**Introduction**

For simplicity this study only looks at the contiguous United States.

**Purpose**

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.

**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.

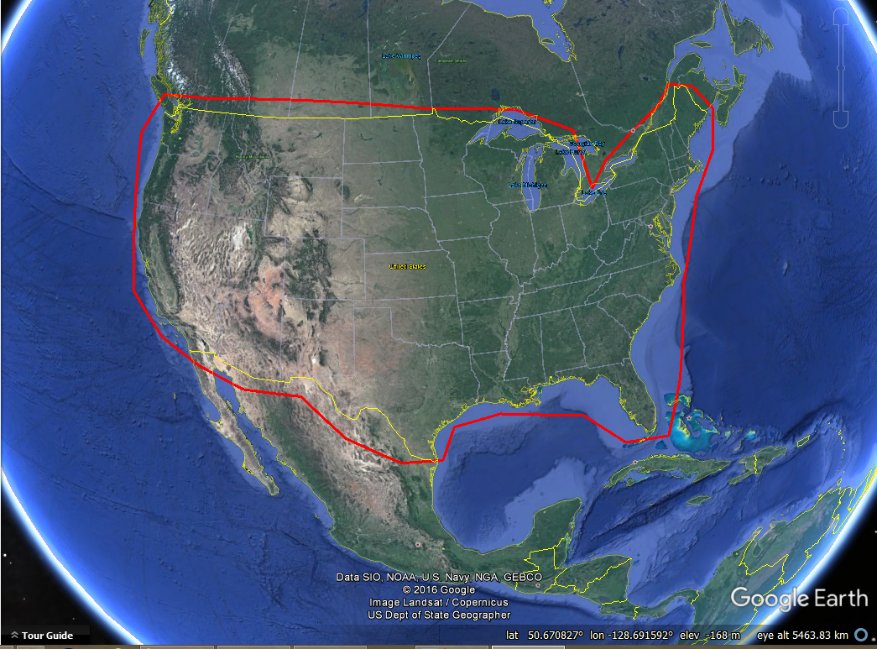


Figure X. Outline of the United States of America.

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).

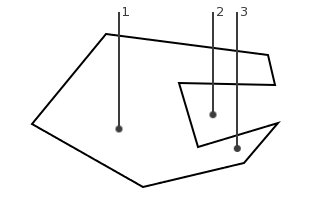


Figure X. An odd number of crossing defines the point as being in the polygon while an even amount defines the point as being outside the polygon.

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) Either northwest, northeast, southwest, or southeast 2) north or southand 3) West or East (Figure X).

