**Where you are in 140 characters: Classifying Tweets using Multinomial Naïve Bayes**

**Jasen Wright & Conner Leverett**

**Abstract**

Twitter is a microblogging site which allows users to post 140 character ‘tweets’. Although there is a vast number of tweets sent per day throughout the globe, a very small amount of them have an attached latitude and longitude. Using a multinomial Naïve Bayes text classifier, we will attempt to classify tweets from the contiguous United States as being in the northwest, northeast, southwest, or southeast.

**Introduction**

Twitter is a microblogging site which allows users to post 140 character ‘tweets’. These tweets can be anything from what the user is cooking for dinner to their thoughts on political events. Although not as popular as in its peak, there are still approximately 300-500 million tweets posted everyday (Edwards, 2016; Twitter, 2015). This provides a wealth of knowledge which can be used to detect earthquakes (Sakai et al, 2010), predict elections (Tumasjan et al, 2010) and much more. However, less than one percent of tweets are geotagged (have an attached latitude and longitude) (Jurgens et al, 2015). This creates a gap in data. Accurate prediction of demographic information is useful for advertising and legal investigation (Burger et al, 2011). For simplicity this study only looks at the contiguous United States. The goal of this study is to use Multinomial Naïve Bayes to attempt to see if tweets can be classified by location solely by the text content of the tweets. Naïve Bayes has been widely used by other researchers in attempt to classify tweets through their textual content (Burger et al, 2011; Liu et al, 2012; Wing & Baldridge, 2011).

**Mining Technique**

A Python script was created to gather tweets. Specifically, the Twitter API was used. Another popular twitter mining Python library is Tweepy. We were unable to use this library as a key word is needed when mining tweets. This studied needed tweets of all kinds, not just ones with specific key words so we used the Twitter API as there is no key word needed. The tweets were stored in a PostgreSQL database for later use. In order to access PostgreSQL, a Python library called psycopg2 was used. This library allows data to be stored and retrieved from the database. Once a tweet was mined the script checked if it had an attached latitude and longitude. If the tweet didn't it was thrown away. However, if the tweet did have a latitude and longitude, it was checked to see if it was within the United States (Figure X). The outline of the United States is a rough outline for simplicity.

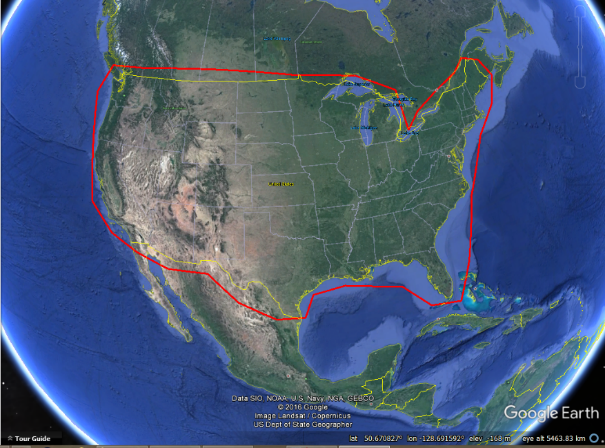


Figure X. Outline of the United States of America.

The point is checked if it is in the polygon using the ray casting method. A line is drawn in any direction from the point in question. The number of times the line crosses the polygon in question is counted. If the count is an odd number, the point is in the polygon. If the count is an even number, the point is outside of the polygon (Figure X).

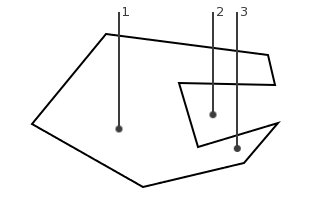


Figure X. An odd number of crossing defines the point as being in the polygon while an even amount defines the point as being outside the polygon.

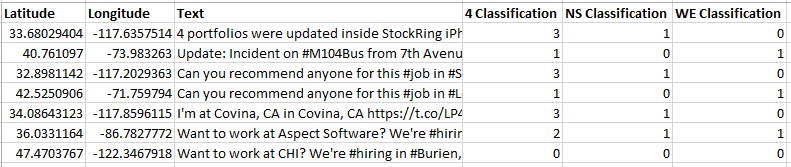
We only collected tweets that were in the United States of America. If the tweet did not have an attached unique ID or text the tweet was thrown away. The twitter script can be seen in appendix X.

**Data Processing**

Data was moved from PostgreSQL to Excel for processing. The first step in processing was removing all new line characters in the tweet. This was done because in the analysis phase each tweet needs to occupy only one line on the document. In order for this to be done the following function was used for each tweet '=SUBSTITUTE(SUBSTITUTE(C2,CHAR(13),""),CHAR(10),"")' where 'C2' is the Excel cell in question. In order for tweets to be consistent, all of the tweets were forced to be lowercase. Stop words, very common words which have no influence on the tweet, were removed to increase analysis speed.

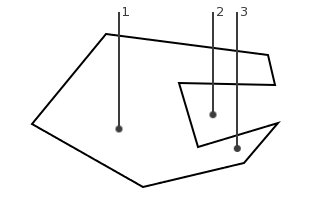
**Truth Classification**

The tweets were classified three different ways. 1) northwest, northeast, southwest, or southeast 2) north or southand 3) West or East (Figure X).



**Further Work**

Including the time of the tweet could be useful in classification as if using PST as the reference, if the tweet is from 5am, it is most likely from the east coast. Using the temporal aspect of tweet has successfully been done by Liu et al (2012).

**References**

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1301-1309). Association for Computational Linguistics.

Edwards, J. (2016). Leaked Twitter API data shows the number of tweets is in serious decline. Retrieved from <http://uk.businessinsider.com/tweets-on-twitter-is-in-serious-decline-2016-2> March 20th, 2017.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. In ICWSM (pp. 188-197).

Liu, H., Luo, B., & Lee, D. (2012). Location type classification using tweet content. In Machine Learning and Applications (ICMLA), 2012 11th International Conference on (Vol. 1, pp. 232-237). IEEE.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web (pp. 851-860). ACM.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM, 10(1), 178-185.

Twitter. (2015). Twitter Support. Retrieved from <https://twitter.com/Support/status/555076845293432834> March 20th, 2017.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. In ICWSM (pp. 188-197).

Wing, B. P., & Baldridge, J. (2011). Simple supervised document geolocation with geodesic grids. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 955-964). Association for Computational Linguistics.