New York Times Articles and Comments:
Text Sentiment Mining and Political Content Prediction
May 8, 2022

GitHub

The following link takes you to our GitHub repository containing the code and data described in the report below: https://github.com/julia-ulziisaikhan/nyt-comments.

Introduction

Are political topics of discussion more conducive to polarization and opposing reactions? We are interested in learning whether the sentiment makeup of comments on New York Times articles can be used to predict whether an article's topic is political, or non-political. The total number of comments on a given article were categorized into three categories of sentiment: positive, neutral, and negative. We used the proportions of comments which fell into the three sentiments in order to engineer a polarization score, a metric aimed to gauge how divisive in user opinion an article's comment section is. We then created a statistical model to predict whether an article is political or not, based on the comments' sentiment analysis and their polarization score.

We identify our audience of the report to be three groups of interest. First, political scientists can use this report to understand to what extent, if any, politics affect the divisiveness of public opinion. Are democratic citizens more united on political or non-political topics? This report would help answer that question of political theory. The second audience would be the financial analysts at the New York Times, as well as other news media companies. It would be valuable for the analysts to understand which articles will receive the most engagement, and thus, price advertising on their pages accordingly. And lastly, we believe our report will provide value to journalists seeking to improve their public image who may want to attract less negative sentiment in the comment body.

The potential for data snooping may be present in two instances. First, in determining whether a comment had positive, negative, or neutral sentiment, we had to establish an arbitrary cut-off as the unigrams, or tokens (text unit of one word length), each had a point value ranging from -5 to 5. We believe this may be a source of data snooping, as who's to say whether a neutral comment should be allowed to have a total sum ranging from -2 to 2, as compared to strictly 0? If a comment's sum was -1 or 1, there's a possibility that we could interpret it as neutral, but because it is not zero, it will get assigned negative or positive. The second potential source for data snooping is the development of our own lexicon of political words, which may be too subjective. Since we each have our own biases, we may exclude words that we may not consider political but others might. For example, topics related to native reservations and oil pipelines may be considered political, but may have not been counted. Another example of this, is that terms relating to reproductive health were not included, leaving the lexicon to be vulnerable to the bias of its creator.

Background

We focus entirely on New York Times article data, and its subsequent comment data. Only New York Times subscribers are able to submit a comment, and some articles prevent users from commenting entirely. We are not sure what percentage of articles allow commenting. According to the NYT website, the comment sections are moderated by humans, such that comments must be individually approved for publication. They claim to not tolerate profanity, attacks, and the like, yet in our analysis we found derogatory curse words in the comments. Comments should add "intelligent and informed commentary that enhances the quality of [their] news and information" (New York Times, 2022). In addition, readers and moderators have the ability to express support for a given comment, through recommendations and Editor's selection, respectively.

A shortcoming of this project is that the readership of the New York Times tends to skew left, older in age, higher in income, more urban in residence, and whiter (Kafka, 2022). As such, whatever findings which may arise cannot be generalized to the U.S. population.

While NYT provides a free API for their comment data, we decided to prioritize our time on improving our algorithms, as such, we opted for already retrieved and scraped comment and article data located on a Kaggle competition, an online community of data scientists and machine learning engineers (Kesarwani, 2018). We understand that the project guidelines insisted on retrieving the data from the original source, because the data would be more robust and up-to-date if we had done so, as we are limited to glimpses of 2017 and 2018 (see: COVID-19 and related news, 2021 U.S. Capitol Insurrection, the U.S. protests of the Summer of 2020, 2022 Invasion of Ukraine, the possible overturning of Roe v. Wade in 2022). In addition, we are not well-versed in Python (querying is done in Python), nor were the available R packages dedicated to the NYT Article Search API fruitful, as they did not readily carry comment data as seen in the Kaggle source.

Methodology

We used R to create all models, figures, and analyses mentioned henceforth. Prior to any text analysis, text data was converted to lowercase, and stripped of punctuation, numbers, special characters, and stopwords, words like 'but' which carries no semantic meaning by its own.

The methodology of this project can be thought of in three main parts.

First, we performed sentiment analysis on multiple levels: at the token level, comment level, and article comment body level. Sentiment analysis is normally used, and was created for, the classification of customer reviews for businesses and apps, but we implemented this style of text analysis into a political context. The lexicon we used for sentiment analysis was "afinn", a popular lexicon used for sentiment analysis developed by Finn Årup Nielsen. Each word in the lexicon has a certain score ranging from -5 to 5, where words of negative sentiment are assigned negative values, words of neutral sentiment are set to zero, and positive sentiment words have positive values. The table below provides a few examples of words of varying sentiment value.

word	**sentiment score**
fraud	-4
deficit	-2
support	2
excellent	3

At the token level, we assigned a sentiment value to each individual word of a given comment body. Then, to determine a comment's overall sentiment, we summed the sentiment values of each word in the given comment. The comment is considered to be of neutral sentiment if the sum of all words' scores were equal to 0, negative sentiment if the sum was a negative number, and positive sentiment if the sum was a positive. This takes us to the final level of sentiment analysis, and to the format of the final processed dataframe, sentiment analysis on an article level. For each article, we created variables *neg*, *pos*, and *neu*, calculated from the proportion of comments of negative sentiment out of the total number of comments. The sum of *neg*, *pos*, and *neu* would thus always be 1.

Next, we engineered the *polarization* score, from the aforementioned mined text data metrics, out of motivation to capture polarization of opinion to a given news article. We utilized sentiment analysis in order to create the polarization score. The polarization score informs us on how divisive an article is. In theory, a highly polarizing article would have a similar share of both negative and positive comments.

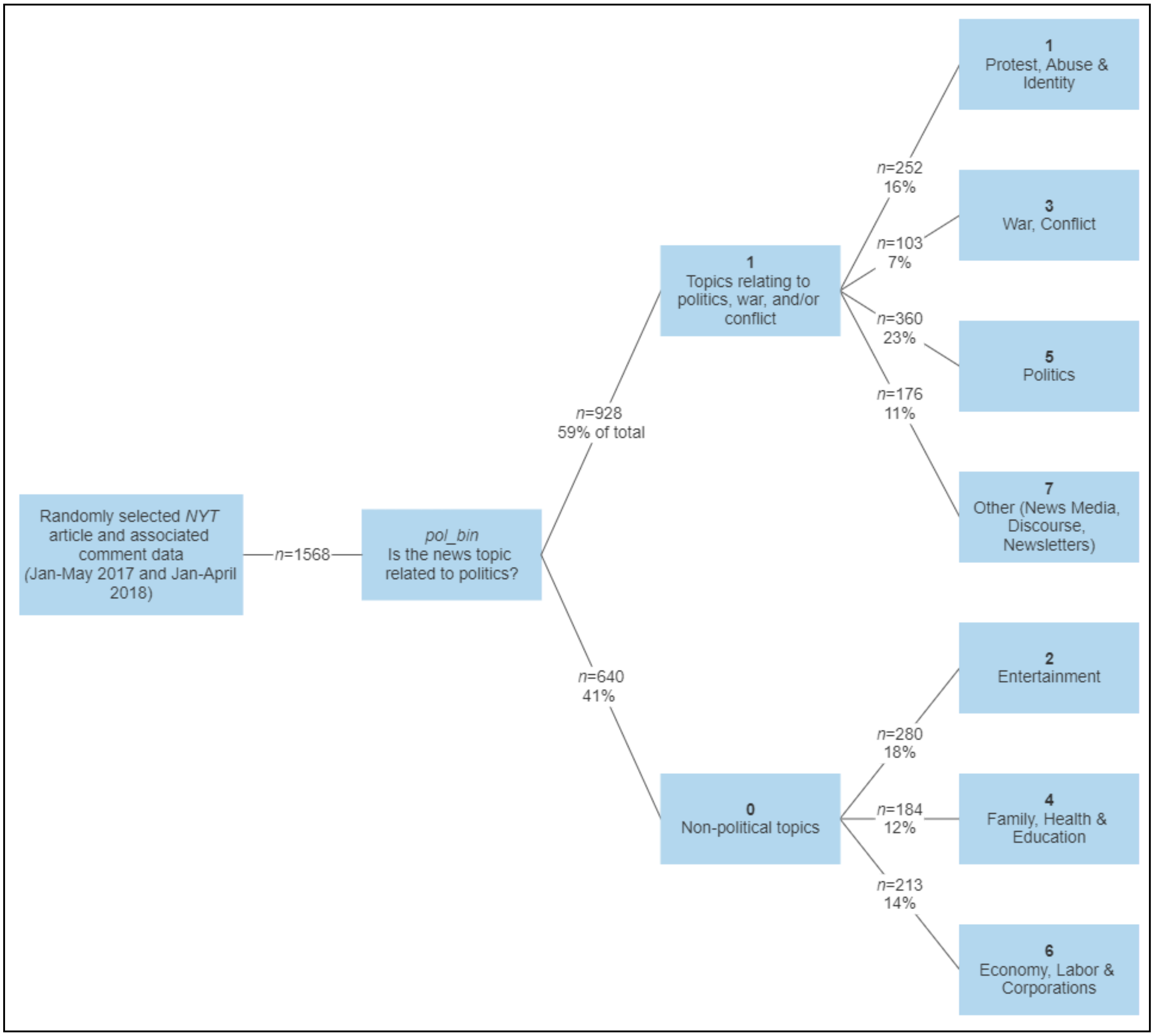
The equation for the polarization score is as follows:

$$polarization \ = \ (\frac{negative}{positive} - 1) + \ neutral$$

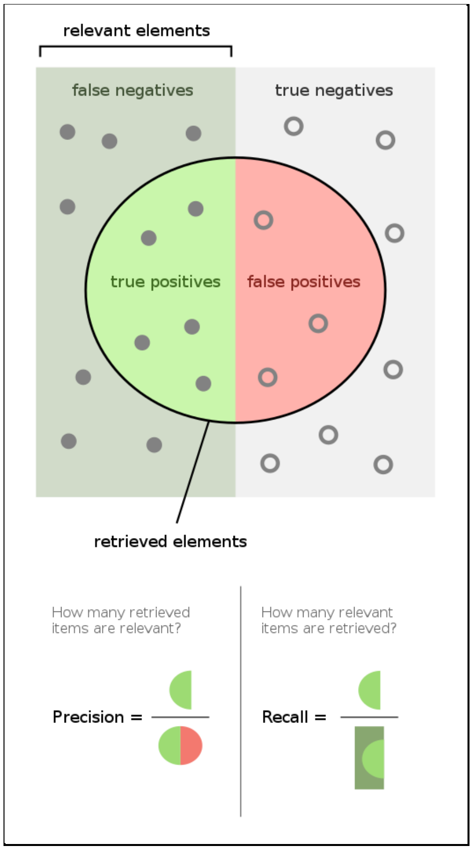
Where *negative* is the proportion of comments with negative sentiment in a given article. Similarly, *positive* is the proportion of comments with positive sentiment and *neutral* is the proportion of comments with neutral sentiment.

An article would be considered polarized when there is a, roughly, equally, high proportion of negative and positive sentiment, in addition to a low proportion of neutral sentiment. So, we expect the polarization score to approach 0, the more polarizing the article is. Be mindful of this: it may seem intuitive that a higher polarization score would mean higher polarization, but we actually formulated the equation so that the closer the score is to 0, the more polarized the article is. This is because the first term in the equation $\frac{negative}{positive}$ would be close to 1, then we subtract 1, then add $neutral$. The polarization score is penalized with a higher proportion of neutral sentiment comments.

The second part of the methodology is devoted to categorizing the news articles as either political or non-political in content. It is a binary categorization task that we have automated through all seven thousand articles. We created a political content lexicon, a list of roughly 100 words of sure fire indication of political content. NYT provides a *keywords* variable, containing keywords to the vast majority of articles (we have lost some rows of articles due to a missing value in *keywords*), so we coded the article as political (*political* = 1) if any of the keywords were contained in the lexicon. If not, the article was filed as non-political (*political* = 0). Some examples of the words in our political content lexicon were ‘politics and government’, ‘international relations’, and ‘trump’. In order to assess the accuracy of this binary categorization, we had manually coded 1,568 articles as either political or non-political, and the proportion of political articles in this sample was 59%. A flow chart of this categorization, and subcategorization, can be found below.


Randomly selected *NYT* article and associated comment data (Jan-May 2017 and Jan-April 2018)
n=1568
pol_bin Is the news topic related to politics?
n=928 59% of total
n=640 41%
1 Topics relating to politics, war, and/or conflict
0 Non-political topics
n=252 16%
n=103 7%
n=360 23%
n=176 11%
1 Protest, Abuse & Identity
3 War, Conflict
5 Politics
7 Other (News Media, Discourse, Newsletters)
n=280 18%
n=184 12%
n=213 14%
2 Entertainment
4 Family, Health & Education
6 Economy, Labor & Corporations

To our great surprise, the automated categorization, to a sample roughly six times larger, yielded a whopping similar *political* proportion of 58.8%. From this, we concluded that our binary categorization had run smoothly with little to no error, and from this, we can proceed on with the confidence in categorization. After having done so, we proceeded to analyze how the sentiment and polarization metrics differed between political and non-political articles. We also have an additional binary marker variable indicating whether the article was posted as an 'Op-Ed' or not. We decided to include this variable as op-eds are, by definition, opinionated and potentially more controversial than a non-op-ed. In sum, this section of methodology is devoted to finding interesting insights about polarization and sentiment among articles of political content and op-ed style, otherwise non-obvious when looking at the un-processed data.



Lastly, we created several regression models, in order to predict whether an article was political or not. By feeding the learning models important information about the engineered sentiment and polarization variables, in addition to article word count, the models are able to learn what differentiates a political article from a non-political one based on those variables, or predictors. We decided to do logistic regression, a model specifically for binary output data, with linear underworkings, and lasso regression, a model that can smartly remove variables that do not help in accurate prediction. We generated metrics for each method's predictions, including random guessing based on the 0.59 political proportion, which were: (1) Classification error, the proportion of articles misclassified; (2) Precision, the number of true positives divided by the sum of true positives and false positives; and (3) Recall , the number of true positives divided by the sum of true positives and false negatives. We aim for a low classification error, and a high precision and recall. This diagram may help illustrate these metrics.

We have also interpreted the coefficients of the aforementioned models, as they may give us insight as to which variables were valuable in predicting political content or not. We also ran the most successful model from above, on a subset of data containing non-op-ed articles. We do so in order to further illustrate the point that our results from our mining is not due to chance, but due to a real pattern in the data. Op-eds are, by nature, more opinionated and likely to attract polarization. By excluding op-eds from our data, we can see that the models still predict political content with adequate accuracy.

Data

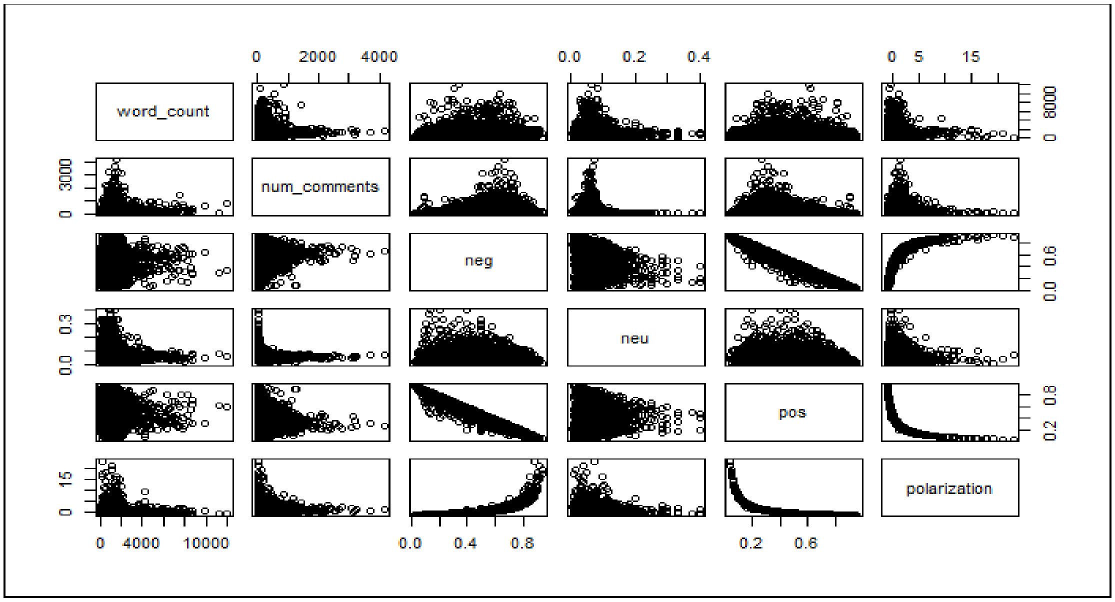
The first part of this section will describe our data and its collection process in more detail, and its relation to our research question. The second part of this section will spell out our exploratory analysis of the data, with figures and tables, to get you familiar with what we are working with.

The timeline of the data spans nine months in total, January through May of 2017, and January through April of 2018. There are 7,808 articles which contain comment data. After processing the data through our algorithms, rows which contained missing data for our necessary variables were omitted, thus reducing the article count to 6,724. In the comment data, the variables of interest are the body text of the comment, and the article it is commenting on, which we can trace to *article_id*, so that joining of the two datatables can be performed. From a theoretical standpoint, the text body of the comment represents the democratic citizen's opinion of the news media. We aim to measure the polarization of mass opinion, by analyzing and summing the sentiment of an article's comment body collection. In the article data, the variables of interest are *article_id*, *article_word_count*, *oped* (is the article an op-ed?), and *keywords*. *Keywords* provide us insight into the content of the article.

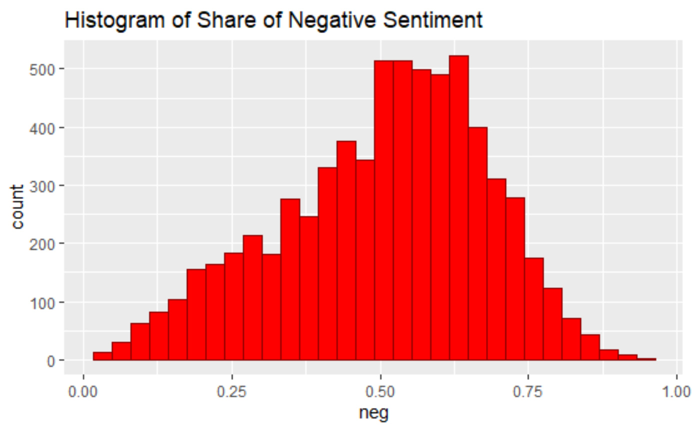
You can download the data for this project from Kaggle here. We randomly selected 1,568 articles and categorized them using our political lexicon, and found that 59% of that sample contained political articles, as described earlier in the Methodology section.

Instead of using the New York Times API, we used the Kaggle data because of time constraints and difficulties in retrieving data using the API. We attempted to use AashitaK's Python package, but it appears that the New York Times API was updated since the creation of the package, as it ran into an error on Python. Since we are not familiar with Python, we could not come up with a solution and looked at solutions using R. There are two R packages dedicated to the New York Times API, but we found that we could not easily obtain comments data using them. Additionally, there is a query limit for the New York Times API of 1,000 articles a day. We decided to prioritize improving our algorithm rather than trying to obtain more data. We made our algorithms and lexicon to be reproducible using the most recent data. For example, our political lexicon includes words related to the current presidency and on-going conflicts across the world, such as in Yemen and Ukraine. We are confident that if you run our code through more recent, but similarly formatted data, the results will be similar.

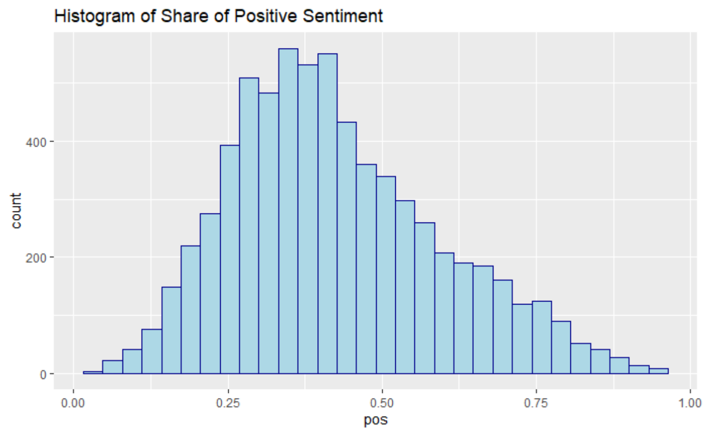
Second, we performed an exploratory data analysis (EDA) of the data. The RMarkdown file contains a more thorough walkthrough of the EDA.



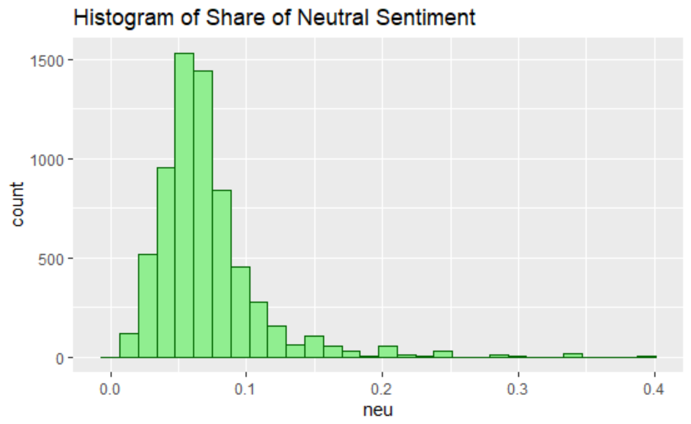
First, we ran a correlation plot in order to catch any unexpected collinearity between variables. The continuous variables do not appear to be correlated, apart from the obviously related variables *neg, neu*, and *pos*. One interesting thing to note is that the correlation between *num_comments* and *neu* suggests that only the articles with the most comments had very little to no proportion of neutral sentiment comments.



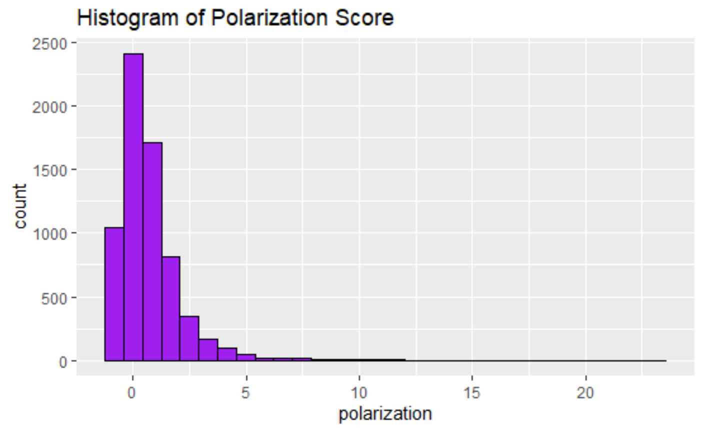
Histogram of *neg* values. A wide normal, almost uniform distribution.



Histogram of *pos* values. A wide normal, almost uniform distribution.

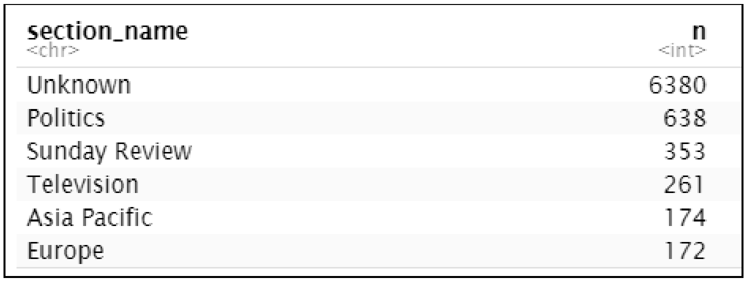


Histogram of *neu* values. Right-tailed distribution.



Histogram of *polarization* values. Right-tailed distribution.

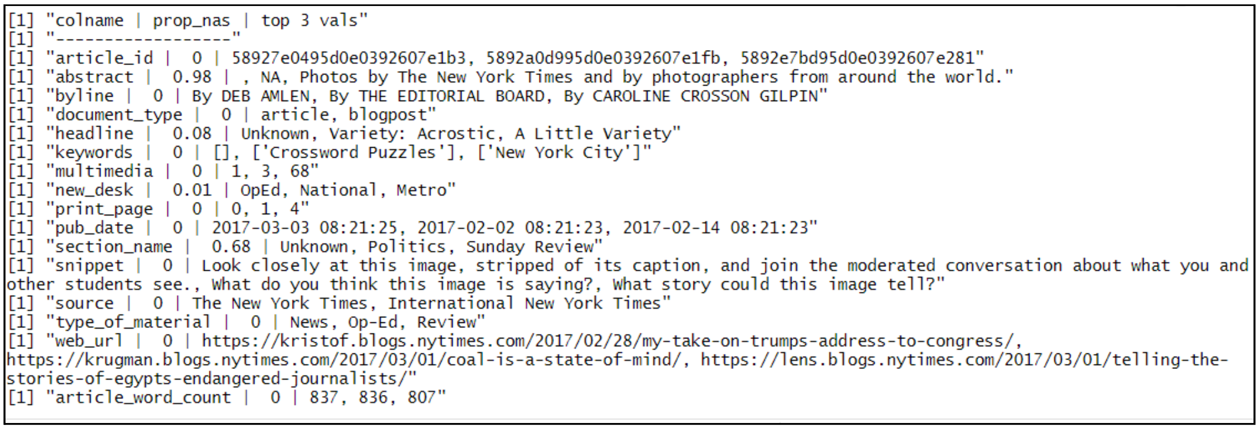
Here we see a histogram for *neg, pos*, *neu*, and *polarization*. *Neg* roughly averages around 0.6, *pos* at 0.3, and *neu* tightly averages around 0.05. This suggests that neutral comments make up very little of the comment section, on average, while negative comments make up around two thirds of the section, on average. The distribution for *polarization* is extremely right-skewed, with values going up as high as 20, but the mode seems to be a little bit above 1. This suggests that most polarization values are around 0, but a good amount go up to 1 or 2.



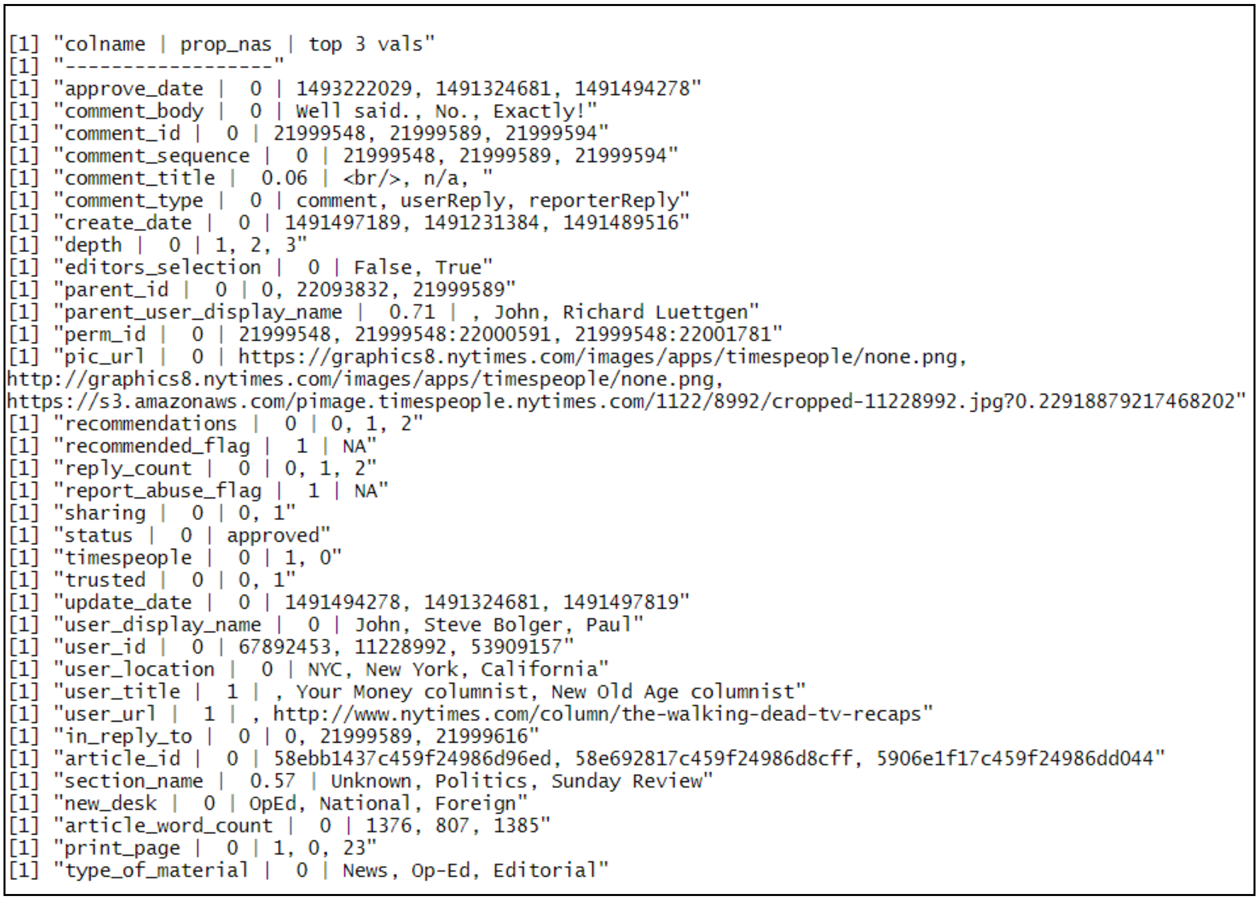
section_name <chr>	n <int>
Unknown	6380
Politics	638
Sunday Review	353
Television	261
Asia Pacific	174
Europe	172

The median article word count is 1,000, with significant outliers. The data does not suffer from missing data issues in our variables of interest, apart from the *section_name* field, in which roughly 80% of the fields are marked as 'Unknown'.

Below are two summaries, for comments data and articles data, respectively, containing the proportion of missing values for each of the columns/variables in the dataset, and the top 3 most common values.



```
[1] "colname | prop_nas | top 3 vals"
[1] "------------------"
[1] "article_id |  0 | 58927e0495d0e0392607e1b3, 5892a0d995d0e0392607e1fb, 5892e7bd95d0e0392607e281"
[1] "abstract |  0.98 | , NA, Photos by The New York Times and by photographers from around the world."
[1] "byline |  0 | By DEB AMLEN, By THE EDITORIAL BOARD, By CAROLINE CROSSON GILPIN"
[1] "document_type |  0 | article, blogpost"
[1] "headline |  0.08 | Unknown, Variety: Acrostic, A Little Variety"
[1] "keywords |  0 | [], ['Crossword Puzzles'], ['New York City']"
[1] "multimedia |  0 | 1, 3, 68"
[1] "new_desk |  0.01 | OpEd, National, Metro"
[1] "print_page |  0 | 0, 1, 4"
[1] "pub_date |  0 | 2017-03-03 08:21:25, 2017-02-02 08:21:23, 2017-02-14 08:21:23"
[1] "section_name |  0.68 | Unknown, Politics, Sunday Review"
[1] "snippet |  0 | Look closely at this image, stripped of its caption, and join the moderated conversation about what you and
other students see., What do you think this image is saying?, What story could this image tell?"
[1] "source |  0 | The New York Times, International New York Times"
[1] "type_of_material |  0 | News, Op-Ed, Review"
[1] "web_url |  0 | https://kristof.blogs.nytimes.com/2017/02/28/my-take-on-trumps-address-to-congress/,
https://krugman.blogs.nytimes.com/2017/03/01/coal-is-a-state-of-mind/, https://lens.blogs.nytimes.com/2017/03/01/telling-the-
stories-of-egypts-endangered-journalists/"
[1] "article_word_count |  0 | 837, 836, 807"
```



```
[1] "colname | prop_nas | top 3 vals"
[1] "------------------"
[1] "approve_date |  0 | 1493222029, 1491324681, 1491494278"
[1] "comment_body |  0 | Well said., No., Exactly!"
[1] "comment_id |  0 | 21999548, 21999589, 21999594"
[1] "comment_sequence |  0 | 21999548, 21999589, 21999594"
[1] "comment_title |  0.06 | <br/>, n/a, "
[1] "comment_type |  0 | comment, userReply, reporterReply"
[1] "create_date |  0 | 1491497189, 1491231384, 1491489516"
[1] "depth |  0 | 1, 2, 3"
[1] "editors_selection |  0 | False, True"
[1] "parent_id |  0 | 0, 22093832, 21999589"
[1] "parent_user_display_name |  0.71 | , John, Richard Luettgen"
[1] "perm_id |  0 | 21999548, 21999548:22000591, 21999548:22001781"
[1] "pic_url |  0 | https://graphics8.nytimes.com/images/apps/timespeople/none.png,
http://graphics8.nytimes.com/images/apps/timespeople/none.png,
https://s3.amazonaws.com/pimage.timespeople.nytimes.com/1122/8992/cropped-11228992.jpg?0.2291887921746820 2"
[1] "recommendations |  0 | 0, 1, 2"
[1] "recommended_flag |  1 | NA"
[1] "reply_count |  0 | 0, 1, 2"
[1] "report_abuse_flag |  1 | NA"
[1] "sharing |  0 | 0, 1"
[1] "status |  0 | approved"
[1] "timespeople |  0 | 1, 0"
[1] "trusted |  0 | 0, 1"
[1] "update_date |  0 | 1491494278, 1491324681, 1491497819"
[1] "user_display_name |  0 | John, Steve Bolger, Paul"
[1] "user_id |  0 | 67892453, 11228992, 53909157"
[1] "user_location |  0 | NYC, New York, California"
[1] "user_title |  1 | , Your Money columnist, New Old Age columnist"
[1] "user_url |  1 | , http://www.nytimes.com/column/the-walking-dead-tv-recaps"
[1] "in_reply_to |  0 | 0, 21999589, 21999616"
[1] "article_id |  0 | 58ebb1437c459f24986d96ed, 58e692817c459f24986d8cff, 5906e1f17c459f24986dd044"
[1] "section_name |  0.57 | Unknown, Politics, Sunday Review"
[1] "new_desk |  0 | OpEd, National, Foreign"
[1] "article_word_count |  0 | 1376, 807, 1385"
[1] "print_page |  0 | 1, 0, 23"
[1] "type_of_material |  0 | News, Op-Ed, Editorial"
```

Of all 6,724 articles, 24.46% of them are op-eds, and 75.54% aren't. In addition, 58.79% of them are political articles, while 41.21% aren't. These proportions are not too surprising, I would have expected roughly half of news articles to be politically related, and a decent amount of articles being op-eds.

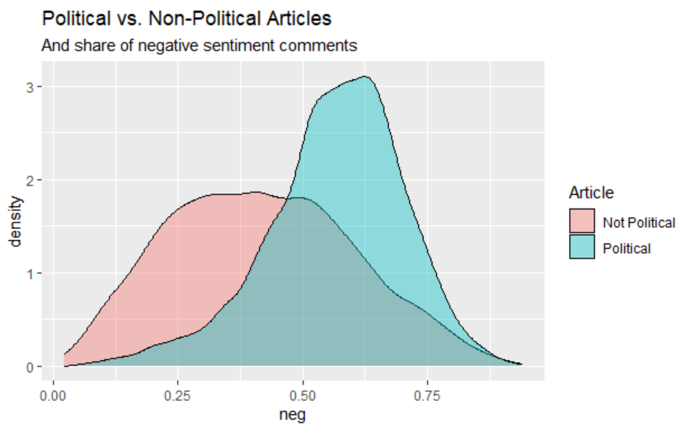
Among political articles, 19.66% are op-eds, 39.13% aren't. Among non-political articles, 4.8% are op-eds, while 36.41% aren't. We can see that a greater proportion of political articles constitue all op-eds, which is as predicted.

Results

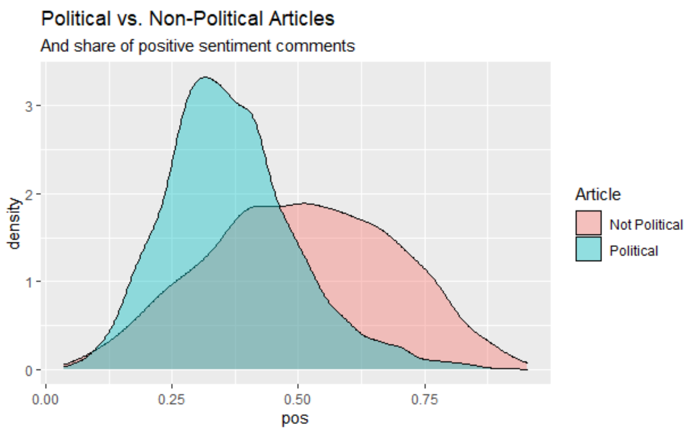
Here we evaluate the usefulness of the sentiment analysis features. We believe it was useful, especially for negative sentiment, because it helped us understand how sentiments varied between political and non-political articles. It was also used to engineer our polarization scores. The polarization score was useful in seeing how divisive the articles were. In the histogram for *polarization* scores, we saw that many political articles were near 0. This means that a good proportion of articles were indeed having polarized opinions in the comments. All in all, these features were used to create the predictive model, which we could have not done with the data alone. Now let us quantitatively determine whether sentiment analysis proved worthy.

Below are density plots, and in different colors are political (teal color) and non-political articles (salmon color) represented. Density plots are merely another form of the all too familiar histogram, its merit is that it can be more visually readable, as it smooths out outliers and extreme skewing.

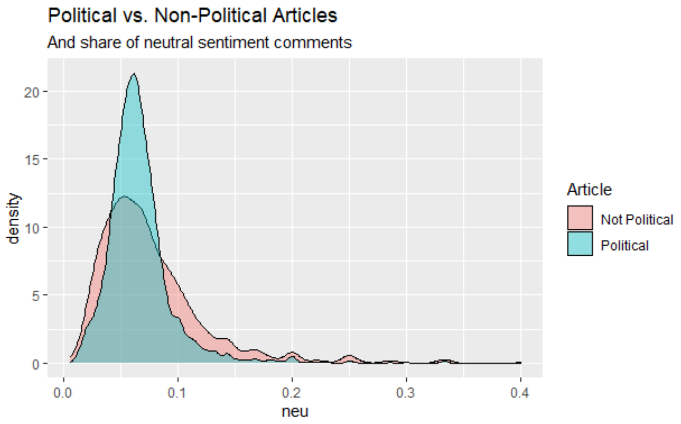
The first plot has *neg*, the proportion of negative comments, plotted on the x-axis. We can see that the distributions for political and non-political articles differ greatly, where political articles appear to have a higher negative makeup, on average. The opposite can be said for *pos*, the proportion of positive comments, as seen in the plot to the right. Political articles seem to have a tighter distribution around its mean, which makes sense, as non-political articles can vary in polarization, much more than political articles can. The density plot for *neu* is quite interesting, there appears to be no significant difference in the proportion of neutral comments between political and non-political articles. And finally, the bottom-right plot depicts a density plot of the polarization score. The result is really strange. While we predicted for political articles to have a lower polarization score, it is actually non-political articles who seem to have a higher polarization score, on average. Next, we will conduct significance tests to make sure the trends we see in the graphs are somewhat real.



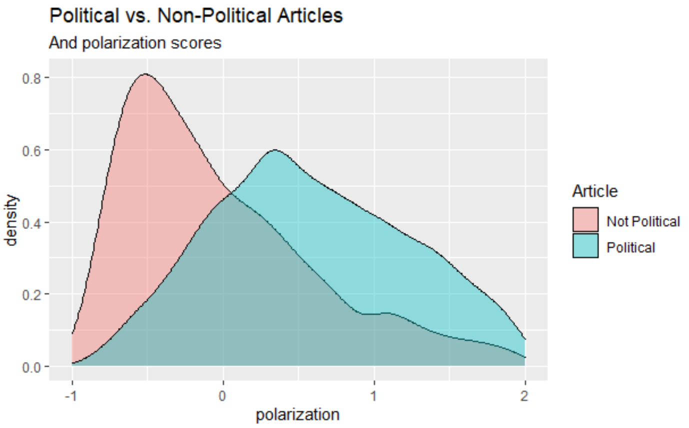
Density plot of negative sentiment comments in political and non-political articles.



Density plot of positive sentiment comments in political and non-political articles.



Density plot of neutral sentiment comments in political and non-political articles.

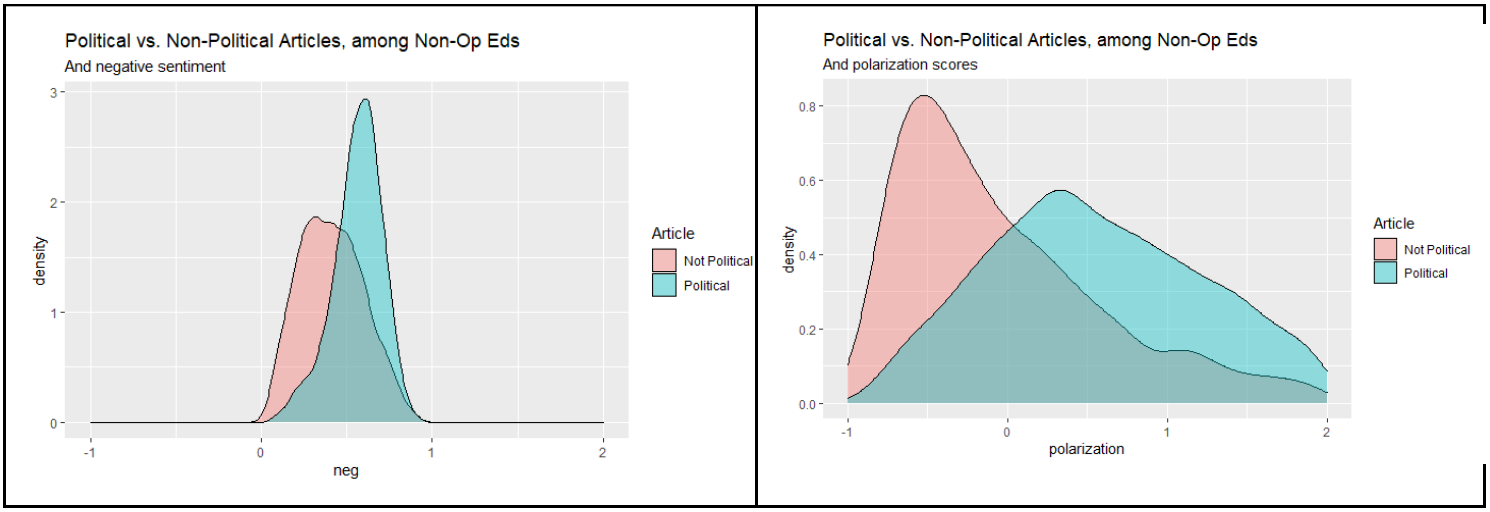


Density plot of polarization scores in political and non-political articles.

We conducted t-tests between political and non-political articles for the values of *neg, neu, pos,* and *polarization*. For *neg*, the mean difference was 0.15, where political articles had a larger mean (t=35.98, df=4833.2, p=0**). For *neu*, the mean difference was -0.01, where non-political articles had the larger mean (t=-9.22, df=4466.2, p= 0**). For *pos*, the mean difference was -0.14 and non-political articles had a higher mean (t= -33.98, df=4669.9, p= 0**). For *polarization*, the mean difference was 0.70 and political articles had a higher mean (t= 18.68, df=5778.6, p= 0**). Negative sentiment and less polarization was greater among political articles, positive sentiment was greater among non-political articles, and neutral sentiment was almost equal, but still significantly different, between both. We then conducted t-tests for *polarization* between op-eds and non-op-eds, the mean difference was 0.14 where op-eds had a higher mean, indicating that they were less polarizing (t= 4.39, df=3501.1, p= 0**). Then, another one for *polarization* between political and non-political articles among op-eds, where the mean difference was 0.61 and political articles had the larger mean, indicating they were less polarizing (t= 8.24, df=518.07, p= 0**). The t-test results for the polarization metric were unexpected, because we would have thought that political articles would have been more polarizing. We have reason to believe that this may have resulted from the '-1' term from our formula, as the vast majority of the polarization scores for the non-political articles (in all subsections of the data) are negative.

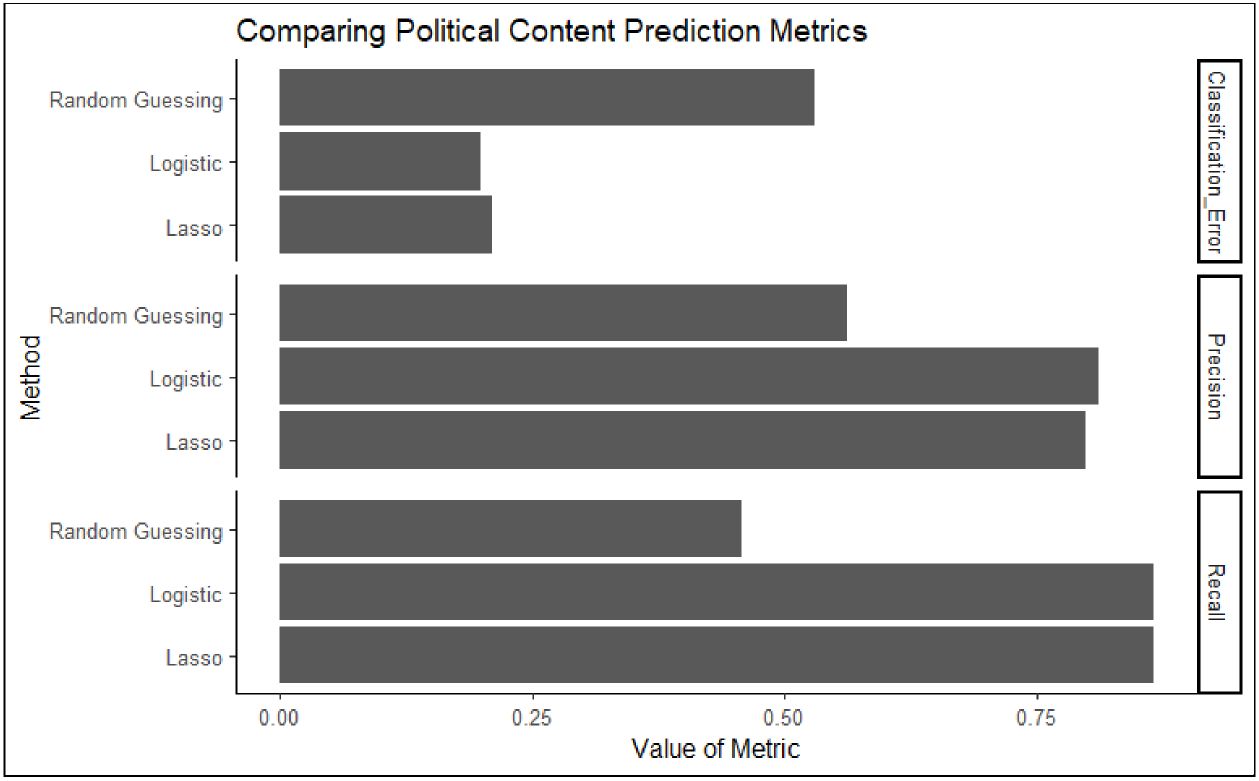
In an effort to show that the results from the sentiment analysis is likely due to a real pattern, we have subsetted the data to exclude all Op-Ed articles. Our reasoning is that, if it is the Op-Ed quality that is solely driving sentiment and polarization, then in the subsetted data, we will see no difference in sentiment and polarization among political and non-political articles. If we still do see a difference, then we can conclude that we are, in fact, seeing a real sentiment pattern in the data among political topics and non-political topics.

Below are two density plots of the non-Op-Ed data. First, we see that the means of distributions for *neg* are considerably closer together as compared to the original data, but we still see a noticeable difference where political articles have higher negative sentiment. A t-test showed that the mean difference was 0.73 where the political articles had the larger mean (t= 16.42, df=5040.2, p= 0**). Second, the density plot for *polarization* is quite similar as it looked when we had all the data, where the political articles have a higher, but more varied, central tendency as compared to non-political. This is still strange, because we would expect political articles to fare as low as 0, but they hover around 0.3 or 0.4.

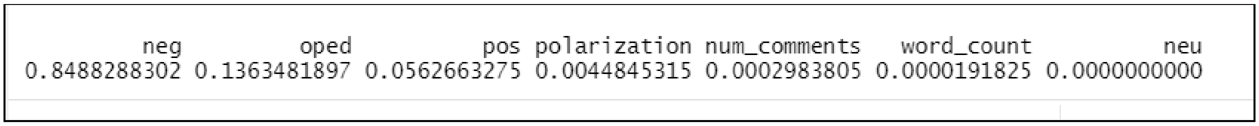


Now, we discuss the performance of the models in predicting binary *political* values. We reserve 100 random rows to serve as the test set, while the remaining 6624 are the training set. Below are how well the predictions fared.

Method	**Classification error**	**Precision**	**Recall**
Random guessing (prop=0.59)	0.53	0.56	0.46
Logistic Regression	0.2	0.81	0.86
Lasso Regression	0.21	0.8	0.86

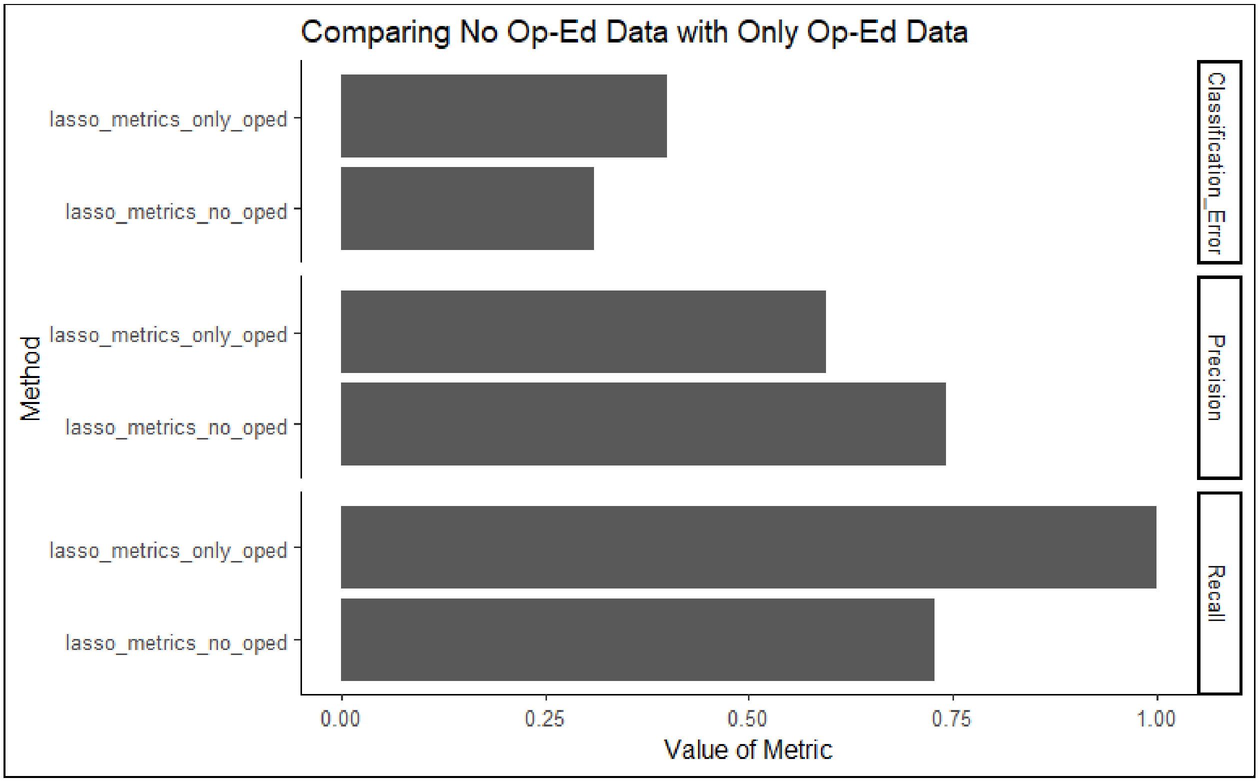


As compared to a dumb model (a model that randomly guesses, based on the proportion), the predictions from logistic regression and lasso regression were fairly similar and accurate. The classification error for both predictions was around 20%, and its precision and recall were around 80% and 86%, respectively. While these predictions were not near perfect, we conclude that the models could predict with good accuracy whether an article was political based only on sentiment and polarization variables of its comment makeup. After tuning the hyperparameter for the lasso model to the best possible one, we find that the model automatically shrunk the *neu* variable to 0. We can conclude that this means the model did not find the proportion of neutral comments in an article's comment section to be useful for predicting if it is a political article. Here is displayed the model's coefficients in descending importance. We can see that the proportion of negative comments and *oped* were the two most important variables for this model's predictions. Surprisingly, the coefficient for *polarization* is quite small, at nearly 0. This is interesting, as I thought the variable would be more important.



```
         neg         oped          pos polarization num_comments   word_count          neu 
0.8488288302 0.1363481897 0.0562663275 0.0044845315 0.0002983805 0.0000191825 0.0000000000 
```

Now, I will display the same performance metrics, after running lasso regression predictions on a subset of the data that only includes op-eds, and on a subset of data that excludes op-eds. The results are quite interesting. The metrics are considerably worse when we do not inform the model about whether the article is an op-ed or not, as the classification error climbs up to around 30%. Interestingly, the recall for only-op-ed data is really good, near 100%.



Conclusion

We have now arrived at the end of our report. We initially began the project with a great interest to derive meaningful insights from comment sentiment on New York Times articles, given that the data is readily available. One of us is a Political Science major, so we were also greatly motivated to tackle a data project from a political science context, specifically in trying to materialize an abstract concept such as polarization into a quantitative and analyze-able metric.

In sum, we found that political articles, as compared to non-political articles, yielded higher proportions of negative comments, lower proportions of positive comments, and interestingly, lower polarization. We also found that Op-Eds saw similar trends, that they tended to have more negative comments, fewer positive comments, and less polarization than non-Op-Eds. After experimenting with subsetting the data in various ways, we concluded that the patterns we saw in the data did not occur due to chance, and that comment sentiment alone is truly able to predict whether an article is political or not. As for binary prediction, our logistic and lasso regression models predicted political articles with good or decent accuracy, with ~20% classification error, ~80% precision, and ~86% recall.

The following paragraph discusses the polarization metric further. The results were the opposite of what we expected: they showed us that political and Op-Ed articles tended to have lower polarization than non-political and non-Op-Ed articles, respectively. We believe this may have risen out of poor metric construction, if not a true reflection of the data. In coming up with this equation, we considered: (1) instead of adding, multiplying *neutral*; and (2) somehow penalizing the score when there is a low number of total comments, so that articles with both a high number of comments and high polarization would receive the lowest score possible. However, we decided to stick with a simpler formula: first, because we were worried about the mathematical implications of a multiplicative effect, and second, *polarization* would be highly correlated with *num_comments* (number of comments on the article), which may be redundant when training the data to a shrinkage model.

Looking to the future, we believe it would be useful to identify sub-categories of topics within political and non-political articles. If successfully executed, we could see which specific political topics bring the most negative sentiment, as well as which topics are the most polarizing, which are extremely useful insights for political science scholars and journalists.

The bias towards NYT subscribers, and thus, towards more affluent demographics, is undeniable as users are only allowed to comment if they pay the monthly subscription. I believe it will be very interesting if one can manage to scrape the text bodies of tweets containing a NYT article URL (URL data is provided by the API) or in response to a tweet containing that URL. I believe exploring that option will be useful, as it will include a wider range of people, including non-White people and more financially diverse people. The same can be said for comments on Facebook posts containing the NYT URL.

Critique

We critiqued our Ailene Torres' project. Her motivation for her project was to determine if student demographics and Regents test scores were important factors in a school receiving a passing rating for student achievement. She used two datasets from NYC Open Data and included data from 2014 to 2018. The data mining aspect of this project was 'mining' a passing or not passing rating for each high school. Something that we would have done differently would be to include data from other New York State schools, because the exam is not only taken by city high schoolers, but also those who live across the state. We would have also included data from middle schools in order to provide more variety and data robustness. Including elementary and middle schools could provide more insight on what impacts a school's student achievement rating, especially since these schools don't partake in the Regents exams.