Whereas traditional natural language processing uses millions of lines of text to train models, Twitter provides a unique advantage to training because each message is limited to a 140 character limit. Though the volume of readily available data than to train model on, but on the other hand each word in a Twitter message has importance and thus the signal to noise ratio should be much higher.

**Dictionary Sentiment Model**

One way to extract information from financial-related text is to perform sentiment or tone analysis. We apply a financial sentiment dictionary using a bag-of-words approach by assigning each word and word vector (tweet) a positive or negative value. Whereas the Harvard Psychosociological Dictionary, or Harvard-IV-4 TagNeg (H4N), is the common source for word classification, Loughran and McDonald [2011] found that it substantially misclassifies words when gauging tone in financial applications. The paper found that 73.8% of negative word counts according to the Harvard list were attributable to words that are not typically not negative in financial context (e.g. *tax, cost, capital, board, liability, foreign*, etc.). Instead, we use the FinNeg and FinPos lists they provide here: <http://www3.nd.edu/~mcdonald/Word_Lists.html>

**Calculations**

Each tweet ***n*** is cleaned and formatted into a list of all-caps words ***wi***. This is used to calculate the single-tweet sentiment score ***Sn***.

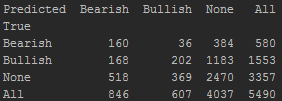
The sentiment score can range between -1 and +1. If a tweet contained three positive financial words and two negative words, then the sentiment score would be +0.2.

Let the number of tweets on a distinct day ***d*** be defined as ***nd***. When we aggregate across tweets for a company on each distinct day ***d***, we calculate the following:

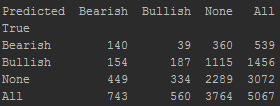
**Confusion Matrix**

**Totals**

Basic Cleaning

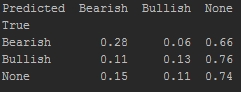


Advanced Cleaning (Bruno’s Method)

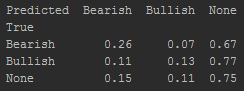


**Percentages**

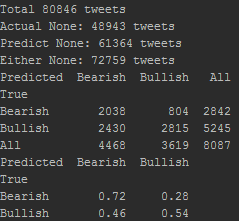
Basic Cleaning



Advanced Cleaning (Bruno’s Method)

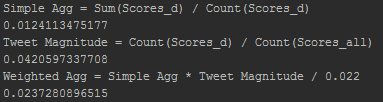


**NEW Confusion Matrix Results**



**Results**

**Single Day Analysis: AAPL 3/28/2017**

****

**TO DO:**

**How big are the positive and negative dictionaries?**

**Reference**

Loughran, T., and McDonald, B. [2011]. “When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks”, SSRN, <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1331573>