

Analysis Written Report

The BEST Fit - Philip, Olivia, Leo, Allison

2025-04-10

Analysis + Peer Review

Draft report

Introduction and Data

Our project utilizes the University of California Irvine Machine Learning Repository's "Online News Popularity" data set. It includes share counts and descriptive characteristics for articles published by Mashable over two years (from 2013 to 2015). Mashable Inc. is a digital media website founded in 2005 and as of November 2015, it has over 6,000,000 Twitter followers and over 3,200,000 fans on Facebook. The data set in total, has 39644 observations, each representing an individual article. Our project motivation is that with the rise of the internet, we're interested in seeing how different factors influence or decide what goes "viral" on social media. For media companies specifically, this could aid in revealing patterns in what attracts readers to certain articles. Thus, our project aims to answer the

Key Research Question:

How do different article attributes (ex. Polarity, Positive/Negative Sentiments, Number of Images, etc.) relate to its virality on social media?

Key Variables:

rate_positive_words - Rate of positive words among non-neutral tokens in the article content. Values range from 0.0 to 1.0, with a mean of 0.6822 and standard deviation of 0.1902. This metric captures the positive emotional tone of the article.

Rate_negative_words - Rate of negative words among non-neutral tokens in the article content. Values range from 0.0 to 1.0, with a mean of 0.2879 and standard deviation of 0.1562. This metric captures the negative emotional tone of the article.

title_sentiment_polarity - Measure of the title's sentiment polarity (positivity/negativity). Values range from -1.0 (extremely negative) to 1.0 (extremely positive), with a mean of 0.0714 and standard deviation of 0.2654. This indicates how emotionally charged article titles are.

n_tokens_content - Number of words in the article content. Values range from 0 to 8,474 words, with a mean of 546.51 and standard deviation of 471.10. This quantifies the overall length of the article.

n_tokens_title - Number of words in the article title. Values range from 2 to 23 words, with a mean of 10.40 and standard deviation of 2.11. This measures length of headlines.

data_channel - Categorical variable denoting article topic, merged from indicators: data_channel_is_lifestyle, data_channel_is_entertainment, data_channel_is_bus, data_channel_is_socmed, data_channel_is_tech, and data_channel_is_world. This classifies content by subject area.

day_published - Categorical variable indicating publication day, merged from indicators: weekday_is_monday, weekday_is_tuesday, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday, weekday_is_saturday, weekday_is_sunday. Additionally includes is_weekend (mean 0.1309) to distinguish weekday from weekend publications.

kw_avg_avg - Average shares of average keywords in the article. Values range from 0.0 to 43,567.66, with a mean of 3,135.86 and standard deviation of 1,318.13. This measures the expected popularity of the article's keyword selection.

global_subjectivity - Measures the overall subjectivity of the article text. Values range from 0.0 (completely objective) to 1.0 (completely subjective), with a mean of 0.4434 and standard deviation of 0.1167. This quantifies how opinion-based versus fact-based the content is.

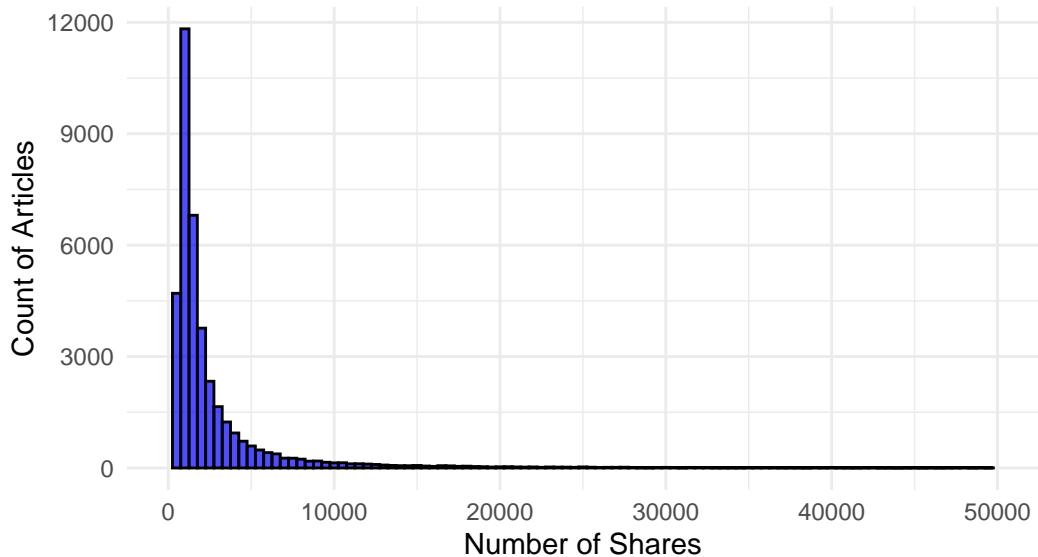
title_sentiment_polarity: A measure of the title's sentiment polarity (positivity/negativity). Values range from -1.0 (extremely negative) to 1.0 (extremely positive), with a mean of 0.0714 and a standard deviation of 0.2654.

Key EDA

Response Variable - our initial EDA of the response variable revealed that it had a heavily right skewed, unimodal distribution. Thus, we imposed a log transformation, which was more symmetric and normally distributed.

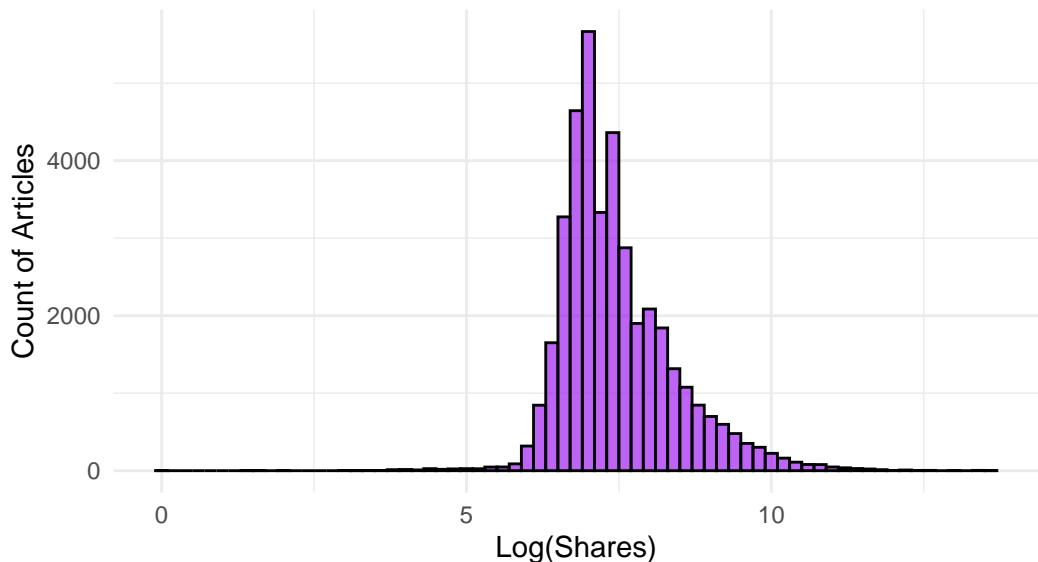
Distribution of Article Shares

Raw Shares Count (Limited to 50,000)



Log-Transformed Dist. of Article Shares

More normal distribution post-transformation



Key Variables - The key predictor variables we found from our initial exploration were Data Channel and Day Published, with the bivariate EDA we performed with our response variable, $\log(\text{shares})$, shown below.

Data Cleaning

(Combining existing weekday and data_channel indicator variables into their respective categorical variables.)

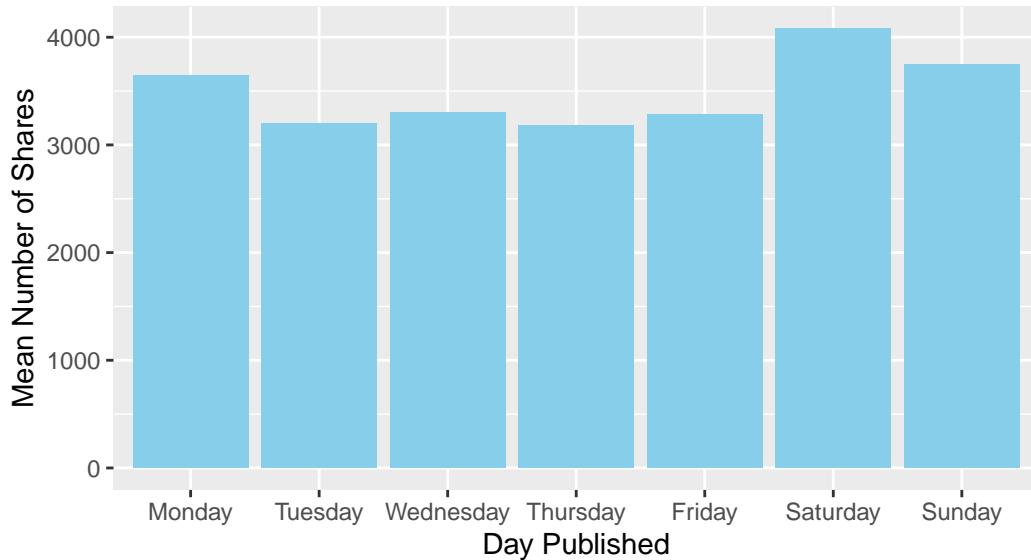
Table 1: Transformed Data

url	day_published	data_channel
http://mashable.com/2013/01/07/amazon-instant-video-browser/	Monday	Entertainment
http://mashable.com/2013/01/07/ap-samsung-sponsored-tweets/	Monday	Business
http://mashable.com/2013/01/07/apple-40-billion-app-downloads/	Monday	Business
http://mashable.com/2013/01/07/astronaut-notre-dame-bcs/	Monday	Entertainment
http://mashable.com/2013/01/07/att-u-verse-apps/	Monday	Technology

```
# A tibble: 6 x 2
  day_published mean_shares
  <fct>           <dbl>
1 Monday            3647.
2 Tuesday           3203.
3 Wednesday         3303.
4 Thursday          3179.
5 Friday            3285.
6 Saturday          4078.
```

Mean Article Shares by Day of Publication

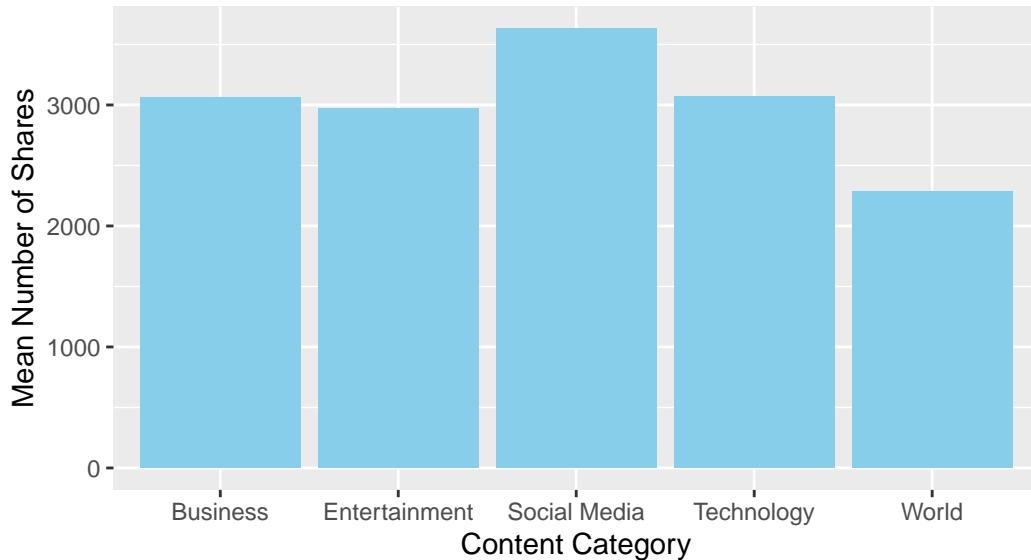
Weekend articles tend to receive more shares



```
# A tibble: 6 x 2
  data_channel  mean_shares
  <fct>          <dbl>
1 Business       3063.
2 Entertainment   2970.
3 Social Media   3629.
4 Technology     3072.
5 World          2288.
6 <NA>           5368.
```

Mean Article Shares by Content Category

Social Media articles more popular than other categories



Methodology

From our EDA, we found that the most significant predictor variables were `data_channel` and `day_published`, however, other predictors we initially viewed (such as `n_tokens_title`, etc.) were not significant. Thus, we chose to fit an initial MLR model using `data_channel` and `day_published`, alongside other predictors that we hadn't attempted before (such as `kw_avg_avg`, `n_tokens_content`, etc.).

This produced said results:

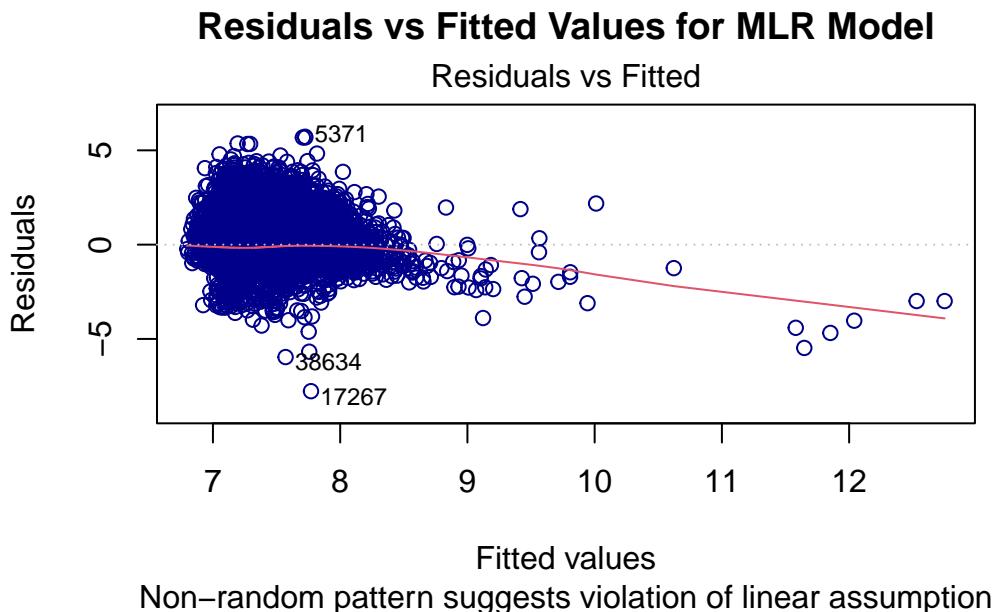
term	estimate	std.error	statistic	p.value
(Intercept)	6.8521	0.0286	239.3774	0.0000
<code>kw_avg_avg</code>	0.0001	0.0000	30.4256	0.0000
<code>n_tokens_content</code>	0.0001	0.0000	9.2061	0.0000
<code>data_channelEntertainment</code>	-0.1485	0.0145	-10.2424	0.0000
<code>data_channelSocial Media</code>	0.3056	0.0203	15.0761	0.0000
<code>data_channelTechnology</code>	0.1816	0.0144	12.6383	0.0000
<code>data_channelWorld</code>	-0.1531	0.0141	-10.8298	0.0000
<code>day_publishedTuesday</code>	-0.0516	0.0156	-3.3116	0.0009
<code>day_publishedWednesday</code>	-0.0590	0.0156	-3.7870	0.0002
<code>day_publishedThursday</code>	-0.0532	0.0157	-3.3904	0.0007
<code>day_publishedFriday</code>	0.0162	0.0168	0.9627	0.3357

term	estimate	std.error	statistic	p.value
day_publishedSaturday	0.2615	0.0224	11.6517	0.0000
day_publishedSunday	0.3113	0.0218	14.2499	0.0000
global_subjectivity	0.2492	0.0491	5.0792	0.0000
title_sentiment_polarity	0.0976	0.0189	5.1778	0.0000

Metric	Value
1 R^2	0.0903
2 Adjusted R-squared	0.0899
3 RMSE	0.8303

We found that all our predictors were statistically significant ($p\text{-value} < 0.05$), with the exception of day_publishedFriday, and thus kept them in this model.

Metric	Value
1 R^2	0.0903
2 Adjusted R-squared	0.0899
3 RMSE	0.8303



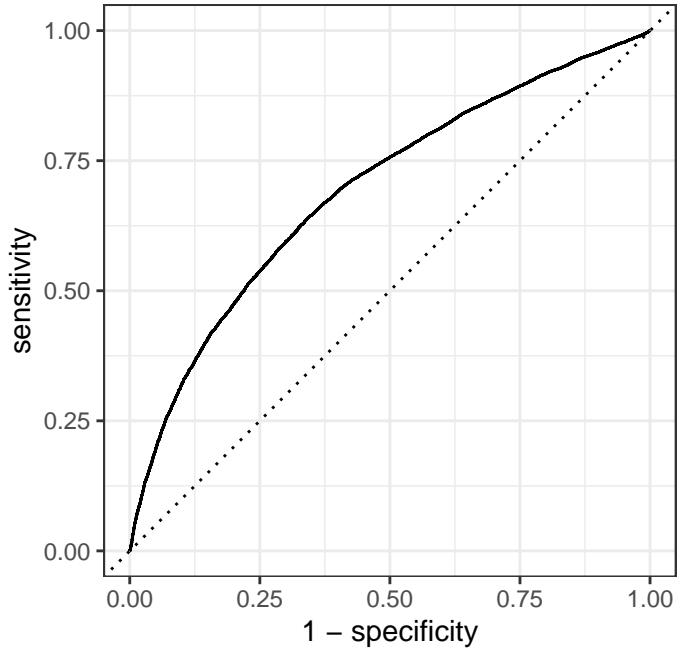
Our MLR had an $RMSE$ of 0.8303 and R^2 value of 0.0903 . Considering that our response variable was $\log(\text{shares})$, this suggests that this MLR model is poorly fit to our data, as it

can only explain about 9.03% of the variability that we can see from our response variable, $\log(\text{shares})$, and moreover, has a very high RMSE value. This is reaffirmed by the residuals plot, which reveals that the residuals are not randomly distributed, meaning that linear regression might not be the ideal model as our data does not satisfy the linear regression criteria.

Given these limitations, we pivoted to use a logistic regression model to classify an article as popular or otherwise. We selected a threshold for popularity of “1400 shares” based on prior literature, transforming shares column into a binary response variable with 1 for articles more popular than 1400 shares and 0 for those with less. Due to the heavy tails of ‘shares’ and non-linear relationship between the predictors and article shares, we would expect the logistic regression to be a more effective choice.

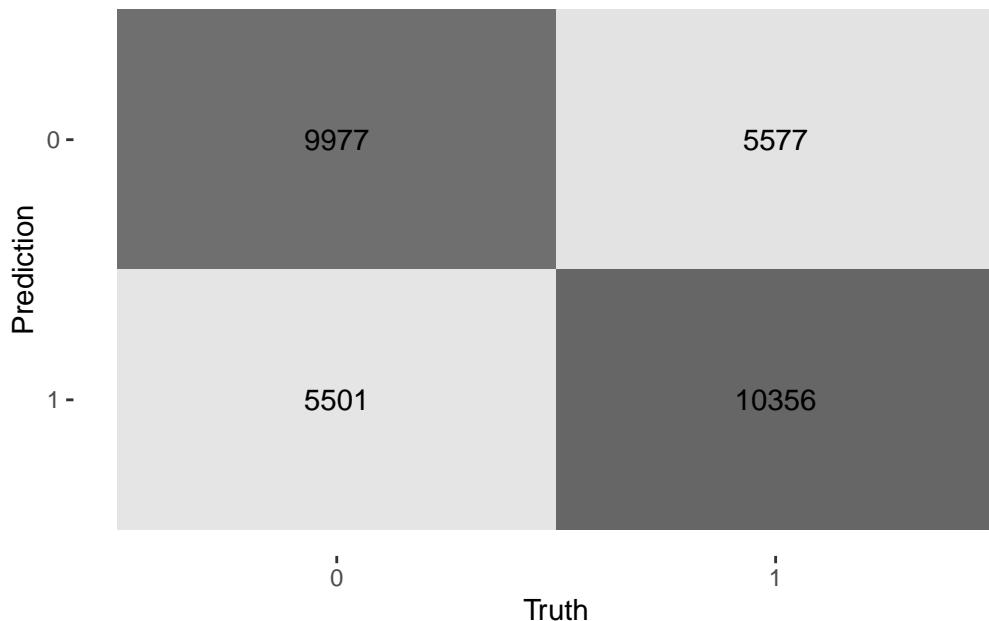
As for predictors, we continued using statistically significant predictors `data_channel`, `day_published`, etc.

term	estimate	std.error	statistic	p.value
(Intercept)	-1.3623	0.0782	-17.4177	0.0000
kw_avg_avg	0.0004	0.0000	23.4937	0.0000
n_tokens_content	0.0003	0.0000	10.1961	0.0000
data_channelEntertainment	-0.6580	0.0367	-17.9427	0.0000
data_channelSocial Media	0.8882	0.0560	15.8688	0.0000
data_channelTechnology	0.4839	0.0362	13.3727	0.0000
data_channelWorld	-0.5048	0.0355	-14.2039	0.0000
day_publishedTuesday	-0.1092	0.0394	-2.7754	0.0055
day_publishedWednesday	-0.1194	0.0393	-3.0376	0.0024
day_publishedThursday	-0.0675	0.0396	-1.7056	0.0881
day_publishedFriday	0.1450	0.0424	3.4206	0.0006
day_publishedSaturday	1.0247	0.0618	16.5937	0.0000
day_publishedSunday	0.8979	0.0580	15.4847	0.0000
global_subjectivity	0.5384	0.1258	4.2801	0.0000
title_sentiment_polarity	0.2244	0.0486	4.6197	0.0000



From the ROC curve, we estimated that the best threshold for our model would be when sensitivity was about 0.65. Thus, we found the closest point on the ROC curve to sensitivity = 0.65, which we found to be when threshold = 0.492. Using this calculation, we were then able to predict the virality categorization for each observation and generate our confusion matrix, as shown below.

5512
0.4916522



While our logistic regression model's ROC curve has an AUC of about 0.693, this is a dramatic improvement from our initial linear regression model, and thus we chose to continue using our logistic model for this data set.

Results

The final model we fitted was:

$$\begin{aligned}
\text{logit}(p_{isViral}) = & -1.3623 \\
& + 0.0004 \times \text{kwAvgAvg} \\
& + 0.0003 \times \text{nTokensContent} \\
& - 0.6580 \times \text{dataChannelEntertainment} \\
& + 0.8882 \times \text{dataChannelSocialMedia} \\
& + 0.4839 \times \text{dataChannelTechnology} \\
& - 0.5048 \times \text{dataChannelWorld} \\
& - 0.1092 \times \text{dayPublishedTuesday} \\
& - 0.1194 \times \text{dayPublishedWednesday} \\
& + 0.1450 \times \text{dayPublishedFriday} \\
& + 1.0247 \times \text{dayPublishedSaturday} \\
& + 0.8979 \times \text{dayPublishedSunday} \\
& + 0.5384 \times \text{globalSubjectivity} \\
& + 0.2244 \times \text{titleSentimentPolarity}
\end{aligned}$$

Table 4: Logistic Model Metrics Summary

Metric	Value
Accuracy	0.647
Misclassification Rate	0.353
Sensitivity (Recall)	0.650
Specificity	0.645
Precision	0.653
False Positive Rate (FPR)	0.355
False Negative Rate (FNR)	0.350
AUC	0.693

	GVIF	Df	GVIF^(1/(2*Df))
kw_avg_avg	1.095946	1	1.046874
n_tokens_content	1.030240	1	1.015007
data_channel	1.174273	4	1.020284
day_published	1.017642	6	1.001458
global_subjectivity	1.069706	1	1.034266
title_sentiment_polarity	1.007743	1	1.003864

Our logistic regression model has an AUC of around 0.693, an accuracy of 0.647, specificity of 0.645, and sensitivity of 0.650. In comparison, the mis-classification rate, FPR, and FNR rates

were 0.353, 0.355, and 0.350 respectively. This suggests that our model is moderately well fit for the data, as while the accuracy, specificity, sensitivity, and precision were relatively high at around 0.650, the FNR, FPR, and mis-classification rates were lower, at around 0.350. This precision means that approximately 65% of articles predicted to be viral were correctly classified, indicating the model performs significantly better than random chance. This relatively low predictive power may also be due to random noise, as many features of each article are likely uncaptured by the dataset and article virality may be influenced by sudden trends.

From our model, we can conclude that several key factors significantly influence article virality:

Content category plays a critical role in determining article popularity. Notably, Social media articles and technology articles are approximately 2.43 and 1.62 times, respectively, more likely than a similar business article to go viral. Conversely, Entertainment and World news are less successful categories, with 48.2% and 39.6% lower odds of going viral, respectively. This suggests that readers are particularly engaged with content about social media and technology innovations, while being less likely to widely share entertainment and world news.

Day of publication is another important factor in article virality. Weekend publications dramatically outperform weekday content, with Saturday articles enjoying 2.79 times higher odds and Sunday articles 2.45 times higher odds of virality compared to Monday publications. This weekend effect likely arises from increased leisure time as people take off from work or school, as weekdays show the opposite effect, with Tuesday and Wednesday's articles being 10.4% and 11.3% less likely to be viral.

While our initial set of “article sentiment” variables had relatively low predictive power, our model predicts a purely subjective article would have 71.3% higher odds of achieving viral status compared to purely objective content. Similarly, an article with purely positive sentiment would have 25.2% higher odds than a similar neutral title. This trend shows that, overall, more emotionally charged and polarizing content tends to be shared more often than neutral reporting.

Finally, while statistically important to the model, article length and keyword popularity have relatively small coefficients, making them less important to practical cases.

Appendix

Exploratory Data Analysis:

Data Set Description:

Our project utilizes the University of California Irvine Machine Learning Repository’s “Online News Popularity” data set. It includes share counts and descriptive characteristics for articles published by Mashable over two years (from 2013 to 2015). Mashable Inc. is a digital media website founded in 2005 and as of November 2015, it has over 6,000,000 Twitter followers and

over 3,200,000 fans on Facebook. The data set in total, has 39644 observations, each representing an individual article. Observations include characteristics such as: Number of Words in Title/Content, Rate of Unique Words, Number of Images, Data Channel, Day Published, Rate of Positive/Negative Words, Polarity, etc. Our intention is to use the data set to predict the number of shares/virality of an article based on different variables.

Key Variables:

rate_positive_words - rate of positive words among non-neutral tokens, which captures how emotionally charged the language is.

Rate_negative_words - rate of negative words among non-neutral tokens, which captures how emotionally charged the language is.

title_sentiment_polarity - A measure of how polarizing the title is

N_tokens_content - A measure of how long the article's content is

N_tokens_title - A measure of how long the article title is

data_channel - a categorical variable denoting article topic merged from: Data_channel_is_entertainment, data_channel_is_bus, data_channel_is_socmed, data_channel_is_tech, and data_channel_is_world.

day_published- a categorical variable indicating publication day merged from indicators: Weekday_is_monday, weekday_is_tuesday, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday, weekday_is_saturday, weekday_is_sunday

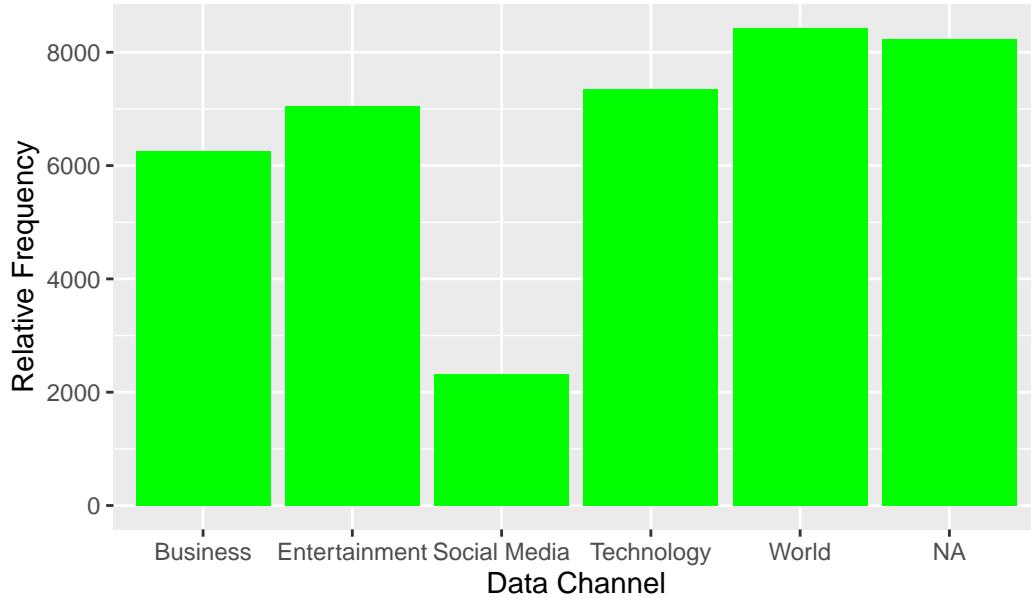
Data Cleaning

First we have to combine the existing weekday and data_channel indicator variables into their respective categorical variables.

Table 5: Transformed Data

url	day_published	data_channel
http://mashable.com/2013/01/07/amazon-instant-video-browser/	Monday	Entertainment
http://mashable.com/2013/01/07/ap-samsung-sponsored-tweets/	Monday	Business
http://mashable.com/2013/01/07/apple-40-billion-app-downloads/	Monday	Business
http://mashable.com/2013/01/07/astronaut-notre-dame-bcs/	Monday	Entertainment
http://mashable.com/2013/01/07/att-u-verse-apps/	Monday	Technology

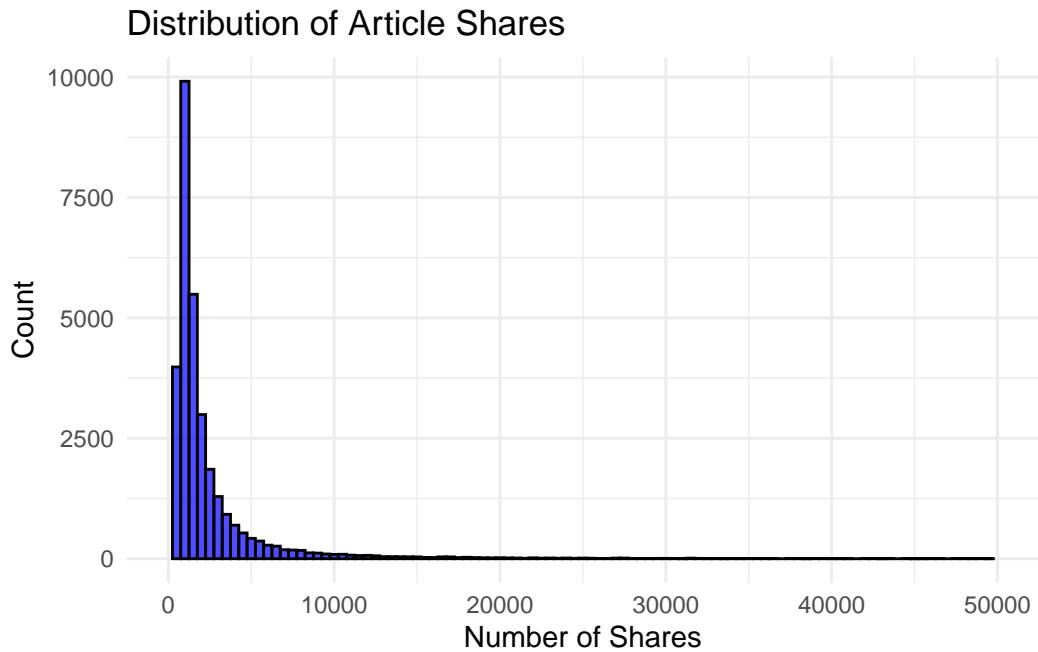
Distribution of article categories



Next, we discovered that approx 8,233 articles are not tagged for a specific data channel. Due to the nature of the dataset, it's unclear if this was because the article was simply missing a tag, it was mis-tagged while being collected, or if it simply doesn't belong in any of these categories. With the relatively large size of our dataset, we decided to exclude entries lacking a

data tag NA's from our data channel analysis altogether. These articles lacking a data channel were filtered out.

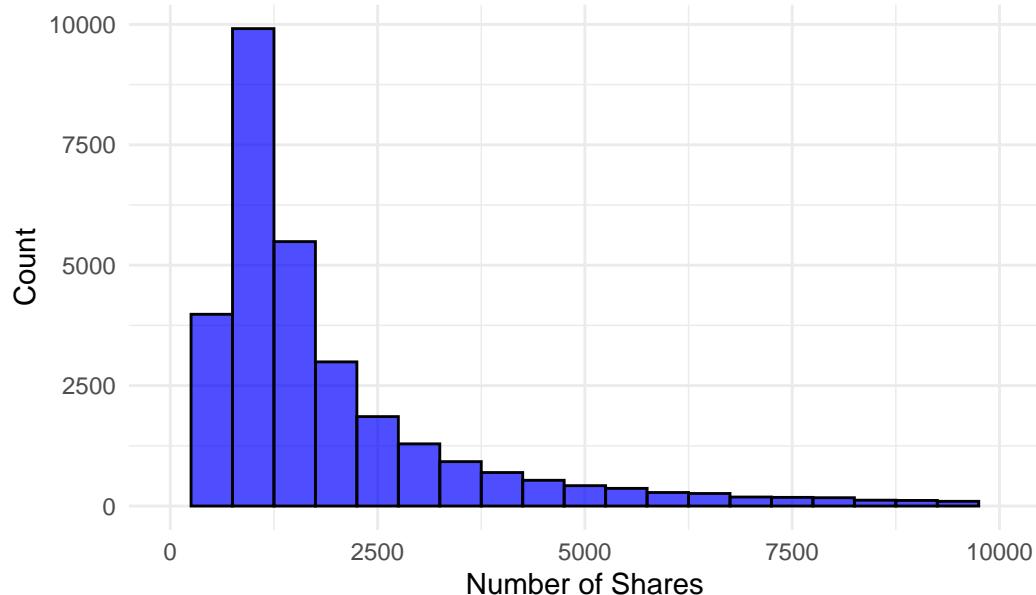
Response Variable/Univariate EDA



```
# A tibble: 1 x 7
  mean_shares median_shares sd_shares min_shares max_shares     q1     q3
    <dbl>        <dbl>      <dbl>      <dbl>      <dbl> <dbl> <dbl>
1     2878.       1400      9506.       1     690400     923   2500
```

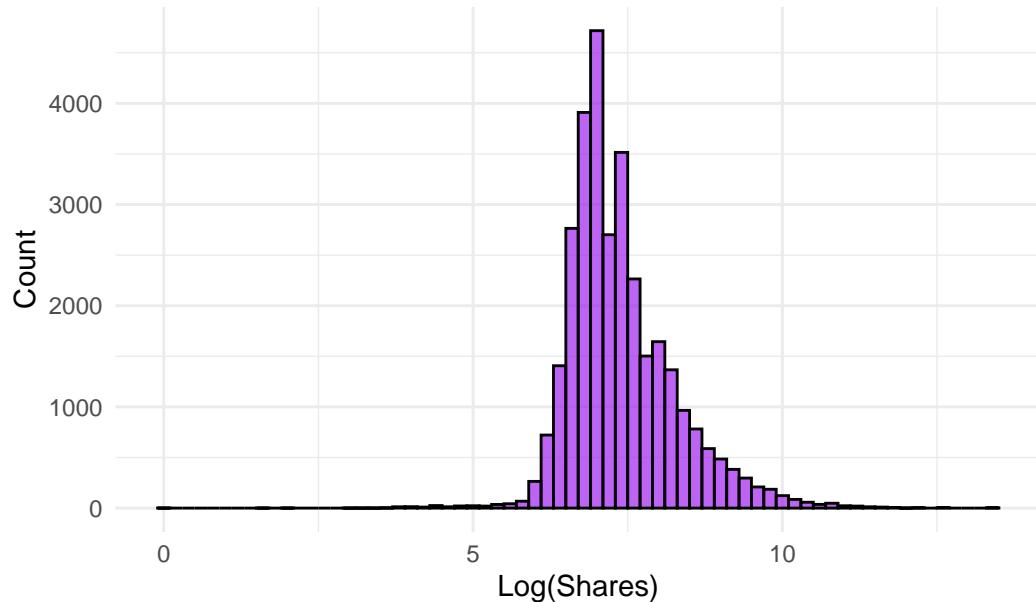
The distribution of # of shares is highly right skewed, a median of 1400 shares and a few highly shared articles. Notably, the mean of 2878 shares is far larger

Distribution of Article Shares (Zoomed In)

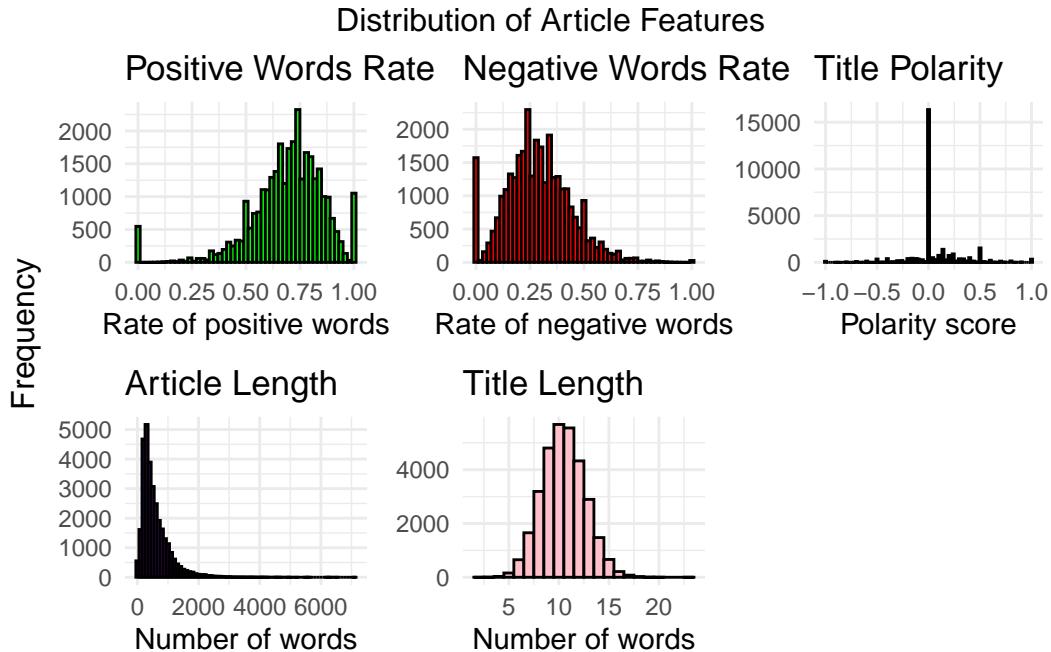


To make deal with this strong right skew, we applied a log transformation to the share variable, yielding a less skewed distribution.

Log–Transformed Distribution of Shares

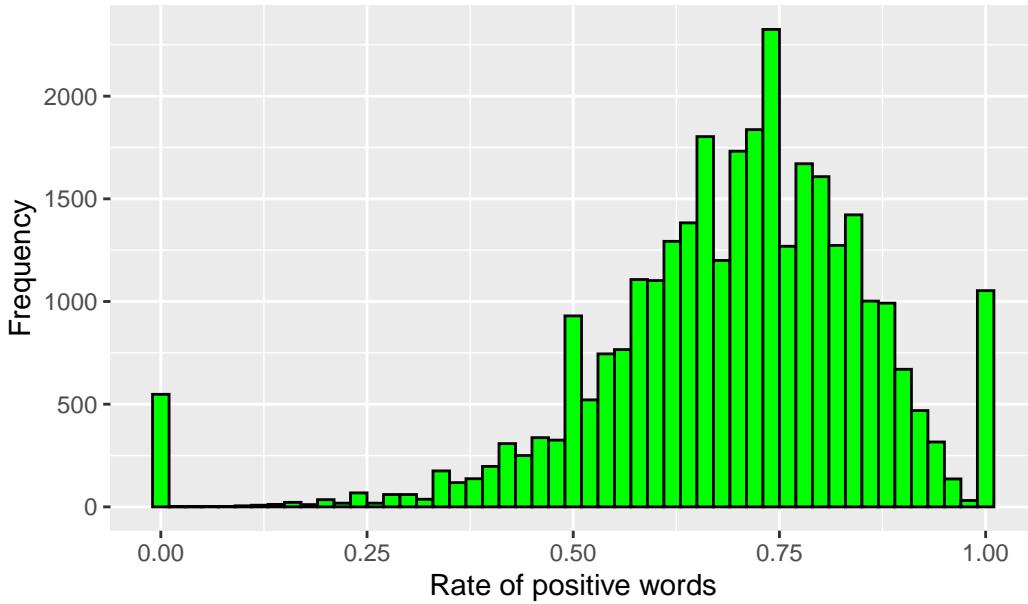


Predictor Variable/Univariate EDA



Examining the rate of positive words in an article, we see a left skewed distribution, with modes at ~ 0 , ~ 0.75 and ~ 1 . The median positivity is approx 0.71 positivity rate, and the range is from 0 to 1.

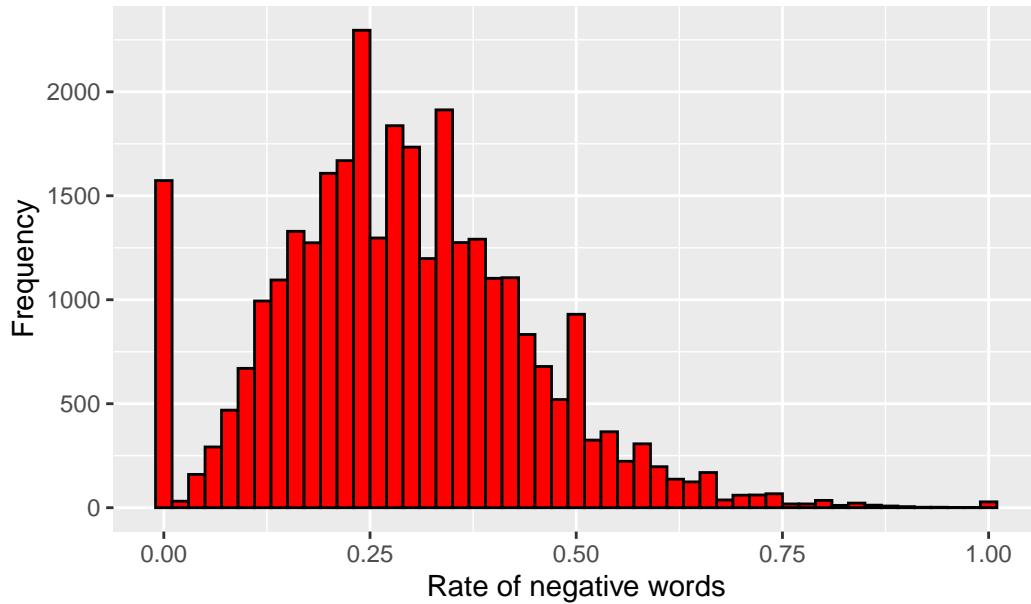
Rel. Freq of Positive Words Rate



```
# A tibble: 1 x 5
  mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1 0.694  0.714   0.172    0     1
```

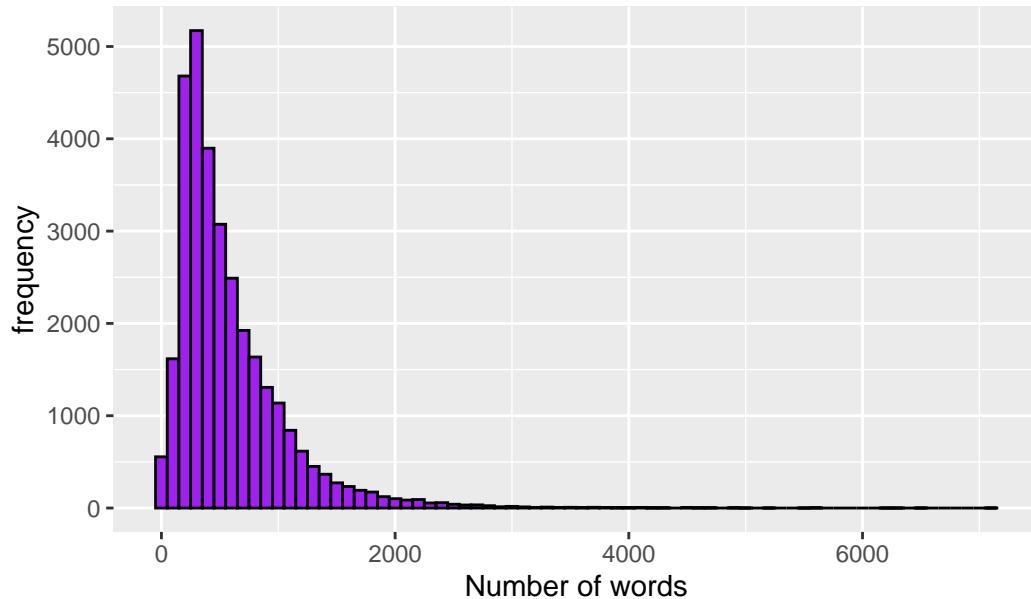
The rate of negative words shows the opposite trend, with a slight right skew. Similarly, there seems to be a second mode at 0 negativity. The median is approx 0.28 negativity rate, with an approximately equal mean.

Rel Freq of Negative Words Rate



```
# A tibble: 1 x 5
  mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1 0.290  0.280   0.152    0     1
```

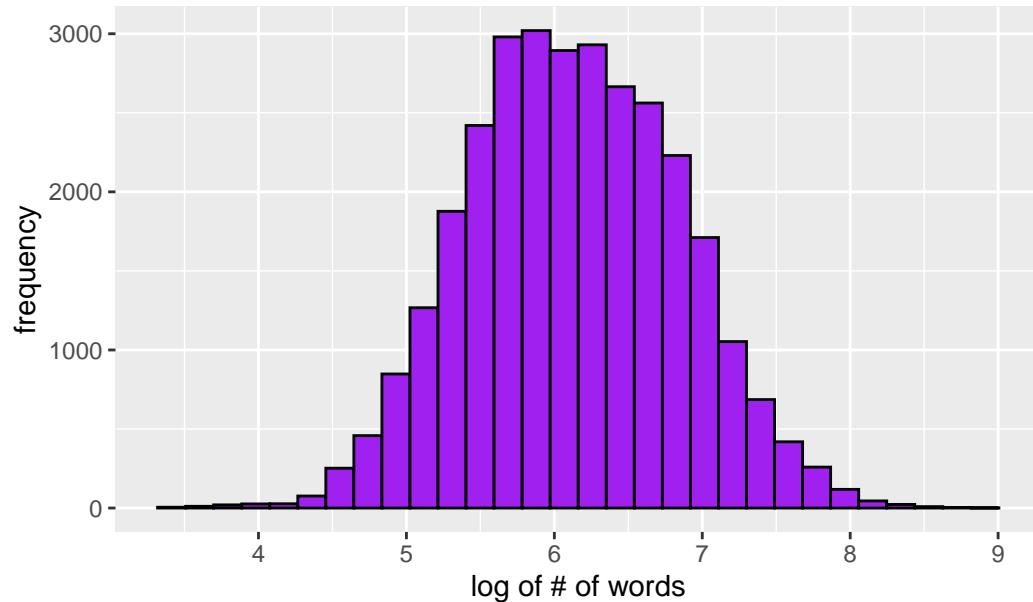
Rel. Freq. of Article Length



```
# A tibble: 1 x 5
  mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1 583.    444    478.     0 7081
```

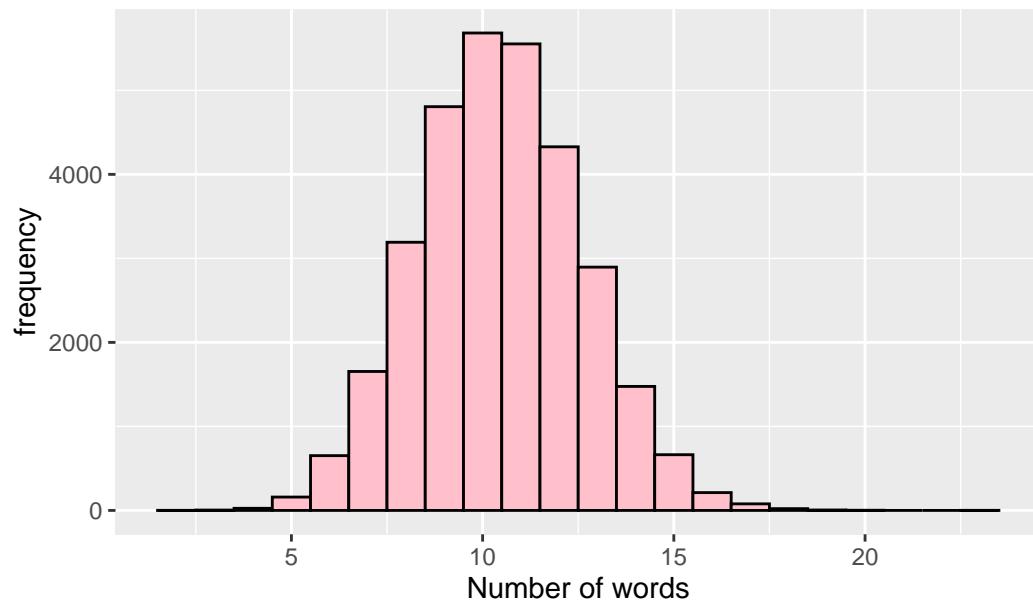
For article length, the graph shows a strongly right-skewed distribution, with a median length of 444 and a high standard deviation of 477 words. To remedy this, we might consider a log transformation which yields a more even distribution.

Rel. Freq of Log transformed Article Length



For the number of tokens in the title, we can see a highly symmetric distribution centered at 10, with a standard deviation of 2.14.

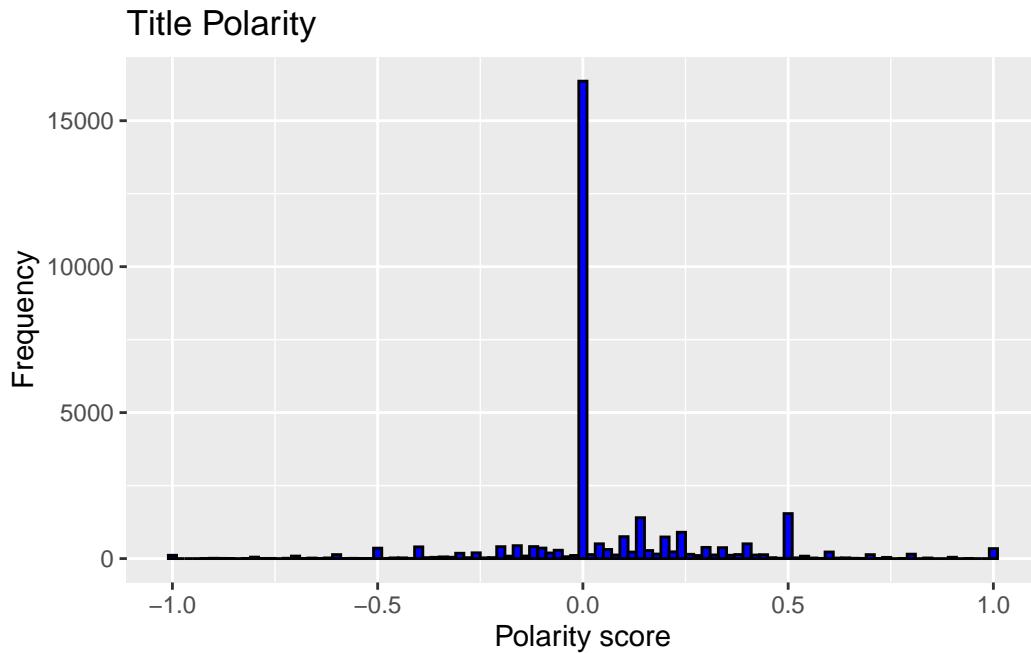
Rel. Freq. of Title Length



```
# A tibble: 1 x 5
```

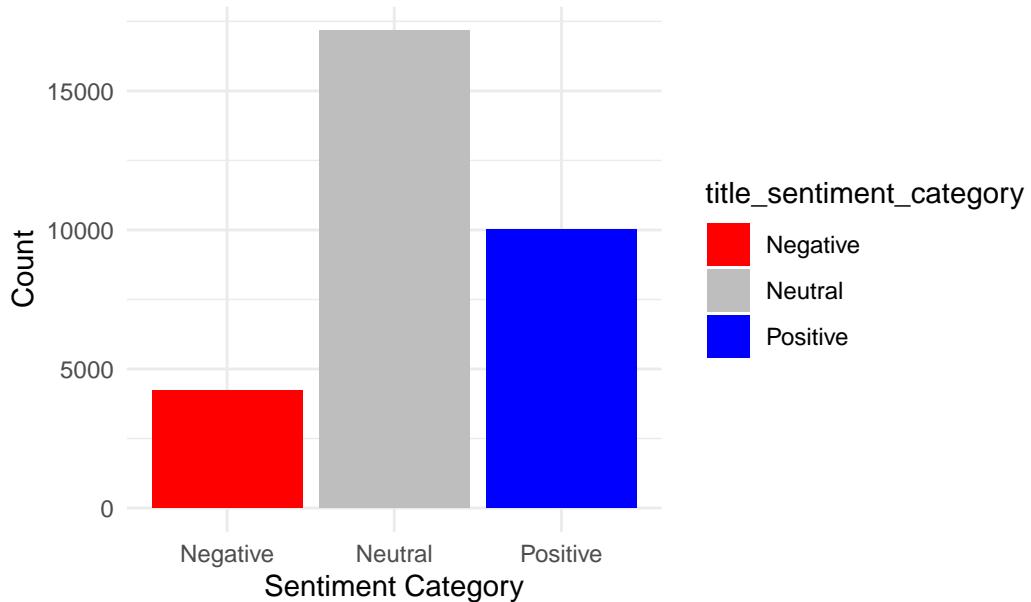
	mean	median	std.dev	min	max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	10.5	10	2.14	2	23

Finally, our initial EDA of title polarity found a massive frequency spike at 0 frequency, which might correspond to failed measurements or the vast majority of our articles not presenting significant title polarity.

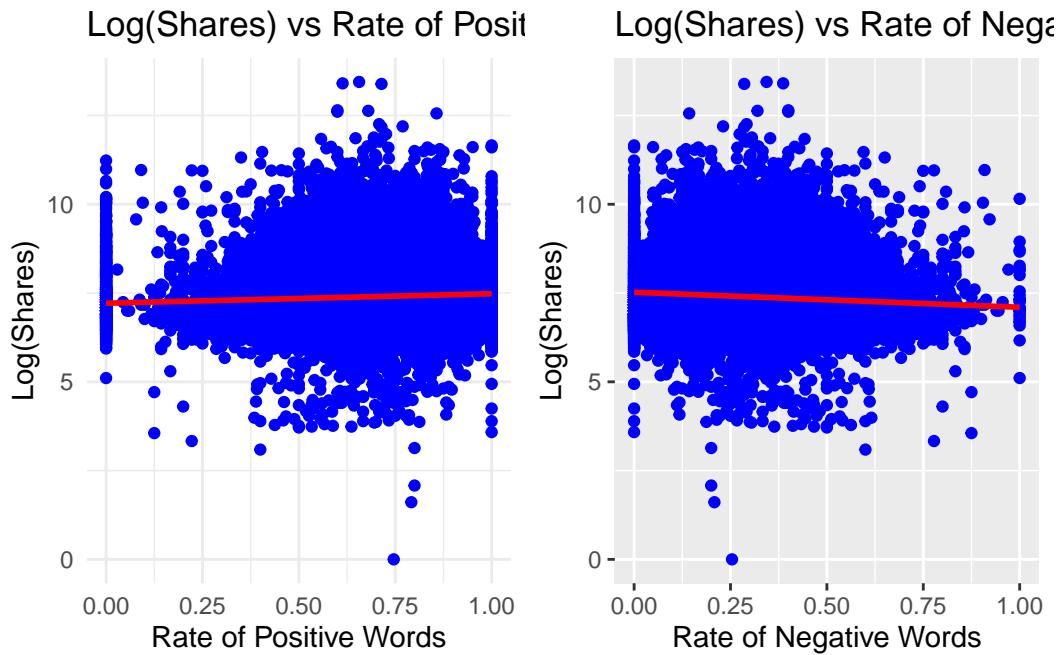


To fix this issue and for ease of use, we categorized articles into negative, neutral and positive polarity, with a threshold of 0 ± 0.05 for neutral. Most titles remain neutral, and there are more positive than negative headlines.

Distribution of Title Sentiment Polarity

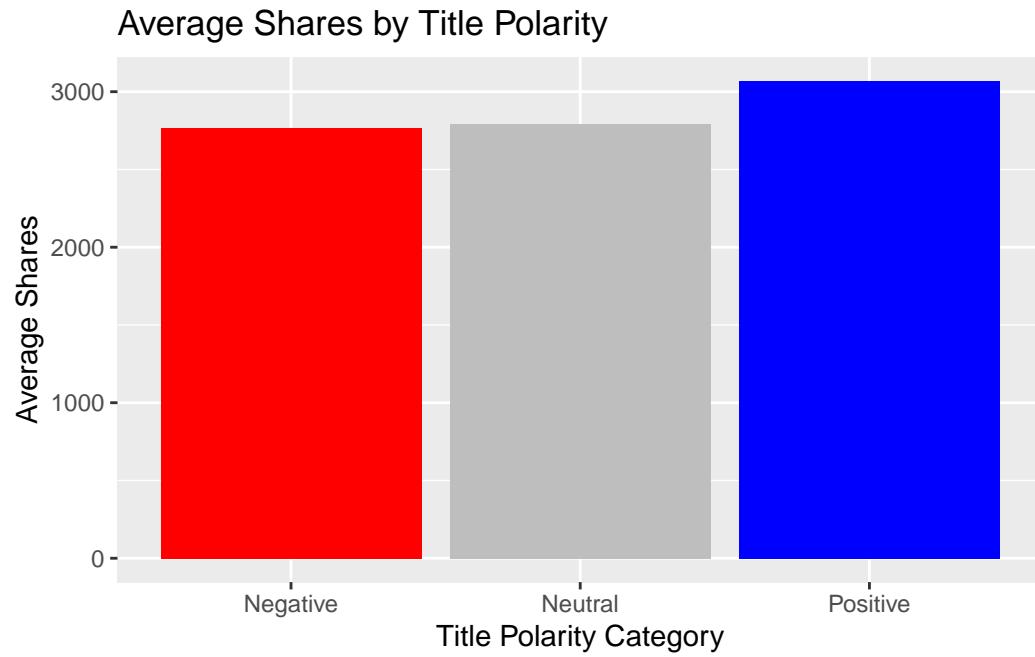


Bi-variate EDA



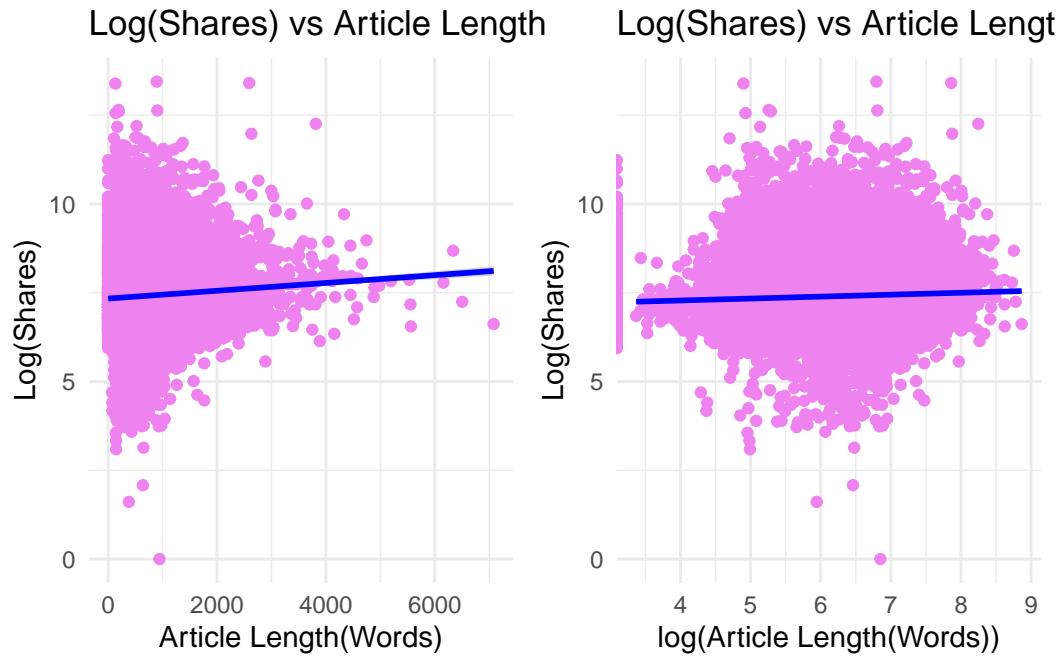
By plotting the rate of positive words and the rate of negative words against the log transformed share, we can see that neither have a particularly strong relationship with how often the article

is shared. The rate of positive words seems to have a weak positive relationship with shares, and the rate of negative words seems to have a weak negative relationship, but both have significant outliers are 1 and 0.



In contrast, there seems to be some relationship between title polarity and the number of shares, with positive polarity associated with greater share counts.

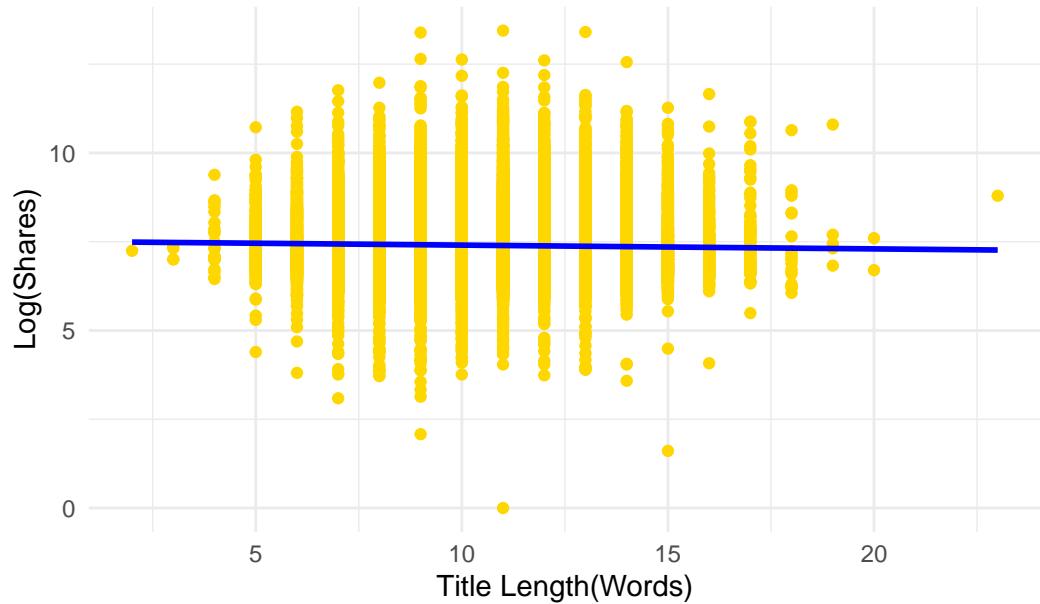
Next,



These graphs show a weak, positive relationship between article length and the log transformed number of shares. Due to the skew, we can apply the log transform to the article length. This shows a more even distribution, with no clear relationship.

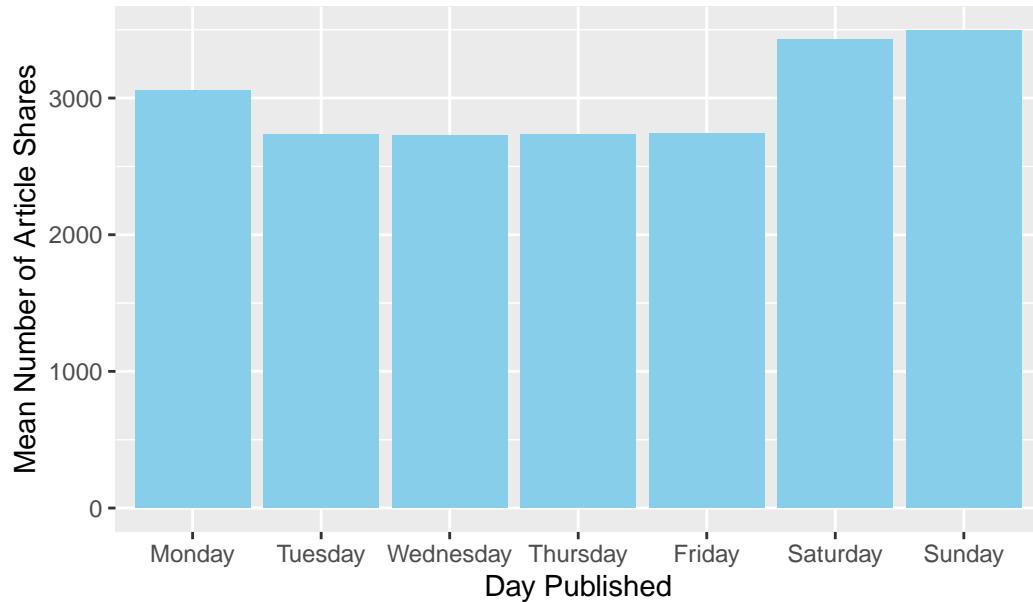
Similarly, there doesn't seem to be any clear relationship between title length and the number of shares in the graph below.

Log(Shares) vs Title Length



```
# A tibble: 6 x 2
  day_published mean_shares
  <fct>           <dbl>
1 Monday          3057.
2 Tuesday         2731.
3 Wednesday       2727.
4 Thursday        2737.
5 Friday          2741.
6 Saturday        3431.
```

Mean Number of Article Shares vs. Day Published



term	estimate	std.error	statistic	p.value
(Intercept)	7.505	0.038	197.123	0.000
rate_negative_words	-0.408	0.045	-9.071	0.000
rate_positive_words	0.017	0.040	0.437	0.662

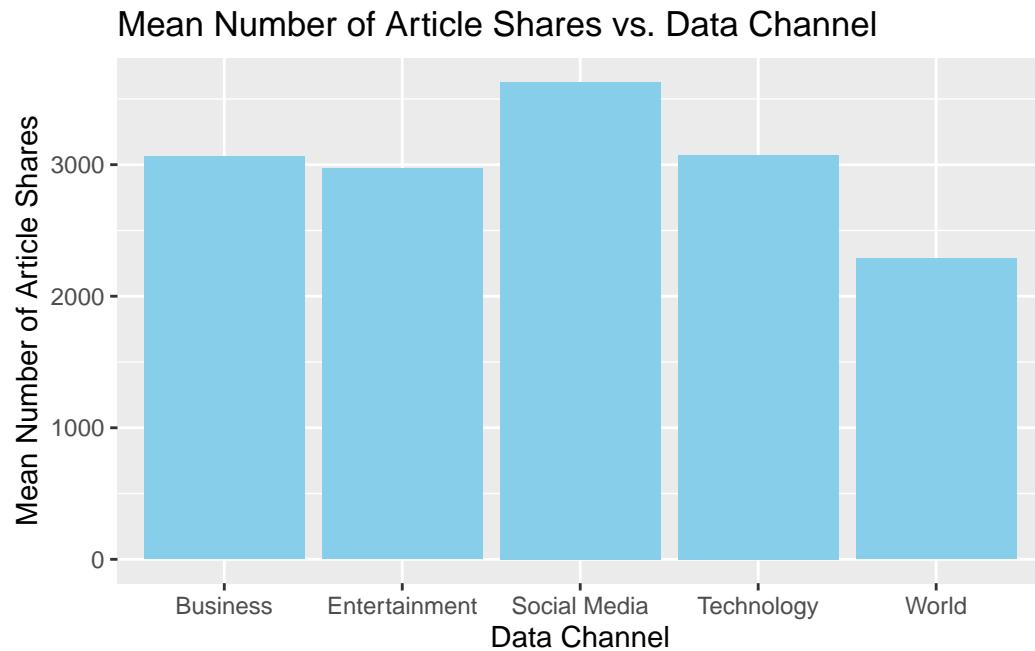
```
[1] "Variance Inflation Factors:"
```

```
rate_negative_words rate_positive_words
1.946291           1.946291
```

This visualization and summarization suggests that the day an article is published does not have a significant impact on the virality of an article, as the mean number of article shares do not differ much between days. Therefore, this predictor may not be as important as others when it comes to predicting the virality of an article.

```
# A tibble: 5 x 2
  data_channel  mean_shares
  <fct>          <dbl>
1 Business       3063.
2 Entertainment   2970.
3 Social Media   3629.
```

4 Technology	3072.
5 World	2288.



From our visualizations, it appears that articles that are in the social media data channel perform the best in terms of average number of shares (~3600), while articles that are in the World data channel perform the worst (~2250). Business, Entertainment, and Technology articles all seem to receive a mean of about 3000 shares. However, from our earlier univariate analysis, we saw that the Social Media Data Channel has the lowest number of articles, and thus there is a possibility that any outliers for this data channel category would have a larger impact in skewing the mean.

Interaction Effects Exploration

Table 7: Regression Coefficients for Both Models

Term	Estimate	Std. Error	t value	p-value	Model
(Intercept)	7.451	0.025	297.679	0.000	Model 1
n_tokens_content	0.000	0.000	10.822	0.000	Model 1
n_tokens_title	-0.011	0.002	-4.886	0.000	Model 1
(Intercept)	7.248	0.038	192.902	0.000	Model 2
n_tokens_content	0.000	0.000	9.410	0.000	Model 2
n_tokens_title	0.008	0.004	2.308	0.021	Model 2
n_tokens_content:n_tokens_title	0.000	0.000	-7.265	0.000	Model 2

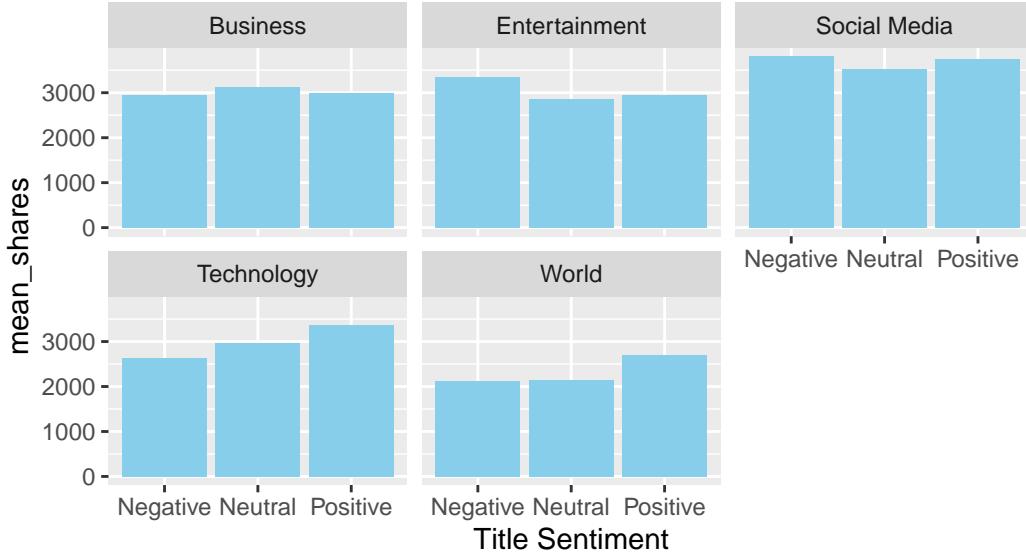
Table 8: ANOVA Comparison of Models

term	df.residual	rss	df	sumsq	statistic	p.value
log(shares) ~ n_tokens_content + n_tokens_title	31408	23701.16	NA	NA	NA	NA
log(shares) ~ n_tokens_content + n_tokens_title + n_tokens_content:n_tokens_title	31407	23661.401	39.762	52.779	0	

When comparing linear models with the title and article length as predictors, we explored to see if an interaction effect would have a meaningful difference. Our results show that the effect of title length depends on article length. IE, for very short articles, we'd expect longer titles to be beneficial and for longer articles, vice versa.

Mean Shares for Title Sentiment Categories

Faceted by Data Channel Type



From this visualization, it's suggested that the title sentiment may have different impacts on the mean number of shares depending on the type of data channel, thus suggesting that there may be a statistically significant interaction effect between the title sentiment and the data channel type. For instance, while for Social Media and Entertainment, it appears that articles with a Negative sentiment have the greatest number of mean shares, for World and Technology articles, Positive sentiment articles had the greatest mean shares, and for Business, Neutral sentiment articles had the greatest number.

term	estimate	std.error	statistic	p.value
(Intercept)	7.341	0.033	220.412	0.000
data_channelEntertainment	-0.032	0.042	-0.768	0.443
data_channelSocial Media	0.418	0.066	6.322	0.000
data_channelTechnology	0.144	0.046	3.111	0.002
data_channelWorld	-0.177	0.040	-4.406	0.000
title_sentiment_categoryNeutral	0.071	0.036	1.969	0.049
title_sentiment_categoryPositive	0.083	0.038	2.169	0.030
data_channelEntertainment:title_sentiment_categoryNeutral	0.046	0.046	-1.863	0.062
data_channelSocial	-0.043	0.072	-0.595	0.552
Media:title_sentiment_categoryNeutral				
data_channelTechnology:title_sentiment_categoryNeutral	0.050	0.405	0.685	
data_channelWorld:title_sentiment_categoryNeutral	-0.059	0.044	-1.326	0.185
data_channelEntertainment:title_sentiment_categoryPositive	0.049	0.049	-1.252	0.211
data_channelSocial	-0.077	0.075	-1.024	0.306
Media:title_sentiment_categoryPositive				
data_channelTechnology:title_sentiment_categoryPositive	0.053	0.929	0.353	
data_channelWorld:title_sentiment_categoryPositive	0.039	0.048	0.810	0.418

However, when looking at the actual fitted model, all the interaction terms between the data channel and the sentiment category have large p-values (greater than 0.05) suggesting that none of the interaction terms are actually statistically significant between data channel and title sentiment category. In the future we could consider looking at the mean of the shares rather than the log, as it appears as though there may be interaction effects for the mean but not the log of the shares.

term	estimate	std.error	statistic	p.value
(Intercept)	7.430	0.052	141.723	0.000
n_tokens_title	-0.002	0.005	-0.417	0.677
data_channelEntertainment	-0.150	0.076	-1.978	0.048
data_channelSocial Media	0.419	0.098	4.270	0.000
data_channelTechnology	0.292	0.072	4.064	0.000
data_channelWorld	-0.415	0.071	-5.827	0.000
n_tokens_title:data_channelEntertainment	0.005	0.007	0.676	0.499
n_tokens_title:data_channelSocial Media	-0.005	0.010	-0.554	0.579
n_tokens_title:data_channelTechnology	-0.012	0.007	-1.712	0.087
n_tokens_title:data_channelWorld	0.020	0.007	2.956	0.003

From this model, we found that while the data channel type at times has a statistically significant linear relationship to log(shares), specifically for Social Media, Technology, and World,

the number of words in the title does not have a linear relationship to $\log(\text{shares})$, based on their respective p-values. For instance, the p-value for `n_tokens_title` is 0.677, thus suggesting that there is not a statistically significant linear relationship between `n_tokens_title` and $\log(\text{shares})$. However, while the number of words in the title does not have a significant linear relationship with $\log(\text{shares})$ directly, its interaction term specifically with when the data channel is World, is significant (as shown by the p-value of 0.003).