A Topic-focused Trust Model for Twitter. Computer Communications, October 2015.