Spring 5-1-2020

Sentiment Analysis on New York Times Articles Data

Gabriel Carvajal
gabriel.carvajal@valpo.edu

Karl Schmitt
Valparaiso University

Gregg B. Johnson
Valparaiso University

Sentiment Analysis in Latino Immigration: Political Science Data Review

Gabriel Carvajal

Introduction

The extant political science literature examines media coverage of immigration and assesses the effect of that coverage on partisanship in the United States. Immigration is believed to be a unique factor that induces large-scale changes in partisanship based on race and ethnicity. The negative tone of media coverage pushes non-Latino Whites into the Republican Party, while Latinos trend toward the Democratic Party. The aim for this project is to look at New York Times data in order to identify how much immigration is covered in newspaper outlets, specifically Latino immigration, and to determine the overall tone of these stories from the years 1981 to 2020.

In this research, we seek to determine whether individual articles take a positive, neutral or negative stance. We achieve this using a dictionary-based approach, meaning we look at individual words to assess if they have a positive, neutral or negative connotation. We train our data using publicly accessible sentiment dictionaries such as VADER (Valence Aware Dictionary and Sentiment Reasoner) and TextBlob. However, this task can be difficult because certain words can be dynamic and may pertain to a positive or negative sentiment in context of the article. In order to resolve this issue, we use reliability measures such as recoding of words to ensure that the words of high frequencies are in the correct sphere of negative, neutral, or positive stance..

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 words were contained in the lead paragraph.

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.

Media Framing and Partisan Identity: The Case of Immigration Coverage and White Macropartisanship

Marisa A. Abrajano and Zoltan Hajnal
University of California

Hans J. G. Hassell
Cornell College

Figure 1

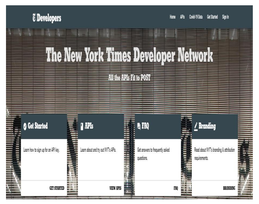

Figure 2

```
Problematic words : ["United States", "Supreme Court", "Homeland Security",
                     "Social Security", "Justice Department", "Great Depression",
                     "United Nations", "Statue of liberty", "Central Intelligence", "Lower East",
                     "Star Wars", "New York"]
```

Figure 3

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 disperse for Vader than Texblob.

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, negative or neutral 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.

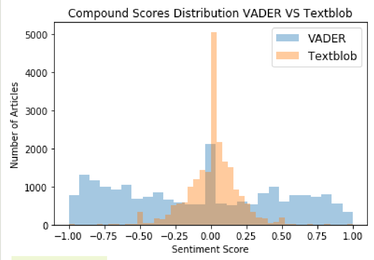

Figure 4

Ranking	Word	Frequency	TAGS	Is a character in word capitalized?
0	United	4508	NNP	TRUE
1	like	1305	IN	FALSE
2	intelligence	814	NN	FALSE
3	legal	764	JJ	FALSE
4	number	671	NN	FALSE
5	help	640	NN	FALSE
6	security	579	NN	FALSE
7	well	551	RB	FALSE
8	party	485	NN	FALSE
9	support	434	NN	FALSE
10	want	395	VBP	FALSE

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 up to the 200 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).

```
Declaring words for retraining

In [92]: list_words_positive_to_neutral = ['like', 'number', 'party', 'want',
                                           'thank', 'parties', 'significant', 'sure', 'playing', 'gain',
                                           'trump', 'certain']
         list_words_negative_to_neutral = ['low', 'drop', 'lower']

In [93]: list_words_postive_to_negative = ['escape','challenge', 'challenges','overwhelmed', 'urgent', 'challenging']

In [94]: list_words_neutral_to_positve = ['home', 'naturalization']
         list_words_neutral_to_negative = ['aliens', 'deportation','alien',
                                           'camps', 'seeking', 'undocumented', 'aid',
                                           'officers', 'troops', 'illegally', 'enforcement', 'fled' ]
```

Figure 6

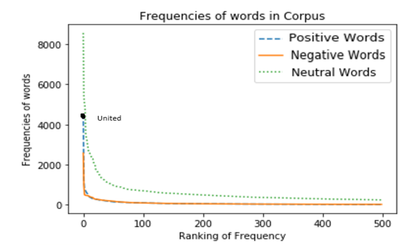

Figure 7

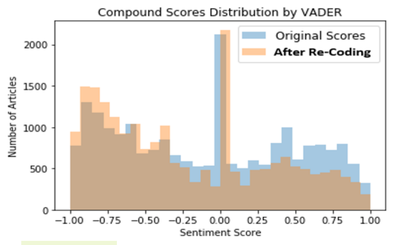

Figure 8

Extract Subset of Latino News

The 21,457 articles refer to immigration articles written all around the globe. We filter on data that pertains to immigration in the United States that comes to a subset of 11,659 articles. Out of those 11659 articles we want to filter further on news articles that pertain to the Latino immigration sphere. The query of words for extracting Latino news words were done on the lead paragraph and are: "Latino", "Mexican", "Mexico", etc. 2299 articles were outputted, which comes to around 20% of articles that pertain to the United States. We expected to output more news that mention Latinos immigration, but the reason why we were not able to fully extract them was because the terms in our query may not be mentioned explicitly in the lead paragraph.

Results

We output the sentiment scores by VADER using the recoded dictionary on the Latino Immigration news that pertain to the United States. We can see that for Latinos the distribution is slightly more negative than the overall tone of immigration across the globe (Figure 10).

We also output a time series analysis on sentiment tone over time (Figure 9). In order to do this, we had to take the scores of -1 to 1 and normalize it on a scale from 0-100 (%). We then computed a yearly average sentiment for the years 1981 to 2020.

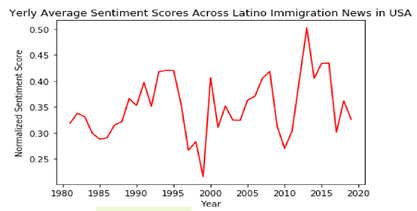

Figure 9

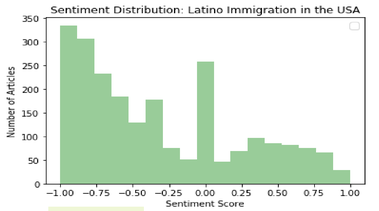

Figure 10

Conclusion

Sentiment analysis is a challenging task to do because it can be subjective. We limited the study to individual words but we know that text analytics can be more complex due to word negations, sarcasm, and different kinds of connotations. Even though our analysis is not perfect, it gives us a good idea on how the immigration and, specially the Latino immigration, sentiment is portrayed in the United States between 1981 – 2020. From our study, we can see that immigration is portrayed in a negative light as well as Latino immigration as a whole.

We now have systematic measures of the frequency and tone of both overall immigration coverage and Latino immigration coverage in the New York Times. Past social science research notes the Times often sets the topics and tone of coverage in other news outlets. Consequently, we can now use this dataset to examine whether the frequency and tone of immigration coverage influences African American, Latino, or White macropartisanship. We can do this using another original dataset constructed by VU faculty and Data Science students.

Acknowledgements:
Faculty Advisors: Professor Karl Schmitt and Professor Gregg B. Johnson

Valparaiso University