Growing Polarity & Sensationalization in the News

Report Summary

Project executed by Jane Shclover
BrainStation, Data Science Diploma
Capstone Project
August 2022

Code: github.com/janeshclover/nyt-nlp

Problem Statement

I am looking to investigate whether or not headline language has become increasingly polarized over the past decades. In the US we talk lot about the polarization of our society and the media as as either a vehicle or reflection of this division. I want to validate/invalidate this by analyzing headline data from recent decades from a leading news source in the US: the New York Times. I believe this analysis has the potential to arm people with vital context when reading the news, in addition to being an asset to existing and future research that considers the evolving trajectory of the American mindset, communication polarity, and its implications.

Background

Idea sensationalization and idea reduction are drivers of polarity, and both can be analyzed through the lens of language analysis, giving everything a positive or negative spin and using increasingly dramatic and charged language. The NLP tool that I employ in this project measures sentiment as well as intensity of language, covering off on both. This problem is well suited to date science / NLP because it offers an opportunity to quantify and properly diagnose something that many people suspect, transforming it from a nebulous communal feeling to something substantive.

Data Source

To execute this analysis I used the NYT API to pull headline data dating back to 1981. By the nature of the API this data has to be ‘grabbed’ one month and year at a time, I imagined there may be differences in data structure over the years so I began by pulling January data for every year. Each month of data had 28 columns including headline metadata, publication date, section metadata, article metadata, content metadata, and format metadata (see individual monthly notebooks for data structure). My final data set covers headlines from 1981 through 2021 filtered to 10 top categories of news, including the headline, section name, publication date, and assigned polarity / sentiment scores.

Data Cleaning / Transformation

Date & Category filtering

- I initially pulled date back to 1950, but saw a significant change in how the API data was structured beginning in 1981 so I limited my analysis to 1981 vs. my original desire to go back to 1950.

  - NYT only added ‘section’ metadata (‘World’, ‘Technology’, ’Blog’, etc.) to their database in 1981.
  - I wanted to investigate top categories and 40 years is more than enough data to work with). For data integrity, it was also important to only use those categories that have had data consistency since 1981 (e.g. if I began comparing headline language from an outdated section to language used today in a complete new section, it would compromise the notion of drawing any parallels and muddy the analysis).
- I created features / columns that grouped publication dates by both 5 and 10-year intervals. The 10-year intervals were useful in looking at overall language trends over time while the 5-year intervals were more useful in a granular look at token & headlines views (after all, the news can change dramatically every year, let alone every 5, or 10).

Removing nulls

- Removing any rows where key values such as headlines, section name, or publication date were missing (I initially remove any rows with any nulls but discovered through this action that prior to 1995 there was essentially no data in some of the content metadata features (snippet, abstract, etc.)

De-duplication

- I removed headlines with any duplicate values, in most cases these were headlines like ‘National Briefing:’ or ‘Lottery Numbers’ -- there are actually group titles that repeat themselves. I initially tried cleaning all of these out using string functions but then realized the de-duping would bring their number down to something insignificant as compared to the larger analysis. There were also some instances of ‘true’ article headlines simply appearing twice that this took care of.

Cleaning by headline length

- In some cases in the data there were rows where the content metadata was all concatenated together in the headlined feature, there were relatively few of these compared to the full data set, and they all had wildly long lengths, so I removed them all.
- Similarly, I removed erroneous headlines like ‘P’, ‘Q’, etc.
- I used histograms to verify that the lion’s share of headline fell into a certain standard / ‘healthy’ length and removed all others.

Removing unnecessary features

- After using some of the metadata to clean the data set, removing unnecessary features re: content, format, article, and headline, keeping only the final headline, section name, and publication date.

Modeling / Polarity Assignment

I leaned on a tool called vaderSentiment for my analysis, which assigns a series of sentiment scores to text. Every headline receives a positive, neutral, and negative score on a scale of 0 to 1, and one compound score which is a combination of all of these, on a scale of -1 to 1. This provides an idea of each headline’s sentiment direction as well as its level of intensity (Is it positive? Neutral? How much so?), helping answer key questions.

The language analysis is directly about the words used in the headline and how charged they are (or are not), *not* the content of the articles, nor the overall tone or spin.

EDA

I took a two-fold approach in investigating the scores, one that looked at net results and one that looked at intensity. In the first case, I used compound scores to identify each headline as either neutral, positive, or negative, then looked at these trends over time (in volume and share of total). This revealed that there is indeed a trend of increasing net positive and negative share of headline over time, while neutral declines. In my second approach I used histograms, seaborn boxplots, and a few other tools to look at how the scores migrated across the scale (0-1 for positive, neutral, negative or, -1 to 1 for compound) over time. I also conducted some work to look at the correlations between the various scores and better understand their relationships / how vaderSentiment assigns and weighs these against each other. See notebook for more.

Analysis & Modeling

My project was not about evaluating vaderSentiment but using it to assign sentiment to headlines and exploring headline language across different categories of news, for that reason I don't havea model / evaluation component.

I continued investigating sentiment trends as above by individual section of news, running linear regressions using seaborn to test for significance, revealing that across many categories my results were completely or partially validated.

My hypothesis was entirely validated in the *Business Day* and *Arts* sections, where net positive and net negative headlines shares have grown significantly each year, while net neutral headline shares declined. In categories like *New York*, *World*, *Movies*, and *Books* results were partially updated, with net negative share growing (at significance). In lifestyle section like *Sports* and *Style*, net positive headline share is actually significantly replacing neutral. *Opinion* and *USA* didn't show any notable trends. (See presentation slides 15 & 16 for figures).

I also grabbed the most frequented tokens in each section, by 5-year interval, from a pool of the lowest-scoring and highest-scoring 1K headlines (by compound score), in addition to an example of a headline in each of these groups. This part of the work is more of a WIP / starting point. It helps paint a picture of the language behind each section over time, but needs more work to become substantive. (See presentation slides 19 & 20).

Final Results & Next Steps

Of 10 categories, 2 validated the hypothesis of sensationalization on both the positive and negative fronts (*Business Day*, *Arts*), 4 validated sensationalizing in one direction or another (*World*, *Sports*, *Style*, *NYC*), and 2 categories where net negative headline share is growing (*Movies*, *Books*), and only 2 without any trends whatsoever (*Opinions*, *USA*).

Overall, there has been a migration from mild to increasingly charged headline language in many categories of the NYT's news. In these instances neutrality has slowly been replaced with both positive, but more often, negative, sentiment.

As next steps, I would like to complete the scorecard concept that I began to flesh out with the *World* section, and I would like to dive much deeper into the actual language and token analysis, answering questions like how the same topic is discussed today vs. years ago or decades ago, and really tying this project into contemporary research around righteousness and sesationalization in the media.