Spatiotemporal Twitter Analysis of the Venezuelan Food Crisis

Análisis Espacio-temporal de Twitter de la Crisis de Comida Venezolana

**Laura H. Kahn**

School of Informatics and Computing

Indiana University

Bloomington, IN 47405

Email: [lkahn@indiana.edu](mailto:lkahn@indiana.edu)

***Abstract***

Social media from countries with limits to free speech is often the most reliable source of event occurrence and is a reliable alternative form of journalism. Spatiotemporal analysis of location-based social media data allows new ways to describe events. Almost 37,000 Spanish geo-tagged 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 over time is explored. The hypotheses of whether certain Tweets are particular to a municipality location is tested using multinomial na**ï**ve Bayes, logistic regression and k-nearest neighbor machine learning classifiers.

Los medios de comunicación social de países con límites a la libertad de expresión es a menudo la fuente más confiable de la ocurrencia del evento y es una forma alternativa confiable de periodismo. Análisis espacio-temporal de datos redes sociales basadas en localización permite nuevas formas de eventos modelo. Casi 37.000 españolas geo-etiquetadas Tweets desde la ciudad de Caracas, Venezuela se utilizaron para observar reacciones a la crisis de escasez de alimentos dentro de cada municipios de la ciudad. El número de Tweets en el tiempo se explora. Las hipótesis de si ciertos Tweets son particulares a un municipio se prueban usando multinomial naïveBayes, regresión logística y vecino k-nearest clasificador de aprendizaje máquinas.

**Keywords** - Twitter, location-based estimation, spatial data mining, text mining, open source indicators, event forecasting

# **1. Introduction**

The high popularity of microblogging services such as Twitter has led to the availability of ever increasing volumes of location and time-referenced data. Analysis of microblogs is interesting for a number of applications, including a form of highly distributed ‘social sensors’ that use Twitter users as potential field reporters of extraordinary events. Although Twitter allows a user to declare their location, such metadata is unstructured and ad hoc.

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. Social media data plays a critical role in Venezuela where there is government censorship in traditional news media reporting. 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 help give insight and context into how citizens express themselves in space and time*.* Reactions to food shortage events were examined and spatiotemporal tweeting patterns were detected using Twitter hashtags and other related Spanish terms within each of Caracas, Venezuela’s five municipalities and all of Caracas from December 2014 - October 2016 (670 days).

Twitter was chosen because according to EModeration\*,   
Venezuela led Latin America in the number of active Twitter users compared to internet users at 14%. Despite having the second lowest internet speed in the region after Cuba and frequent electricity outages, there are approximately four million Twitter users in Venezuela. Therefore using Twitter social media mining is an appropriate approximation of to describe citizens’ reactions to the event.

+[www.el-nacional.com](http://www.el-nacional.com) (Accessed 10 Nov 2016)

\*[www.emoderation.com](http://www.emoderation.com) (Accessed 6 Oct 2016)

Tweets were limited to geocoded Tweets from the city of Caracas in Spanish in accordance with Tobler’s first law of geography [27], which states that *near things are more related than distant things.*

In Section 2, related research is discussed, in section 3 the raw data is discussed, section 4 is a discussion of the data mining techniques, section 5 is a discussion of the machine learning methods, section 6 is a detailed discussion and section 7 is the conclusion.

# **2. Related work**

Korkmaz uses Twitter and other heterogeneous data sources from 3,072 Venezuela events and a logistic regression model to predict where civil unrest activities will occur with a 71% precision and 97% recall within a city [17]. He uses a space time permutation scan and average nearest neighbor to predict hotspots of Twitter activity for protests in St. Louis, Missouri by neighborhood [15]. Han discusses the likelihood of geotagging Tweets by city [14]. Zhao describes a model where 75-91% of Venezuelan Tweets are trustworthy content [29]. Bora looks at social network mobility patterns of gangs by neighborhood of Los Angeles, California using a DBSCAN clustering algorithm [5]. Cranshaw studies the social dynamics of a city at the neighborhood level using geo-tagged Tweets [9]. Kounadi uses Twitter to analyze how the public responds to London crime events and finds that over half of the Tweet authors live within 10km of the event [18]. Cheng applied a type of naïve Bayes classifier to classify Twitter messages by city [6]. Priedhorsky uses a relatively small number of Tweets (about 30,000) to infer user location with a 67.5% accuracy [25].

Peca defines an event ‘as any physical or abstract discrete object having a particular position in space and time’ [23]. Examples of events include geo-referenced photos, occurrence of earthquakes, disease cases or in this case, the food shortage crisis in Caracas, Venezuela.

Richer, more densely populated states have more Tweets [3]which was another reason for selecting Caracas as the geolocation of the Tweet messages since it is a very densely populated area with about 5.5 million people.

Tweets retrieved without the use of hashtags present a much higher signal-to-noise ratio than tweets archived using hashtags though the original sample used 35 hashtags and keywords [3]. Bastos also states that the number of tweets in Sao Paulo was 4 times higher as the relative number of protestors at these locations, which indicates that the number of Tweets will probably only be an approximation of the local event. 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 [15,22], it is appropriate to use only geocoded Tweets and only Tweets originating in Caracas.

# **3. RAW DATA**

**3.1 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 personal information about the user.

###### 3.2 Data Extraction Steps

The data used in my research comes from a corpus of Spanish Tweets collected from December 2014 to October 2016. 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 is disregarded. Even though it is not included in this text mining analysis, all of the URLs including pictures were extracted and are included in the Appendix.

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.   


**Figure 1**. Caracas with each of its five municipalities.

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 [17], the radius ranges were chosen because the entire size of Caracas is 4.714 sq km. 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 collection period, approximately 37,000 geotagged tweets were retrieved.

Each Tweet is considered to be a data point. The approximate number of tweets about all subjects in Caracas from the same time period was calculated by multiplying the average daily tweets by the total number of days in the period. There were approximately 1.32 million Tweets from Caracas from December 2014 to October 2016.

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.

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. context of caracas tweets counts

|  |  |
| --- | --- |
| 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)

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.

###### 3.3 Phrase Filtering for Search Terms

Since the project uses text-centered data mining and no subject matter expert was available, search terms were developed using phrase filtering [23]. 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. A manual search of these sources using food scarcity terms yielded 38 initial search terms. Most 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.) or because they would have duplicated results from the final phrases (e.g. - comida/food, inseguridad de alimentos/food insecurity).

Final search terms included adjectives and nouns since nouns are most expressive [20]. The top three most popular hashtags reported by traditional media outlets were also chosen.

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, I decided to remove this term from the analysis.

# **4. DATA MINING**

**4.1 Implicit Assumptions**Interpreting the data implies background knowledge, assumptions and prior research. It is quite plausible to find apparent connections or false positives in the data. By using social media data, interviewer effects of bias that could alter the data were avoided [12].

Because about 14% of the population in Venezuela uses Twitter, the SMM methodology has limited analytic power and excludes certain demographics not using Twitter [28]. Selecting Twitter search criteria on a dependent variable introduces new uncertainties to the analysis. Note 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. Looking at this subset of data presents a snapshot of spatiotemporal Tweeting patterns from December 2014 - September 2016 but does not disclose the entire socio, economic or political story of the food shortage crisis in Venezuela.

It is also important to note that only about 0.833% of all Tweets are tagged with geographical coordinates [15].Prior research also shows that a user group with geocoded Tweets has a larger proportion of people that initiate original tweet content as compared to a user group without geo-tagged Tweets. It is therefore appropriate to use geocoded Tweets as a representative sampling of Tweets [22].

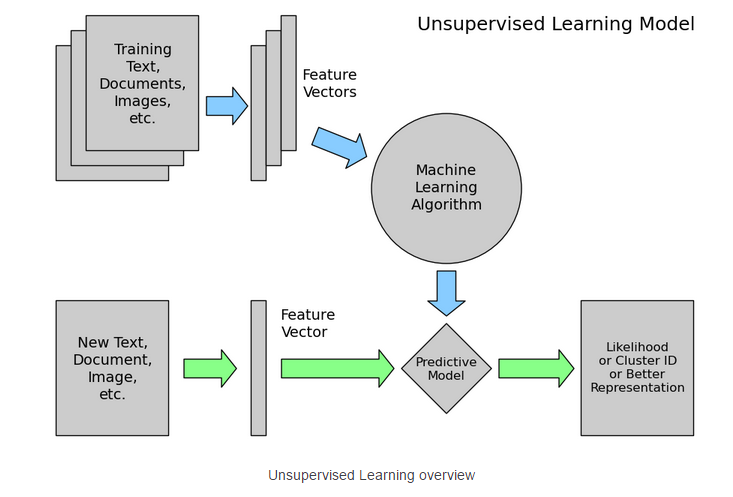
Tweets retrieved without the use of hashtags have a higher signal-to-noise ratio than Tweets archived with hashtags [3].

A user’s reported location in their Twitter profile was assumed to be their accurate location for purposes of analysis. 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.

The analysis reported in this paper were performed from Tweets originating in Caracas, Venezuela. For the reasons detailed in the Methods section, all approximately 37,000 Twitter messages were sampled to perform the naïve Bayes, logistic regression and k-nearest neighbor algorithm with the municipality as the feature label.

**4.2 Machine Learning Models**



**Figure 2.** Unsupervised learning Model

Unsupervised learning predicts the output of new data based on past data and detects any patterns. Selecting relevant features and choosing how to encode them for an unsupervised learning model has a big impact on the machine learning method’s ability to extract a good model. naïve Bayes, logistic regression and k-nearest neighbor algorithms were chosen for the research in this paper to predict whether certain words are more common in certain Caracas municipalities.

**4.3 Multinomial Naïve Bayes Classifier**

The naïve Bayes classifier attempts to simplify the problem by computing the probability of some value given the set of all attributes. The model assumes all the features are conditionally independent so if some of the features are dependent on each other, the prediction might be poor.

A multinomial naïve Bayes (MNB) classifier describes the probability that an event occurs and is represented by Equation 1 [16]:

 (1)

The class prior Pr(*c*) can be estimated by dividing the number of documents belonging to class *c* by the total number of documents. Pr(ti|*c*) is the probability of obtaining a document like ti in class *c* and is calculated as:

 (2)

Where fni is the count of word n in our test document ti and Pr(wn|c) the probability of word n in given class c. The latter probability is estimated from the training documents as:

 (3)

Where *F*xc is the count of word *x* in all the training documents belonging to class *c*, and the Laplace estimator is used to prime each word’s count with one to avoid the zero-frequency problem. The normalization factor Pr(*ti)* in Equation 1 can be computed using

 (4)

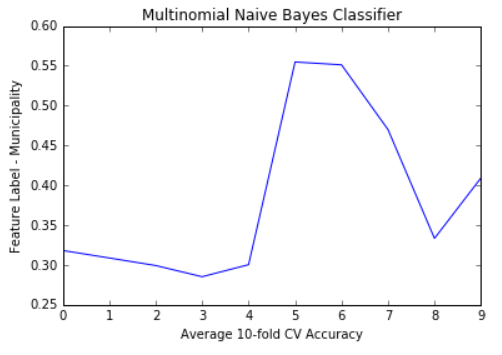
Note that the computationally expensive terms (Ʃn*fni*)! and nfni! in Equation 2 can be deleted without any change in the results, because neither depends on the class c, and Equation 2 can be written as:

(5)

where ɑ is a constant that drops out because of the normalization step.

Expressed simply, the naïve Bayes classifier estimates a target value by assuming that the data to be evaluated belongs to a category and then estimating the probability of the given attributes being present. The value with the highest probability is chosen as the resulting estimate.

The simplest metric that can be used to evaluate a classifier, accuracy, measures the percentage of inputs in the test set that the classifier correctly labeled.

The average 10-fold cross validation accuracy for the Multinomial naïve Bayes classifier was 0.383. The signal relationship of the Tweet text and the municipality will be described in the Discussion section.

**Figure 3.** Multinomial naïve Bayes accuracy

**4.4 Logistic Regression Classifier**Logistic regression (LR) estimates the probability(y/x) directly from the training data by minimizing error [21]. The model assumes feature spaces are split linearly. Limitations to this model is that with a small training data set, the model estimates may over fit the data.

Logistic regression is a method for analyzing data in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (True) or 0 (False).

The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a *logit transformation* of the probability of presence of the characteristic of interest in equations 6-8:

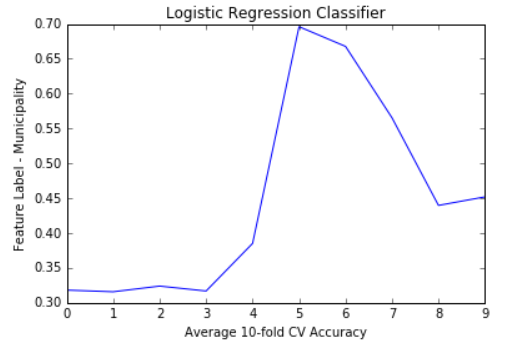
(6)

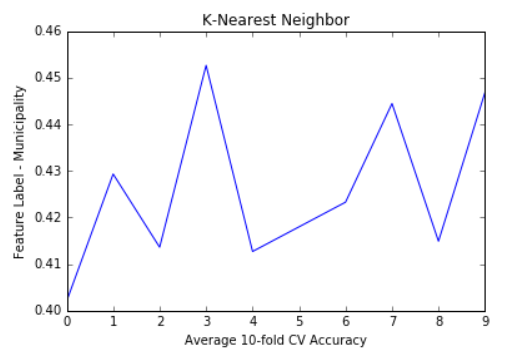
where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

(7)

and

(8)

The average 10-fold cross validation accuracy for the Logistic Regression classifier was 0.448. The signal relationship of the Tweet text and the Caracas municipality will be described in the Discussion section.



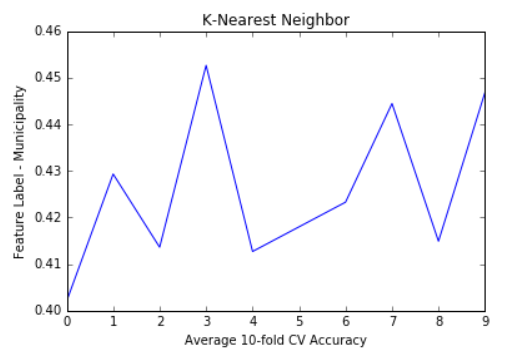
**Figure 4.** Logistic regression accuracy

**4.5 *k*-Nearest Neighbor Classifier**

The *k*-nearest neighbors (k-NN) algorithm is a method used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) where the input consists of the *k* closest training examples in the feature space [19].

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors. If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.

Again the feature label of municipality was used in the algorithm. The mean cross-validation scores with an optimized k-value of 20 was 0.422 as shown in Figure 5. The signal relationship of the Tweet text and the Caracas municipality will be further described in the Discussion section.



**Figure 5.** k-nearest neighbor accuracy

**5. METHODS**

The analyses reported in this paper were performed using Python. I began natural language processing (NLP) by stemming, tokenizing and punctuation. The Snowball Stemmer python package was used to remove stop words (“de”/of, “que”/that, “y”/and, etc).

The top 30 most frequently used words in all of the processed text was plotted using the NumPy Python package. The following top words that were not included in the five original search terms included ‘puebl’ (‘pueblo’/people), ‘distribu’ (‘distribución’/distribution), , ‘col’ (‘cola’/line), ‘medicin’ (‘medicina’/medicine), ‘compr’ (‘compra’/buy) and ‘gobiern’ (‘gobierno’/government).

|  |  |  |
| --- | --- | --- |
| **Spanish Word** | **English Translation** | **Frequency** |
| Pueblo | People | 304 |
| Distribución | Distribution | 291 |
| Cola | Line (of people) | 252 |
| Compra | Buy | 204 |
| Gobierno | Government | 113 |

**Table 2**. Most Frequently Used Stemmed Words

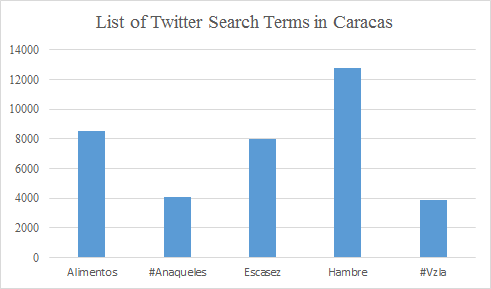
Next a Spanish corpus CESS-ESP was used to train the data, which is normally the second step in the typical NLP pipeline. Then naive Bayes, logistic regression and k-nearest neighbor ML classifiers were run to produce evaluation metrics.

The CESS-ESP is the largest annotated corpus of Spanish with 500,000 words coming mostly from newspapers [26]. Although it is unclear how many of the newspapers used to develop the corpus were Venezuelan, the CESS-ESP corpus is adequate for the research since the desired search terms (hunger, scarcity, food) and their usage do not differ significantly among different Spanish-speaking countries. For certain locally-used words, this corpus may be less accurate. The CESS-ESP was automatically annotated using the SKLEARN Python library.

In order to evaluate the selected models, a portion of the annotated data must be used for the test data set. There is a balance between a test set that is too small or too large to make the evaluation the most accurate possible. Cross-validation allows us to solve this problem by performing multiple evaluations on different test sets and then combine the scores from those evaluations [4]. A cross-validation of 10 was used for all three classifiers.

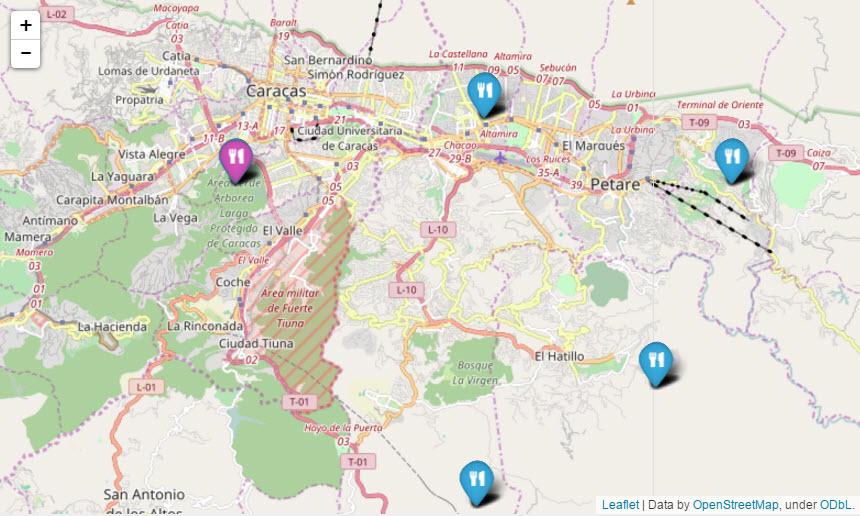
**6. DISCUSSION**

All the search terms (“alimentos”, “escasez”, “hambre”) were two to three times more popular than the hashtags cited in the traditional media outlets in each municipality. The term “hambre” (hungry) was the most frequently used term, followed by “alimentos” (foods), then “escasez” (shortage), then “#AnaquelesVaciosenVenezuela” (#EmptyShelvesin Venezuela) and then “#VzlaTieneHambre” (#VenezuelaisHungry). All of the search terms had the same frequency of use in the Chacao municipality.



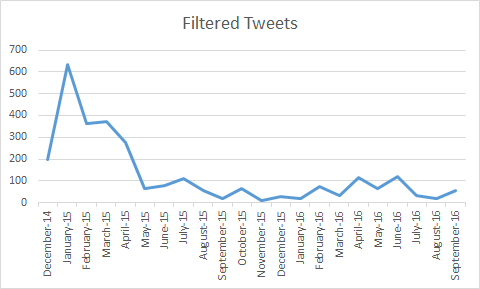
**Figure 6**. Total Number of Twitter Search Terms

The next layer of analysis is which of the search terms is used most often in each of the five municipalities and all of Caracas. The word “Hambre” (hunger) was the most popular in all municipalities except the Libertador municipality which had the same popularity as “Alimentos”, “#AnaquelesVaciosEnVenezuela”, and “VzlaTieneHambre.” Each of those 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.



**Figure 7.** Most popular search term by municipality

In addition to looking at the total number of search terms from the Tweet corpus, a temporal analysis was performed on the total number of filtered Tweets by month from the studied period. Recall that the filtered Tweets were Tweets with at least five Retweets and / or five Favorites. The highest number of filtered Tweets in a month occurred in January 2015 with 632. The number of Tweets drops considerably to 66 in May 2015 and then fluctuates between 11 and 122 for the remaining 16 months. The highest number of Tweets occurs during the second month of the studied time period. The drop in Tweets is consistent with prior research into how information is diffused as time elapses after the initial start of the event - also known as *Immediacy* [28].



**Figure 8.**  Filtered Tweets by month

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). The training data set labels are shown in Table 3.

|  |  |  |
| --- | --- | --- |
| **Label** | **Municipality** | **# Filtered Tweets** |
| 1 | Baruta | 65 |
| 2 | Caracas - all | 367 |
| 3 | Chacao | 900 |
| 4 | Hatillo | 101 |
| 5 | El Libertador | 749 |
| 6 | Sucre | 638 |

**Table 3.** Training data labels for the classifiers

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 MNB accuracy is 0.383, the LR accuracy is 0.448 and the k-NN accuracy is 0.422. The MNB accuracy outperforms the baseline by 16.7%, the LR outperforms it by 28.8% 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 of these models perform better than the baseline classifier, a reasonable conclusion to the initial hypothesis 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.

|  |  |  |
| --- | --- | --- |
| **Classifier** | **Accuracy** | **% Improvement** |
| Baseline | 0.319 | n/a |
| Multinomial Naïve Bayes | 0.383 | 16.7 |
| Logistic Regression | 0.448 | 28.8 |
| k-NN | 0.422 | 24.4 |

**Table 4**. Machine learning classifier results

**7. 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.

Reactions to food shortage events were examined and spatiotemporal 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 from over 37,000 Spanish language Tweets.

The hypothesis whether certain words are particular to a municipality was tested using naïve Bayes, logistic regression and k-nearest neighbor machine learning classifier. Words used in a Tweet in Caracas are signals of their municipality location. Overall reactions to the crisis in how the number of Tweets changed over time was also explored.

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 and included in the Appendix.

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.

**Acknowledgment**

The author would like to thank 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] Bird, S., E. Klein and E. Loper. Natural Language Processing with Python. 2009. O’Reilly Media.

[5] Bora, N., Zaytsev, V., Chang, Y.H. and R. Maheswaran. Spatiotemporal   
Patterns in Social Networks. 2013 AAAI Fall Symposium Series.

[6] Cheng, Z., J. Caverlee, K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. 2010. Proceedings of 19th ACM International Conference on Information and Knowledge Management.

[7] Chesley, P., B. Vincent, L.Xu , R. Srihari. Using Verbs and Adjectives to Automatically Classify Blog Sentiment. Dataset self-published online. Retrieved from the University of Minnesota Digital Conservancy, <http://hdl.handle.net/11299/127160>. Accessed 22 November 2016.

[8] Compton, R. D. Jurgens, D. Allen. Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization. Big Data (Big Data), 2014 IEEE International Conference on, pp. 393-401.

[9] 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.

[10] Ester, M., H. Kriegel, J. Sander, X. Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd.

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

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

[13] Hai-Jew, S. *Social Media Data Extraction and Content Analysis*. Hershey Global: IGI Global, 2016.

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

[15] 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.

[16] Kibriya, A., E. Frank, B. Pfahringer and G. Holmes. Multinomial Naïve Bayes for Text Categorization Revisited. <http://www.cs.waikato.ac.nz/~eibe/pubs/kibriya_et_al_cr.pdf> . Accessed 29 November 2016.

[17] 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.

[18] 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.

[19] Larose, Daniel T. k-Nearest Neighbor Algorithm, chapter 5: *Discovering Knowledge in Data: An Introduction to Data Mining.* 2005.

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

[21] MedCalc Manual. Logistic Regression. <https://www.medcalc.org/manual/logistic_regression.php> Accessed 27 November 2016.

[22] 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.

[23] 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.

[24] 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.

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

[26] Recasens, M., M. Martí and M. Taulé. Where Anaphora and Coreference Meet. Annotation in the Spanish CESS-ECE Corpus. 2007. Proceedings of the International Conference on Recent Advances in Natural Language Processing.

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

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

[29] Von Steinkirch, M. Information Diffusion in Twitter. (June 29, 2012).

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

APPENDIX

**Table 1** List of Twitter Hashtags and Keywords Associated with the Food Shortage Crisis

|  |  |
| --- | --- |
| **Search Term** | **Total Number of Raw Tweets** |
| Alimentos (Foods) | 8,558 |
| #AnaquelesVaciosEnVenezuela (Empty Shelves in Venezuela) | 4,067 |
| Escasez (Shortage) | 7,963 |
| Hambre (Hungry) | 12,762 |
| #VzlaTieneHambre (Venezuela Is Hungry) | 3,866 |

**Table 2** List of Twitter Hashtags and Keywords Associated with the Food Shortage Crisis by Municipality, Raw versus Filtered

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Caracas - All** | **Raw** | **Filtered** | **El Hatillo** | **Raw** | **Filtered** |
| Alimentos | 1,974 | 180 | Alimentos | 553 | 24 |
| #AnaquelesVaciosEnVenezuela | 92 | 8 | #AnaquelesVaciosEnVenezuela | 37 | 3 |
| Escasez | 1,758 | 192 | Escasez | 502 | 26 |
| Hambre | 3,583 | 169 | Hambre | 1,084 | 47 |
| #VzlaTieneHambre | 14 | 2 | #VzlaTieneHambre | 2 | 1 |
| *All Tweets* | 7,421 | 551 | *All Tweets* | 2,178 | 101 |
|  |  |  |  |  |  |
| **Baruta** | **Raw** | **Filtered** | **Libertador** | **Raw** | **Filtered** |
| Alimentos | 311 | 22 | Alimentos | 1,917 | 180 |
| #AnaquelesVaciosEnVenezuela | 20 | 3 | #AnaquelesVaciosEnVenezuela | 1,917 | 180 |
| Escasez | 252 | 9 | Escasez | 1,697 | 105 |
| Hambre | 635 | 27 | Hambre | 1,697 | 105 |
| #VzlaTieneHambre | 3 | 1 | #VzlaTieneHambre | 1,917 | 180 |
| *All Tweets* | 1,221 | 62 | *All Tweets* | 9,145 | 750 |
|  |  |  |  |  |  |
| **Chacao** | **Raw** | **Filtered** | **Sucre** | **Raw** | **Filtered** |
| Alimentos | 1,917 | 180 | Alimentos | 1,886 | 178 |
| #AnaquelesVaciosEnVenezuela | 1,917 | 180 | #AnaquelesVaciosEnVenezuela | 84 | 8 |
| Escasez | 1,917 | 180 | Escasez | 1,837 | 103 |
| Hambre | 1,917 | 180 | Hambre | 3,846 | 178 |
| #VzlaTieneHambre | 1,917 | 180 | #VzlaTieneHambre | 13 | 2 |
| *All Tweets* | 9,585 | 900 | *All Tweets* | 7,666 | 469 |
|  |  |  |  |  |  |
| *All Neighborhoods + Caracas* | 37216 | 2833 |  | | |

**Figure 1**  Pictures from the Extracted Twitter URLs.

For all 292 of the unique URLs, visit: <https://github.com/thedatalass/IUSocialMediaMining/blob/master/allURLS-cleaned.xlsx>



