Characterizing and Predicting Article Virality based off of Metadata and 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, we are interested in exploring what exactly impacts an article's virality and can these factors aid news agencies when publishing new articles. Our research question is as follows: 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 representing an individual article and its associated textual and metadata features.

Key Variables:

is_viral - Binary response variable we created to evaluate whether an article is considered viral or not, based on whether total shares is greater/less than 1,400 based on past literature (Fernandes Et. al 2015) (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.

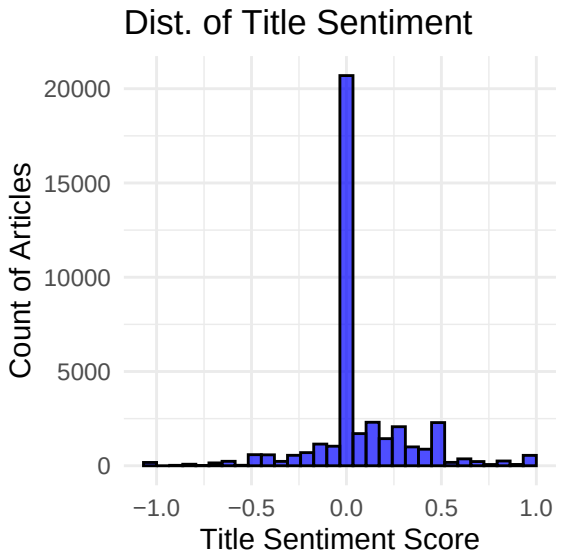
n_tokens_content - Number of words in the article content. Values range from 0 to 8,474 words, with a mean of 546.51 and standard deviation of 471.10. This quantifies the overall length of the article.

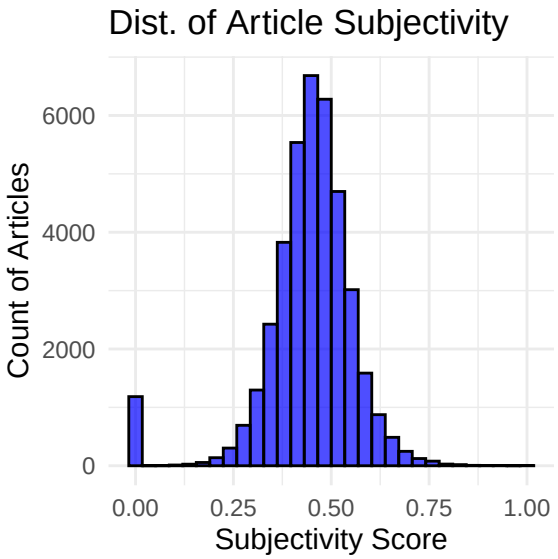
kw_avg_avg - Average shares of average keywords in the article. Values range from 0.0 to 43,567.66, with a mean of 3,135.86 and standard deviation of 1,318.13. This measures the expected popularity of the article's keyword selection.

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.

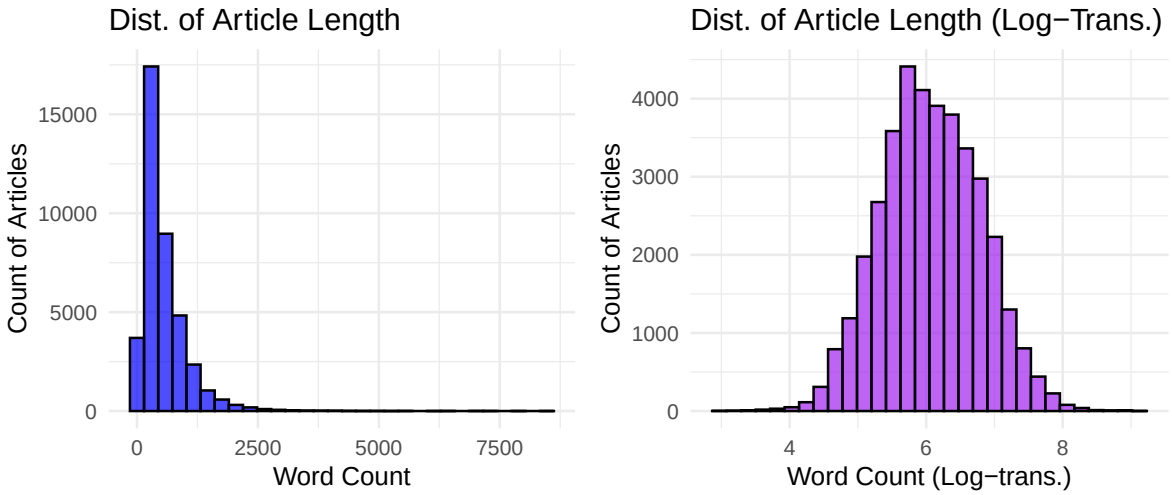
Univariate EDA

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.



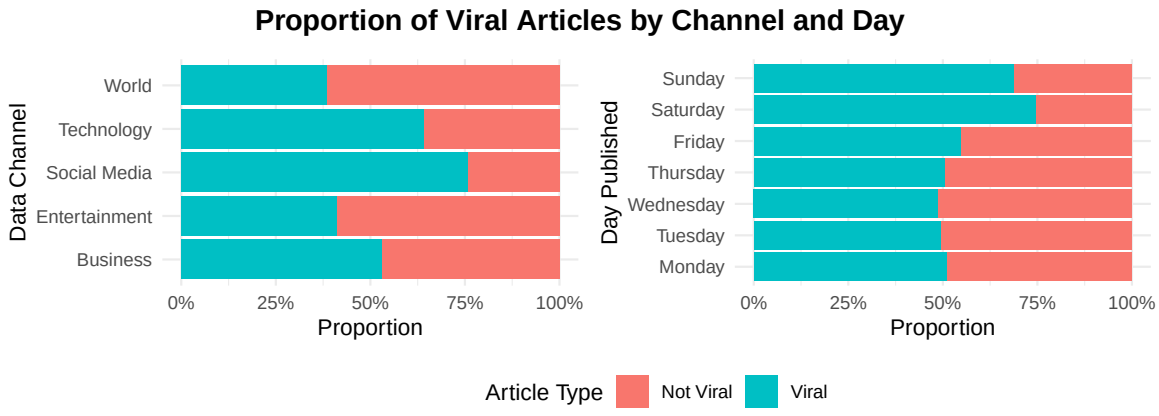


`title_sentiment_polarity`'s distribution suggests that many titles are emotionally neutral. The `global_subjectivity` variable has a roughly symmetric distribution, suggesting most articles contained a balanced mix of factual and opinion-based language. Both display relatively balanced distributions that do not require transformation. `kw_avg_avg` also has a relatively symmetric distribution, and was left untransformed (re: fig 1, appendix).



`n_tokens_content` displayed a heavily right-skewed distributions that warranted log transformations. To correct for this skewness and reduce the influence of extreme values, we log-transformed it, resulting in a more centered and symmetric distribution. better aligning with modeling assumptions.

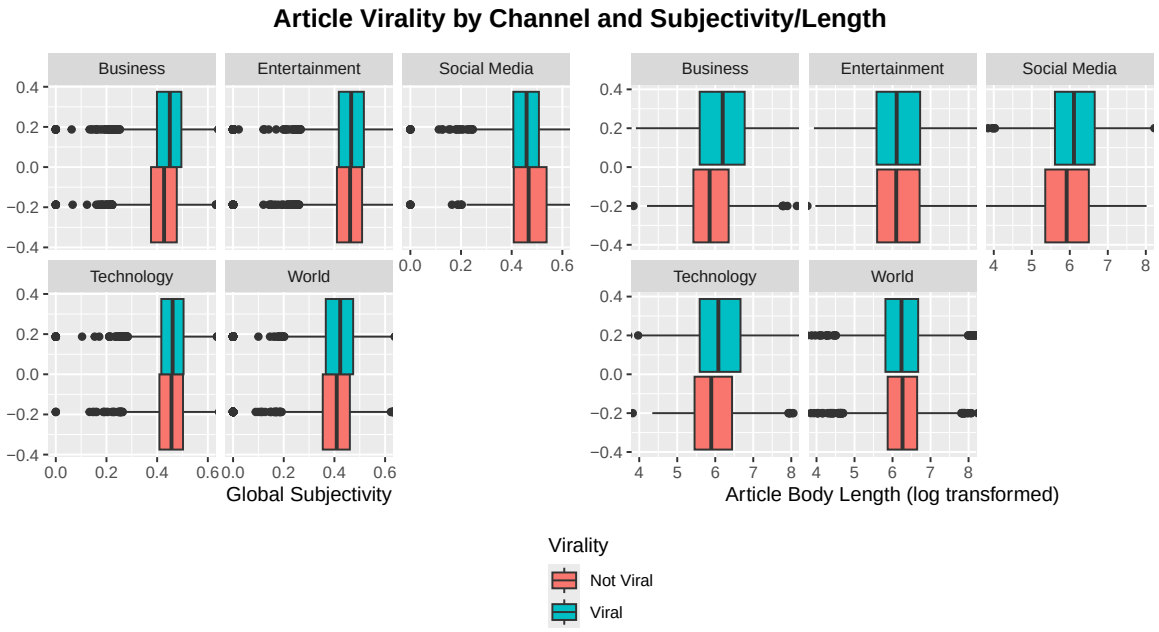
Bivariate EDA



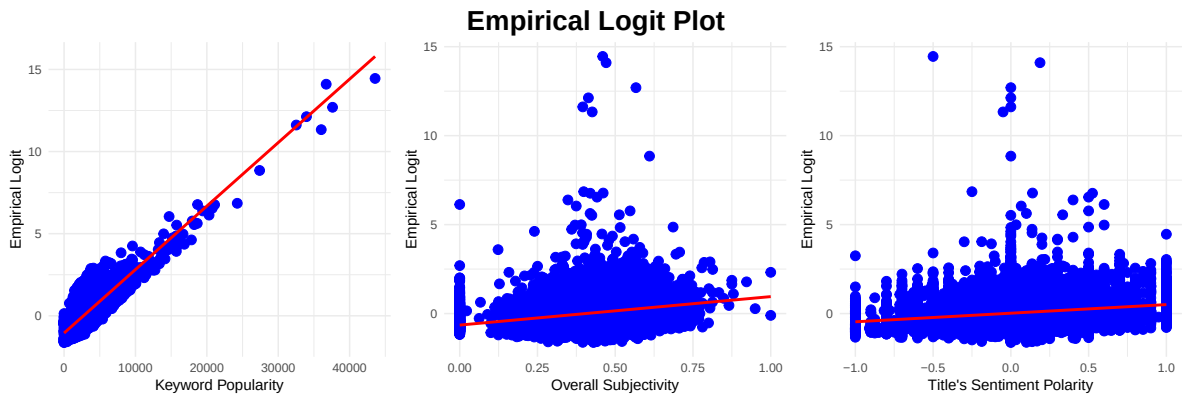
The first graph shows how virality varies across different content categories, described by data_channel. Articles published in the Social Media category have the highest proportion of viral articles, compared to other categories. This suggests that content tailored for or about

social platforms may be particularly beneficial for engagement. Articles that are categorized under Technology and Business also show strong performance, with over half of the articles in each channel categorized as viral. On the other hand, articles in the Entertainment and World categories had a lower proportion of viral articles, falling below the 50% mark. This graph underscores how topic area potentially influences content reach due to 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 proportion of viral content, followed closely by Sunday. In contrast, weekday articles 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.



From visualizations of global subjectivity/(log) article body length compared to virality proportions and proportioned by topic area, we also found two potential interaction effects we wanted to consider when building our model, as the relationship between global subjectivity/(log) article length and virality differed across data channel type.



The empirical logit plots offer further insight into the relationship between key predictors and the binary response, `is_viral`. The plot for average shares of average keywords reveals a linear positive relationship with log odds of virality. As the average popularity of an article's keywords increases, the log odds of the article going viral increases, with the empirical logit showing an upward trend. This suggests that keyword selection plays a key role in driving article engagement and affirms kw_avg_avg is a key predictor in modeling vitality. In contrast, the plot of overall subjectivity indicated a slight linear positive relationship. Although the empirical logit fit line trends upward, the data points are widely dispersed, and the effect appears weak and inconsistent, implying subjectivity alone is not a reliable driver of viral outcomes. Similarly, the plot for title sentiment polarity shows a marginal positive slope, implying that more positive article titles may be associated with slightly higher chances of virality. However, the relationship is again weak and shows considerable variability. Overall, the empirical logit plots highlight meaningful differences in predictive strength across variables and suggest that keyword popularity metrics are stronger predictors of article virality than emotional tone or subjectivity.

Methodology

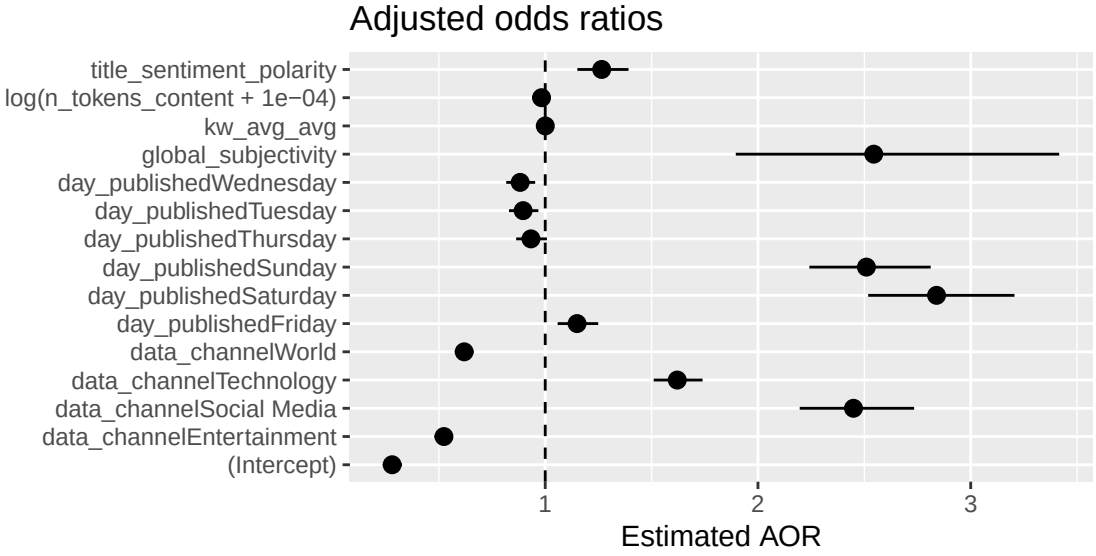
While we initially fitted a linear regression model, using log(shares) as our response variable (re: fig 2, appendix), we found that it violated the linear regression assumption of constant variance (re: fig 3, appendix). Thus, based on our initial EDA and empirical logit visualization, we selected data channel, day published, article subjectivity, title sentiment polarity, average keyword popularity, and (log) article content length to fit an initial logistic model instead, shown below:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \text{kw avg avg} + \beta_2 \log(\text{n tokens content} + 10^{-4}) + \beta_3 \text{global subjectivity}$$
$$+ \beta_4 \text{title sentiment polarity}$$
$$+ \left(\sum_{i=1}^{4} \beta_{4+i}\ \mathbf{1}\{\text{data channel} = \text{level}_i\}\right)$$
$$+ \left(\sum_{j=1}^{6} \beta_{8+j}\ \mathbf{1}\{\text{day published} = \text{level}_j\}\right)$$

All the predictors in our initially fit model possess significant p-values ($p<0.05$) and most predictors have relatively high magnitude z-statistics, indicating that all variables in the model have statistically significant relationships with the likelihood of content going viral (re: fig 4, appendix).

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 did not include 1, suggesting that the majority of our predictors significantly contributed to the model fit. While a couple predictors (such as day_published_Thursday and log(n_tokens_content) were close to 1, we kept them in the model due to the overall significance of the categorical variable (ex. day_published) or their role in potential interaction effects.

Interaction Effects

Next, we considered the addition of potential interaction effects between article length and data channel, and between global subjectivity and data channel based on our EDA. The hypothesis for this test was:

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

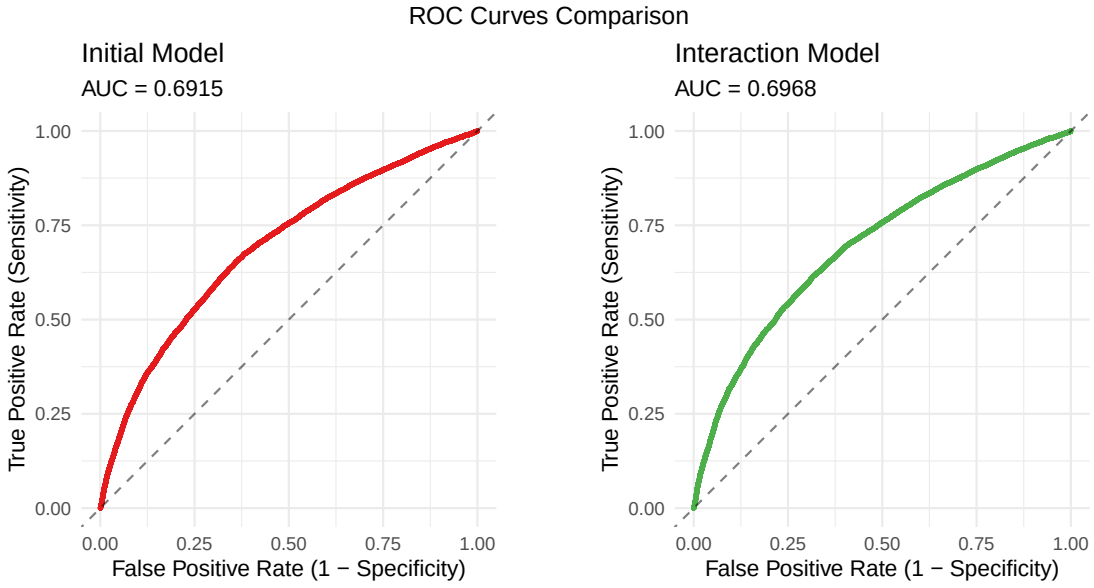
Table 1: Drop in Deviance Test Results

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 keyword popularity, and main body length for a given article. This suggests that these interaction effects add to our final model.

Model Evaluation and Comparison

To reaffirm this model selection, we further compared our two models' ROC curves and AUCs.



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 = 0.6968) demonstrates marginally better predictive performance than the Initial Model (AUC = 0.6915), 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 0.493, to be used when evaluating the final model (view appendix).

Results

The final model we fit was:

$$\begin{aligned}
\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 &+ \beta_1 \text{ kw avg avg} + \beta_2 \ \log(\text{n tokens content} + 10^{-4}) + \beta_3 \text{ global subjectivity} \\
&+ \beta_4 \text{ title sentiment polarity} \\
&+ \left(\sum_{i=1}^{4} \beta_{4+i} \ \mathbf{1}\{\text{data channel} = \text{level}_i\}\right) + \left(\sum_{j=1}^{6} \beta_{8+j} \ \mathbf{1}\{\text{day published} = \text{level}_j\}\right) \\
&+ \left(\sum_{i=1}^{4} \beta_{14+i} \ \log(\text{n tokens content} + 10^{-4}) \ \mathbf{1}\{\text{data channel} = \text{level}_i\}\right) \\
&+ \left(\sum_{i=1}^{4} \beta_{18+i} \text{ global subjectivity } \mathbf{1}\{\text{data channel} = \text{level}_i\}\right)
\end{aligned}$$

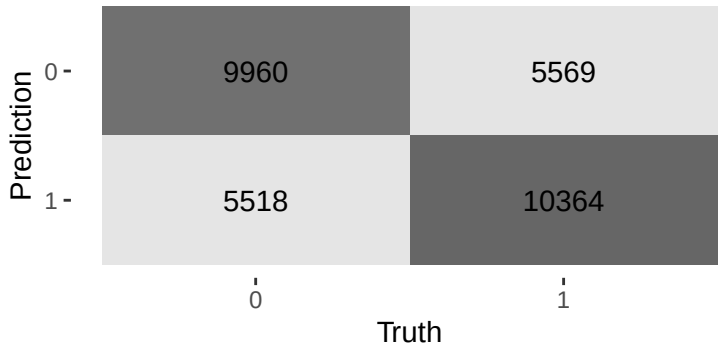


Figure 1: Confusion Matrix for the Interaction Model

When evaluating our final model, we see it has an AUC of ~ 0.697, an accuracy of $\sim 64.7\%$, specificity of $\sim 64.3\%$, and sensitivity of $\sim 65.0\%$ (re: fig 5). If we consider the misclassification

rate of only $\sim 35.3\%$, evaluation suggests that our model is moderately well-fit for the data. This precision means that $\sim 65.3\%$ of articles predicted to be viral were correctly classified, indicating that the model performs substantially better than random chance.

If we assess the assumptions of logistic regression underlying our model, we see approximately linear relationships randomly dispersed around the trend line. Notably, vertical scatter of points around the trend line for subjectivity and title polarity presents a high degree of noise that might be impacting model performance.

Interpreting the coefficients of the model, we can conclude:

Article category, or 'Data Channel' the most strongly correlated with virality, with 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, with weekend publications significantly outperforming 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 than similar Monday articles, with odds ratios of 0.883 and 0.865, respectively. This weekend effect may arise from increased leisure time, as people take off from work or school during the weekend.

Article subjectivity and sentiment likely also influence 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.

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 were not statistically significant, suggesting that the impact of Article Length on probability of virality does not differ greatly for Social Media and Technology compared to the baseline catagory, Business.

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, each one unit subjectivity increase reduces the 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.

Discussion + Conclusion

Our findings support the idea that Article Length, Category, Keyword Popularity, Date of publication, subjectivity and title polarity all contribute to an online article's success, while other factors like the rate of positive/negative words were less reliable predictors. Through the empirical logit plot and the drop-in-deviance test of the interaction terms, we determined that the effects of article length and global subjectivity both vary based on article topics.

Since the Mashable repository was collected over a multi-year time period, one limitation of our approach is the distinct possibility of outside temporal factors. This poses a potential violation of the independent observation assumption, with major events potentially shaping article performance on shorter time scales.

Another limitation to our findings is the poor interpretability of the NLP derived predictors, such as `title-sentiment-polarity` and `keyword average average`, which were derived in an extensive process by the UCI researchers who curated the dataset.

Finally, our model only contains observations from Mashable, so extrapolation to other news/article sites may be limited.

To reduce the influence of these factors in future work, we might control for time period in the data and compare articles between different sites for better generalization. Beyond that, utilizing our own NLP metrics might give us finer grain control, allowing us to consider better predictors for our model.

Citations

Fernandes, K., Vinagre, P., Cortez, P., & Sernadela, P. (2015). Online News Popularity [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5NS3V.

Fernandes, Kelwin et al. "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News." Portuguese Conference on Artificial Intelligence (2015).

Obiedat, R. (2020). Predicting the popularity of online news using classification methods with feature filtering techniques. Journal of Theoretical and Applied Information Technology, 98(8), 1163–1172. http://www.jatit.org

Appendix

Figure 1: Average Keyword Distribution

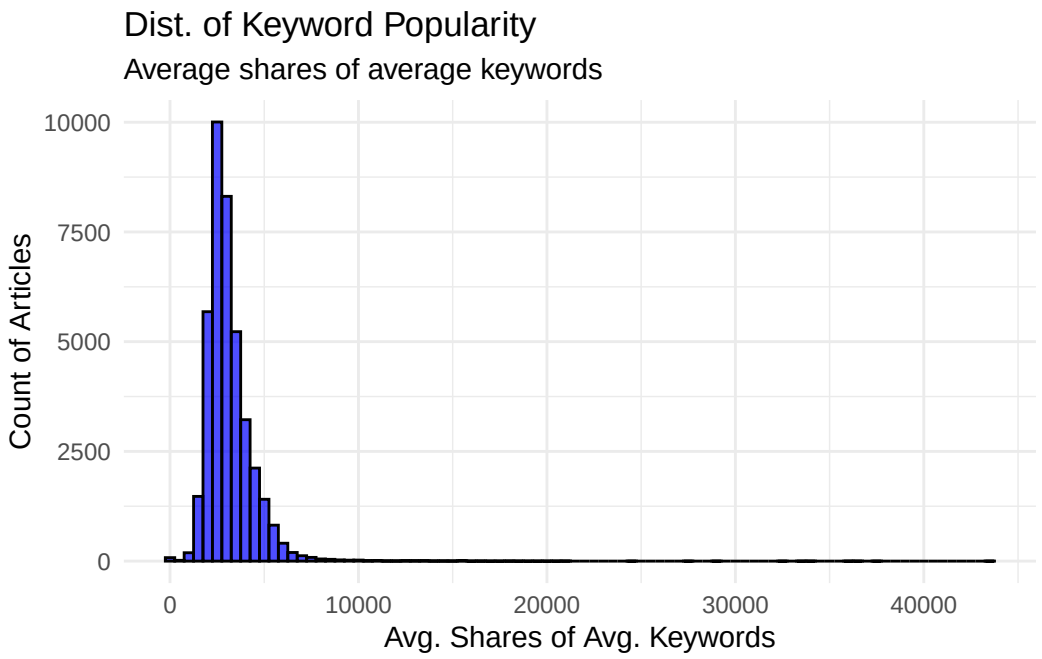


Figure 2: Initial MLR Model

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

term	estimate	std.error	statistic	p.value
day_publishedFriday	0.0162	0.0168	0.9627	0.3357
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

Figure 3: MLR Residual Plot

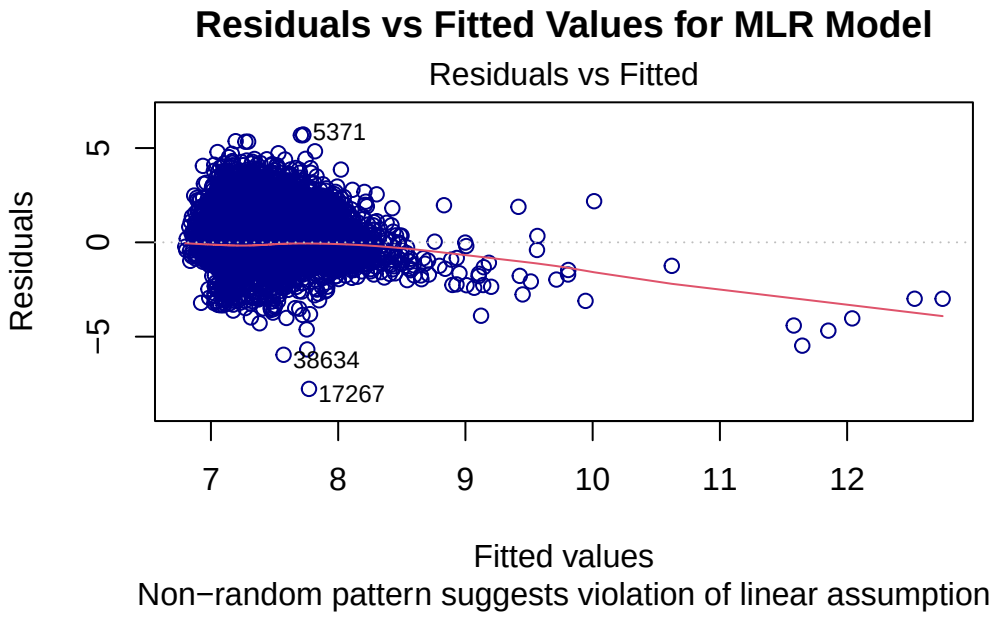


Figure 4: Initial Logistic Regression Table

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

Term	Estimate	Std.Error	z-statistic	p-value
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

Figure 5: Final Logistic Model Metrics

Table 4: Logistic Model Metrics Summary

Metric	Value
Accuracy	0.647
Misclassification Rate	0.353
Sensitivity (Recall)	0.650
Specificity	0.643
Precision	0.653
False Positive Rate (FPR)	0.357
False Negative Rate (FNR)	0.350

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