Online News Popularity Analysis

Alice Huang

10/12/2022

Part 1

Objective

In the age of the Internet, many people get their news from articles shared on social media. Social media and online news websites disseminate information around the world at a speed never seen before in human history. Sharing news articles on social media helps raise awareness and change public opinion about important social, economic and political issues (Bhagat & Kim, 2022), and helps instigate discourse around how we can make our world a better place. As people spend more and more time on social media feeds curated towards their individual interests, browsing news articles shared on these feeds helps shape their worldviews. However, there have been increasing concerns that frequent use of social media and consumption of negative news content may be associated with negative mental health outcomes (Davey, 2020).

Given concerns about the role of online news media in our society, it is of interest to determine how the emotions of texts, subjectivity of texts, and other measures play a role in how viral articles become. In this paper, we attempt to build a regression model that predicts the number of shares based on various features of news articles. We attempt to identify which features are the most significant in modelling the number of article shares.

Data Collection

This dataset was downloaded from Kaggle but was originally featured in Fernandes et al (2015) paper. The creators of the dataset pulled all the articles published on Mashable from January 7 2013 to January 7 2015. Mashable is a popular digital media, online news, and entertainment company. They post news articles covering entertainment, tech, world issues, and more. Fernandes et al measured various features of each article, for example, the subjectivity of the title, the rate of negative words in the text, and so on. Each row in the dataset corresponds to a news article from Mashable and its various features. The response variable is the number of times an online news article posted on Mashable was shared via Facebook, Twitter, Google+, LinkedIn, Stumble-Upon and/or Pinterest.

The original dataset had 58 variables. Since we are primarily interested in polarity and subjectivity, we only consider the 16 polarity and subjectivity related explanatory variables. Polarity is a measure of how negative or positive the emotions conveyed in a text are. Polarity is measured using a score that takes on values in the interval [-1,1]. A text with polarity -1 is considered to evoke strong negative emotions. A text with polarity 0 is considered neutral. A text with polarity 1 is considered to evoke strong positive emotions.

Subjectivity is a measure of how much a text reflects personal opinion/feelings/beliefs. In contrast, objectivity measures how factual a text is. Subjectivity is measured on a score that takes on values in the interval [0,1]. 0 is considered very objective and 1 is considered very subjective. Values close to 0 indicate more objective texts, and values close to 1 indicate more subjective/opinionated texts.

For example, the sentence "I won the lottery" would have low subjectivity and high positive polarity. The sentence "I think pineapple on pizza is nasty" would have high subjectivity and low negative polarity.

To compute the subjectivity and polarity sentiment analysis, the researchers adopted the Pattern web mining module (Smedt et al, 2014). Here positive words refer to words that evoke positive emotions in the reader, such as joy, excitement, and hope. Negative words refer to words that evoke negative emotions in the reader, such as sadness, fear, and anxiety. We use "words" and "tokens" interchangeably.

Here are the potential explanatory variables we consider in this paper:

- `global_subjectivity`: Text subjectivity (float)
- `global_sentiment_polarity`: Text sentiment polarity (float)
- `global_rate_positive_words`: Rate of positive words across all words in the content (float)
- `global_rate_negative_words`: Rate of negative words across all words in the content (float)
- `rate_positive_words`: Rate of positive words among non-neutral tokens (float)
- `rate_negative_words`: Rate of negative words among non-neutral tokens (float)
- `avg_positive_polarity`: Average polarity of positive words (float)
- `min_positive_polarity`: Minimum of positive words' polarity scores (float)
- `max_positive_polarity`: Maximum polarity of positive words (float)
- `avg_negative_polarity`: Average polarity of negative words (float)
- `min_negative_polarity`: Minimum polarity of negative words (float)
- `max_negative_polarity`: Maximum polarity of negative words (float)
- `title_subjectivity`: Title subjectivity (float)
- `title_sentiment_polarity`: Title polarity (float)
- `abs_title_subjectivity`: Absolute value of subjectivity level minus 0.5 (float)
- `abs_title_sentiment_polarity`: Absolute value of polarity level (float)

Data Processing

Since the step() function in R only runs on dataset with maximum size 5000, we randomly sample 5000 news articles from articles published in 2014 instead of using the original dataset which had 39644 observations.

Upon further inspection of the dataset, it appears that several hundred articles had all the in-text polarity and subjectivity related variables entered as 0. After visiting the links of those articles and reading the texts, we noticed those articles did contain positive and negative words, so labelling those articles as having a rate of 0 positive words out of the entire text did not seem appropriate. For example, one article entitled "500 Migrant Workers Feared Dead After Human Traffickers Ram Their Boat" (url: https://mashable.com/archive/500-migrant-workers-feared-dead) contained many negative words, so the global rate of negative words should not have been 0. Thus I removed the observations containing mostly 0's from consideration, as those were likely data entry errors.

Preliminary Description of Data

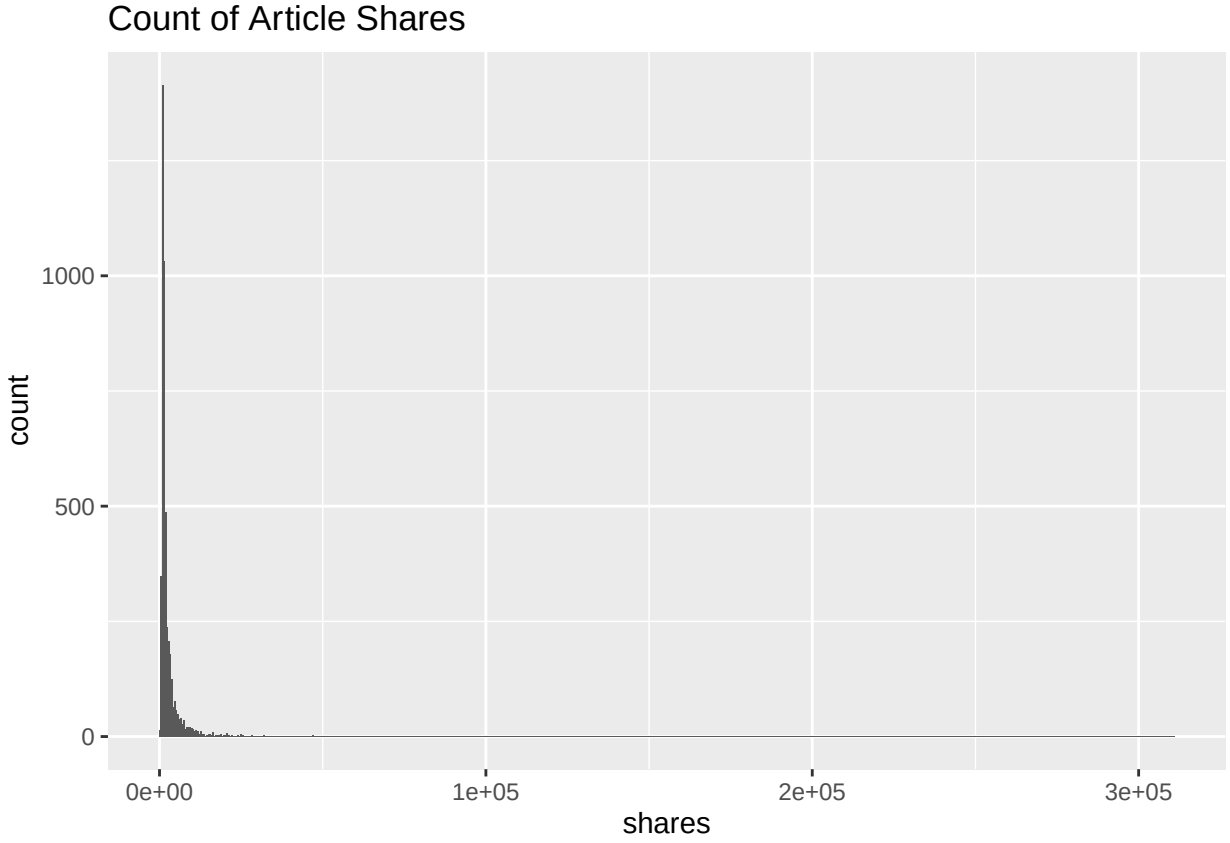
The cleaned dataset has 4771 observations, each corresponding to a news article. We can see that the range for the number of times an article was shared is very large, with the minimum value being 22 and the maximum value being 310800. In contrast, the ranges for all the explanatory variables are within [-1,1]. All the variables take on positive values, except the ones involving negative polarity and polarity in general. Transformation of the variables may be needed, so that the response and explanatory variables are on more similar scales.

Summary Plots of Response Variable

The response variable, the number of times an article was shared, seems to follow a right-skewed distribution with a very long right tail. From the summary statistic, we can see that the mean (3221) is much larger

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
global_subjectivity	0.08949	0.39723	0.45364	0.45352	0.50760	1.00000
global_sentiment_polarity	-0.38021	0.05192	0.11017	0.11020	0.16743	0.61389
global_rate_positive_words	0.00000	0.02725	0.03696	0.03827	0.04753	0.13158
global_rate_negative_words	0.00000	0.01077	0.01616	0.01749	0.02232	0.10112
rate_positive_words	0.0000	0.5909	0.6988	0.6837	0.7857	1.0000
rate_negative_words	0.0000	0.2143	0.3012	0.3163	0.4091	1.0000
avg_positive_polarity	0.0000	0.3096	0.3586	0.3617	0.4080	1.0000
min_positive_polarity	0.00000	0.05000	0.10000	0.09645	0.10000	1.00000
max_positive_polarity	0.0000	0.6000	0.8000	0.7734	1.0000	1.0000
avg_negative_polarity	-1.0000	-0.3361	-0.2646	-0.2742	-0.2000	0.0000
max_negative_polarity	-1.0000	-0.1250	-0.1000	-0.1091	-0.0500	0.0000
min_negative_polarity	-1.0000	-0.8000	-0.5000	-0.5663	-0.4000	0.0000
title_subjectivity	0.0000	0.0000	0.1667	0.2844	0.5000	1.0000
title_sentiment_polarity	-1.00000	0.00000	0.00000	0.06301	0.13636	1.00000
abs_title_subjectivity	0.0000	0.1667	0.5000	0.3411	0.5000	0.5000
abs_title_sentiment_polarity	0.0000	0.0000	0.0000	0.1559	0.2500	1.0000
shares	22	968	1400	3221	2600	310800

than the median (1400). This makes sense as there were a small proportion of articles that went viral and were shared tens of thousands of times, but 75% of articles were shared less than 2600 times.



Non-technical Summary

We tried fitting various regression models (Linear, log-linear, Poisson, Quasi-Poisson, Negative Binomial) to predict the number of times an article is shared based on the previously described features measuring polarity and subjectivity. See Appendix for codes and summaries of the models. It turned out that the Negative Binomial model gave the best fit. From the summary of our best negative binomial model, it appears that the most significant features in predicting how many times articles get shared in a year were `global_subjectivity`, `global_rate_positive_words`, `global_rate_negative_words`, `min_positive_polarity`, `max_positive_polarity`, `max_negative_polarity`, `title_subjectivity`, and `avg_positive_polarity:max_positive_polarity`. It appeared that articles with subjective titles and texts got shared significantly more often than more objective articles. Perhaps they were controversial, and more likely to elicit emotional responses from readers. As expected, articles with high polarity were shared more often. Positive articles got shared somewhat more often than neutral articles. Highly negative articles got shared significantly more often than neutral articles. Perhaps highly negative articles are more shocking and induce more fear in readers. They may depict rare tragedies of historical importance so readers may feel greater need to share with their social network.

Part 2

Correlation of Variables

First we explore the correlation between different pairs of variables that may be related. This dataset had lots of variables that were related to each other, as they were measuring polarity and subjectivity in different ways. For example, the creators of the dataset measured `global_rate_positive_words`, the ratio of positive words to all words in the text (positive, negative, neutral), and also `rate_positive_words`, the ratio of positive words to just the non-neutral words in the text. We examined the correlation matrix for the explanatory variables and made scatterplots for the pairs with correlations greater than 0.5. See Appendix for more code and plots. After creating scatterplots for pairs of explanatory variables, we believe the following pairs are moderately or strongly correlated, and suggest ways to account for their correlations while building regression models:

- `rate_positive_words` , `rate_negative_words`
  - Rate of positive words is 1 minus rate of negative words.
  - Pick one variable to include in the model but not both.
- `rate_positive_words`, `global_rate_positive_words`
  - Both involve counting the number of positive words in the text.
  - Pick one, but not both.
- `rate_negative_words`, `global_rate_negative_words`
  - Both involve counting the number of negative words in the text.
  - Pick one but not both.
- `global_sentiment_polarity`, `global_subjectivity`
  - Consider interactions.
- `title_sentiment_polarity`, `title_subjectivity`
  - Consider interactions.
- `abs_title_sentiment_polarity`, `title_subjectivity`
  - Consider interactions.
- `title_subjectivity`, `abs_title_subjectivity`

  - `abs_title_subjectivity` is the absolute value of `title_subjectivity`'s distance from 0.5.
  - Pick one but not both.

- `title_sentiment_polarity`, `abs_title_sentiment_polarity`
  - `abs_title_sentiment_polarity` is the absolute value of the `title_sentiment_polarity`.
  - Pick one but not both.

- `global_rate_positive_words`, `global_sentiment_polarity`
  - Polarity measures how positive or negative a text is so it involves measuring positive words.
  - Pick one but not both.

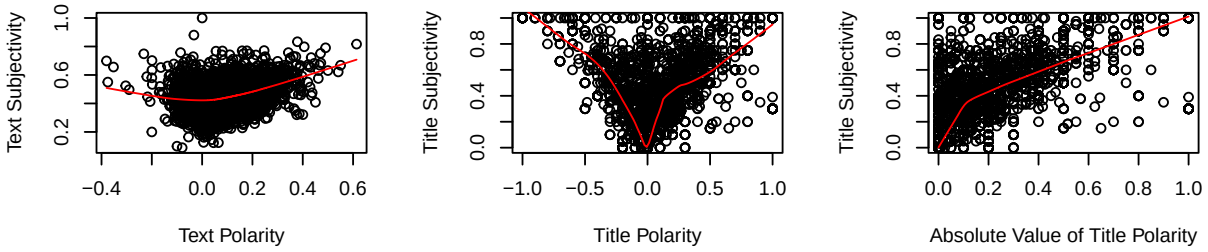
- `global_rate_negative_words`, `global_sentiment_polarity`
  - Polarity measures how positive or negative a text is so it involves measuring negative words.
  - Pick one but not both.

- `avg_positive_polarity`, `max_positive_polarity`
  - Consider interactions.

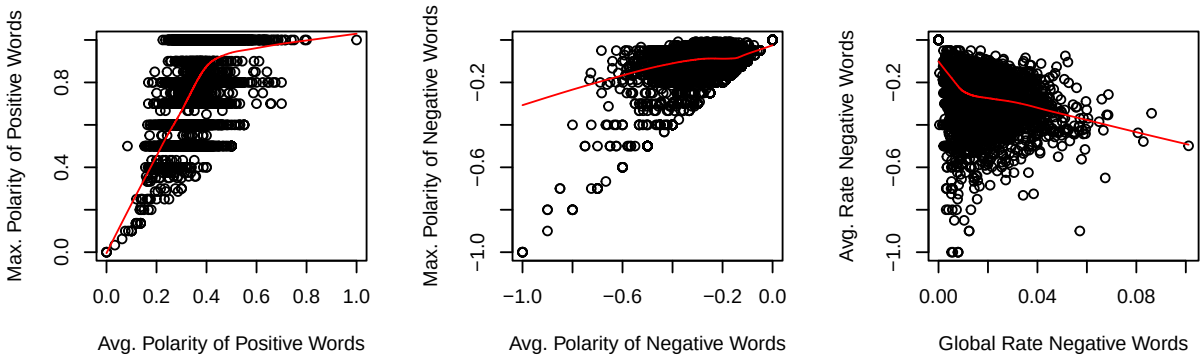
- `avg_negative_polarity`, `max_negative_polarity`
  - Consider interactions.

- `global_rate_negative_words`, `avg_negative_polarity`
  - Consider interactions.

There did not appear to be strong correlations in the other pairs of variables. The variables measuring text polarity and text subjectivity seemed to be weakly correlated. One may guess that very opinionated pieces may have more emotionally charged words, but the scatterplot did not seem to suggest strong evidence for such findings. Interestingly, there was strong correlation between absolute value of polarity and subjectivity in the title rather than the text. Perhaps the text had a lot of noise, due to being significantly longer than the title.



There was some moderate correlation between the average polarity of positive words and the maximum polarity of positive words. It may be necessary to consider interactions between `avg_positive_polarity` and `max_positive_polarity`. It seems like most articles have the polarity of negative words being around (-0.4, 0), so there aren't many articles that are very negative. There was some moderate correlation between the average polarity of negative words and the maximum polarity of negative words. It may be necessary to consider interactions between `avg_negative_polarity` and `max_negative_polarity`.



Trying Poisson Regression

Given the shape of the histogram for the number of article shares, we guessed that the number of times an article was shared in a year may follow a Poisson regression. We tried fitting a Poisson regression on the number of shares and running stepAIC to get the Poisson regression model with the best fit (see this model's summary below). However, it turns out a Poisson regression fit is unsuitable because the mean and the variance are drastically different, as evidenced by the Residual deviance being 25333822 on 4753 degrees of freedom. Also the rate at which people share articles throughout a year may not be constant, because people are less likely to share old articles. Here we use an identity link rather than a log link because the number of shares was collected, so the response data is not binary. A quasi-Poisson model also yielded similar high Residual deviance. See Appendix for more code and model outputs.

```
##                                                    Estimate  Std. Error
## (Intercept)                                      1485.7640     10.8889
## global_subjectivity                              5099.8806     13.9785
## global_sentiment_polarity                       14759.9282     52.6513
## global_rate_positive_words                      -9128.3863     93.8797
## global_rate_negative_words                      54046.2078    225.3397
## avg_positive_polarity                           -9872.6784     36.1187
## min_positive_polarity                            3323.9689     16.0289
## max_positive_polarity                           -3333.2901     15.0525
## avg_negative_polarity                           -1986.7521     19.7414
## max_negative_polarity                            -616.8472     25.3008
## min_negative_polarity                             382.3171      5.5063
## title_subjectivity                                787.6058      2.7134
## title_sentiment_polarity                          764.2252      8.8199
## global_subjectivity:global_sentiment_polarity   -18513.1758     87.1200
## title_subjectivity:title_sentiment_polarity       -857.3756     12.0296
## avg_positive_polarity:max_positive_polarity     11735.5869     41.9706
## avg_negative_polarity:max_negative_polarity      2330.8556     42.8310
## global_rate_negative_words:avg_negative_polarity 24174.7747    698.1495
##                                                   z value  Pr(>|z|)
## (Intercept)                                       136.448 < 2.2e-16
## global_subjectivity                               364.838 < 2.2e-16
## global_sentiment_polarity                         280.334 < 2.2e-16
## global_rate_positive_words                        -97.235 < 2.2e-16
## global_rate_negative_words                        239.843 < 2.2e-16
## avg_positive_polarity                            -273.340 < 2.2e-16
```

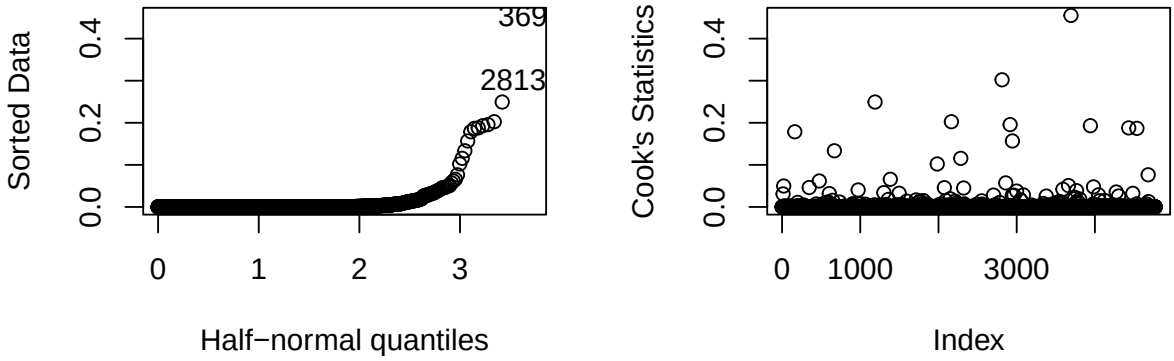
```
## min_positive_polarity                                207.373 < 2.2e-16
## max_positive_polarity                               -221.444 < 2.2e-16
## avg_negative_polarity                               -100.639 < 2.2e-16
## max_negative_polarity                                -24.381 < 2.2e-16
## min_negative_polarity                                 69.433 < 2.2e-16
## title_subjectivity                                   290.264 < 2.2e-16
## title_sentiment_polarity                              86.647 < 2.2e-16
## global_subjectivity:global_sentiment_polarity       -212.502 < 2.2e-16
## title_subjectivity:title_sentiment_polarity          -71.272 < 2.2e-16
## avg_positive_polarity:max_positive_polarity          279.614 < 2.2e-16
## avg_negative_polarity:max_negative_polarity           54.420 < 2.2e-16
## global_rate_negative_words:avg_negative_polarity      34.627 < 2.2e-16
##
## n = 4771 p = 18
## Deviance = 25333821.73781 Null Deviance = 26495939.63305 (Difference = 1162117.89524)
```

Trying a Negative Binomial Regression

We consider a negative binomial regression, which is similar to the Poisson regression but the variance is not required to be equal to the mean. The model assumes the dispersion parameter takes on the same value at all parameter values. We ran stepAIC to get the negative binomial regression model with the best fit. We first tried a log link, then an identity link. The residual deviance and AIC for the model with the log link were lower than the negative binomial model with the identity link, so we kept the negative binomial model with the log link.

Removing Outliers

We use Cook statistics plots and half-normal jackknife residual plots (as described in Extending the Linear Model with R, Faraway), to check if there are outliers with unusually high influence and leverage over the model's fit.



It seems like some observations, including observations 2813, 3694 have very high Cook Statistics compared to the rest, and they don't seem to follow the trend on the half-normal jackknife residual plots. They seem to be outliers. If we remove them, most coefficients change significantly and the residual deviance decreases.

Statistical Summary

Here is the output of our best negative binomial regression model.

```
##
## Call:
## MASS::glm.nb(formula = shares ~ global_subjectivity + global_rate_positive_words +
##     global_rate_negative_words + avg_positive_polarity + min_positive_polarity +
##     max_positive_polarity + max_negative_polarity + min_negative_polarity +
##     title_subjectivity + title_sentiment_polarity + title_subjectivity:title_sentiment_polarity +
##     avg_positive_polarity:max_positive_polarity, data = data2014sampled2,
##     init.theta = 1.037160862, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6879  -0.9723  -0.6716  -0.1420   9.3011
##
## Coefficients:
##                                             Estimate Std. Error z value
## (Intercept)                                 7.165260   0.191992  37.321
## global_subjectivity                         1.228266   0.190662   6.442
## global_rate_positive_words                  3.212161   1.079196   2.976
## global_rate_negative_words                  2.863379   1.609873   1.779
## avg_positive_polarity                      -0.676314   0.611498  -1.106
## min_positive_polarity                       0.625652   0.266896   2.344
## max_positive_polarity                      -0.398468   0.252063  -1.581
## max_negative_polarity                      -0.635489   0.162382  -3.914
## min_negative_polarity                       0.002051   0.060089   0.034
## title_subjectivity                          0.192911   0.045840   4.208
## title_sentiment_polarity                    0.162245   0.144695   1.121
## title_subjectivity:title_sentiment_polarity -0.116890   0.193175  -0.605
## avg_positive_polarity:max_positive_polarity  1.692944   0.697849   2.426
##                                               Pr(>|z|)
## (Intercept)                                  < 2e-16 ***
## global_subjectivity                         1.18e-10 ***
## global_rate_positive_words                   0.00292 **
## global_rate_negative_words                   0.07530 .
## avg_positive_polarity                        0.26873
## min_positive_polarity                        0.01907 *
## max_positive_polarity                        0.11392
## max_negative_polarity                       9.09e-05 ***
## min_negative_polarity                        0.97277
## title_subjectivity                          2.57e-05 ***
## title_sentiment_polarity                     0.26216
## title_subjectivity:title_sentiment_polarity  0.54511
## avg_positive_polarity:max_positive_polarity  0.01527 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0372) family taken to be 1)
##
##     Null deviance: 5753.6  on 4768  degrees of freedom
## Residual deviance: 5480.6  on 4756  degrees of freedom
## AIC: 86081
```

```
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0372
##           Std. Err.:  0.0188
##
##  2 x log-likelihood:  -86053.4720
```

The residual deviance of 5479.7 and the AIC of 86045 are still high, but they are still much better than those of the Poisson model, which had residual deviance of 25333822 on 4753 degrees of freedom, and AIC 25378267.

From the summary of our best negative binomial model, it appears that the most significant features in predicting how many times articles get shared in a year were `global_subjectivity`, `global_rate_positive_words`, `global_rate_negative_words`, `min_positive_polarity`, `max_positive_polarity`, `max_negative_polarity`, `title_subjectivity`, and `avg_positive_polarity:max_positive_polarity`.

`title_subjectivity` was one of the most significant features in predicting the number of shares. Perhaps articles with titles featuring controversial opinions grabbed readers attention more quickly and readers were more likely to click on them and share them on social media. `global_subjectivity` was another subjectivity-related feature that was significant in predicting shares. Perhaps readers find subjective articles more fun to read, especially if they include controversial opinions. In general, it appears that subjective articles get shared more often.

It appeared that various features measuring positive polarity and the rate of positive words out of the entire text were somewhat significant in modelling the number of shares. Perhaps some Mashable users, especially those who are younger, like to browse Mashable for entertainment, so they prefer sharing funny, light-hearted articles. An older user browsing a more serious, established news site may display different behaviour. Overall, it seems that positive articles get shared somewhat more often than neutral articles.

Interestingly, `max_negative_polarity` was highly significant, but `avg_negative_polarity` and `min_negative_polarity` were not. Note that `max_negative_polarity` is measured on a scale from -1 to 0, so having a more negative article would mean having the negative polarity decrease to -1. Thus a negative coefficient for `max_negative_polarity` would correspond with an increased number of shares on average. Perhaps articles with very negative words are more shocking and induce more fear in readers. Perhaps they may depict rare tragedies of historical importance (e.g. a pandemic), so they get shared more often. Readers may believe it's their responsibility to share frightening, shocking, tragic news with their social network.

Appendix

References

Bhagat, S. & Kim, D. J. (2022). Examining users' news sharing behaviour on social media: role of perception of online civic engagement and dual social influences. Behaviour & Information Technology. DOI: 10.1080/0144929X.2022.2066019

Davey, G. C. L. (2020, September 21). The psychological impact of Negative News. Psychology Today. Retrieved December 19, 2022, from https://www.psychologytoday.com/us/blog/why-we-worry/202009/the-psychological-impact-negative-news

De Smedt, T., Nijs, L., Daelemans, W. (2014). Creative web services with pattern. In: Proceedings of the Fifth International Conference on Computational Creativity. https://computationalcreativity.net/iccc2014/wp-content/uploads/2014/06/13.5_DeSmedt.pdf

Fernandes, K., Vinagre, P., & Cortez, P. (2015). A proactive intelligent decision support system for predicting the popularity of online news. In Portuguese Conference on Artificial Intelligence (pp. 535-546). Springer, Cham. http://archive.ics.uci.edu/ml/datasets/Online%20News%20Popularity

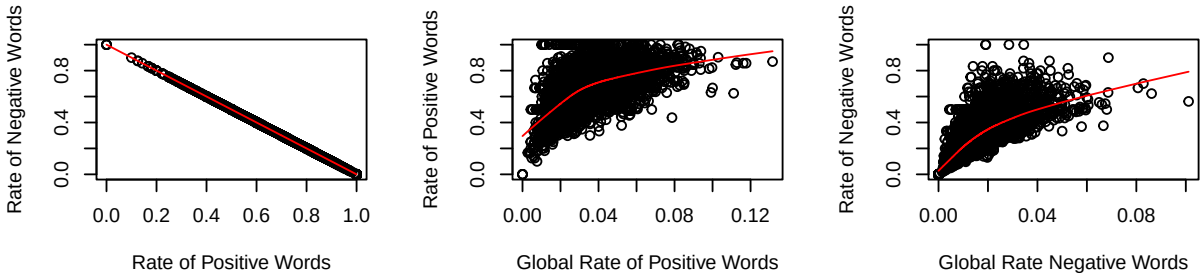
R codes

More on Correlation of Variables

```
# Code for correlation matrix
data2014sampled2 %>% select(-c(url, shares)) -> exp_vars
cor(exp_vars)
```

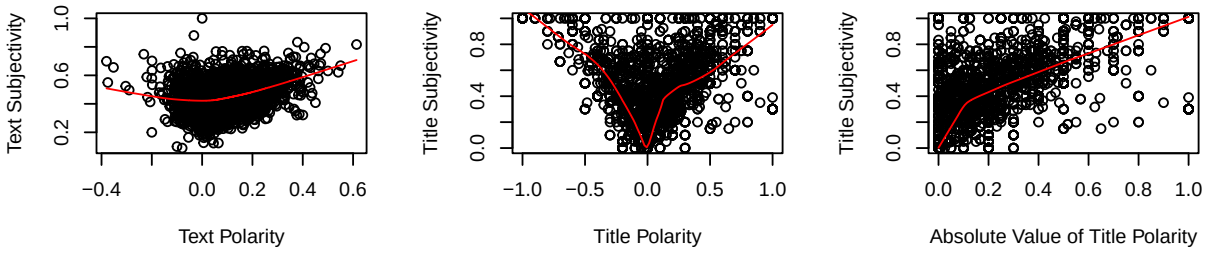
We can see that `rate_positive_words` seems to be strongly negatively correlated with `rate_negative_words`. This is expected because the rate of positive words and rate of negative words should sum up to 1. There is correlation between `rate_positive_words`, the rate of positive words among non-neutral words, and `global_rate_positive_words`, the rate of positive words across all words in the text. This is expected because both involve counting the number of positive words in the text. While building models, I will consider interactions between `rate_positive_words`, `rate_negative_words`. I would pick one of `rate_positive_words` or `global_rate_positive_words`, but not both.

```
par(mfrow=c(1,3))
# main = "Rate of Negative Words vs Rate of Positive Words"
 plot(data2014sampled2$rate_positive_words, data2014sampled2$rate_negative_words,
      xlab = "Rate of Positive Words", ylab = "Rate of Negative Words")
lines(lowess(data2014sampled2$rate_positive_words, data2014sampled2$rate_negative_words), col = "red")

plot(data2014sampled2$global_rate_positive_words, data2014sampled2$rate_positive_words,
     xlab="Global Rate of Positive Words", ylab = "Rate of Positive Words")
lines(lowess(data2014sampled2$global_rate_positive_words, data2014sampled2$rate_positive_words), col =

plot(data2014sampled2$global_rate_negative_words, data2014sampled2$rate_negative_words,
     xlab="Global Rate Negative Words", ylab="Rate of Negative Words")
lines(lowess(data2014sampled2$global_rate_negative_words, data2014sampled2$rate_negative_words), col =
```



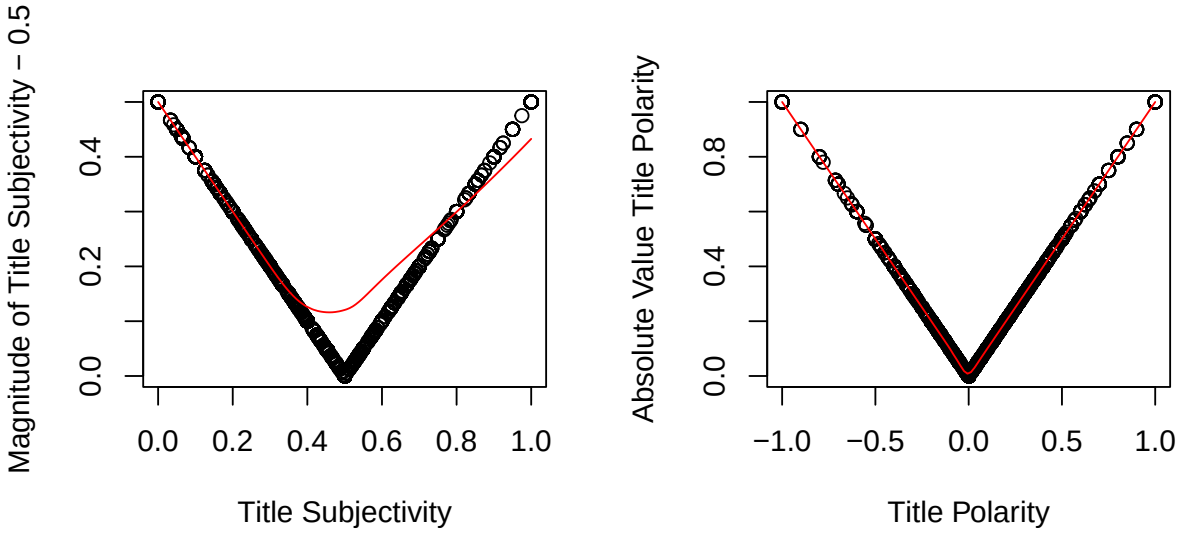
The variables measuring text polarity and text subjectivity seemed to be weakly correlated. One may guess that very opinionated pieces may have more emotionally charged words, but the scatterplot did not seem to suggest strong evidence for such findings. Interestingly, there was strong correlation between absolute value of polarity and subjectivity in the title rather than the text. Perhaps the text had a lot of noise, due to being significantly longer than the title. It may be necessary to consider interactions between title polarity and subjectivity while building models.

```
par(mfrow=c(1,3))
plot(data2014sampled2$global_sentiment_polarity, data2014sampled2$global_subjectivity,
     xlab = "Text Polarity", ylab="Text Subjectivity")
lines(lowess(data2014sampled2$global_sentiment_polarity, data2014sampled2$global_subjectivity), col = "
# cor(data2014sampled2$global_sentiment_polarity, data2014sampled2$global_subjectivity)

plot(data2014sampled2$title_sentiment_polarity, data2014sampled2$title_subjectivity,
     ylab = "Title Subjectivity", xlab="Title Polarity")
lines(lowess(data2014sampled2$title_sentiment_polarity, data2014sampled2$title_subjectivity), col = "re

plot(data2014sampled2$abs_title_sentiment_polarity, data2014sampled2$title_subjectivity,
     xlab = "Absolute Value of Title Polarity", ylab = "Title Subjectivity")
lines(lowess(data2014sampled2$abs_title_sentiment_polarity, data2014sampled2$title_subjectivity), col =
```



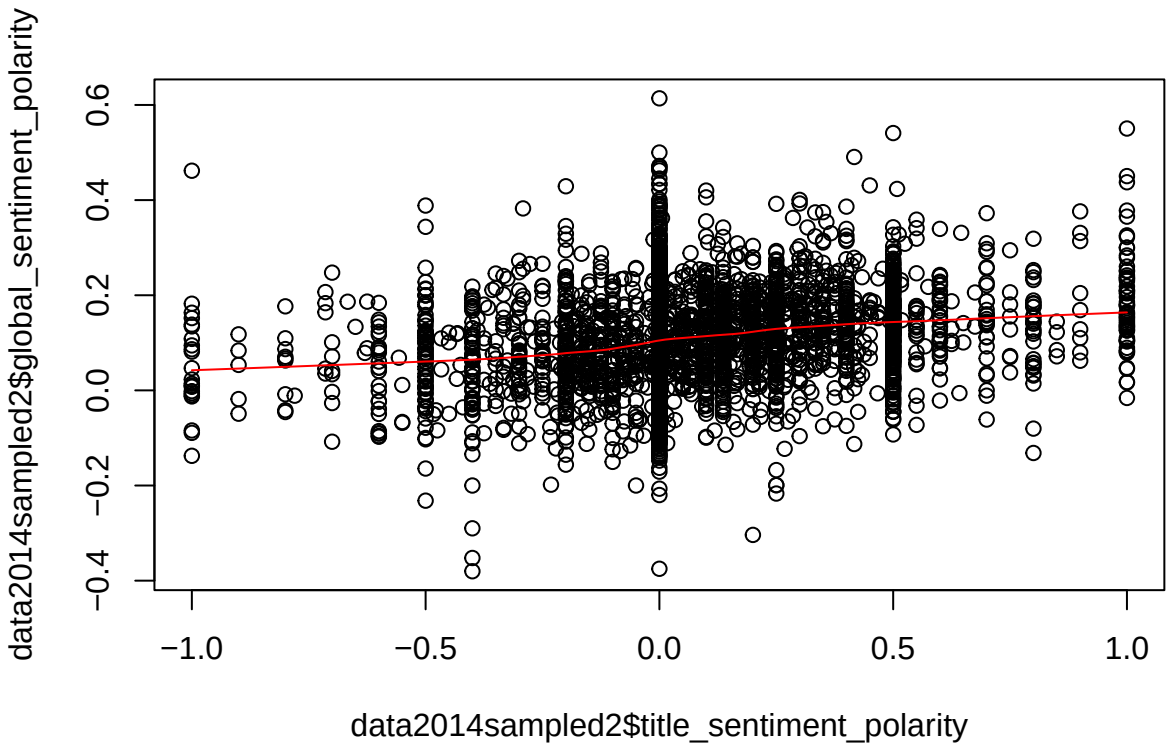
The title subjectivity and the absolute title subjectivity are highly correlated, because the absolute title subjectivity is the absolute value of the subjectivity level minus 0.5. While building regression models, we will just keep one or the other. We will compare to see which one is more significant. Similarly, the title polarity and absolute title polarity were also highly correlated, because the absolute title polarity is the absolute value of the title polarity. While building regression models, we will include one of the two, but not both.

```
par(mfrow = c(1,2))

plot(data2014sampled2$title_subjectivity, data2014sampled2$abs_title_subjectivity,
     xlab = "Title Subjectivity", ylab = "Magnitude of Title Subjectivity - 0.5")
lines(lowess(data2014sampled2$title_subjectivity, data2014sampled2$abs_title_subjectivity), col = "red")

plot(data2014sampled2$title_sentiment_polarity, data2014sampled2$abs_title_sentiment_polarity,
     xlab = "Title Polarity", ylab = "Absolute Value Title Polarity")
lines(lowess(data2014sampled2$title_sentiment_polarity, data2014sampled2$abs_title_sentiment_polarity),
```



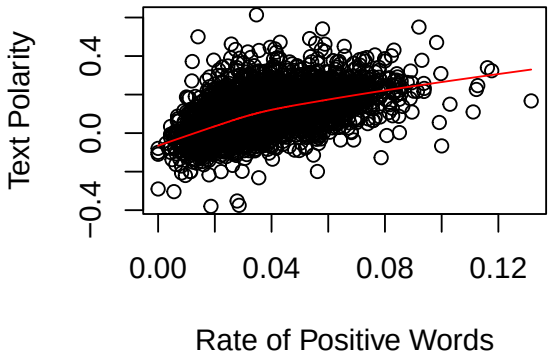
Polarity in the title and article did not seem to be correlated.

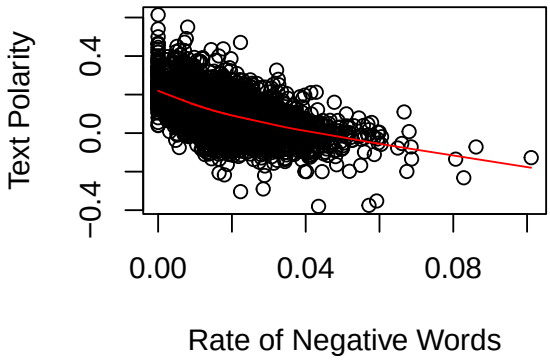
```
plot(data2014sampled2$title_sentiment_polarity, data2014sampled2$global_sentiment_polarity)
lines(lowess(data2014sampled2$title_sentiment_polarity, data2014sampled2$global_sentiment_polarity), co
```



Polarity and the rate of positive words among all words in the text were somewhat correlated, probably because polarity is a measure of how positive or negative the emotions conveyed in a text are. Similarly, polarity and the rate of negative words among all words in the text were somewhat correlated. While choosing variables for a regression model, I'd probably opt for choosing the variables measuring positive and negative words, rather than polarity in general, which seems to have less detailed information about the emotions of the text.

```
par(mfrow=c(1,2))
plot(data2014sampled2$global_rate_positive_words, data2014sampled2$global_sentiment_polarity,
     xlab="Rate of Positive Words", ylab="Text Polarity")
lines(lowess(data2014sampled2$global_rate_positive_words, data2014sampled2$global_sentiment_polarity),
#cor(data2014sampled2$global_rate_positive_words, data2014sampled2$global_sentiment_polarity)
plot(data2014sampled2$global_rate_negative_words, data2014sampled2$global_sentiment_polarity,
     xlab = "Rate of Negative Words", ylab="Text Polarity")
lines(lowess(data2014sampled2$global_rate_negative_words, data2014sampled2$global_sentiment_polarity),
```

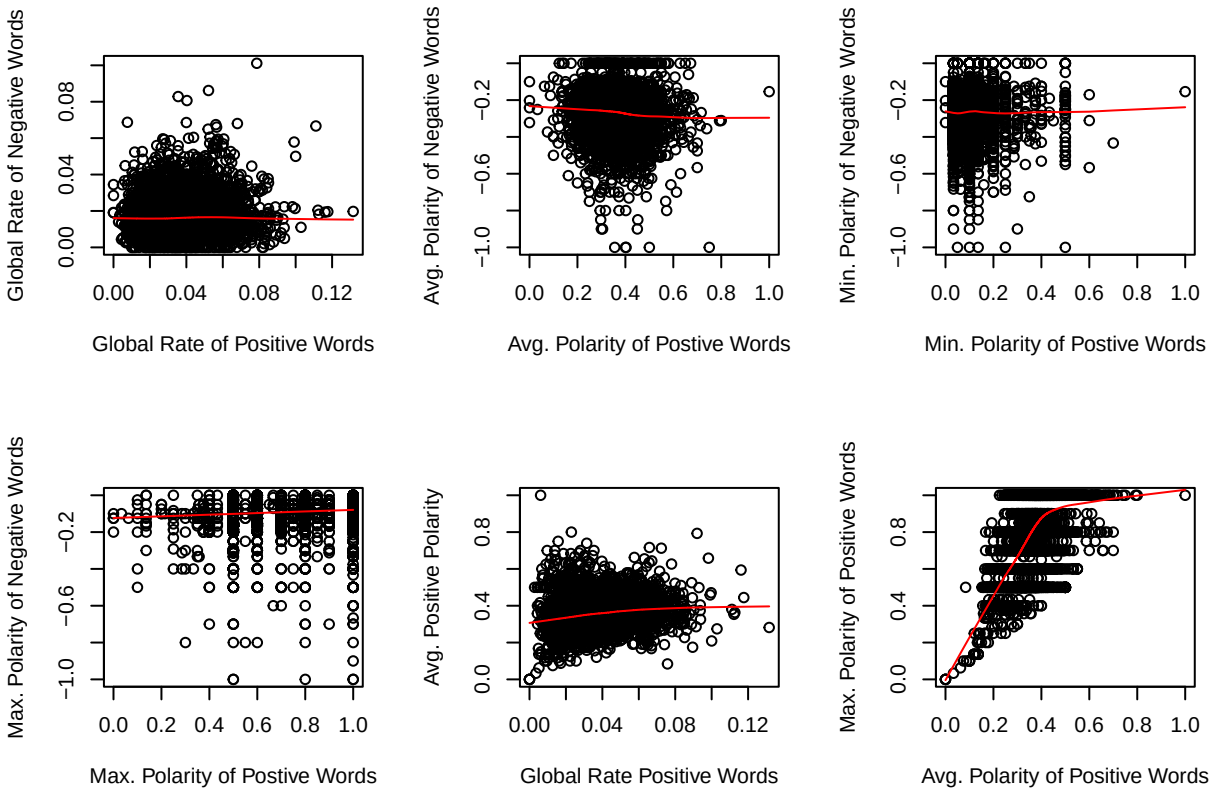





```
#cor(data2014sampled2$global_rate_negative_words, data2014sampled2$global_sentiment_polarity)
```

The positive and negative polarities did not appear to be correlated. The rate of positive words out of all words and the rate of negative words out of all words did not seem to be correlated. It seemed that in most articles, having lots of positive words did not come at the expense of having no negative words. The dataset seemed to have a balance of articles with varying levels of positive, negative, and neutral words. That is, there didn't seem to be a large portion of articles that had mostly negative language with little positive language, or mostly positive language with little negative language.

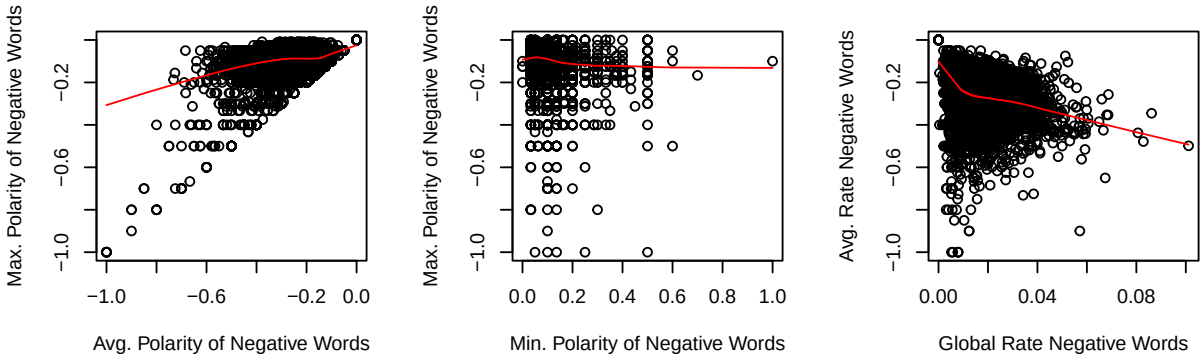
```
par(mfrow=c(2,3))
plot(data2014sampled2$global_rate_positive_words, data2014sampled2$global_rate_negative_words,
     xlab="Global Rate of Positive Words", ylab="Global Rate of Negative Words")
lines(lowess(data2014sampled2$global_rate_positive_words, data2014sampled2$global_rate_negative_words),

plot(data2014sampled2$avg_positive_polarity, data2014sampled2$avg_negative_polarity,
     xlab="Avg. Polarity of Postive Words", ylab="Avg. Polarity of Negative Words")
lines(lowess(data2014sampled2$avg_positive_polarity, data2014sampled2$avg_negative_polarity), col = "red

plot(data2014sampled2$min_positive_polarity, data2014sampled2$avg_negative_polarity,
     xlab="Min. Polarity of Postive Words", ylab="Min. Polarity of Negative Words")
lines(lowess(data2014sampled2$min_positive_polarity, data2014sampled2$avg_negative_polarity), col = "red

plot(data2014sampled2$max_positive_polarity, data2014sampled2$max_negative_polarity,
     xlab="Max. Polarity of Postive Words", ylab="Max. Polarity of Negative Words")
lines(lowess(data2014sampled2$max_positive_polarity, data2014sampled2$max_negative_polarity), col = "red

plot(data2014sampled2$global_rate_positive_words, data2014sampled2$avg_positive_polarity,
     xlab="Global Rate Positive Words", ylab = "Avg. Positive Polarity")
lines(lowess(data2014sampled2$global_rate_positive_words, data2014sampled2$avg_positive_polarity), col =

plot(data2014sampled2$avg_positive_polarity, data2014sampled2$max_positive_polarity,
     xlab="Avg. Polarity of Positive Words", ylab="Max. Polarity of Positive Words")
lines(lowess(data2014sampled2$avg_positive_polarity, data2014sampled2$max_positive_polarity), col="red"
```



There is insignificant correlation between the rate of positive words across all words in the text, and the average polarity of those positive words. There is weak correlation between the rate of negative words across all words in the text, and the average polarity of those negative words.

The minimum, average and maximum polarity of positive words did not appear to be strongly correlated with the minimum, average and maximum polarity of negative words.

```
par(mfrow = c(1,3))

plot(data2014sampled2$avg_negative_polarity, data2014sampled2$max_negative_polarity,
     xlab="Avg. Polarity of Negative Words", ylab="Max. Polarity of Negative Words")
lines(lowess(data2014sampled2$avg_negative_polarity, data2014sampled2$max_negative_polarity), col = "red

plot(data2014sampled2$min_positive_polarity, data2014sampled2$max_negative_polarity,
     xlab="Min. Polarity of Negative Words", ylab="Max. Polarity of Negative Words")
lines(lowess(data2014sampled2$min_positive_polarity, data2014sampled2$max_negative_polarity), col = "red

plot(data2014sampled2$global_rate_negative_words, data2014sampled2$avg_negative_polarity,
     xlab = "Global Rate Negative Words", ylab = "Avg. Rate Negative Words")
lines(lowess(data2014sampled2$global_rate_negative_words, data2014sampled2$avg_negative_polarity), col =
```



There was some moderate correlation between the average polarity of positive words and the maximum polarity of positive words. It may be necessary to consider interactions between `avg_positive_polarity` and `max_positive_polarity`. It seems like most articles have the polarity of negative words being around (-0.4, 0), so there aren't many articles that are very negative. There was some moderate correlation between the average polarity of negative words and the maximum polarity of negative words. It may be necessary to consider interactions between `avg_negative_polarity` and `max_negative_polarity`.

Additional Models

Code for Poisson Regression Model

```
pois_model <- glm(shares ~ .-url- rate_positive_words - rate_negative_words
                  - abs_title_subjectivity - abs_title_sentiment_polarity +
                      global_subjectivity*global_sentiment_polarity +
                      `title_sentiment_polarity`*`title_subjectivity` +
                      `avg_positive_polarity`*`max_positive_polarity` +
                      `avg_negative_polarity`*`max_negative_polarity` +
                      `global_rate_negative_words`*`avg_negative_polarity`, family = poisson(link = "id
step(pois_model, trace = 0)
```

```
##
## Call:  glm(formula = shares ~ (url + global_subjectivity + global_sentiment_polarity +
##     global_rate_positive_words + global_rate_negative_words +
##     rate_positive_words + rate_negative_words + avg_positive_polarity +
##     min_positive_polarity + max_positive_polarity + avg_negative_polarity +
##     max_negative_polarity + min_negative_polarity + title_subjectivity +
##     title_sentiment_polarity + abs_title_subjectivity + abs_title_sentiment_polarity) -
##     url - rate_positive_words - rate_negative_words - abs_title_subjectivity -
##     abs_title_sentiment_polarity + global_subjectivity * global_sentiment_polarity +
##     title_sentiment_polarity * title_subjectivity + avg_positive_polarity *
##     max_positive_polarity + avg_negative_polarity * max_negative_polarity +
##     global_rate_negative_words * avg_negative_polarity, family = poisson(link = "identity"),
##     data = data2014sampled2)
##
## Coefficients:
##                                     (Intercept)
##                                          1176.6
```

```
##                              global_subjectivity
##                                           4312.9
##                           global_sentiment_polarity
##                                          12268.5
##                          global_rate_positive_words
##                                          -5340.3
##                          global_rate_negative_words
##                                          30146.7
##                               avg_positive_polarity
##                                          -6364.6
##                               min_positive_polarity
##                                           2697.1
##                               max_positive_polarity
##                                          -2416.0
##                               avg_negative_polarity
##                                          -1728.1
##                               max_negative_polarity
##                                            468.3
##                               min_negative_polarity
##                                            259.1
##                                  title_subjectivity
##                                            605.8
##                            title_sentiment_polarity
##                                            424.6
##    global_subjectivity:global_sentiment_polarity
##                                         -15102.4
##      title_subjectivity:title_sentiment_polarity
##                                           -172.2
##      avg_positive_polarity:max_positive_polarity
##                                           8209.8
##      avg_negative_polarity:max_negative_polarity
##                                           3584.5
## global_rate_negative_words:avg_negative_polarity
##                                         -10453.6
##
## Degrees of Freedom: 4768 Total (i.e. Null);  4751 Residual
## Null Deviance:       23700000
## Residual Deviance: 22770000  AIC: 22810000
```

```
poisfit <- glm(formula = shares ~ (url + global_subjectivity + global_sentiment_polarity +
    global_rate_positive_words + global_rate_negative_words +
    rate_positive_words + rate_negative_words + avg_positive_polarity +
    min_positive_polarity + max_positive_polarity + avg_negative_polarity +
    max_negative_polarity + min_negative_polarity + title_subjectivity +
    title_sentiment_polarity + abs_title_subjectivity + abs_title_sentiment_polarity) -
    url - rate_positive_words - rate_negative_words - abs_title_subjectivity -
    abs_title_sentiment_polarity + global_subjectivity * global_sentiment_polarity +
    title_sentiment_polarity * title_subjectivity + avg_positive_polarity *
    max_positive_polarity + avg_negative_polarity * max_negative_polarity +
    global_rate_negative_words * avg_negative_polarity, family = poisson(link = "identity"),
    data = data2014sampled2)
summary(poisfit)
```

```
##
```

```
## Call:
## glm(formula = shares ~ (url + global_subjectivity + global_sentiment_polarity +
##     global_rate_positive_words + global_rate_negative_words +
##     rate_positive_words + rate_negative_words + avg_positive_polarity +
##     min_positive_polarity + max_positive_polarity + avg_negative_polarity +
##     max_negative_polarity + min_negative_polarity + title_subjectivity +
##     title_sentiment_polarity + abs_title_subjectivity + abs_title_sentiment_polarity) -
##     url - rate_positive_words - rate_negative_words - abs_title_subjectivity -
##     abs_title_sentiment_polarity + global_subjectivity * global_sentiment_polarity +
##     title_sentiment_polarity * title_subjectivity + avg_positive_polarity *
##     max_positive_polarity + avg_negative_polarity * max_negative_polarity +
##     global_rate_negative_words * avg_negative_polarity, family = poisson(link = "identity"),
##     data = data2014sampled2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -118.57   -44.33   -30.50    -6.74   785.64
##
## Coefficients:
##                                                  Estimate Std. Error z value
## (Intercept)                                      1176.559     10.410  113.02
## global_subjectivity                              4312.888     13.912  310.02
## global_sentiment_polarity                       12268.515     52.453  233.90
## global_rate_positive_words                      -5340.305     93.490  -57.12
## global_rate_negative_words                      30146.700    222.699  135.37
## avg_positive_polarity                           -6364.558     34.820 -182.78
## min_positive_polarity                            2697.149     15.899  169.64
## max_positive_polarity                           -2415.976     14.421 -167.53
## avg_negative_polarity                           -1728.128     19.823  -87.18
## max_negative_polarity                             468.269     25.240   18.55
## min_negative_polarity                             259.053      5.476   47.31
## title_subjectivity                                605.755      2.663  227.50
## title_sentiment_polarity                          424.570      8.666   48.99
## global_subjectivity:global_sentiment_polarity   -15102.448     86.651 -174.29
## title_subjectivity:title_sentiment_polarity        -172.238     11.783  -14.62
## avg_positive_polarity:max_positive_polarity       8209.828     40.123  204.62
## avg_negative_polarity:max_negative_polarity       3584.462     43.240   82.90
## global_rate_negative_words:avg_negative_polarity -10453.634    692.061  -15.11
##                                                  Pr(>|z|)
## (Intercept)                                        <2e-16 ***
## global_subjectivity                                <2e-16 ***
## global_sentiment_polarity                          <2e-16 ***
## global_rate_positive_words                         <2e-16 ***
## global_rate_negative_words                         <2e-16 ***
## avg_positive_polarity                              <2e-16 ***
## min_positive_polarity                              <2e-16 ***
## max_positive_polarity                              <2e-16 ***
## avg_negative_polarity                              <2e-16 ***
## max_negative_polarity                              <2e-16 ***
## min_negative_polarity                              <2e-16 ***
## title_subjectivity                                 <2e-16 ***
## title_sentiment_polarity                           <2e-16 ***
## global_subjectivity:global_sentiment_polarity      <2e-16 ***
## title_subjectivity:title_sentiment_polarity        <2e-16 ***
```

```
## avg_positive_polarity:max_positive_polarity        <2e-16 ***
## avg_negative_polarity:max_negative_polarity        <2e-16 ***
## global_rate_negative_words:avg_negative_polarity   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 23697576  on 4768  degrees of freedom
## Residual deviance: 22769531  on 4751  degrees of freedom
## AIC: 22813948
##
## Number of Fisher Scoring iterations: 11
```

Code for Negative Binomial Regression

```
negbinfit <- MASS::glm.nb(shares ~ .-url- rate_positive_words - rate_negative_words -
                            abs_title_subjectivity - abs_title_sentiment_polarity -
                            global_sentiment_polarity +
                      `title_sentiment_polarity`*`title_subjectivity` +
                      `avg_positive_polarity`*`max_positive_polarity` +
                      `avg_negative_polarity`*`max_negative_polarity` +
                      `global_rate_negative_words`*`avg_negative_polarity`,
                      data = data2014sampled2, link = log)
step(negbinfit, trace=0)
```

```
##
## Call:  MASS::glm.nb(formula = shares ~ global_subjectivity + global_rate_positive_words +
##     global_rate_negative_words + avg_positive_polarity + min_positive_polarity +
##     max_positive_polarity + max_negative_polarity + title_subjectivity +
##     title_sentiment_polarity + avg_positive_polarity:max_positive_polarity,
##     data = data2014sampled2, init.theta = 1.037091806, link = log)
##
## Coefficients:
##                                  (Intercept)
##                                      7.16216
##                           global_subjectivity
##                                      1.22958
##                      global_rate_positive_words
##                                      3.26130
##                      global_rate_negative_words
##                                      2.81129
##                           avg_positive_polarity
##                                     -0.67119
##                           min_positive_polarity
##                                      0.63263
##                           max_positive_polarity
##                                     -0.39117
##                           max_negative_polarity
##                                     -0.63547
##                              title_subjectivity
##                                      0.19191
```

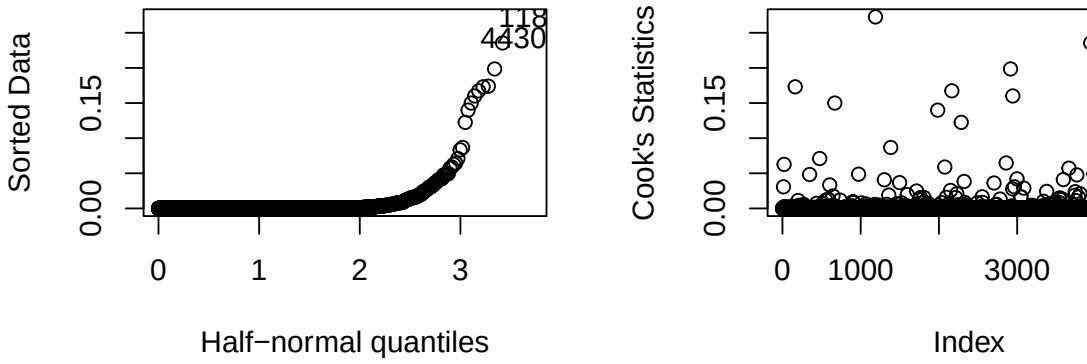
```
##                   title_sentiment_polarity
##                                    0.07923
## avg_positive_polarity:max_positive_polarity
##                                    1.67119
##
## Degrees of Freedom: 4768 Total (i.e. Null);  4758 Residual
## Null Deviance:       5753
## Residual Deviance: 5481  AIC: 86080
```

```
negbinfitbest <- MASS::glm.nb(formula = shares ~ global_subjectivity + global_rate_positive_words +
        global_rate_negative_words + avg_positive_polarity + min_positive_polarity +
        max_positive_polarity + max_negative_polarity + min_negative_polarity +
        title_subjectivity + title_sentiment_polarity + title_subjectivity:title_sentiment_polarity +
        avg_positive_polarity:max_positive_polarity, data = data2014sampled2,
    init.theta = 1.005901769, link = log)
summary(negbinfitbest)
```

```
##
## Call:
## MASS::glm.nb(formula = shares ~ global_subjectivity + global_rate_positive_words +
##     global_rate_negative_words + avg_positive_polarity + min_positive_polarity +
##     max_positive_polarity + max_negative_polarity + min_negative_polarity +
##     title_subjectivity + title_sentiment_polarity + title_subjectivity:title_sentiment_polarity +
##     avg_positive_polarity:max_positive_polarity, data = data2014sampled2,
##     init.theta = 1.037160862, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6879  -0.9723  -0.6716  -0.1420   9.3011
##
## Coefficients:
##                                             Estimate Std. Error z value
## (Intercept)                                 7.165260   0.191992  37.321
## global_subjectivity                         1.228266   0.190662   6.442
## global_rate_positive_words                  3.212161   1.079196   2.976
## global_rate_negative_words                  2.863379   1.609873   1.779
## avg_positive_polarity                      -0.676314   0.611498  -1.106
## min_positive_polarity                       0.625652   0.266896   2.344
## max_positive_polarity                      -0.398468   0.252063  -1.581
## max_negative_polarity                      -0.635489   0.162382  -3.914
## min_negative_polarity                       0.002051   0.060089   0.034
## title_subjectivity                          0.192911   0.045840   4.208
## title_sentiment_polarity                    0.162245   0.144695   1.121
## title_subjectivity:title_sentiment_polarity -0.116890   0.193175  -0.605
## avg_positive_polarity:max_positive_polarity  1.692944   0.697849   2.426
##                                            Pr(>|z|)
## (Intercept)                                 < 2e-16 ***
## global_subjectivity                        1.18e-10 ***
## global_rate_positive_words                  0.00292 **
## global_rate_negative_words                  0.07530 .
## avg_positive_polarity                       0.26873
## min_positive_polarity                       0.01907 *
## max_positive_polarity                       0.11392
## max_negative_polarity                      9.09e-05 ***
```

```
## min_negative_polarity                          0.97277
## title_subjectivity                            2.57e-05 ***
## title_sentiment_polarity                       0.26216
## title_subjectivity:title_sentiment_polarity    0.54511
## avg_positive_polarity:max_positive_polarity    0.01527 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0372) family taken to be 1)
##
##     Null deviance: 5753.6  on 4768  degrees of freedom
## Residual deviance: 5480.6  on 4756  degrees of freedom
## AIC: 86081
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0372
##           Std. Err.:  0.0188
##
##  2 x log-likelihood:  -86053.4720
```

```
par(mfrow = c(1,2))
faraway::halfnorm(cooks.distance(negbinfitbest))
plot(cooks.distance(negbinfitbest), ylab = "Cook's Statistics")
```



Removing Outliers Code

Trying a Linear Model

We can see that a linear model is not a good fit for the data because the R^2 value is close to 0, and the Shapiro-Wilk test shows strong evidence that the residuals are not normal.

```
fit <- lm(shares ~ .-url- rate_positive_words - rate_negative_words -
              abs_title_subjectivity - abs_title_sentiment_polarity +
                      `title_sentiment_polarity`*`title_subjectivity` +
                      `avg_positive_polarity`*`max_positive_polarity` +
                      `avg_negative_polarity`*`max_negative_polarity` +
                      `global_rate_negative_words`*`avg_negative_polarity`,
          data=data2014sampled2)
step(fit, trace = 0)
```

```
##
## Call:
## lm(formula = shares ~ global_subjectivity + avg_positive_polarity +
##     max_positive_polarity + avg_negative_polarity + title_subjectivity +
##     title_sentiment_polarity + avg_positive_polarity:max_positive_polarity,
##     data = data2014sampled2)
##
## Coefficients:
##                                   (Intercept)
##                                        1466.3
##                           global_subjectivity
##                                        3950.4
##                         avg_positive_polarity
##                                       -2551.5
##                         max_positive_polarity
##                                       -2894.5
##                         avg_negative_polarity
##                                       -1544.8
##                            title_subjectivity
##                                         747.4
##                       title_sentiment_polarity
##                                         606.4
## avg_positive_polarity:max_positive_polarity
##                                        8158.3
```

```
fitlmbest <- lm(formula = shares ~ global_subjectivity + global_sentiment_polarity +
    global_rate_positive_words + global_rate_negative_words +
    avg_negative_polarity + title_subjectivity, data = data2014sampled2)
summary(fitlmbest)
```

```
##
## Call:
## lm(formula = shares ~ global_subjectivity + global_sentiment_polarity +
##     global_rate_positive_words + global_rate_negative_words +
##     avg_negative_polarity + title_subjectivity, data = data2014sampled2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5543  -2159  -1441   -362 124391
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -68.78     524.59  -0.131  0.89569
```

```
## global_subjectivity          3328.01    1322.02   2.517  0.01186 *
## global_sentiment_polarity    6237.85    1930.81   3.231  0.00124 **
## global_rate_positive_words -14273.40    8707.22  -1.639  0.10123
## global_rate_negative_words  30069.89   13560.13   2.218  0.02663 *
## avg_negative_polarity       -2844.01     994.50  -2.860  0.00426 **
## title_subjectivity            866.44     302.35   2.866  0.00418 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6605 on 4762 degrees of freedom
## Multiple R-squared:  0.012,  Adjusted R-squared:  0.01075
## F-statistic: 9.639 on 6 and 4762 DF,  p-value: 1.441e-10
```

```
shapiro.test(resid(fitlmbest))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fitlmbest)
## W = 0.3659, p-value < 2.2e-16
```

Trying a Log-linear model

We also tried a linear model on the log of the shares, but that was also a poor fit, with the residuals not being normal, and the R^2 value also being close to 0.

```
fit2 <- lm(log(shares) ~ .-url - rate_positive_words - rate_negative_words -
               abs_title_subjectivity - abs_title_sentiment_polarity +
                      `title_sentiment_polarity`*`title_subjectivity` +
                      `avg_positive_polarity`*`max_positive_polarity` +
                      `avg_negative_polarity`*`max_negative_polarity` +
                      `global_rate_negative_words`*`avg_negative_polarity`,
           data=data2014sampled2)
step(fit2, trace = 0)
```

```
##
## Call:
## lm(formula = log(shares) ~ global_subjectivity + global_rate_positive_words +
##     global_rate_negative_words + avg_positive_polarity + max_positive_polarity +
##     avg_negative_polarity + title_subjectivity + title_sentiment_polarity +
##     avg_positive_polarity:max_positive_polarity + global_rate_negative_words:avg_negative_polarity,
##     data = data2014sampled2)
##
## Coefficients:
##                                       (Intercept)
##                                           6.83233
##                                 global_subjectivity
##                                           1.12861
##                           global_rate_positive_words
##                                           3.88964
##                           global_rate_negative_words
##                                           3.61113
```

```
##                             avg_positive_polarity
##                                           -0.32438
##                             max_positive_polarity
##                                           -0.30217
##                             avg_negative_polarity
##                                           -0.13778
##                                title_subjectivity
##                                            0.09225
##                           title_sentiment_polarity
##                                            0.13266
##      avg_positive_polarity:max_positive_polarity
##                                            0.90322
## global_rate_negative_words:avg_negative_polarity
##                                           13.96483
```

```
loglm <- lm(formula = log(shares) ~ global_subjectivity + global_rate_positive_words +
               global_rate_negative_words + avg_positive_polarity + max_positive_polarity +
               avg_negative_polarity + title_subjectivity + title_sentiment_polarity +
               avg_positive_polarity:max_positive_polarity +
               global_rate_negative_words:avg_negative_polarity,
            data = data2014sampled2)
summary(loglm)
```

```
##
## Call:
## lm(formula = log(shares) ~ global_subjectivity + global_rate_positive_words +
##     global_rate_negative_words + avg_positive_polarity + max_positive_polarity +
##     avg_negative_polarity + title_subjectivity + title_sentiment_polarity +
##     avg_positive_polarity:max_positive_polarity + global_rate_negative_words:avg_negative_polarity,
##     data = data2014sampled2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0243 -0.5733 -0.2034  0.3990  4.2797
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                       6.83233    0.17509  39.022
## global_subjectivity                               1.12861    0.17502   6.448
## global_rate_positive_words                        3.88964    0.92938   4.185
## global_rate_negative_words                        3.61113    3.10835   1.162
## avg_positive_polarity                            -0.32438    0.51576  -0.629
## max_positive_polarity                            -0.30217    0.22036  -1.371
## avg_negative_polarity                            -0.13778    0.16730  -0.824
## title_subjectivity                                0.09225    0.04088   2.257
## title_sentiment_polarity                          0.13266    0.04981   2.663
## avg_positive_polarity:max_positive_polarity       0.90322    0.61567   1.467
## global_rate_negative_words:avg_negative_polarity 13.96483    9.32028   1.498
##                                                  Pr(>|t|)
## (Intercept)                                       < 2e-16 ***
## global_subjectivity                              1.24e-10 ***
## global_rate_positive_words                       2.90e-05 ***
## global_rate_negative_words                        0.24539
## avg_positive_polarity                             0.52943
```

```
## max_positive_polarity                                0.17037
## avg_negative_polarity                                0.41024
## title_subjectivity                                   0.02408 *
## title_sentiment_polarity                             0.00776 **
## avg_positive_polarity:max_positive_polarity          0.14242
## global_rate_negative_words:avg_negative_polarity  0.13411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8731 on 4758 degrees of freedom
## Multiple R-squared:  0.03357,    Adjusted R-squared:  0.03154
## F-statistic: 16.53 on 10 and 4758 DF,  p-value: < 2.2e-16
```

```
shapiro.test(resid(loglm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(loglm)
## W = 0.91582, p-value < 2.2e-16
```

Trying a Quasi-Poisson Regression

```
quaspoisfit <- glm(shares ~.-url- rate_positive_words - rate_negative_words -
                      abs_title_subjectivity - abs_title_sentiment_polarity +
                     global_subjectivity*global_sentiment_polarity +
                     `title_sentiment_polarity`*`title_subjectivity` +
                     `avg_positive_polarity`*`max_positive_polarity` +
                     `avg_negative_polarity`*`max_negative_polarity` +
                     `global_rate_negative_words`*`avg_negative_polarity`,
                   family = quasipoisson(link = "identity"), data = data2014sampled2)
summary(quaspoisfit)
```

```
##
## Call:
## glm(formula = shares ~ . - url - rate_positive_words - rate_negative_words -
##     abs_title_subjectivity - abs_title_sentiment_polarity + global_subjectivity *
##     global_sentiment_polarity + title_sentiment_polarity * title_subjectivity +
##     avg_positive_polarity * max_positive_polarity + avg_negative_polarity *
##     max_negative_polarity + global_rate_negative_words * avg_negative_polarity,
##     family = quasipoisson(link = "identity"), data = data2014sampled2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -118.57   -44.33   -30.50    -6.74   785.64
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                      1176.6     1181.9   0.995
## global_subjectivity                              4312.9     1579.4   2.731
## global_sentiment_polarity                       12268.5     5955.1   2.060
```

```
## global_rate_positive_words                          -5340.3    10614.2  -0.503
## global_rate_negative_words                          30146.7    25283.9   1.192
## avg_positive_polarity                               -6364.6     3953.2  -1.610
## min_positive_polarity                                2697.1     1805.1   1.494
## max_positive_polarity                               -2416.0     1637.3  -1.476
## avg_negative_polarity                               -1728.1     2250.5  -0.768
## max_negative_polarity                                 468.3     2865.6   0.163
## min_negative_polarity                                 259.1      621.7   0.417
## title_subjectivity                                    605.8      302.3   2.004
## title_sentiment_polarity                              424.6      983.9   0.432
## global_subjectivity:global_sentiment_polarity     -15102.4     9837.8  -1.535
## title_subjectivity:title_sentiment_polarity         -172.2     1337.8  -0.129
## avg_positive_polarity:max_positive_polarity         8209.8     4555.3   1.802
## avg_negative_polarity:max_negative_polarity         3584.5     4909.2   0.730
## global_rate_negative_words:avg_negative_polarity -10453.6    78572.3  -0.133
##                                                    Pr(>|t|)
## (Intercept)                                         0.31957
## global_subjectivity                                 0.00634 **
## global_sentiment_polarity                           0.03944 *
## global_rate_positive_words                          0.61490
## global_rate_negative_words                          0.23319
## avg_positive_polarity                               0.10747
## min_positive_polarity                               0.13520
## max_positive_polarity                               0.14012
## avg_negative_polarity                               0.44260
## max_negative_polarity                               0.87020
## min_negative_polarity                               0.67692
## title_subjectivity                                  0.04515 *
## title_sentiment_polarity                            0.66611
## global_subjectivity:global_sentiment_polarity       0.12481
## title_subjectivity:title_sentiment_polarity         0.89756
## avg_positive_polarity:max_positive_polarity         0.07157 .
## avg_negative_polarity:max_negative_polarity         0.46534
## global_rate_negative_words:avg_negative_polarity    0.89416
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 12889.9)
##
##     Null deviance: 23697576  on 4768  degrees of freedom
## Residual deviance: 22769531  on 4751  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 11
```

Trying a Negative Binomial Model with Identity Link

We also fit a negative binomial model using the identity link, and tried using stepwise selection to come up with a smaller model, but its residual deviance and AIC were higher than the negative binomial model with the log link, so we keep the negative binomial model with the log link.

```
negbinfit_id <- MASS::glm.nb(shares ~ .-url- rate_positive_words - rate_negative_words - abs_title_subj
                               `title_sentiment_polarity`*`title_subjectivity` +
```

```
                      `avg_positive_polarity`*`max_positive_polarity` +
                      `avg_negative_polarity`*`max_negative_polarity` +
                      `global_rate_negative_words`*`avg_negative_polarity`,
                      data = data2014sampled2, link = identity)
summary(negbinfit_id)
```

```
##
## Call:
## MASS::glm.nb(formula = shares ~ . - url - rate_positive_words -
##     rate_negative_words - abs_title_subjectivity - abs_title_sentiment_polarity +
##     title_sentiment_polarity * title_subjectivity + avg_positive_polarity *
##     max_positive_polarity + avg_negative_polarity * max_negative_polarity +
##     global_rate_negative_words * avg_negative_polarity, data = data2014sampled2,
##     link = identity, init.theta = 1.041764441)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6563  -0.9746  -0.6688  -0.1366   9.2216
##
## Coefficients:
##                                                  Estimate Std. Error z value
## (Intercept)                                        1253.0      492.1   2.546
## global_subjectivity                                2625.3      593.5   4.423
## global_sentiment_polarity                          4663.4     1351.8   3.450
## global_rate_positive_words                          596.6     4998.3   0.119
## global_rate_negative_words                        16402.5    11497.8   1.427
## avg_positive_polarity                             -4200.3     1586.2  -2.648
## min_positive_polarity                              1871.7      838.1   2.233
## max_positive_polarity                             -1502.8      683.4  -2.199
## avg_negative_polarity                              -779.9     1060.6  -0.735
## max_negative_polarity                               594.3     1458.9   0.407
## min_negative_polarity                               335.5      281.3   1.193
## title_subjectivity                                  494.7      144.0   3.435
## title_sentiment_polarity                            708.2      481.2   1.472
## title_subjectivity:title_sentiment_polarity        -757.5      665.5  -1.138
## avg_positive_polarity:max_positive_polarity        6113.0     1918.1   3.187
## avg_negative_polarity:max_negative_polarity        5240.3     2962.4   1.769
## global_rate_negative_words:avg_negative_polarity -58822.5    39895.1  -1.474
##                                                  Pr(>|z|)
## (Intercept)                                      0.010883 *
## global_subjectivity                              9.72e-06 ***
## global_sentiment_polarity                        0.000561 ***
## global_rate_positive_words                       0.904993
## global_rate_negative_words                       0.153701
## avg_positive_polarity                            0.008098 **
## min_positive_polarity                            0.025539 *
## max_positive_polarity                            0.027871 *
## avg_negative_polarity                            0.462133
## max_negative_polarity                            0.683724
## min_negative_polarity                            0.232936
## title_subjectivity                               0.000593 ***
## title_sentiment_polarity                         0.141106
## title_subjectivity:title_sentiment_polarity      0.255036
```

```
## avg_positive_polarity:max_positive_polarity      0.001438 **
## avg_negative_polarity:max_negative_polarity      0.076907 .
## global_rate_negative_words:avg_negative_polarity 0.140366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0418) family taken to be 1)
##
##     Null deviance: 5779.1  on 4768  degrees of freedom
## Residual deviance: 5477.8  on 4752  degrees of freedom
## AIC: 86062
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0418
##           Std. Err.:  0.0189
##
##  2 x log-likelihood:  -86026.4210
```

```
negbinfit_id <- MASS::glm.nb(shares ~ .-url- rate_positive_words - rate_negative_words -
                               abs_title_subjectivity - abs_title_sentiment_polarity +
                               `title_sentiment_polarity`*`title_subjectivity` +
                     `avg_positive_polarity`*`max_positive_polarity`
                      - global_rate_positive_words - min_negative_polarity - avg_negative_polarity,
                     data = data2014sampled2, link = identity)
summary(negbinfit_id)
```

```
##
## Call:
## MASS::glm.nb(formula = shares ~ . - url - rate_positive_words -
##     rate_negative_words - abs_title_subjectivity - abs_title_sentiment_polarity +
##     title_sentiment_polarity * title_subjectivity + avg_positive_polarity *
##     max_positive_polarity - global_rate_positive_words - min_negative_polarity -
##     avg_negative_polarity, data = data2014sampled2, link = identity,
##     init.theta = 1.039566197)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6251  -0.9745  -0.6650  -0.1360   9.1421
##
## Coefficients:
##                                                Estimate Std. Error z value
## (Intercept)                                       931.9      471.1   1.978
## global_subjectivity                              3532.2      528.4   6.685
## global_sentiment_polarity                        3617.8      753.2   4.803
## global_rate_negative_words                      26924.7     6007.4   4.482
## avg_positive_polarity                           -4753.0     1589.4  -2.990
## min_positive_polarity                            2054.0      826.0   2.487
## max_positive_polarity                           -1487.0      678.7  -2.191
## max_negative_polarity                           -2831.8      597.9  -4.736
## title_subjectivity                                513.9      143.8   3.575
## title_sentiment_polarity                          746.3      479.9   1.555
## title_subjectivity:title_sentiment_polarity      -822.1      661.8  -1.242
```

```
## avg_positive_polarity:max_positive_polarity   6662.8     1935.0   3.443
##                                                Pr(>|z|)
## (Intercept)                                    0.047908 *
## global_subjectivity                            2.32e-11 ***
## global_sentiment_polarity                      1.56e-06 ***
## global_rate_negative_words                     7.40e-06 ***
## avg_positive_polarity                          0.002785 **
## min_positive_polarity                          0.012898 *
## max_positive_polarity                          0.028458 *
## max_negative_polarity                          2.18e-06 ***
## title_subjectivity                             0.000351 ***
## title_sentiment_polarity                       0.119925
## title_subjectivity:title_sentiment_polarity    0.214159
## avg_positive_polarity:max_positive_polarity    0.000575 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0396) family taken to be 1)
##
##     Null deviance: 5766.9  on 4768  degrees of freedom
## Residual deviance: 5479.2  on 4757  degrees of freedom
## AIC: 86065
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0396
##           Std. Err.:  0.0188
##
##  2 x log-likelihood:  -86039.3200
```