Training the Data

After standardizing the data, we do a first run of dictionary level sentiment with VADER and TextBlob. Normalized compound scores for both dictionaries run from -1 (most negative) to 1 (most positive). We can safely assume that VADER is performing better than TextBlob in our dataset as it is finding words that TextBlob does not contain in its dictionary. We see in Figure 4 that the scores are more diverse for Vader than Textblob.

In order to understand our first sentiment run, we focus on the dictionary output of individual words as to whether they are trained in the positive, neutral or negative sphere. This is because we later want to recode words that were not in the correct sphere. A subset of the output for the words that VADER trained as positive is in Figure 5. We stuck with the definition in the VADER documentation of Positive is >0.05, Negative is <0.05, and the rest of the words belong to a neutral sphere.

Extracting the Data

One of the most time-consuming parts of this project was extracting the data from the New York Times databases in order to do analysis. We tried replicating the data extraction from a past research conducted by Rivera, Abrajano, and Hassell (2017), as well as many other sources. We then focused on getting our data directly from the New York Times, which we did via public API’s (Figure 2). However, the information that was available was only the lead paragraph and not the entire article. What we queried was also difficult to determine because we wanted to get all the articles that pertain to immigration. We investigated individual words that pertain to the immigration sphere. The final query was the following: “migrant OR immigration OR immigrant OR migration OR refugee OR alien OR undocumented OR asylum”. The API had limited amount of requests, so we were able to get the data from 1981 to 2020 chunks at a time. We ended up with 21,457 rows of data. We also focused our search so that those phrases did not skew our results (Figure 3). In order to achieve this we use Part-of-speech Tagging for our words in the corpus. 

Data Standardization

In order to standardize the data, we had to think about our end goal. We focused on the lead paragraph for every article that we had and we wanted to understand how our data was structured. We had to normalize our text data. We removed punctuation and stop words, such as “the,” “is,” and “and.” However, we did not convert everything to lower case. The reason behind this was because we wanted to identify certain proper nouns that had very high frequencies which could be embedded as a positive, neutral and negative words. We wanted to make sure that these phrases did not skew our results (Figure 3). In order to achieve this we use Part-of-speech Tagging for our words in the corpus.

Figure 4

Figure 5: Positive words by VADER

Recode Keywords and Pronouns

In order for us to look at individual words, we need to focus on the frequency levels. Figure 7 displays a sharp drop off in distribution levels at the 20-50 word range for the positive words displayed in figure 5 (blue line). To be safe we hand coded more words than our visualization tries to tell us. In the positive and negative words corpus we looked at the words in up to the 300 ranking frequency. However, given the high number of neutral words we looked until the 300 ranking frequency for words in this category. We conducted inter reliability coding for the new training/recoding of problematic classifications of words.

We found were able to find problematic pronouns by using a TF-IDFs and its n-grams counter parts. For example, the word “United” was appearing as a positive word but we really knew that the word was appearing in the context of “United States,” which belongs in the neutral sphere as the word per se does not belong to any sentiment. What was done in order to recode these bigrams of words was to replace the word with a dummy variable called “n, gram removal.” This is due to the fact that the lower case word “united” actually belongs to a positive sphere, so it was not safe to recode all of the words “united” into a neutral classification. Other words are highlighted in red in Figure 3.

We also recoded individual words in their appropriate sphere after the first sentiment run. In order to recode new words, we used the mean scores for words that pertained to a positive and negative light (-0.5 and 0.5) in the original VADER dictionary. Below are the individual keywords that we recoded if the word pertains to a sentiment in the context of immigration (Figure 6) Ranking of frequencies helped us looking at the words that were the most frequent for each spheres and recode them if necessary. Comparing our new recoded dictionary with our first run, we can see that the new recoding skewed the overall scores to a more negative light (Figure 8)