Sentiment Analysis and Predictive Modeling for Titles of Online News Articles

Hongye Wu, James Wu
MDML Final Project, Fall 2018

Introduction

News outlets play a crucial role in the formation of ideologies and particularly political leanings of their audiences. In the U.S. bipartisan context, news media would even go as far as to openly endorse certain parties or candidates over others. Over the past decades, people have increasingly recognized the existence of media bias.[1] According to a Gallup poll, a high of 69% respondents say that news outlet owners attempting to influence the way stories are reported is a major problem, while 66% believe that news organizations are attracting viewers by being too dramatic or sensational. 73% admit that the spread of false information is a major problem today.[2]

Unlike traditional web-based text analysis, our research specifically targets the titles of news articles online and their sentimental and political impacts. We hypothesize that a news outlets may cover events differently, whether their usage of words to draw emotions or just the quantity of coverage on certain topics. Fox News is known to favor Republican candidates and issues, while other liberal sources could likely focus on more progressive agenda and perspectives. There has been plenty of research on news bias and text-based sentiment analysis respectively, but our research bridges the gap between the two by analyzing and quantifying the sentimental impact of news titles.[3] We also built a machine-learning model attempting to predict the political ideology of the news source based on news titles in hope of shedding light on the divisive and divided nature of digital news media.

Framing the Context: a snapshot of the evolution of media bias

The evolution of media bias and the recognition of that bias started with television news. Surrounding highly divisive issues, news media grow to employ "positive" tone towards the party they align with, and "negative" tone towards those on the opposite side. Coverage of U.S. invasion of Iraq in 2003 and the subsequent National Conventions of Bush and Kerry race are among the examples of media influences on how people vote. Swing voters who watched Fox

[1] See Thomas B. Christie, (2009) "*Using the internet for news and perceptions of news organization bias*", Competitiveness Review: An International Business Journal, Vol. 19 Issue: 1, pp.17-25, https://doi.org/10.1108/10595420910929031

[2] Pew Research Center, "*Ideological Placement of Each Source's Audience*", http://www.pewresearch.org/pj_14-10-21_mediapolarization-08-2/, access on 12/13/2018

[3] For research on news bias, see William P. Eveland Jr. and Dhavan V. Shah, *The Impact of Individual and Interpersonal Factors on Perceived News Media Bias*. For traditional sentiment analysis, see E. Cambria, B. White, "*Jumping NLP curves: A review of natural language processing research [review article]*", IEEE Comput. Intell. Mag., vol. 9, no. 2, pp. 48-57, May 2014.

News during that time were more likely to vote for Bush despite voting for Al Gore in the previous election. Interestingly, liberal stations were not able to have the same impact.[4]

Along with the rise of internet since early 2000s, the ways in which people absorb news have transformed into a more immediate, clickbait fashion. Internet news outlets are seen as the most promising segment of the news industry as usage of some traditional news sources decline.[5] Emails, Twitter, Facebook, and even Instagram become just a few in a million ways people can easily access news these days. As our society becomes seriously concerned with the effect of information overload, readers seem to be increasingly engrossed in their respective echo chambers due to algorithm that will personalize their newsfeed with articles that best suit their interests. Whether media becomes more polarizing as a result of the divided audience, or vice versa, is beyond the scope of this paper. But the result is a media sphere more divisive than ever. So on one hand, media in the digital age needs to come up with more sensational titles to combat readers' fatigue from information overload. On the other hand, news need to be in line with the interests, and sometimes fanning the passion, of their audiences to play along the algorithmic game.

Refining Our Research: what can we do to understand media bias manifested through sentimental and political influence

The immediacy and abundance of news articles online suggest that not everyone has the time to read through an article whenever it pops up on their feed. More likely, people are only reading the headlines of the news that are being tailored to create sensational effects. This is why we decided to build our research upon analyzing news titles. We focused on the language in news headlines and quantified the use of certain words related to evoking emotions like fear, sadness, anger, and other inflammatory sentiments within each title. We hypothesize that there will be differences regarding the usage and the types of sentiments based on ideology. The sentiment attributes will then be used to train our predictive model, which will be explained in details in the method section below. We also specifically looked at some of the more salient issues which allowed us to see the differences of sentiments around the more controversial and divisive topics.

In our analysis, we included an array of news sources that represent diverse political leanings: CNN, The New York Times, The Washington Post, The Huffington Post, ABC, Fox News, Breitbart, The American Conservative, and The Hill. To better illustrate the ideological differences among them, we allocated these sources into three categories which eventually become the dependent variable in our predictive model-

1) Liberal: The New York Times, The Washington Post, The Huffington Post

[4] Jonathan S. Morris, "*Slanted Objectivity? Perceived Media Bias, Cable News Exposure, and Political Attitudes",* SOCIAL SCIENCE QUARTERLY, Volume 88, Number 3, September 2007

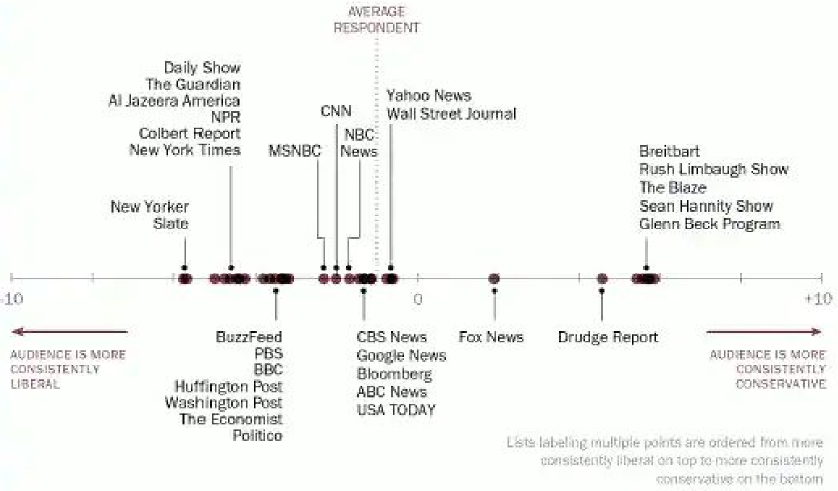
[5] Thomas B. Christie, (2009) "*Using the internet for news and perceptions of news organization bias*", Competitiveness Review: An International Business Journal, Vol. 19 Issue: 1, pp.17-25, https://doi.org/10.1108/10595420910929031

2) Moderate: CNN, ABC, The Hill
3) Conservative: Fox News, Breitbart, The American Conservative

The categorizations were created based on the survey conducted by the Pew Research Center, which took the average of viewers' ratings on all the media outlets and plotted them on a continuum in an effort to delineate the political spectrum reflected in today's media in the U.S.[6]

Ideological Placement of Each Source's Audience

Average ideological placement on a 10-point scale of ideological consistency of those who got news from each source in the past week...



Lists labeling multiple points are ordered from more consistently liberal on top to more consistently conservative on the bottom

American Trends Panel (wave 1). Survey conducted March 19-April 29, 2014. Q22. Based on all web respondents. Ideological consistency based on a scale of 10 political values questions (see About the Survey for more details.) ThinkProgress, DailyKos, Mother Jones, and The Ed Schultz Show are not included in this graphic because audience sample sizes are too small to analyze.

PEW RESEARCH CENTER

Data Collection and Manipulation

Our data was pulled from News API (newsapi.org). The 30 most popular (i.e. "...popular sources and publishers come first" as the site describes it) articles was pulled from the 9 news sources each day from 2018-11-11 UTC to 2018-12-10 UTC when available. Since only the news articles that are less than a month old are free to download, our research is limited to this specific timeframe. The resulting JSON had the following variables:

a. An id for each news source
b. The name of the source
c. The author of article
d. The headline or title
e. A short description from the article
f. The URL of the article

[6] Pew Research Center, "*Ideological Placement of Each Source's Audience*", http://www.pewresearch.org/pj_14-10-21_mediapolarization-08-2/, access on 12/13/2018

g. The URL to a relevant image for the article, if any
h. The date and time that the article was published, in UTC (+000)
i. An unformatted content of the article, where available truncated to 260 characters.

We were able to obtain 7465 observations with the 9 above-mentioned variables from News API. We then removed the mention of news stations' names from the text data to make sure our analysis capture the core content of the news titles.

Due to the limitation in timeframe (1 month), in addition to the nature of mainstream news reporting which usually clusters around the same events, we recognize that our training data is very specific to time- and place-wise, and thus our model might not be applicable in another setting. As a result, we put special emphasis on the process to generate the predictive model, which might have more generalizing power when confronted with new data.

Methods and Techniques

Part 1. Bag-of-words exploration

After collecting the data, we wanted to explore what the most used words in headlines were. Before we broke down all the titles into single words (*unigrams*), two word combinations (*bigrams*), and three word combinations (*trigrams*), we lemmatized the words and removed all stop words so that the remaining words could be blocked together in meaning and had more significance. We counted all of these n-grams and found certain words (e.g. trump, fire, etc.) appeared much more than others, and later on we decided to build them into our classifier as topical features. Additionally, we calculated the tf_idf and document term matrix (dtm) for top unigrams. The resulting words largely overlapped with our topical features, so we decided not to include either tf_idf or dtm in our predictive model.

Part 2. Sentiment Analysis

Since we theorized that there might be some fundamental differences between how news outlets use words for their headlines, the next step was to attempt to quantify it with a sentiment analysis. We used the NRC Word-Emotion Association lexicon, which is a list of English words and their associations with eight basic emotions (*anger, fear, anticipation, trust, surprise, sadness, joy, and disgust*) and two sentiments (*positive* and *negative*) manually done through crowdsourcing, that is available from the tidytext package. We added some words that we saw often in our bag-of-words exploratory analysis that we believed may be useful for our analysis. You can find a table of our additional words in *Appendix 1*.

One of the largest challenges we encountered (and exists in sentiment analysis in general) was the existence of negation words such as *not*, *no*, *n't*, and *never* which could flip an entire sentence sentiment to the opposite spectrum. Our solution was whenever a negation word was used in a title, we would add a *not_* to the next word. For example, *we are not good* would become *we are not not_good*. Therefore, instead of *good* getting tokenized, *not_good* gets tokenized and is matched to the opposite sentiment of *good*.

The counts of each of the 10 sentiments for each title is used as features in our data. We also added frequently talked about topics, such as *mueller*, *migrant*, *fire*, *climate*, *trump*, *border*, and *ocasio-cortez* as binary features. If a title included one of those terms, the feature would represent it with a 1. We added this because we theorize that news outlets with different political ideologies may cover these subjects at different rates and therefore could add to the predictive power of our classifier. We opted to keep the observations with 0s across the board because we thought that these observations may represent less "sensationalized" or more mundane articles which certain news outlets may be more prone in publishing. These constructed features along with the day of week was used as input variables for our Random Forest model.

Part 3. Random Forest

We chose the Random Forest method to conduct classification on the three ideological levels which we constructed to categorize the political leanings of various news sources. We used all the features listed above and 200 bootstrapping replicates (*ntrees* in the *randomForest* command), and all the default settings in the package. To evaluate the model performance, we calculated the accuracy, precision, recall, in addition to the confusion matrix. We did not include the Area Under the Curve (AUC) in our evaluation due to limitation of *ROCR* package on non-binary outcome variable.

Limitations

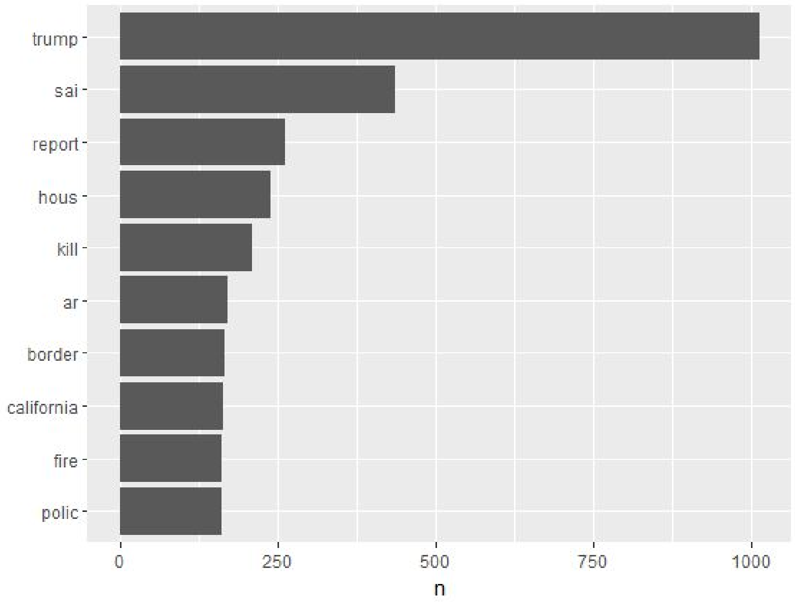
Sentiment analysis is a bag-of-limitations as we implemented it. Besides the fact that many words don't exist in the NRC lexicon, news titles oftentimes employ language that is colloquial, sensationalized, and hard to categorize. Words in a sentence are naturally correlated with each other, by tokenizing it into unigram we lose that interaction. As we saw with negation terms, these interactions could determine the overall sentiment of a sentence. We attempted to minimize some of these opposite sentiment effects by merging in negation words with the word after and adding opposite sentiments in the lexicon but that is far from perfect.

Results

1. Exploratory Analysis

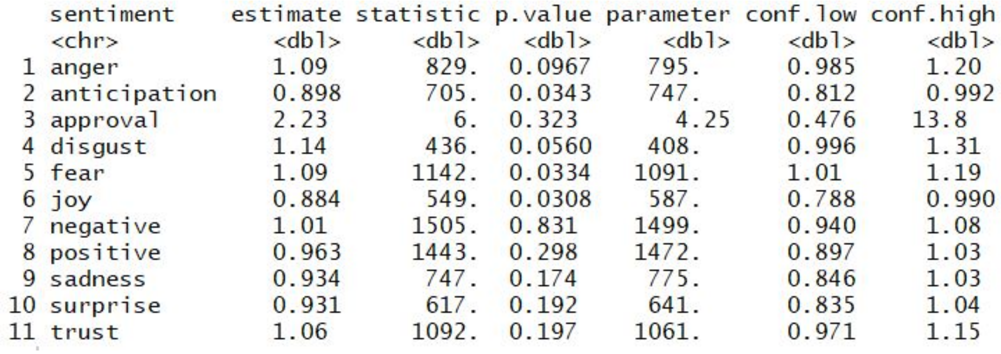
The top 10 most frequent words in titles is shown in Figure 1. Unsurprisingly, Trump is a news maker - he made it in 16.34% of all the headlines. Many of the headlines also quoted other people evidenced by the second most common word used, "said". The rest of the words found seemed to be of topics the particular month was entrenched in (e.g. california fire, kill, police, etc.).

Figure 1. Top 10 Most Frequently Used Words



We ran a Poisson count test (*Figure 2*) to see whether there were any differences in the usage of words between the media we labelled as “conservative” and the media we labelled as “liberal”. There were statistically significant differences between usage of anticipation, fear, and joy words. To interpret, at a 95% CI level, conservative media uses 81.2% to 99.2% times the anticipation words liberal media uses. Similarly at a 95% CI level, conservative media uses 1% to 19% more words associated with the fear sentiment.

Figure 2. Poisson Test between Conservative and Liberal Usage of sentiments



```
   sentiment    estimate statistic p.value parameter conf.low conf.high
   <chr>           <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl>
 1 anger           1.09       829.  0.0967    795.      0.985      1.20
 2 anticipation    0.898      705.  0.0343    747.      0.812      0.992
 3 approval        2.23         6.  0.323       4.25    0.476     13.8
 4 disgust         1.14       436.  0.0560    408.      0.996      1.31
 5 fear            1.09      1142.  0.0334   1091.      1.01       1.19
 6 joy             0.884      549.  0.0308    587.      0.788      0.990
 7 negative        1.01      1505.  0.831    1499.      0.940      1.08
 8 positive        0.963     1443.  0.298    1472.      0.897      1.03
 9 sadness         0.934      747.  0.174     775.      0.846      1.03
10 surprise        0.931      617.  0.192     641.      0.835      1.04
11 trust           1.06      1092.  0.197    1061.      0.971      1.15
```

2. Sentiment Analysis

We conducted sentiment analysis by both news source and ideology (see *Figure 3* and *4*). In general, the news sources try to evoke trust, positivity, negativity, and fear the most out of all the sentiments. This finding is consistent with the nature of news reporting, which most frequently touches on some sort of value judgement or emergency that is related to the language of fear. The American Conservative stood out for using the highest levels of words related to trust, positivity, and negativity. However, we believed this is partly due to the small amount of samples we were able to collect for the American Conservative compared to the other sources.

We saw similar trends in most commonly used sentiments when new sources are broken down by ideology. Moderate and conservative sources, however, tap into the language of fear, anger, and trust more than their liberal counterparts. The conservative side also uses slightly more words evoking disgust than the other two categories. All in all, the results don't differ by much.

Figure 3. Top 10 Sentiments by News Source

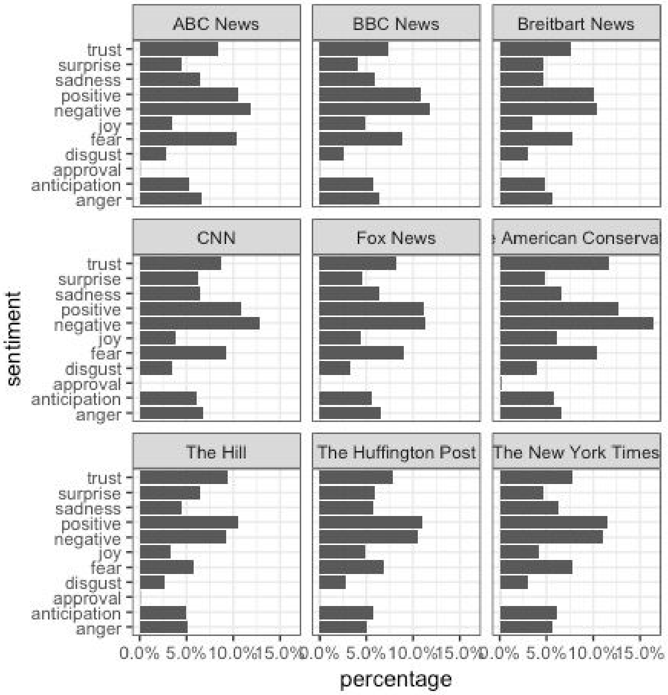
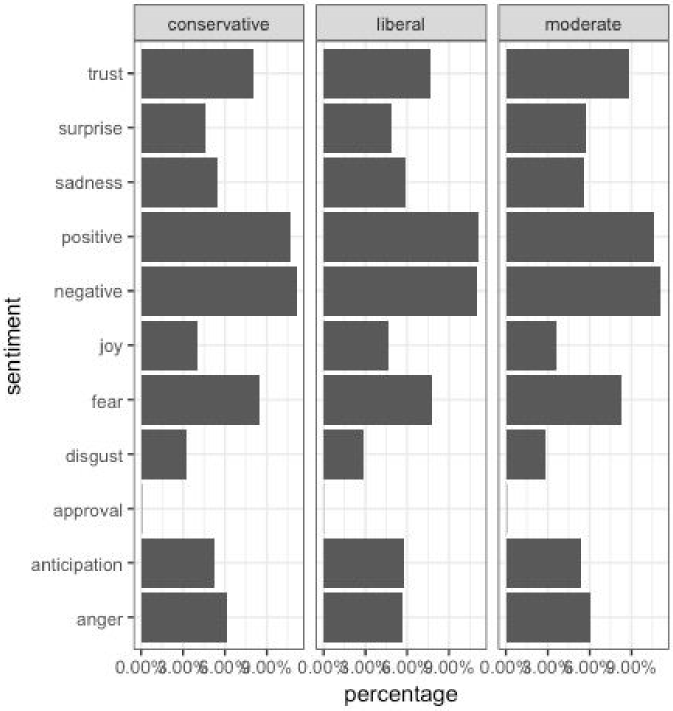


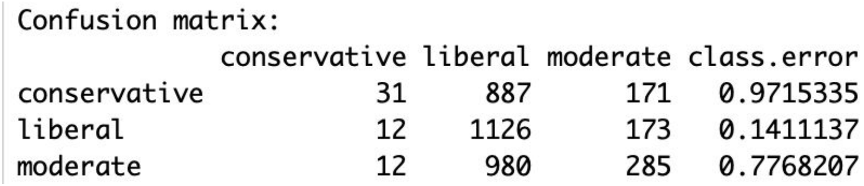
Figure 4. Top 10 Sentiments by Ideology



3. Random Forest Model

The model tried 4 variables at each split, which gave us the following confusion matrix (*Figure 5*). The Random Forest did pretty well in predicting liberal outlets, but did particularly bad in predicting conservative ones.

Figure 5. Random Forest Confusion Matrix



```
Confusion matrix:
             conservative liberal moderate class.error
conservative           31     887      171   0.9715335
liberal                12    1126      173   0.1411137
moderate               12     980      285   0.7768207
```

The accuracy is 73.2% for conservative media, 45.4% for the liberal, and 62.7% the moderate. The high accuracy for conservative and moderate predictions is due to the high level of true negatives that were classified by our model for these two categories, which didn’t tell us much about our model performance. The precision is 45.9% for conservative media, 39.4% for the liberal, and 46.1% for the moderate. Again, the conservative and moderate media were predicted with higher percentage of precision because the numbers of resulting predictions in these two categories are much smaller than the ones in the liberal category- only 55 titles were predicted to be coming from conservative media and 629 from the moderate, whereas 2993 were predicted to be coming from the liberal media. The recall is 3.47% for the conservative media, 85.3% for the liberal, 21.8% for the moderate. This means that out of all the titles from news sources labeled as conservative, only 3.47% of them were correctly classified as so. The recall shows that our model performs pretty well when it comes to classifying titles from liberal media. For the purpose of this research, we want the model to predict as many true positives as possible from each category, which makes recall the most important measure out of the three.

Since the results of the sentiment analysis are fairly similar across ideology, we think either the topical features and/or the interaction between topical and sentiment features make our model especially good at predicting liberal sources. Unfortunately, we are not able to further delineate the relations between the variables due to the limited interpretability of Random Forest techniques.

Lessons For Future Research

We believe our classifier can be calibrated by adding more relevant features. According to the categorizations done by the Pew Research center, true moderate news outlets (score of 0) doesn't exist, so it may be possible that there is some overlap between the languages used by moderate media with media on either side of the spectrum. If the categories of the outcome variable have characteristics that are very distinct from each other, the classification method might have yielded better performance. One extra step we could take is to look at only the polarized opposites, liberal and conservative media. We found that there were statistically significant differences in the number of times conservative and liberal media used *anticipation*, *fear*, and *joy* vocabularies.

In addition, we could further refine our lexicon to include words commonly used in news headlines. This may improve both our features for the classifier and our sentiment analysis results. In order to make our results more generalizable, we will need to pull more data than a month's worth. It may also be better to focus the research to specific topics, as the topic perception between the different news outlets may show to have greater variance than just the most popular headlines.

Implications

The Social Identity Theory developed by Henri Tajfel caution us of potential impact of one's group membership in the process of media use for social identity gratification.[7] Group

[7] Mei-Chen Lin, Paul M. Haridakis, and Gary Hanson, "*The Role of Political Identity and Media Selection on Perceptions of Hostile Media Bias During the 2012 Presidential Campaign*", 2016 Broadcast Education Association Journal of Broadcasting & Electronic Media 60(3), 2016, pp. 425–447, DOI: 10.1080/08838151.2016.1203316

memberships affect our attitude towards and perceptions of others within and outside those groups. If media intentionally taps into and instigates hostile and negative sentiments against the opposing (political) groups, then the audiences are likely to become more divided and less understanding of different viewpoints.

This is why research regarding online information and its impact should have a very special place in today's discussion. Our research introduces the ways to apply sentiment analysis and machine learning method around online news content, despite some obvious data limitations. It also illustrates the potential of more fruitful explorations in the future by learning the lessons of engineering features and building predictive models. What we have accomplished is just a small drop in the ocean of natural language processing research, with the hope that our elementary efforts can contribute to the general knowledge of this very field.

Appendix

Appendix 1: Extra words added to the NRC lexicon

Word	Sentiment
Die	Sadness
Dying	Sadness
Death	Sadness
Died	Sadness
Dying	Sadness