Evaluating Article Virality based off of Article Attributes

The BEST Fit - Philip, Olivia, Leo, Allison

2025-04-28

Introduction and Data

Project Motivation and Research Question

The ways in which people interact with media and discover news have dramatically shifted in recent years, with social media often displacing traditional news outlets. The decentralized nature of social media means the reach of each article is largely dependent on its individual merits, rather than the popularity of the publication it belongs to.

Thus a question arises: What article attributes are associated with social media virality?

Dataset and Key Variables

In this report we investigate the effects of different article features on social media success using 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, a digital media website, over two years (from 2013 to 2015). The data has 39,644 entries in total, with each repsenting an individual article and its associated textual and metadata features.

Key Variables:

is_viral - Binary response variable created that evaluates whether an article is considered viral or not, based off of whether total shares is greater/less than 1,400 [INSERT CITATION HERE] (0: FALSE, 1: TRUE).

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.

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.

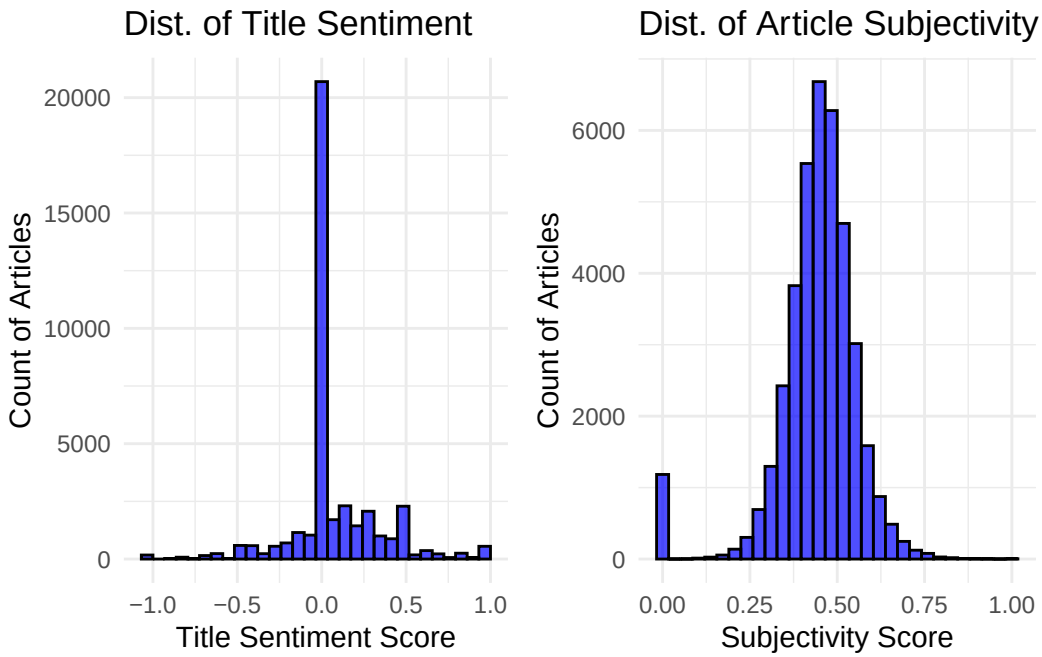
n_tokens_content - Number of words in the article content. Values range from 0 to 8,474 words, with a mean of 546.51 and standard deviation of 471.10. This quantifies the overall length of the article.

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.

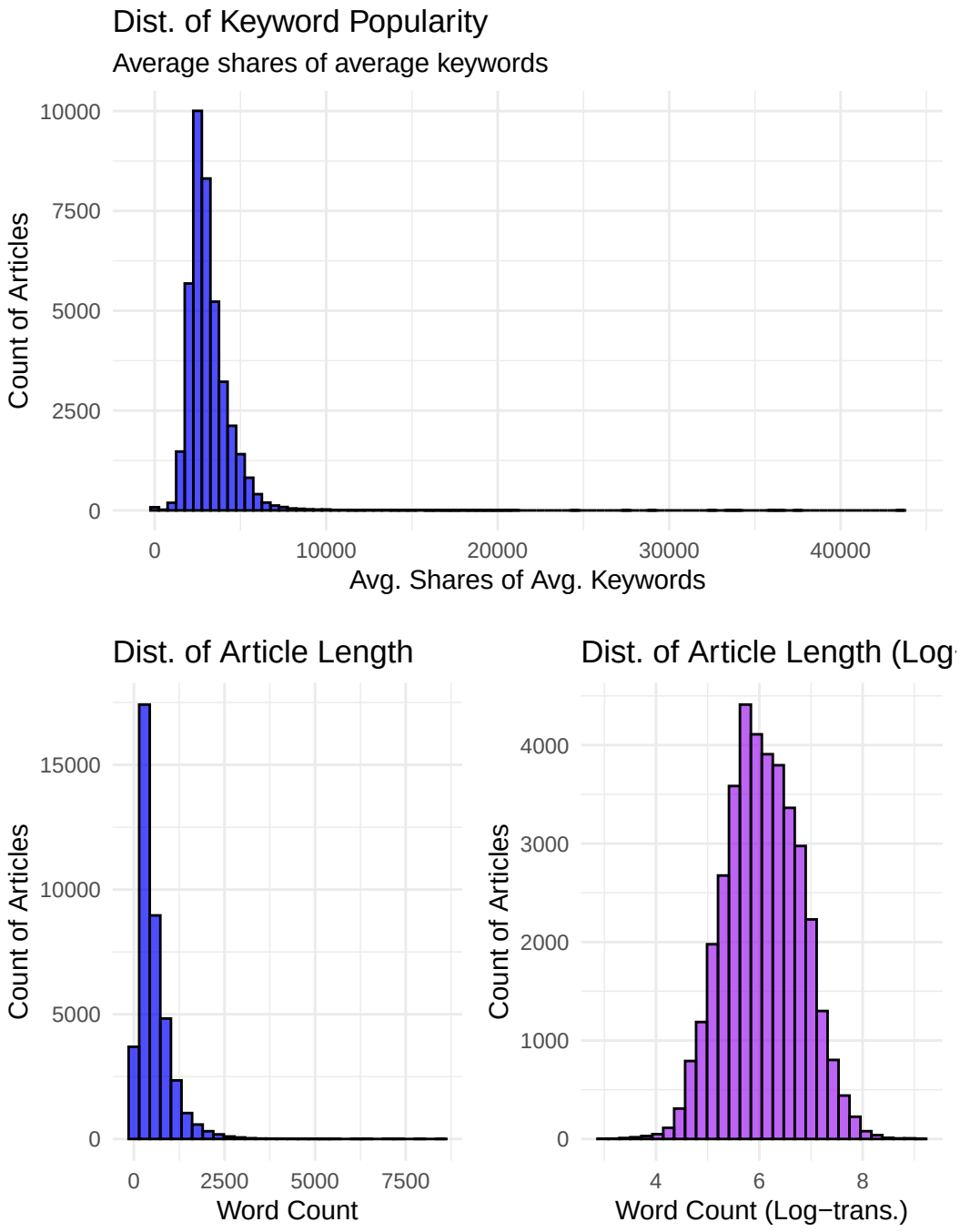
global_subjectivity - Measures the overall subjectivity of the article text. Values range from 0.0 (completely objective) to 1.0 (completely subjective), with a mean of 0.4434 and standard deviation of 0.1167. This quantifies how opinion-based versus fact-based the content is.

Univariate EDA

To understand our response and predictor variables more deeply, we first looked at their individual distributions. We found that while some variables are relatively approximately symmetric, others are heavily skewed and required log transformations to better meet modeling assumptions.

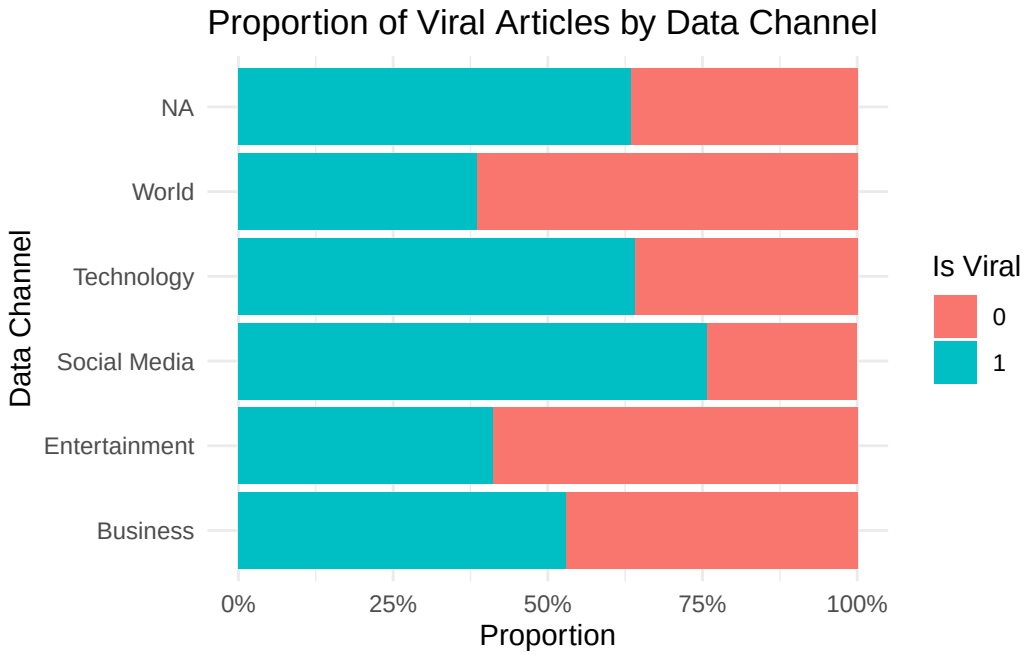


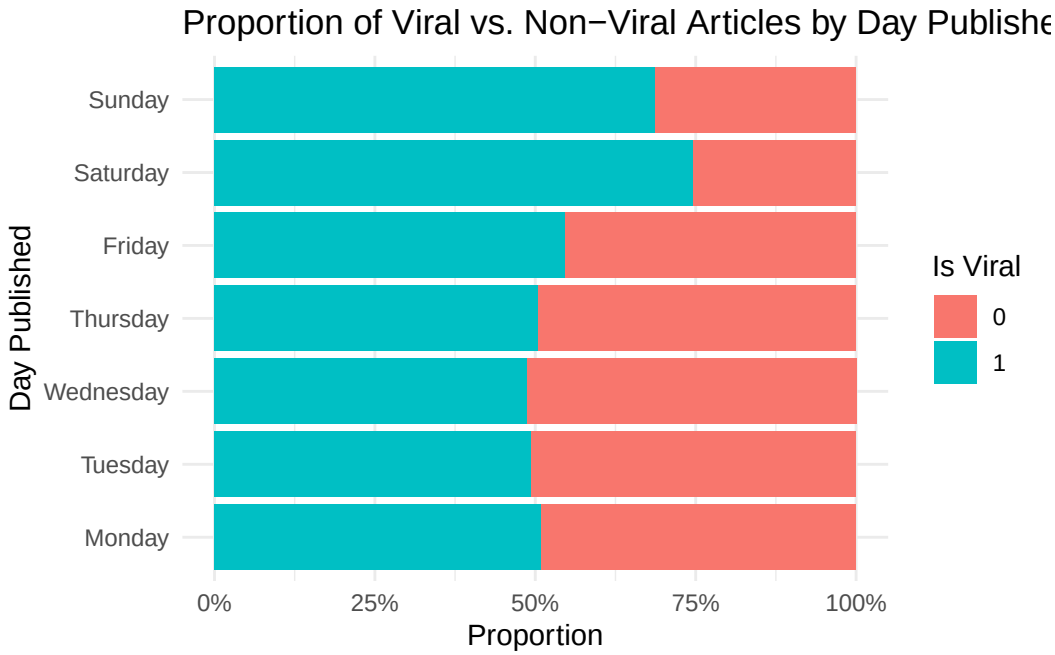
The title_sentiment_polarity variable quantifies the emotional change of article titles. The distribtuion suggests that many titles are emotionally neutral. The global_subjectivity variable is rougly symmetric with its distribution suggesting that most articles contained a balanced mix of factual and opinion-based language. Both global_subjectivity and title_sentiment_popularity display relatively balanced distributions that do not require transformation.



Both `n_tokens_content` and `kw_avg_avg` displayed heavily right-skewed distributions that warranted log transformations.To correct for this skewness and reduce the influence of extreme values, we log-transformed both variables. Post-transformation, the distributions became more symmetric and centered, better aligning with modeling assumptions.

Bivariate EDA

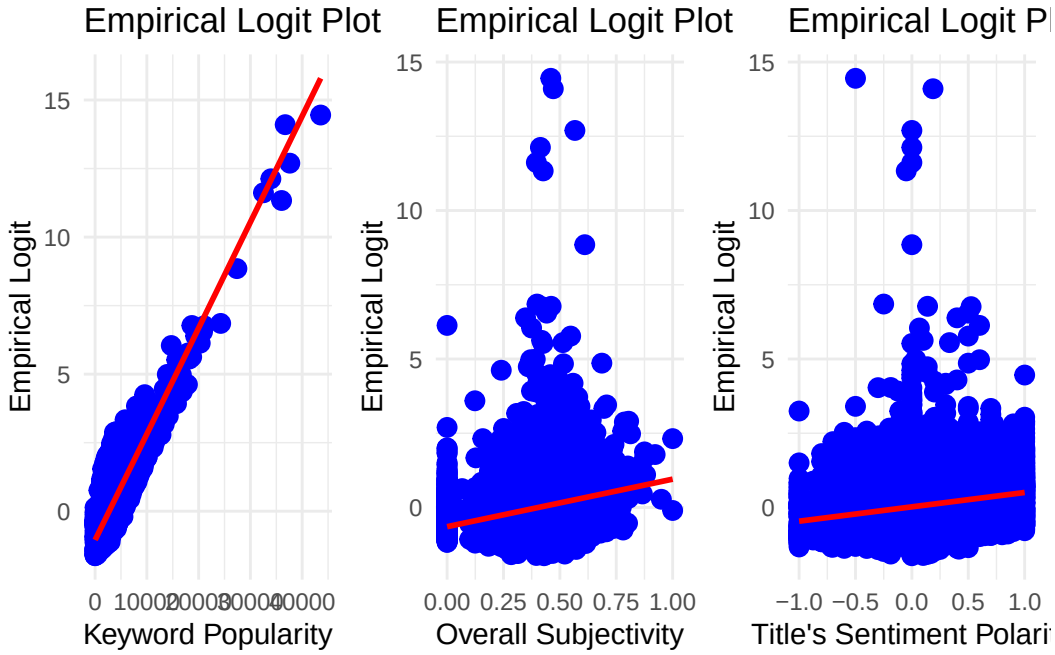




The first graph explores how virality varies across different content categories, as captured by the data_channel variable. Article published under the Social Media category have the highest

proportion of viral outcomes, with many reaching viral status, This suggest that content tailored for or about social platforms may be particularly good for engagement. Articles that are categoized under Technology and Business also show strong performance, with over half of the articles in each category going viral, On the other hand, articles in the Entertainment and World categories are less likely to go viral, falling below the 50% mark. This graph underscores how topic area influences content reach potentially because of the differences in audience behavior or platform algorithms.

The second graph examines the relationship between the day an article is published and its likelihood of going viral. Articles published on weekends are substantially more likely to be viral compared to those published during the week. Saturday stands out with the highest porportion of viral content, followed closely by Sunday. In contrast, weekday article tend to have lower virality rates, with viral and non-viral occurring in nearly equal proportions. These findings suggest that timing plays a role in determining an article's reach, likely reflecting differences in user engagement patterns across the week.



The empirical logit plots offer further insight into the relationship between key predictors and the binary outcome is_viral. The plot for average shares of average keywords reveals a positive relationship with virality. As the average popularity of an article's keywords increases, the likelihood of the article going viral rises sharply, with the empirical logit showing an upward linear trend. This suggest that keyword selection plays a key role in driving article engagement and affirms kw_avg_avg is a key predictors in modeling vitality. In contrast, the plot of overall subjectivity indicated on a slight positive relationship. Although the fitted line trends upward, the data are widely dispersed, and the effect appears weak and inconsistent,

implying subjectivity alone is not a reliable driver of viral outcomes. A similar conclusion can be drawn from the plot of title sentiment polarity. While there is a marginal positive slope, suggesting that more positive titles generate a slightly greater chance of going viral, the overall relationship is weak. The plots highlight a clear distinction in predictive power among the variables and support prioritizing keyword metrics over emotional tone or subjectivity when modeling article vitality.

Methodology

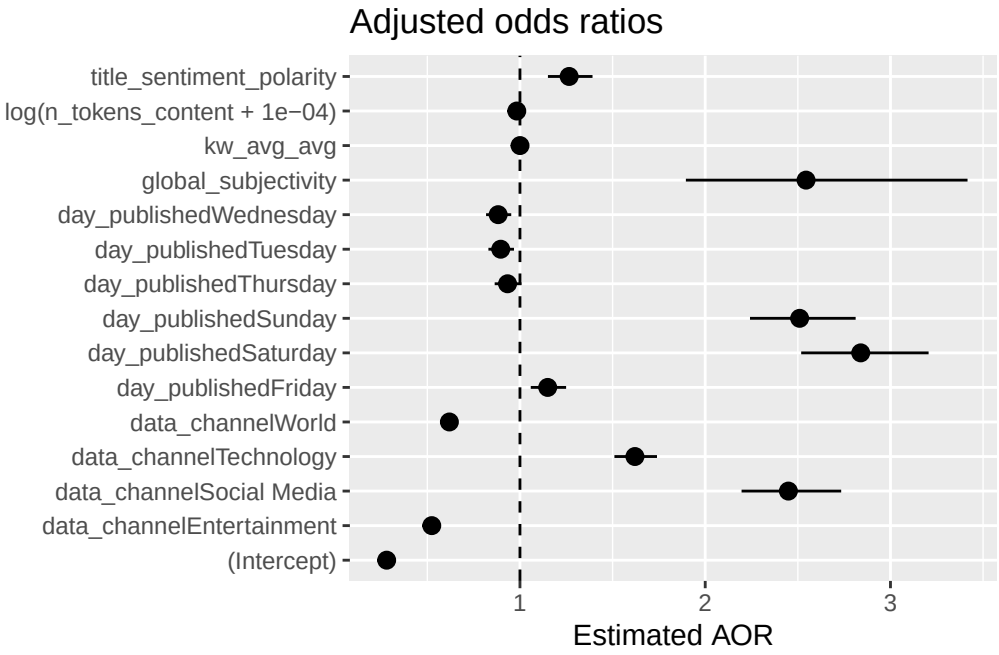
Based on our initial EDA and empirical logit visualization, we selected data channel, day published, article subjectivity, title sentiment polarity, average keyword popularity, and (log) article content length to fit an initial logistic model. Our model attributes are shown below:

Term	Estimate	Std.Error	z-statistic	p-value
(Intercept)	-1.2713	0.0782	-16.2534	0.0000
kw_avg_avg	0.0004	0.0000	23.0515	0.0000
log(n_tokens_content + 1e-04)	-0.0173	0.0070	-2.4803	0.0131
data_channelEntertainment	-0.6464	0.0366	-17.6423	0.0000
data_channelSocial Media	0.8957	0.0559	16.0287	0.0000
data_channelTechnology	0.4828	0.0361	13.3593	0.0000
data_channelWorld	-0.4796	0.0354	-13.5503	0.0000
day_publishedTuesday	-0.1097	0.0393	-2.7913	0.0052
day_publishedWednesday	-0.1253	0.0393	-3.1910	0.0014
day_publishedThursday	-0.0694	0.0395	-1.7574	0.0788
day_publishedFriday	0.1395	0.0423	3.2957	0.0010
day_publishedSaturday	1.0436	0.0616	16.9403	0.0000
day_publishedSunday	0.9202	0.0578	15.9141	0.0000
global_subjectivity	0.9338	0.1502	6.2160	0.0000
title_sentiment_polarity	0.2354	0.0484	4.8601	0.0000

Our initial fit gives all of the predictors significant p-values ($p<0.05$) and most predictors relatively high magnitude z-statistics, indicating that all variables in the model have statistically significant relationships with the likelihood of content going viral.

Coefficient Analysis

When fitting our initial model, we also visualized the adjusted odds ratios to ensure that all predictors were statistically significant.



From this initial visualization, most of the 95% confidence intervals for our predictor coefficients included 1, suggesting that the majority of our predictors added to the model. While a couple predictors (such as day_published_Thursday and log(n_tokens_content) were close to 1, we decided to still keep them in the model as they were either one level of several of a categorical variable, or because we found a possible interaction effect from the earlier EDA and felt it was at least a valuable predictor to consider in our model.

Interaction Effects

Next, we considered the addition of potential interaction effects between article length and data channel, and between global subjectivity and data channel. The hypothesis for this experiment were:

$H_o : B_{n-tokens-content*data-channel} = B_{global-subectivity*data-channel} = 0 H_A : B_j \neq 0$ for atleast one j

Table 2: Drop in Deviance Test Results

Model	Log-Likelihood	Deviance Statistic (G)	df	p-value
Null Model	-20023.03	NA	NA	NA
Interaction Model	-19927.15	191.76	5	0

Examining the output of the deviance test, the p-value is very low, at around 0. This indicates that the data provides sufficient evidence that at-least one of the newly added interaction terms is a statistically significant predictor in whether an article will go viral or not, after accounting for data channel, day published, global subjectivity, title sentiment polarity, average key word popularity, and main body length for a given article. Therefore, we will keep the interaction effects in the final model.

Model Evaluation and Comparison

We further compared our two models through their ROC and AUCs.

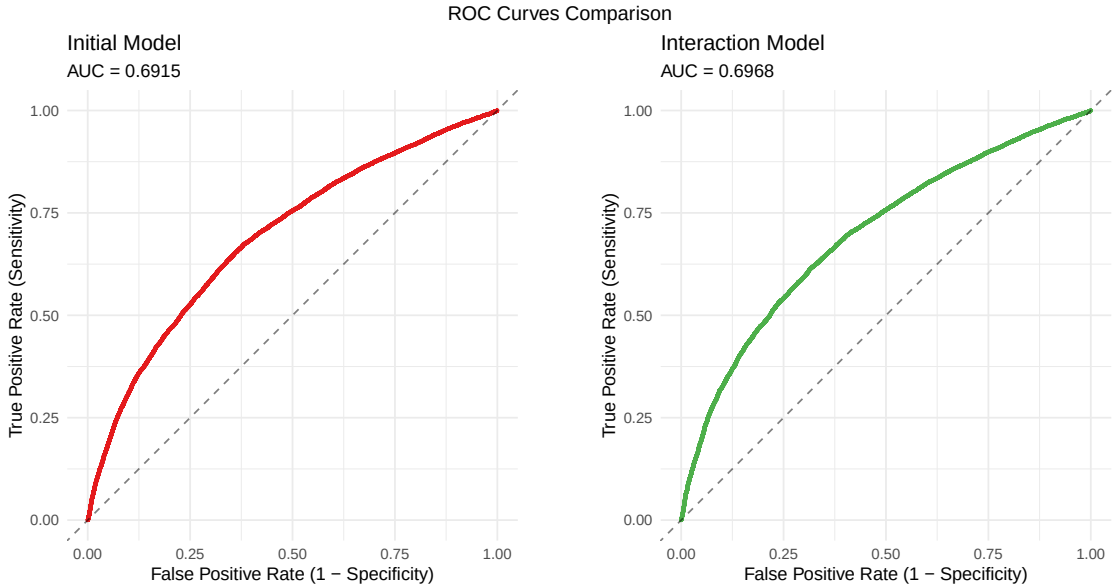


Table 3: AUC Values for Both Models

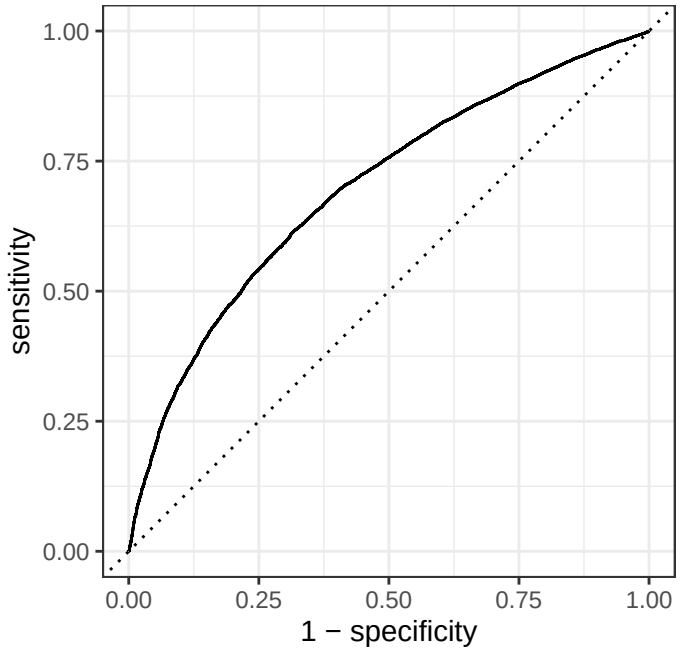
.metric	.estimator	.estimate	model
roc_auc	binary	0.6915	Initial Model
roc_auc	binary	0.6968	Interaction Model

From the ROC curves, we can see that 1) Both models have an ROC curve above the random threshold, approaching the top left corner, indicating some predictive power in classifying an article and 2) The Interaction Model (AUC = 0.6902) demonstrates marginally better predictive performance than the Initial Model (AUC = 0.681), confirming our belief that the interaction effects are meaningful predictors. 3) Based on the curve, the optimal threshold for our model should target sensitivity ~ 0.65.

Selecting the point closest to the ROC curve to sensitivity 0.65 yields a threshold of approximately 0.464.

```
Optimal threshold for classification: 0.493
```

Model Performance




```
    35211
0.4932468
```

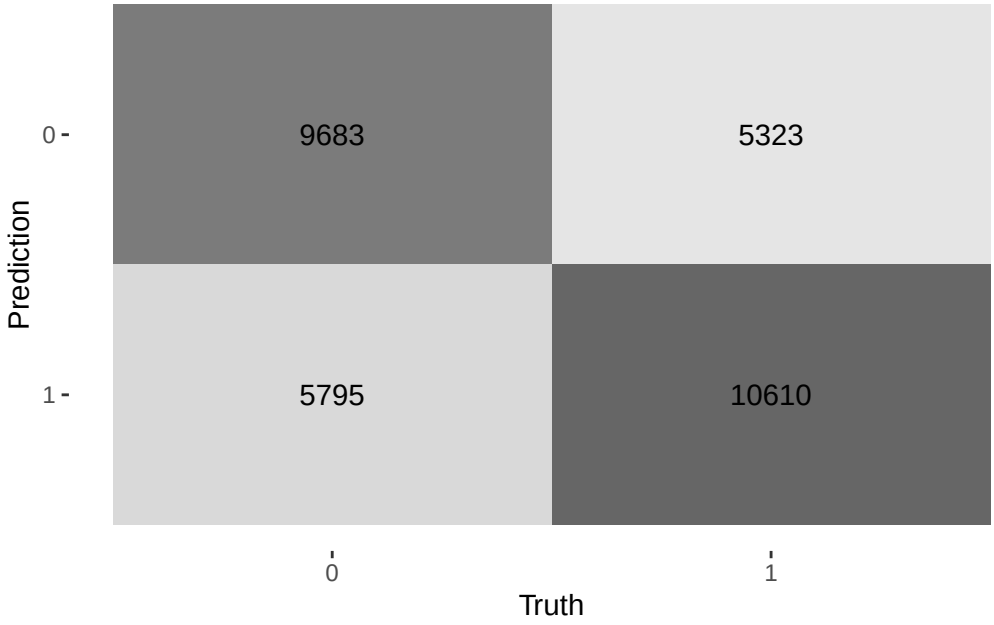


Figure 1: Confusion Matrix for the Interaction Model

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.697
```

Analysis + Peer Review

Draft report

Introduction

The ways in which people interact with media and discover news have dramatically shifted in recent years, with social media often displacing traditional news outlets. The decentralized nature of social media means the reach of each article is largely dependent on its individual merits, rather than the popularity of the publication it belongs to.

Thus a question arises: What article attributes are associated with social media virality?

In this report we will investigate the effects of different article features on social media success using 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, a digital media website, over two years (from 2013 to 2015). The data has

39,644 entries in total, with each repsenting an individual article and its associated textual and metadata features.

Key Variables:

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.

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.

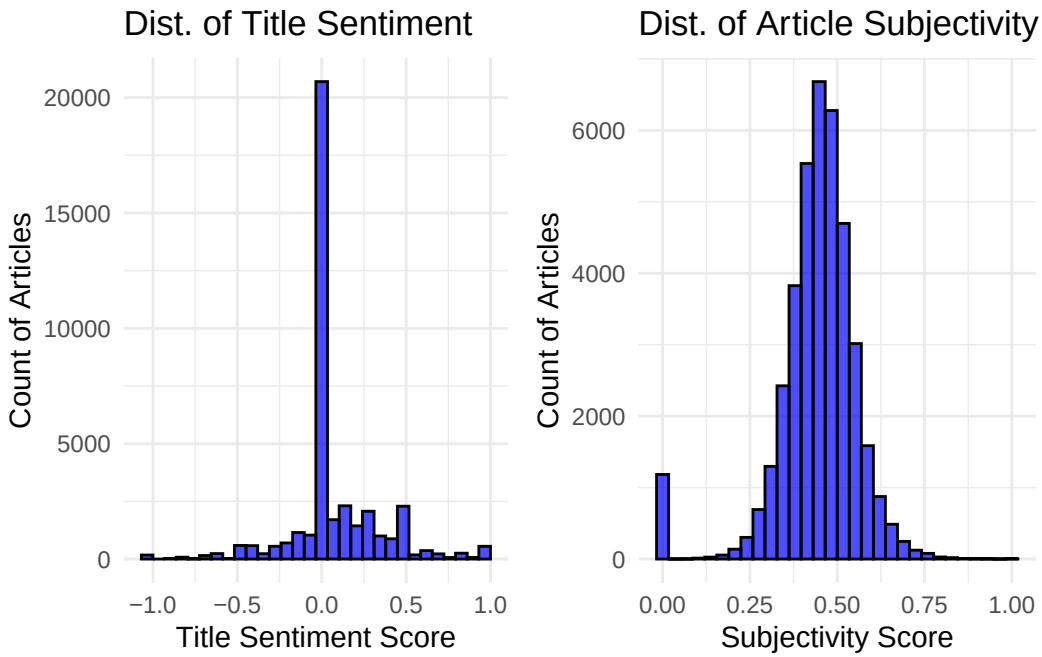
day_published - Categorical variable indicating publication day, merged from indicators: weekday_is_monday, weekday_is_tuesday, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday, weekday_is_saturday, weekday_is_sunday. Additionally includes is_weekend (mean 0.1309) to distinguish weekday from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

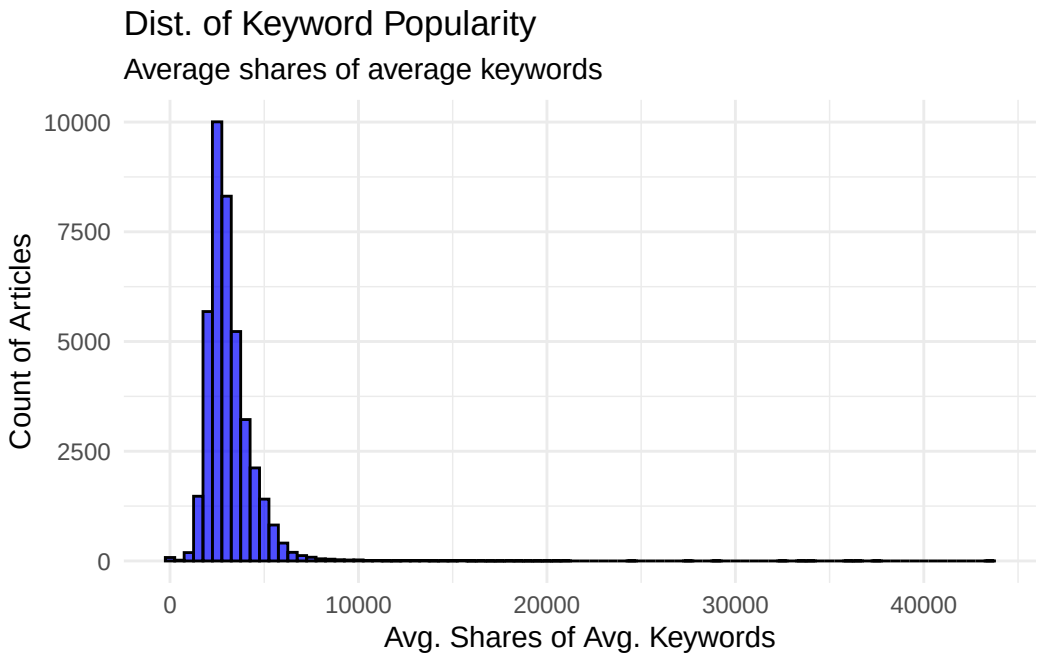
global_subjectivity - Measures the overall subjectivity of the article text. Values range from 0.0 (completely objective) to 1.0 (completely subjective), with a mean of 0.4434 and standard deviation of 0.1167. This quantifies how opinion-based versus fact-based the content is.

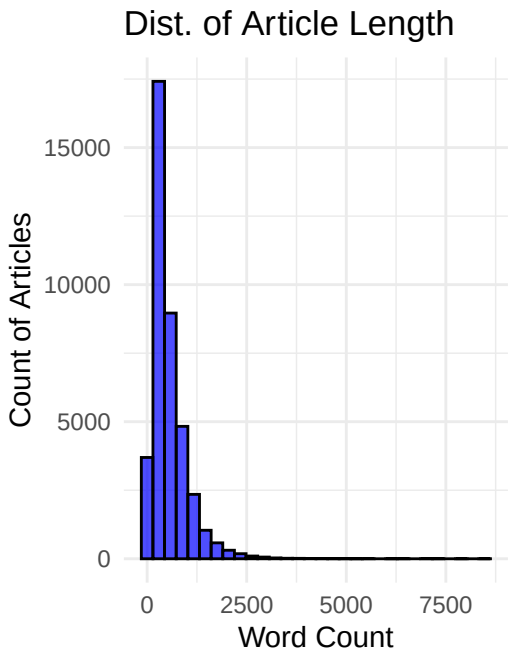
Univariate Exploratory Data Analysis

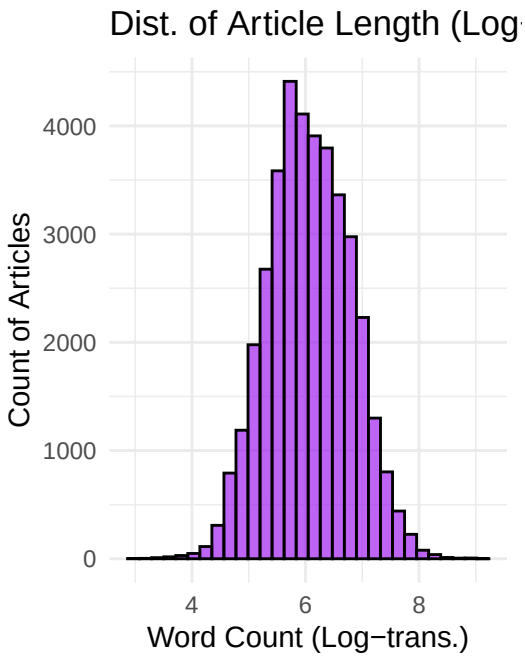
To understand our response and predictor variables more deeply, we first looked at their individual distributions. We found that while some variables are relatively approximately symmetric, others are heavily skewed and required log transformations to better meet modeling assumptions.

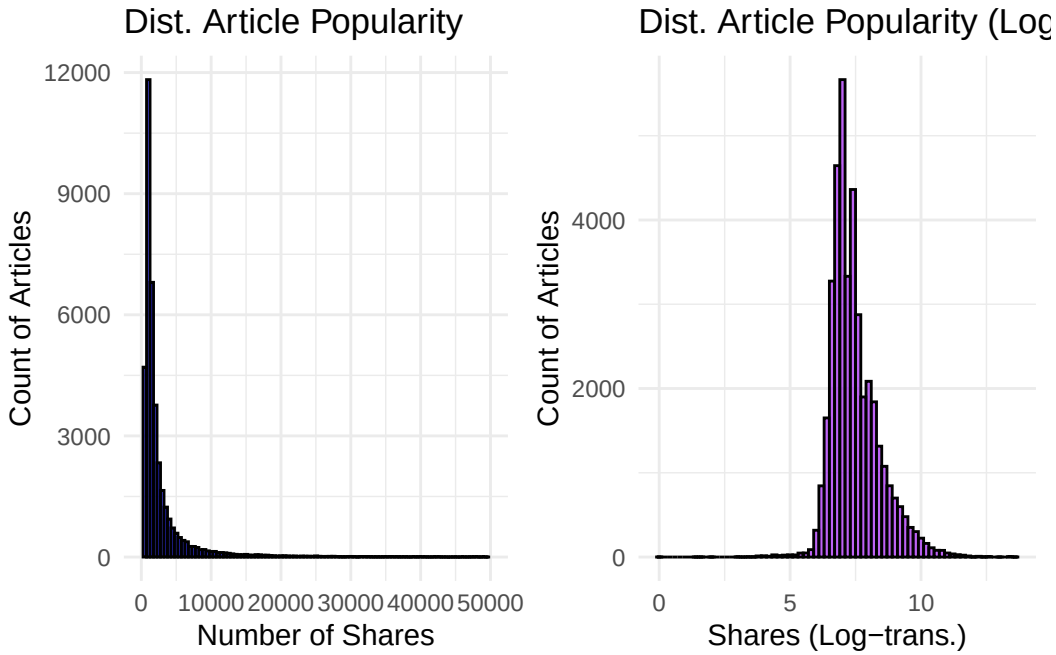


The title_sentiment_polarity variable quantifies the emotional change of article titles. The distribtuion suggests that many titles are emotionally neutral. The global_subjectivity variable is rougly symmetric with its distribution suggesting that most articles contained a balanced mix of factual and opinion-based language. Both global_subjectivity and title_sentiment_popularity display relatively balanced distributions that do not require transformation.


Dist. of Keyword Popularity
Average shares of average keywords
Count of Articles
10000
7500
5000
2500
0
0
10000
20000
30000
40000
Avg. Shares of Avg. Keywords


Dist. of Article Length
Count of Articles
15000
10000
5000
0
0
2500
5000
7500
Word Count


Dist. of Article Length (Log
Count of Articles
4000
3000
2000
1000
0
4
6
8
Word Count (Log−trans.)



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.

Both `n_tokens_content` and `kw_avg_avg` displayed heavily right-skewed distributions that warranted log transformations.To correct for this skewness and reduce the influence of extreme values, we log-transformed both variables. Post-transformation, the distributions became more symmetric and centered, better aligning with modeling assumptions.

The original distribution of the `shares` variable was extremely right-skewed, with the majority of articles receiving low engagement and a small number going viral. This made it difficult to model using linear approaches due to non-normal residuals and unequal variance.

We applied a log transformation to address this skewness. The transformed distribution is notably more symmetric and bell-shaped, making it more appropriate for regression analysis and enabling clearer interpretation of the effects of predictor variables on article popularity.

Data Cleaning

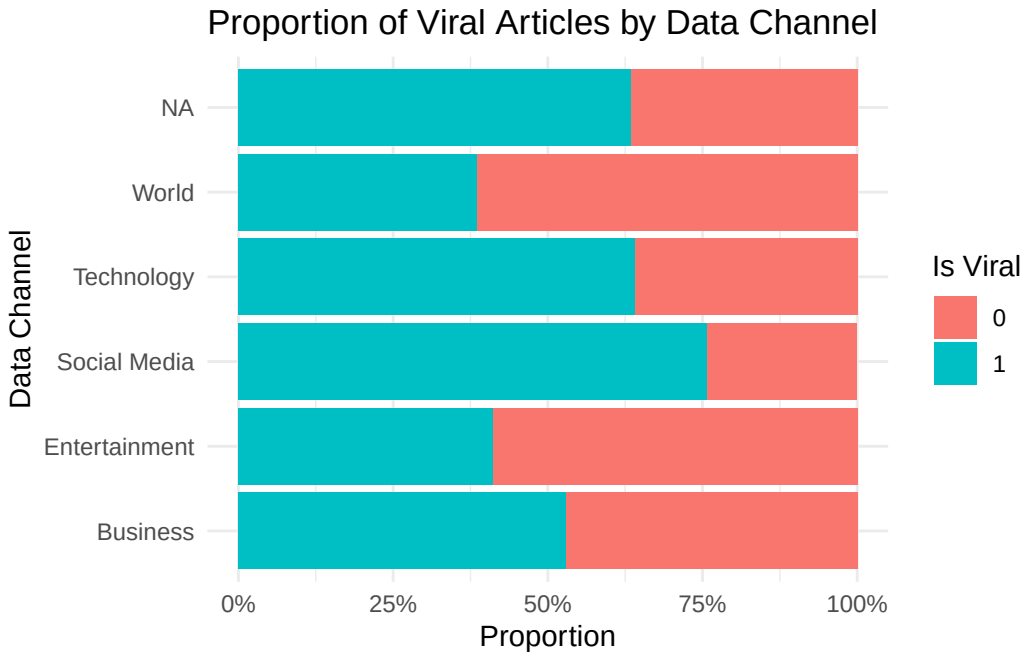
Data Cleaning

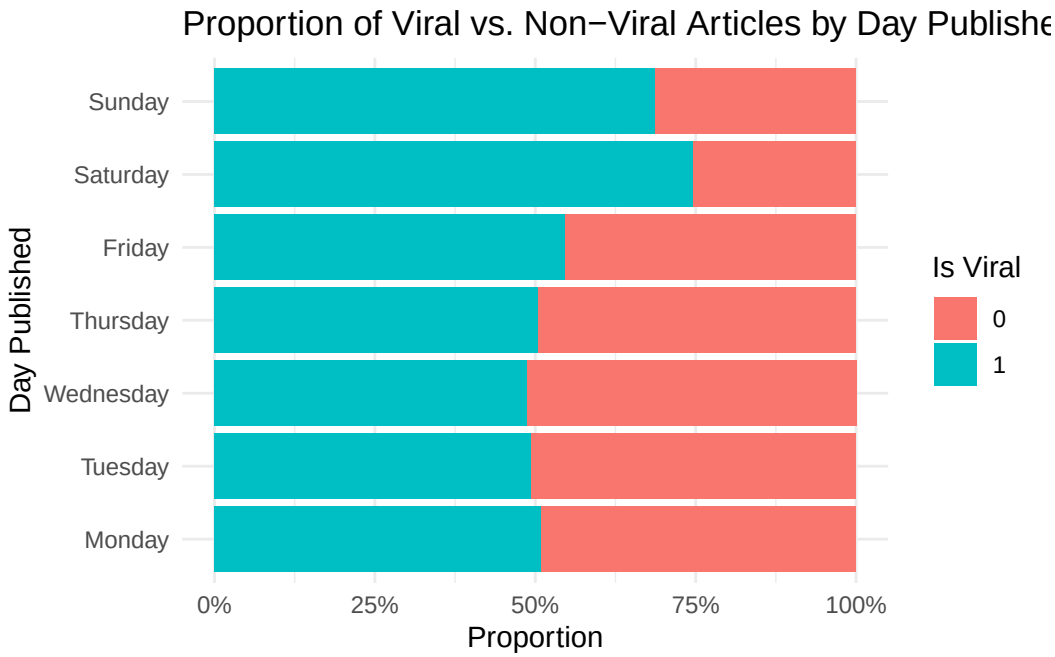
This data cleaning step restructures the dataset by converting multiple binary indicator columns into more interpretable categorical variables.

Table 4: 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

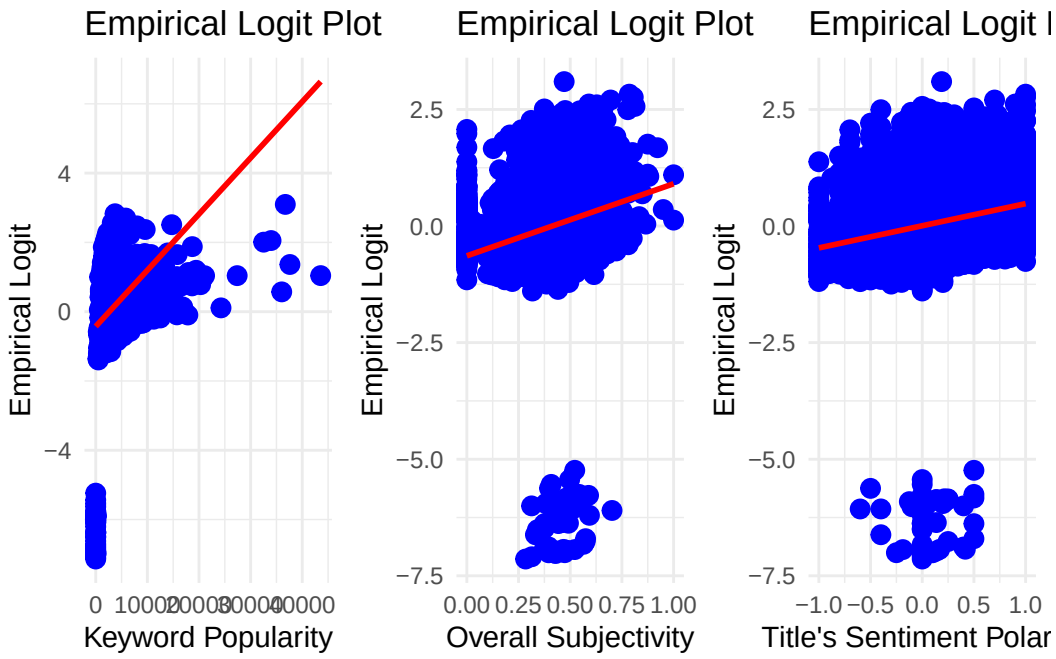
Bivariate Exploratory Data Analysis





The first graph explores how virality varies across different content categories, as captured by the data_channel variable. Article published under the Social Media category have the highest proportion of viral outcomes, with many reaching viral status, This suggest that content tailored for or about social platforms may be particularly good for engagement. Articles that are categoized under Technology and Business also show strong performance, with over half of the articles in each category going viral, On the other hand, articles in the Entertainment and World categories are less likely to go viral, falling below the 50% mark. This graph underscores how topic area influences content reach potentially because of the differences in audience behavior or platform algorithms.

The second graph examines the relationship between the day an article is published and its likelihood of going viral. Articles published on weekends are substantially more likely to be viral compared to those published during the week. Saturday stands out with the highest porportion of viral content, followed closely by Sunday. In contrast, weekday article tend to have lower virality rates, with viral and non-viral occurring in nearly equal proportions. These findings suggest that timing plays a role in determining an article's reach, likely reflecting differences in user engagement patterns across the week.



The empirical logit plots offer further insight into the relationship between key predictors and the binary outcome is_viral. The plot for average shares of average keywords reveals a positive relationship with virality. As the average popularity of an article's keywords increases, the likelihood of the article going viral rises sharply, with the empirical logit showing an upward linear trend. This suggest that keyword selection plays a key role in driving article engagement and affirms kw_avg_avg is a key predictors in modeling vitality. In contrast, the plot of overall subjectivity indicated on a slight positive relationship. Although the fitted line trends upward, the data are widely dispersed, and the effect appears weak and inconsistent, implying subjectivity alone is not a reliable driver of viral outcomes. A similar conclusion can be drawn from the plot of title sentiment polarity. While there is a marginal positive slope, suggesting that more positive titles generate a slightly greater chance of going viral, the overall relationship is weak. The plots highlight a clear distinction in predictive power among the variables and support prioritizing keyword metrics over emotional tone or subjectivity when modeling article vitality.

Methodology

Since initial EDA revealed a heavy right skew in the distribution of article shares, as well as a poential non-linear relationship, we elected to use a logistic regression model with a transformed binary response variable.

[INCLUDE ATTEMPT AT LINEAR referencing appendix]

Based on our initial EDA and empirical logit visualization, we selected data channel, day published, article subjectivity, title sentiment polarity, log transformed 'avg keyword popularity', and log transformed article content length to fit an initial logistic model.

For the response variable, we constructed "is_viral" by transforming 'share' count into a binary response variable, with 1 for articles more popular than 1400 shares and 0 for those with less. We selected this threshold of 1400 shares based on prior literature[CITE HERE] and the recommendation of the data set curator. ## Model Specification

$$\begin{aligned}\text{logit}(P(\text{is\_viral} = 1)) = \beta_0 \\ &+ \beta_1 \times \log(\text{kw\_avg\_avg} + 0.0001) \\ &+ \beta_2 \times \log(\text{n\_tokens\_content} + 0.0001) \\ &+ \beta_3 \times \text{data\_channel} \\ &+ \beta_4 \times \text{day\_published} \\ &+ \beta_5 \times \text{global\_subjectivity} \\ &+ \beta_6 \times \text{title\_sentiment\_polarity}\end{aligned}$$

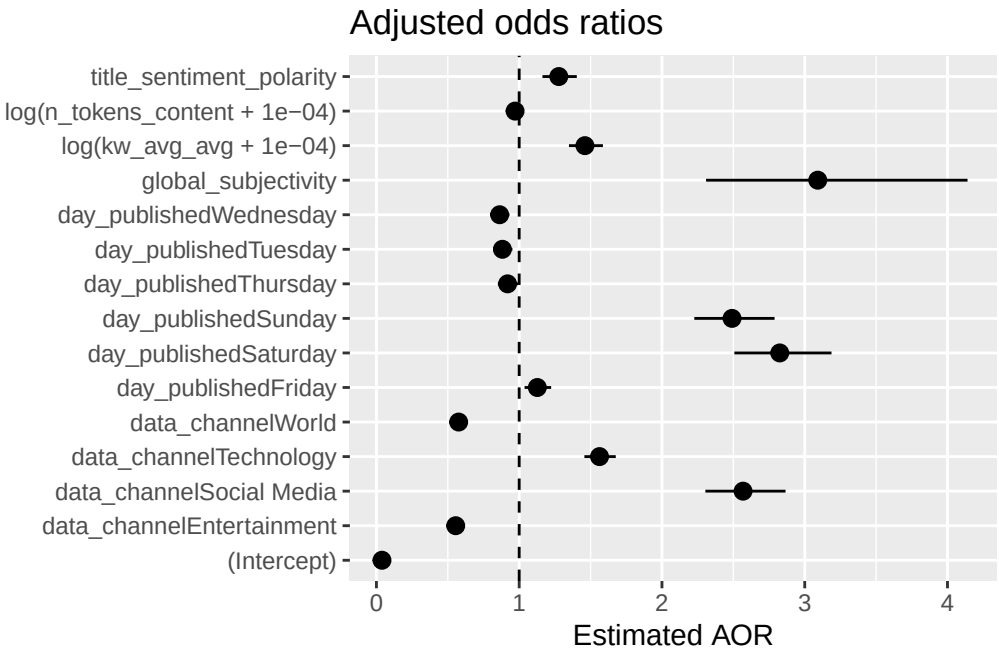
[can explain output/visualizations for log transform in appendix]

Term	Estimate	Std.Error	z-statistic	p-value
(Intercept)	-3.2464	0.3361	-9.6589	0.0000
log(kw_avg_avg + 1e-04)	0.3787	0.0417	9.0807	0.0000
log(n_tokens_content + 1e-04)	-0.0293	0.0069	-4.2364	0.0000
data_channelEntertainment	-0.5881	0.0361	-16.2849	0.0000
data_channelSocial Media	0.9431	0.0556	16.9573	0.0000
data_channelTechnology	0.4464	0.0359	12.4408	0.0000
data_channelWorld	-0.5515	0.0351	-15.6925	0.0000
day_publishedTuesday	-0.1251	0.0391	-3.2013	0.0014
day_publishedWednesday	-0.1472	0.0390	-3.7701	0.0002
day_publishedThursday	-0.0849	0.0393	-2.1606	0.0307
day_publishedFriday	0.1195	0.0421	2.8384	0.0045
day_publishedSaturday	1.0386	0.0612	16.9624	0.0000
day_publishedSunday	0.9126	0.0574	15.8857	0.0000
global_subjectivity	1.1286	0.1490	7.5759	0.0000
title_sentiment_polarity	0.2449	0.0480	5.1035	0.0000

Our initial fit gives all of the predictors significant p-values ($p<0.05$) and most predictors relatively high magnitude z-statistics, indicating that all variables in the model have statistically significant relationships with the likelihood of content going viral.

Coefficent Analysis

When fitting our model, we also visualized the adjusted odds ratios to ensure that all predictors were statistically significant.



From this initial visualization, none of the 95% confidence intervals for our predictor coefficients included 1, suggesting that they were all statistically significant. While we had a couple predictors whose confidence intervals were close to 1, we decided to still keep them in the model as day_published_Thursday is one of the factors of the day_published variable, and thus it's acceptable that some of the levels for day_published were not necessarily significant because the other levels were. To add, we decided to keep log(n_tokens_content) as we both saw a possible interaction effect in the earlier EDA and we felt that it was at least a valuable predictor to consider in our model.

Interaction Effects

Next, we considered the addition of poetential interaction effects between article length and data channel, and between global subjectivity and data channel. The hypothesis for this experiment were:

$$H_o : B_{n-tokens-content*data-channel} = 0 H_A : B_{n-tokens-content*data-channel} \neq 0$$

Table 6: Drop in Deviance Test Results

Model	Log-Likelihood	Deviance Statistic (G)	df	p-value
Null Model	-20241.27	NA	NA	NA
Interaction Model	-20144.56	193.414	5	0

Examining the output of the deviance test, the p-value is very low, at around 0. This indicates that the data provides sufficient evidence that at-least one of the newly added interaction terms is a statistically significant predictor in whether an article will go viral or not, after accounting for data channel, day published, global subjectivity, title sentiment polarity, average key word popularity, and main body length for a given article. Therefore, we will keep the interaction effects in the final model.

Model Evaluation and Comparison

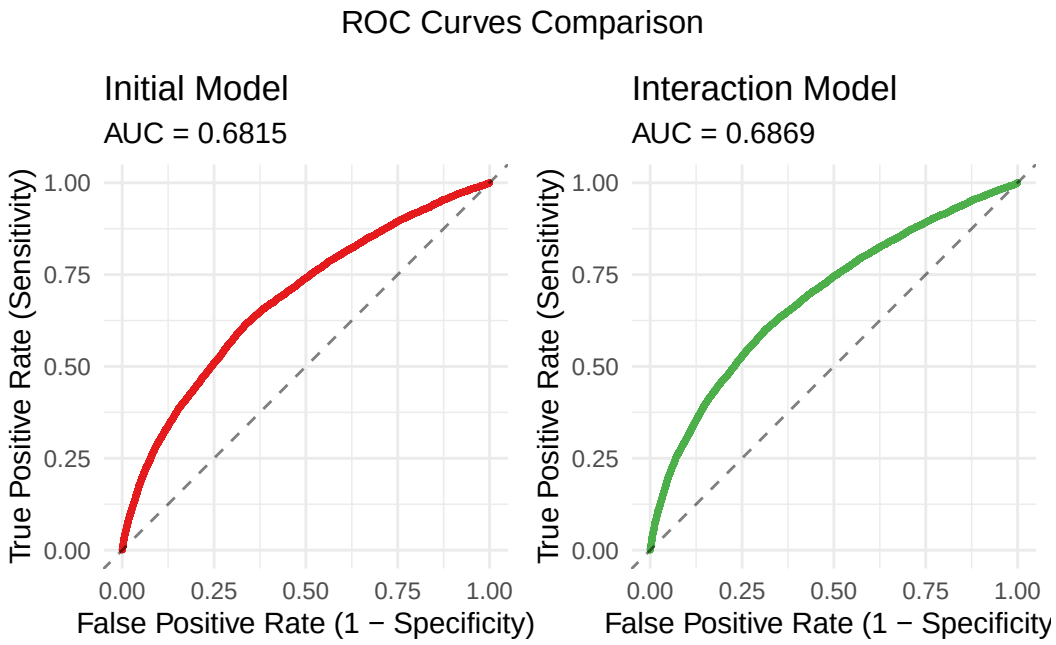


Table 7: AUC Values for Both Models

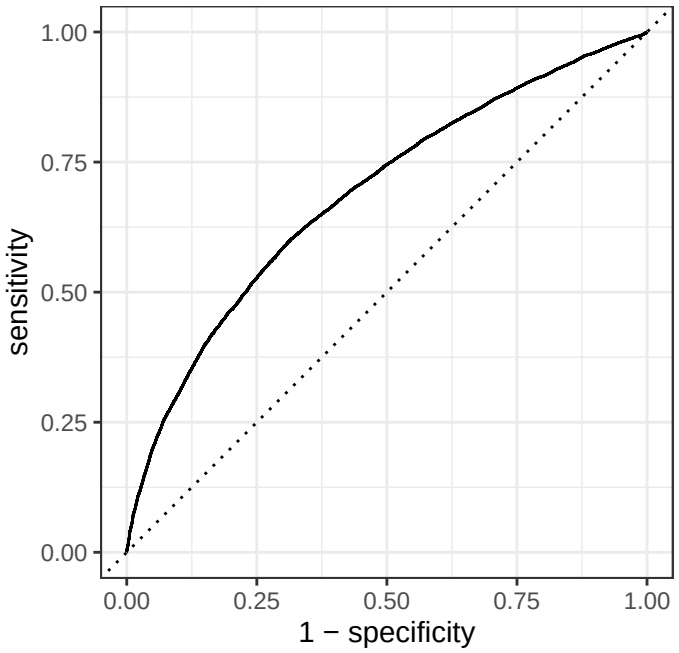
.metric	.estimator	.estimate	model
roc_auc	binary	0.6815	Initial Model
roc_auc	binary	0.6869	Interaction Model

From the ROC curves, we can see that 1) Both models have an ROC curve above the random threshold, approaching the top left corner, indicating some predictive power in classifying an article 2) The Interaction Model (AUC = 0.6902) demonstrates marginally better predictive performance than the Initial Model (AUC = 0.681), confirming our belief that the interaction effects are meaningful predictors. 3) Based on the curve, the optimal threshold for our model should target sensitivity ~ 0.65.

Selecting the point closest to the ROC curve to sensitivity 0.65 yields a threshold of approximately 0.464.

```
Optimal threshold for classification: 0.482
```

Model Performance




```
   16669
0.4821143
```

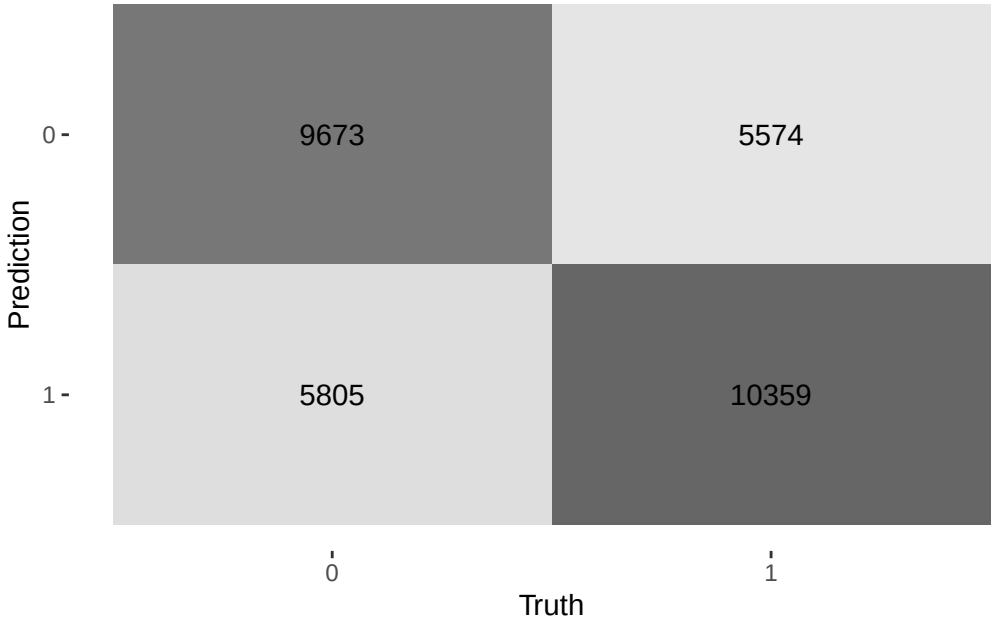


Figure 2: Confusion Matrix for the Interaction Model

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.687
```

Results

The final model we fitted was:

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

term	estimate	std.error	statistic	p.value
(Intercept)	-4.628	0.390	-11.876	0.000
log(kw_avg_avg + 1e-04)	0.371	0.042	8.893	0.000
log(n_tokens_content + 1e-04)	0.162	0.032	4.996	0.000
data_channelEntertainment	0.891	0.243	3.668	0.000
data_channelSocial Media	2.751	0.348	7.913	0.000
data_channelTechnology	1.560	0.286	5.454	0.000
data_channelWorld	0.789	0.237	3.330	0.001
day_publishedTuesday	-0.124	0.039	-3.171	0.002
day_publishedWednesday	-0.145	0.039	-3.709	0.000
day_publishedThursday	-0.081	0.039	-2.054	0.040
day_publishedFriday	0.128	0.042	3.026	0.002
day_publishedSaturday	1.033	0.062	16.772	0.000
day_publishedSunday	0.906	0.058	15.670	0.000
global_subjectivity	1.811	0.331	5.465	0.000
title_sentiment_polarity	0.236	0.048	4.917	0.000
log(n_tokens_content + 1e-04):data_channelEntertainment	-0.200	0.035	-5.768	0.000
log(n_tokens_content + 1e-04):data_channelSocial Media	0.009	0.048	0.179	0.858
log(n_tokens_content + 1e-04):data_channelTechnology	-0.056	0.041	-1.363	0.173

term	estimate	std.error	statistic	p.value
log(n_tokens_content + 1e-04):data_channelWorld	-0.263	0.034	-7.664	0.000
data_channelEntertainment:global_subjectivity	-0.647	0.443	-1.463	0.143
data_channelSocial Media:global_subjectivity	-4.062	0.634	-6.405	0.000
data_channelTechnology:global_subjectivity	-1.750	0.476	-3.677	0.000
data_channelWorld:global_subjectivity	0.578	0.436	1.326	0.185

Table 9: Logistic Model Metrics Summary

Metric	Value
Accuracy	0.638
Misclassification Rate	0.362
Sensitivity (Recall)	0.650
Specificity	0.625
Precision	0.641
False Positive Rate (FPR)	0.375
False Negative Rate (FNR)	0.350
AUC	0.687

```
                                GVIF Df GVIF^(1/(2*Df))
log(kw_avg_avg + 1e-04)       1.095840  1        1.046824
log(n_tokens_content + 1e-04) 1.488499  1        1.220040
data_channel                  1.165365  4        1.019313
day_published                 1.016600  6        1.001373
global_subjectivity           1.545659  1        1.243245
title_sentiment_polarity      1.007553  1        1.003769
```

Our logistic regression model has an AUC of around 0.687, an accuracy of 0.638, specificity of 0.625, and sensitivity of 0.650. In comparison, the misclassification rate, FPR, and FNR were 0.362, 0.375, and 0.350, respectively. This suggests that our model is moderately well-fit for the data. While the accuracy, specificity, sensitivity, and precision were relatively high at around 0.640, the FNR, FPR, and misclassification rates were lower at around 0.360. This precision means that approximately 64% of articles predicted to be viral were correctly classified, indicating that the model performs substantially better than random chance. The relatively low predictive power may also be due to random noise, as many features influencing article virality are likely uncaptured by the dataset, and virality itself may be shaped by sudden trends.

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

Data Channel category plays a critical role. Notably, the odds of articles in the Social Media and Technology categories to go viral are approximately 15.66 times and 4.76 times that of a similar article in the Business category (reference group). Similarly, Entertainment and World news articles also show significantly higher odds of going viral than similar Business news articles, with odds ratios of 2.44 and 2.20, respectively. This suggests that readers are particularly engaged with content about social media and technology innovations, and also tend to share Entertainment and World news more than articles about Business news, however, not to the degree of Social Media and Technology articles.

Day of publication is another important factor. Weekend publications significantly outperform weekday content. Compared to Monday, Saturday articles have 2.81 times the odds , and Sunday articles have 2.47 times the odds of going viral, holding all else constant. In contrast, the odds of Tuesday and Wednesday articles going viral are 11.7% and 13.5% lower, with odds ratios of 0.883 and 0.865, respectively, than similar Monday articles. This weekend effect likely arises from increased leisure time, as people take off from work or school during the weekend.

Article subjectivity and sentiment also significantly impact virality. A fully subjective article (global subjectivity = 1) has 6.12 times the odds of going viral compared to a fully objective article (global subjectivity = 0), keeping all else constant. Similarly, a one-unit increase in title sentiment polarity (a neutral article compared to strongly positive) increases the odds of virality by 26.6%. This trend supports the idea that emotionally charged or opinionated content tends to be shared more frequently than neutral content.

Keyword popularity also plays a role in determining the odds of an article going viral. Specifically, every time keyword popularity is doubled, the odds of an article going viral increase by about 29.1%, holding all else constant. While article length is a statistically significant predictor of an article's virality, its impact is minimal. Specifically, every 10% increase in article length increases the odds of virality by approximately 1.55%, holding all else constant. Thus, while article length still adds to our model, it doesn't add as much as predictors like day published or data channel.

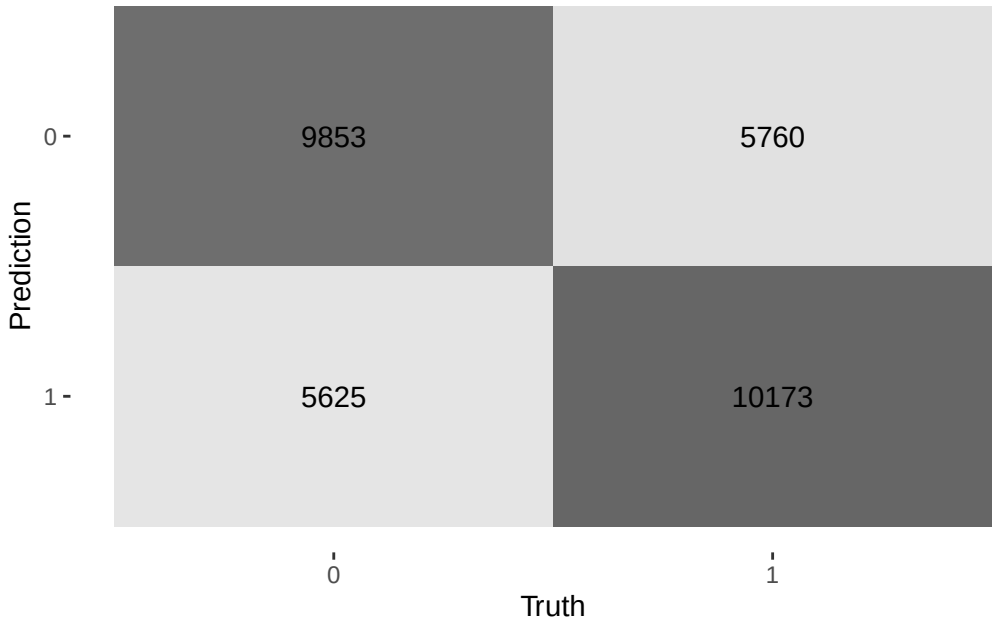
Through our empirical logit graphs, visualizations, and drop in deviance test we also found statistically significant interaction effects for both article length and global subjectivity with data channel category.

For Entertainment and World news, the benefit of article length is diminished or even reversed. For Entertainment articles, every 10% increase in article length leads to a 1.9% decrease in the odds of going viral compared to a similar Business article, while for World news, a 10% increase in article length results in a 2.5% decrease in odds. The interaction terms between article length and Social Media or Technology is not statistically significant, suggesting the main effect of article length dominates in those categories.

In Social Media articles, a fully subjective tone actually reduces the odds of virality by 98.3%, despite the strong main effect of subjectivity. This suggests that objective tone may be more prone to virality for Social Media. In comparison, for Technology articles, subjectivity decreases odds by 82.6%. However, for both Entertainment and World news, these interaction

effects are not statistically significant, which suggests that subjectivity could still hold a positive or neutral effect.

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.

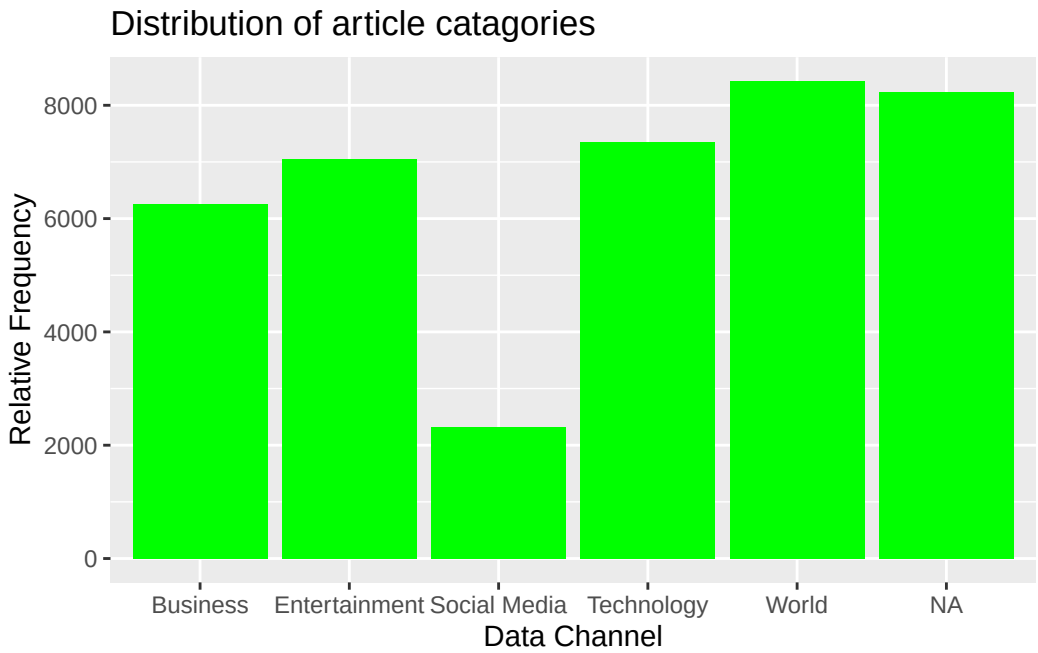
day_published- a categorical variable indicating publication day merged from indicators: Weekday_is_monday, weekday_is_tuesday, weekday_is_wednesday, weekday_is_thursday, weekday_is_friday,weekday_is_saturday, weekday_is_sunday

Data Cleaning

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

Table 10: Transformed Data

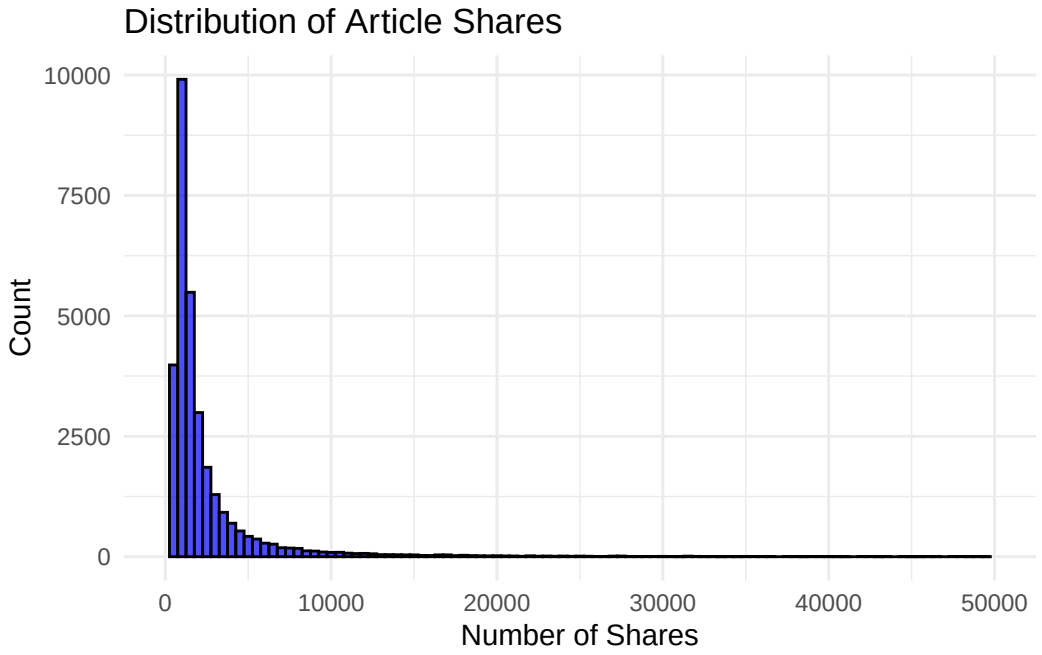
url	day_published	data_channel
http://mashable.com/2013/01/07/amazon-instant-video-browser/	Monday	Entertainment
http://mashable.com/2013/01/07/ap-samsung-sponsored-tweets/	Monday	Business
http://mashable.com/2013/01/07/apple-40-billion-app-downloads/	Monday	Business
http://mashable.com/2013/01/07/astronaut-notre-dame-bcs/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



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

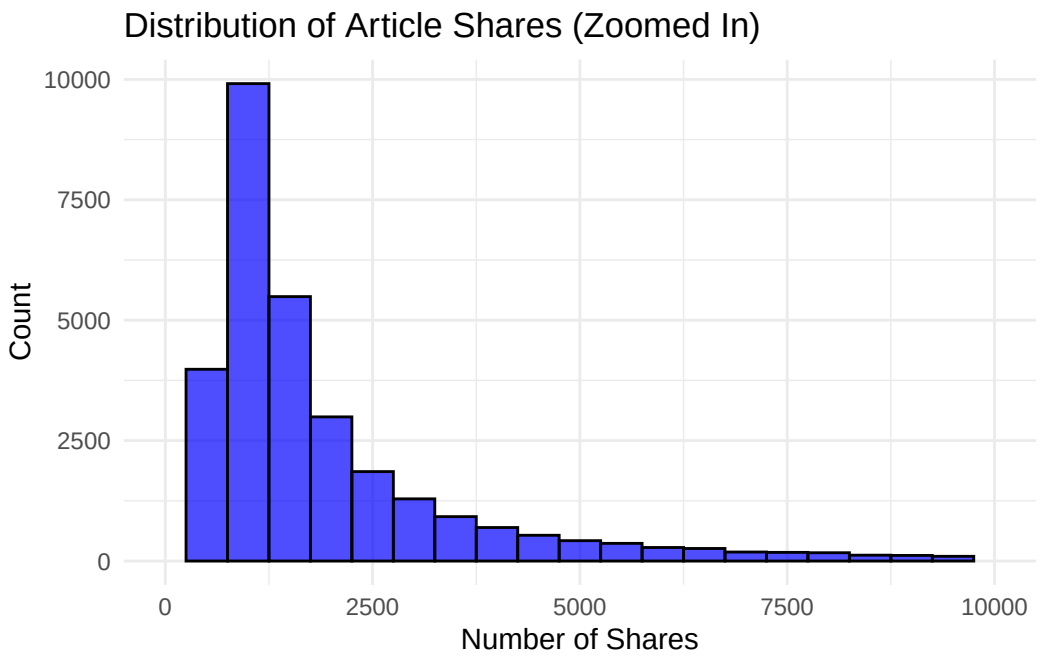
data tag NA's from our data channel analysis altogether. These articles lacking a data channel were filtered out.

Response Variable/Univariate EDA

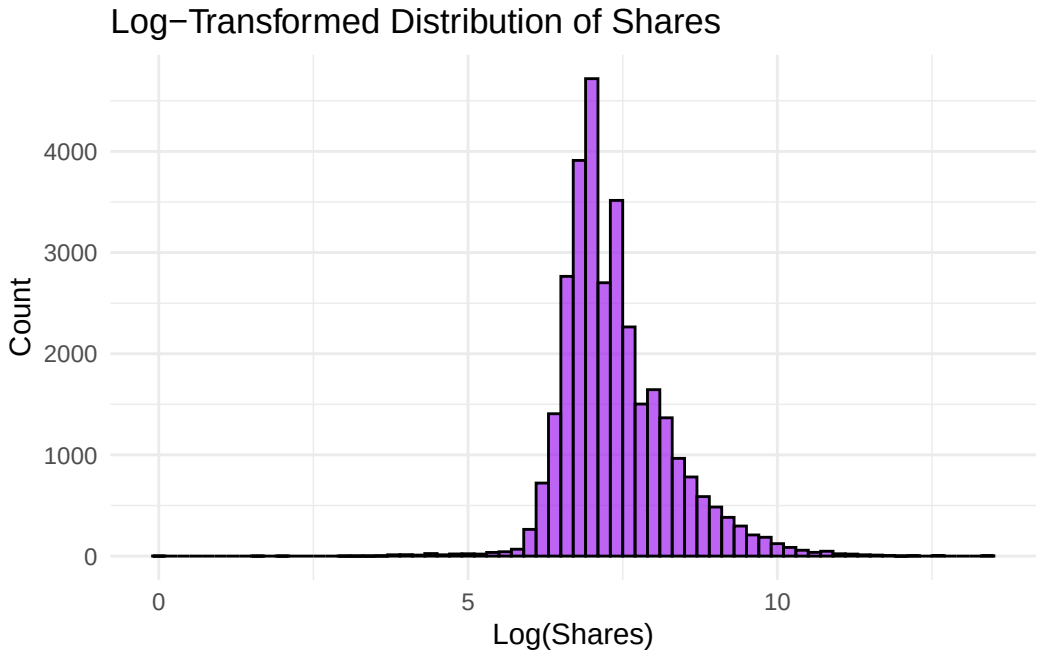



```
# A tibble: 1 x 7
  mean_shares median_shares sd_shares min_shares max_shares    q1    q3
        <dbl>         <dbl>     <dbl>      <dbl>      <dbl> <dbl> <dbl>
1       2878.          1400     9506.          1     690400   923  2500
```

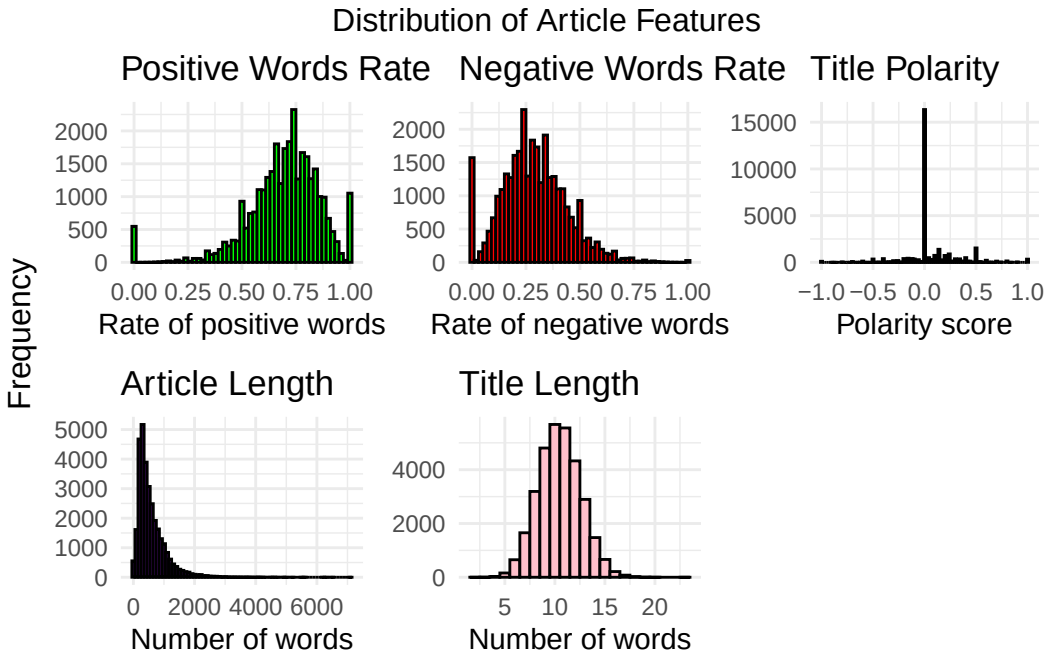
The distribution of # of shares is highly right skewed, a median of 1400 shares and a few highly shared articles. Notably, the mean of 2878 shares is far larger



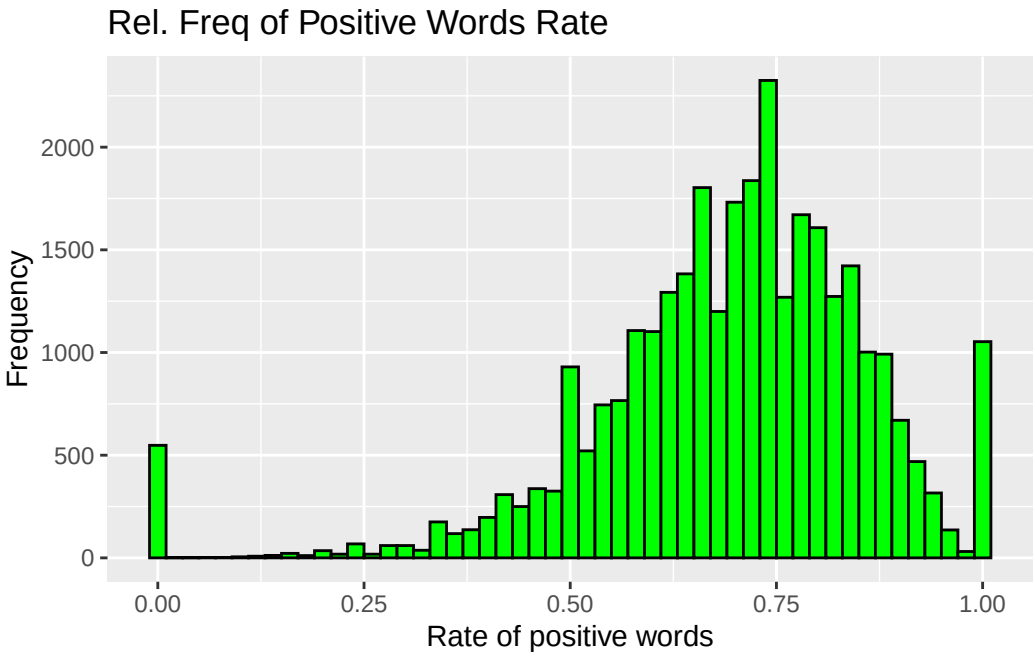
To make deal with this strong right skew, we applied a log transformation to the share variable, yielding a less skewed distribution.



Predictor Variable/Univariate EDA

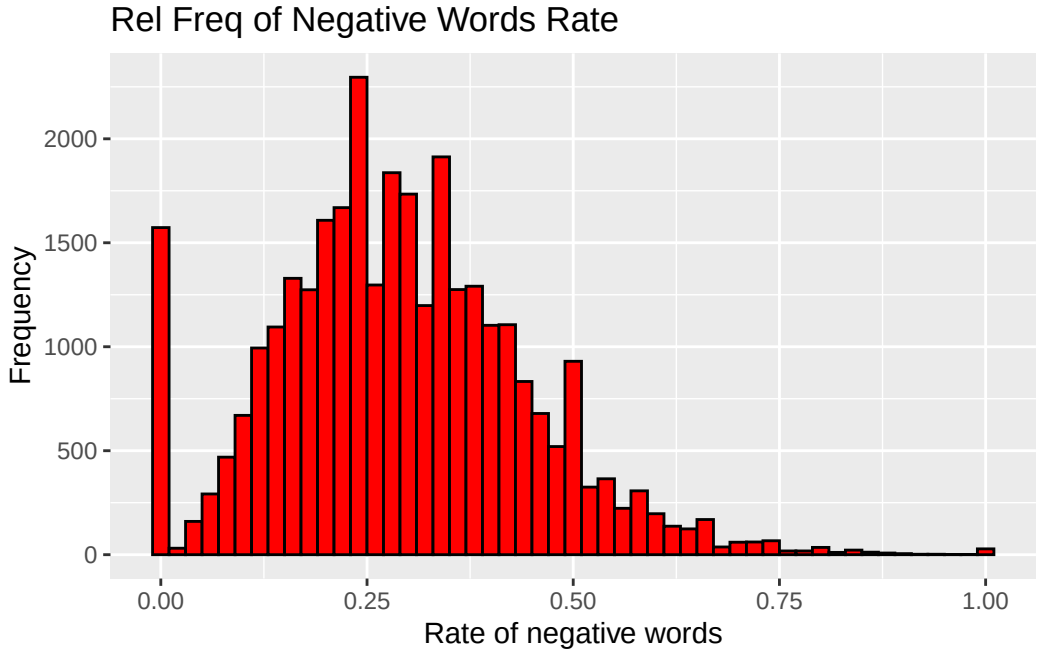


Examining the rate of positive words in an article, we see a left skew distribution, with modes at ~0, ~0.75 and ~1. The median positivity is approx 0.71 positivity rate, and the range is from 0 to 1.

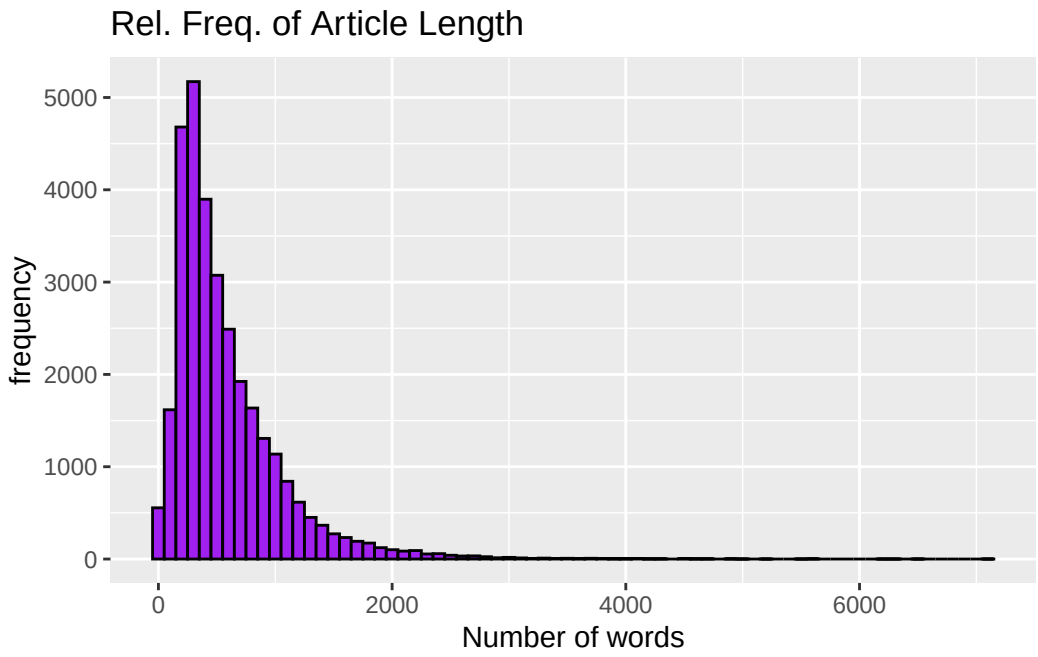


```
# A tibble: 1 x 5
   mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1 0.694  0.714   0.172     0     1
```

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.

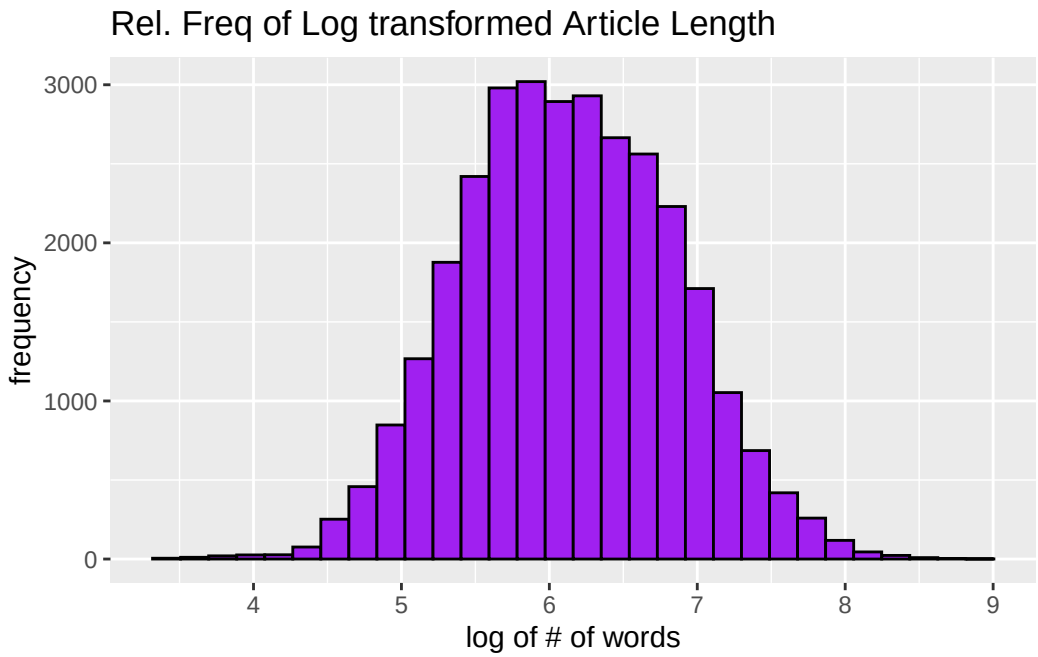



```
# A tibble: 1 x 5
   mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1 0.290  0.280   0.152     0     1
```

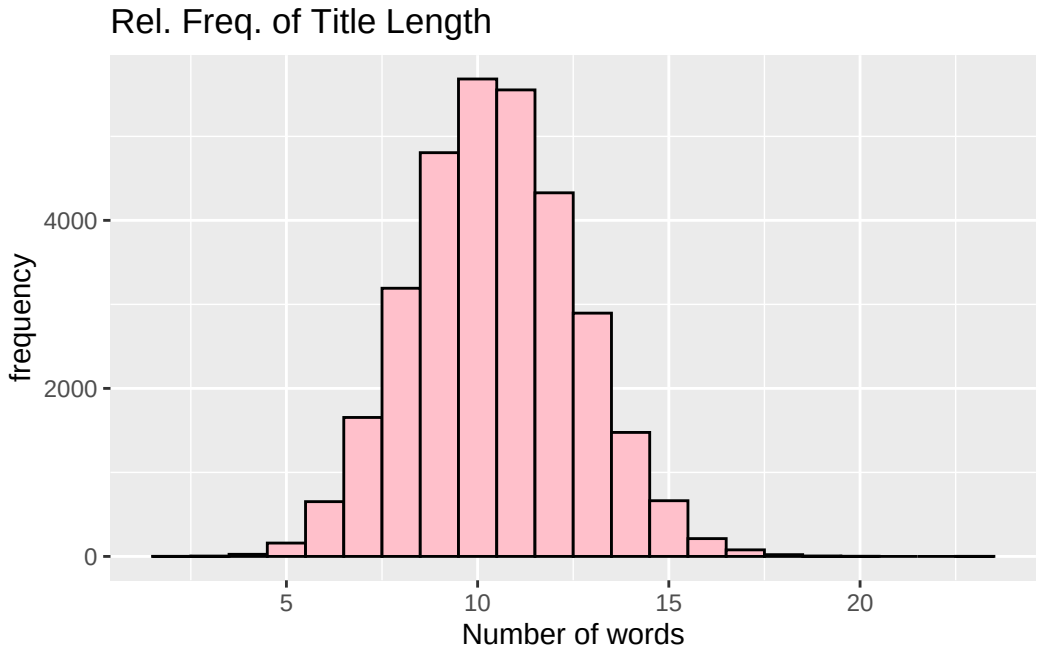



```
# A tibble: 1 x 5
   mean median std.dev   min   max
  <dbl>  <dbl>   <dbl> <dbl> <dbl>
1  583.    444    478.     0  7081
```

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.



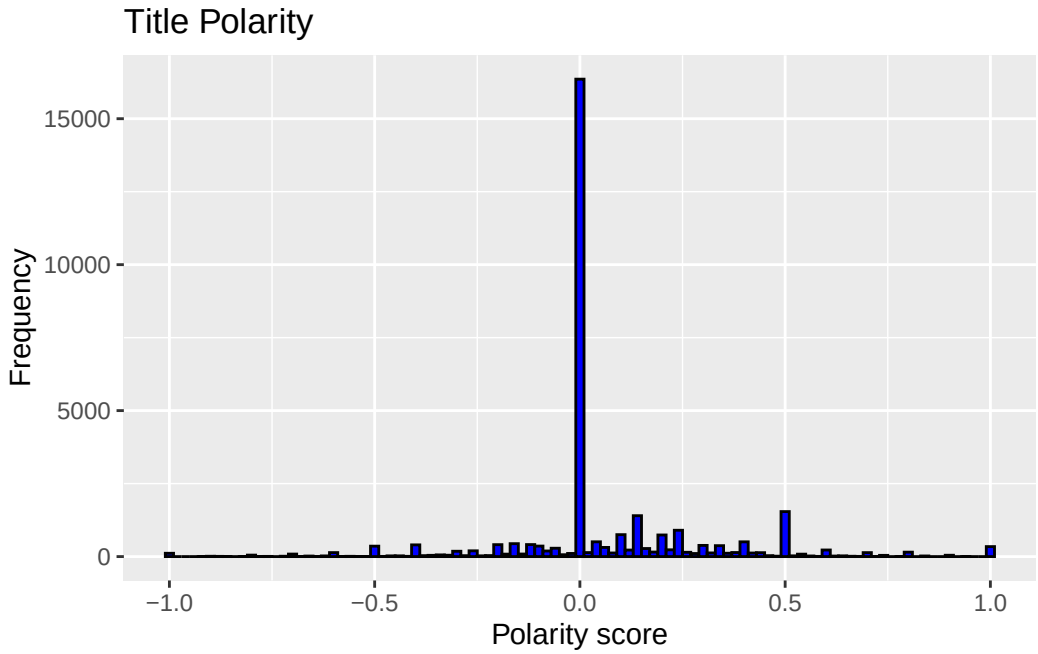
For the number of tokens in the title, we can see a highly symetric distribution centered at 10, with a standard deviation of 2.14.



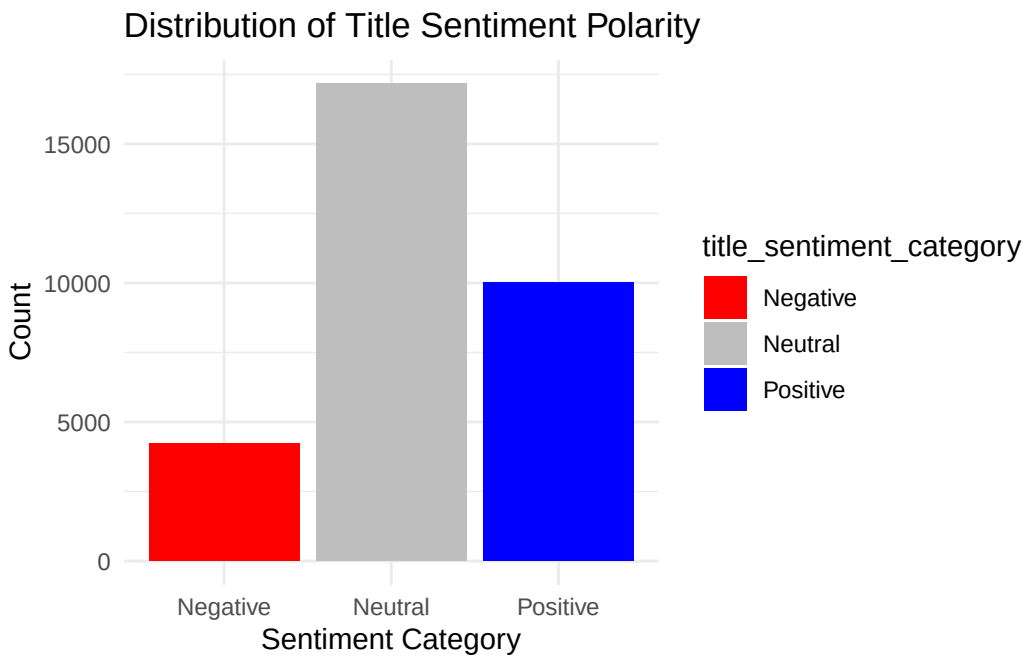

```
# A tibble: 1 x 5
```

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

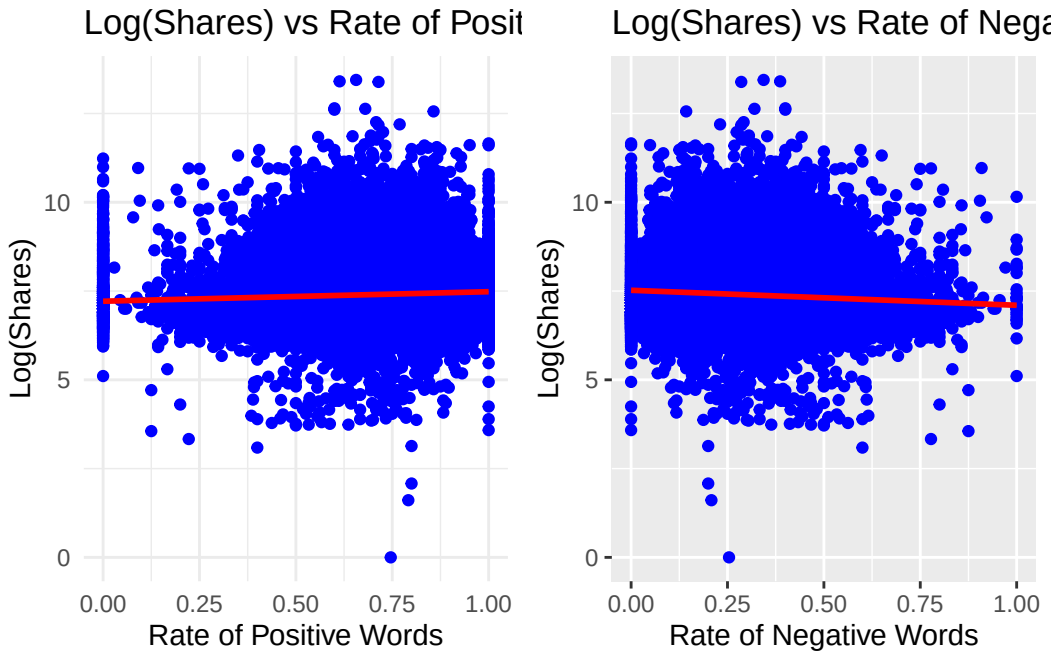
Finally, our initial EDA of title polarity found a massive frequency spike at 0 frequency, which might correspond to failed measurements or the vast majority of our articles not presenting signifigant title polarity.



To fix this issue and for ease of use, we catagorized 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.

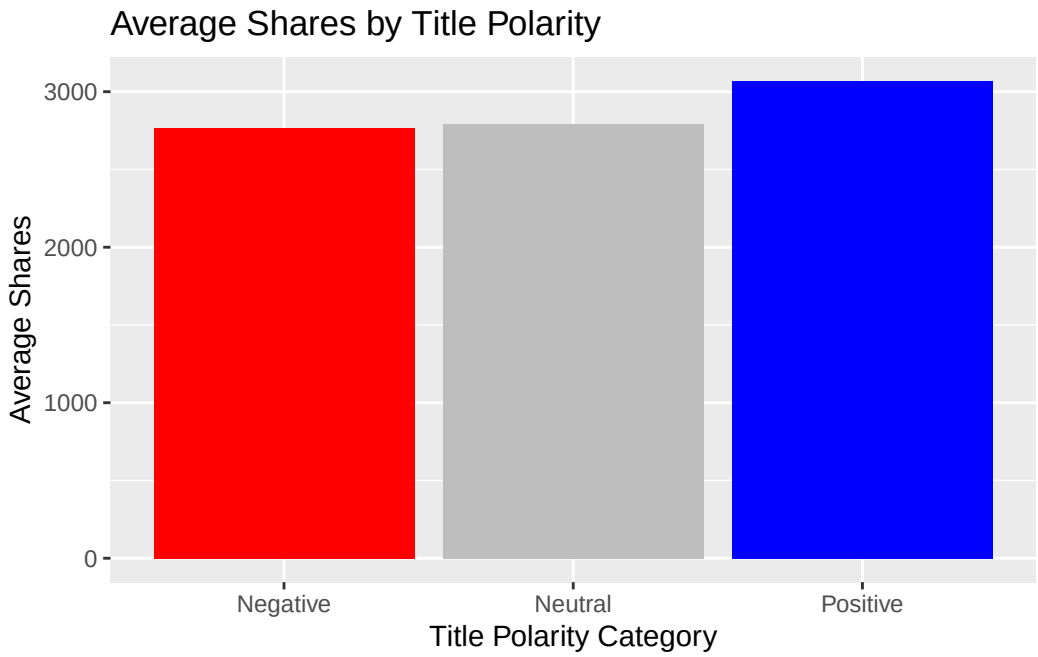


Bi-variate EDA



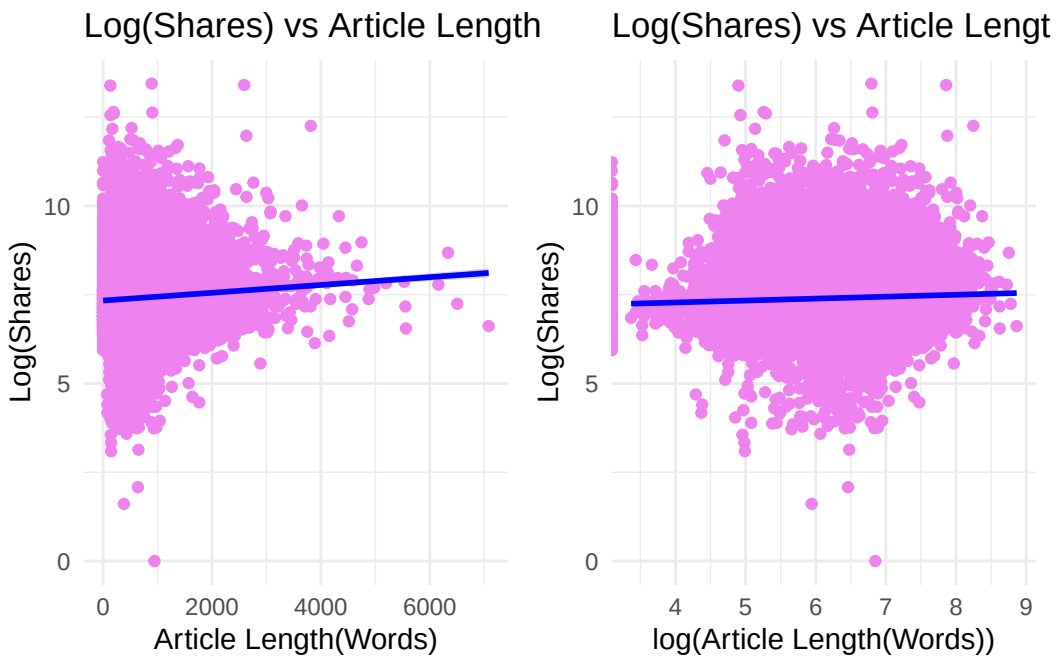
By plotting the rate of positive words and the rate of negative words against the log transformed share, we can see that neither have a particularly strong relationship with how often the article

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.



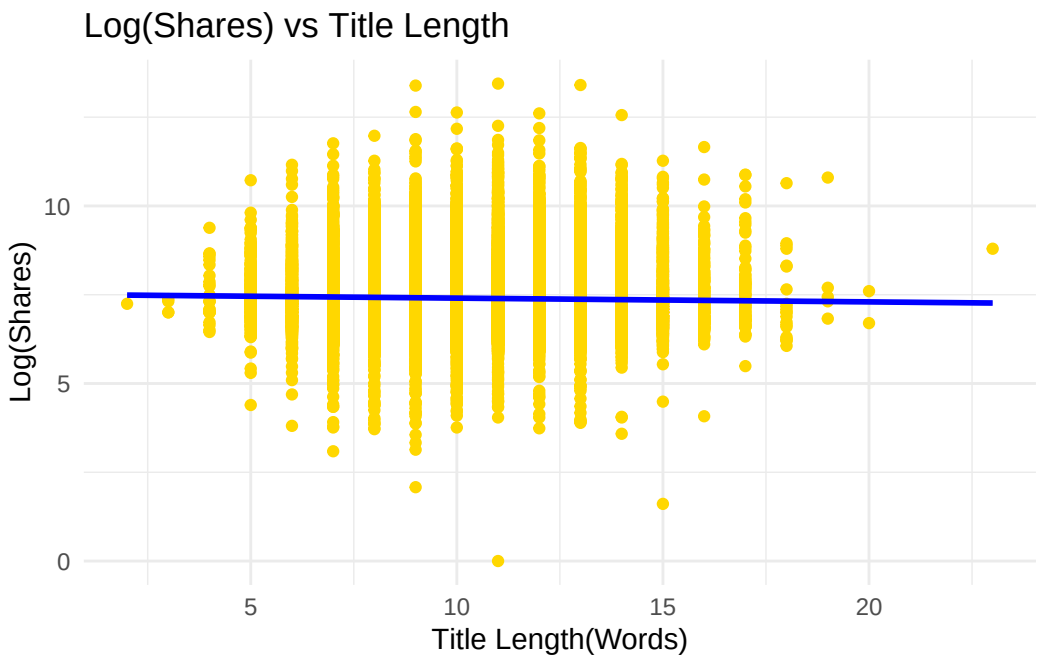
In contrast, there seems to be some relationship between title polarity and the number of shares, with positive polarity associated with greater share counts.

Next,

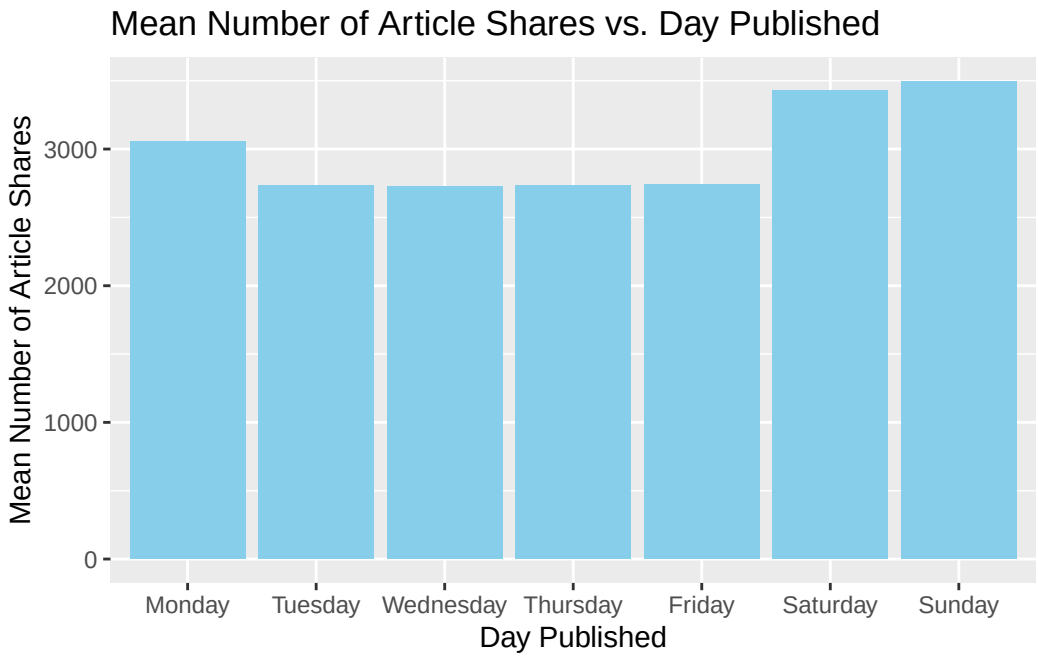


These graphs show a weak, positive relationship between article length and the log transformed number of shares. Due to the skew, we can apply the log transform to the article length. This shows a more even distribution, with no clear relationship.

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




```
# 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.
```



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

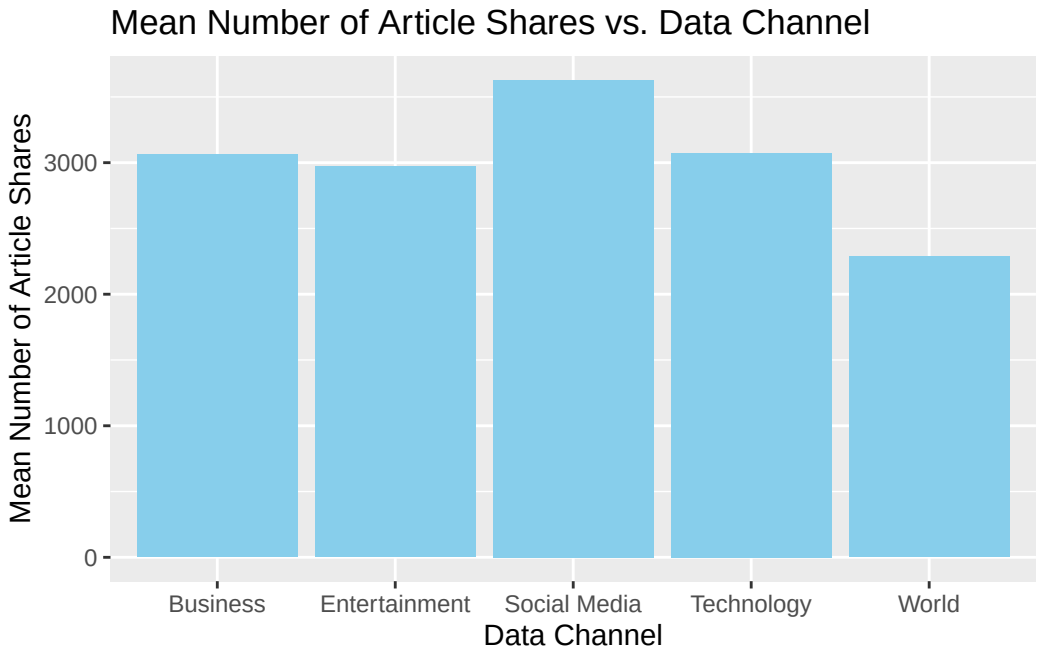
```
[1] "Variance Inflation Factors:"
```

```
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

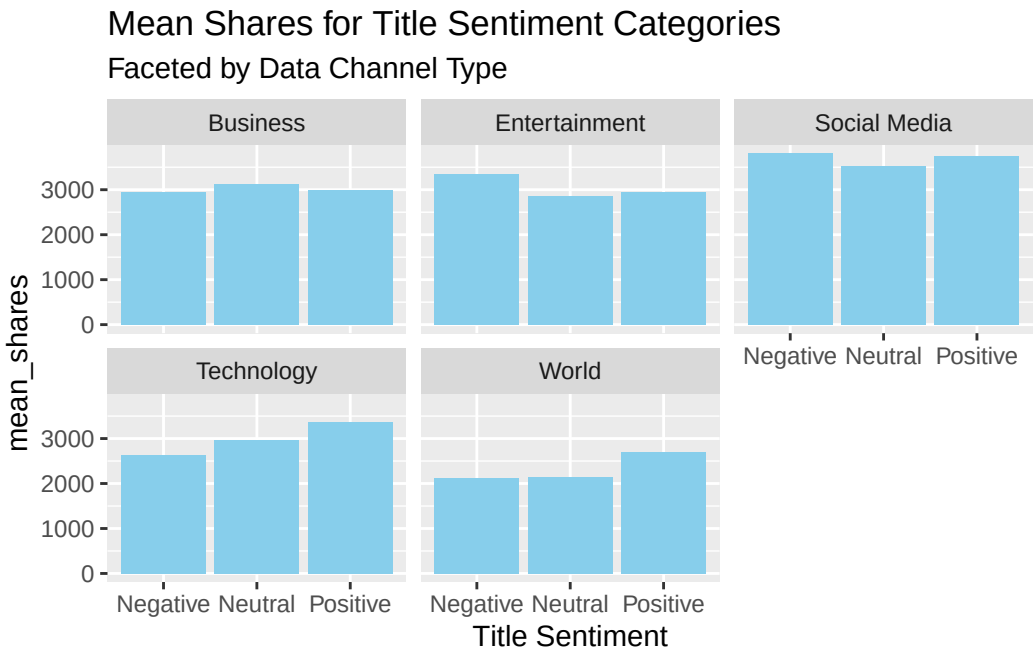
Table 12: Regression Coefficients for Both Models

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 13: 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.



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	-0.086	0.046	-1.863	0.062
data_channelSocial Media:title_sentiment_categoryNeutral	-0.043	0.072	-0.595	0.552
data_channelTechnology:title_sentiment_categoryNeutral	0.020	0.050	0.405	0.685
data_channelWorld:title_sentiment_categoryNeutral	-0.059	0.044	-1.326	0.185
data_channelEntertainment:title_sentiment_categoryPositive	-0.061	0.049	-1.252	0.211
data_channelSocial Media:title_sentiment_categoryPositive	-0.077	0.075	-1.024	0.306
data_channelTechnology:title_sentiment_categoryPositive	0.049	0.053	0.929	0.353
data_channelWorld:title_sentiment_categoryPositive	0.039	0.048	0.810	0.418

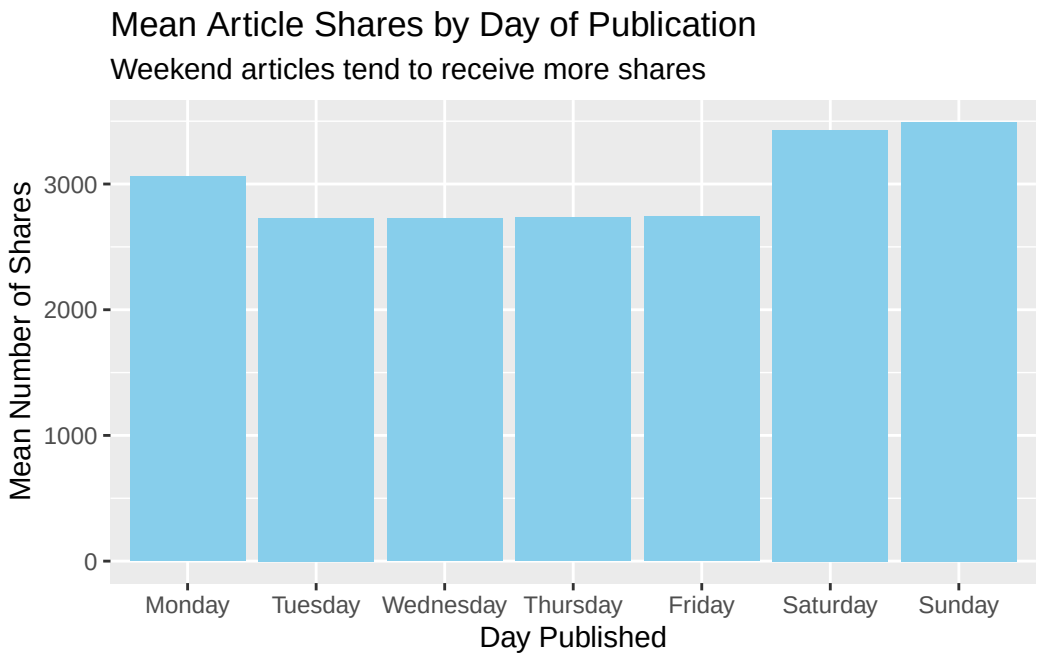
However, when looking at the actual fitted model, all the interaction terms between the data channel and the sentiment category have large p-values (greater than 0.05) suggesting that none of the interaction terms are actually statistically significant between data channel and title sentiment category. 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(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(shares). However, while the number of words in the title does not have a significant linear relationship with log(shares) directly, its interaction term specifically with when the data channel is World, is significant (as shown by the p-value of 0.003).

```
# 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.
```




```
# 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.
```