DATASCI 203, Lab 2 Report

Project Repository

Spring Sec 5: Kobby Hanson, Saranya Tadikonda, Yoko Morishita

4/17/25

Online News Popularity: Analysis of Key Features

Introduction

In today's attention economy, a single news article can reach millions or vanish within hours. What determines this fate? This question is fascinating enough to drive our analysis. Every day, we scroll past many headlines but only a few capture our interest enough to share. This sharing behavior more than anything determines what survives in this digital era where fast news consumption and competition dominates.
In the past publishers relied on their editorial intuition to write the best stories, but modern media brings new opportunities to quantify these relationships through data. This analysis investigates the relationship between characteristics of online news articles and their popularity, as measured by social media shares. Our focus is on quantifying how decision points like the publication timing, multimedia usage, title length, and overall sentiment correlate with the number of shares. Researching the topic of online news is guided by:

- What factors are associated with the social media shares of online news articles?
- How strongly are these factors related to article share volume?

Understanding these specific relationships would be valuable for publishers seeking to potentially increase audience engagement in the digital space by building data-driven growth strategies.This study uses the same dataset as Fernandes et al. (2015), who focused on predictive classification using machine learning models. Our approach differs significantly by concentrating on descriptive analysis and statistical inference using linear regression.

Data

This analysis draws on the "Online News Popularity" dataset from the UCI Machine Learning Repository, sourced from articles published by Mashable.[1] The dataset contains 39,797 unique news articles described by 61 attributes (58 predictive, 2 non-predictive, 1 target). The non-predictive variables includes time delta which describes how long since the article had been published. The target variable shares, a continuous measure of social media shares for each article.
The analysis uses a subset of the 58 predictive attributes which include text length metrics such as word counts and average word length, counts of images and videos, binary indicators for content channels and sentiment polarity derived by Pattern web mining. This dataset contains no missing values, allowing us to use all the information directly without needing to fill in any gaps. However, as discussed in the next section, some feature values raise concerns about the quality and were carefully filtered out during the wrangling process to better interpret the relationships between the factors and the number of shares.

[1] Fernandes, K., Vinagre, P., Cortez, P., & Sernadela, P. (2015). Online News Popularity. Dataset in UCI Machine Learning Repository.

Data Wrangling & Operationalization

Prior to modeling, the raw dataset was transformed following steps in the wrangling.Rmd file. The first key consideration behind was ensuring a consistent observation window for article popularity. Since the dataset was collected on January 8, 2015, we kept only those from 2013, which cut the dataset from about 39k to 18k articles. This choice to exclude 2014 articles helps guarantee that each article had at least one full year of full exposure for shares to stabilize before the data collection. Then a focused subset of descriptors were chosen to set up the descriptive models based on their statistical properties and our judgment as authors, guided by real-world relevance:

- Day of Week Factors: The initial seven binary indicators (weekday_is_) were consolidated into a single binary variable (is_weekend) to distinguish between weekend and weekday publications, simplifying interpretation.
- Structural Features: Although we removed articles that have zero in n_tokens_content (body word count) to draw meaningful insight, the variable was dropped from model due to its high variability for data quality concerns. Instead, n_tokens_title (word count of the title) was retained as it ranged from 2 to 23 words, aligning more realistically with expected title lengths.
- Sentiment Indicators: Although title_sentiment_polarity was considered, the global variable was favored. This decision reduced multicollinearity, as the title sentiment was highly correlated with global sentiment, which reflects the overall tone of the article.
- Multimedia Contents: The variable number of images and videos are summed to reflect the overall effect of multi-media contens on article popularity.
- Content Category: Multiple binary variables (data_channel_is_) representing the article's content category (lifestyle, entertainment, business, social media, technology, world) were merged into one categorical variable. This approach streamlined the model while still capturing the article's theme.

Lastly, the dataset was randomly divided into a 30% exploration set and a 70% confirmation set. This allowed us to examine variable distributions and relationships, and test linear regression models with multiple sets of variables. We selected the full model (described later) because it showed the best residual distribution and statistically significant results. Finally, we applied this model to the confirmation set without adjustments to avoid bias.

Model Specification

Our null hypothesis states that the day of the week an article is published has no effect on its number of shares. The full model was added to test whether the features significantly explained the variance in shares. Given that the shares distribution was heavily right-skewed, we applied a base-10 logarithm to better satisfy the assumptions of linear regression.
Simply, the reduced model is: The relationship between transformed shares and timing.

$$\log_{10}(\text{shares}) = \beta_0 + \beta_1 \times \text{is\_weekend} + \epsilon$$

The full model: The combined effects of timing, number of multimedia (images and videos), title length, overall sentiment polarity, and category on share counts.

$$\log_{10}(\text{shares}) = \beta_0 + \beta_1 \times \text{is\_weekend} + \beta_2 \times \text{num\_digital\_media} + \beta_3 \times \text{n\_tokens\_title} + \beta_4 \times \text{global\_sentiment\_polarity} + \beta_5 \times \text{data\_channel} + \epsilon$$

Model Assumptions

IID: The articles are assumed to be independent and identically distributed, but this raises some concerns. Since all articles come from Mashable within a 1–2 year span, they might be influenced by common factors such as internal links, site-wide trends, or external events. Additionally, articles within the same category or different content types like tech versus lifestyle may behave similarly and attract shares in different ways. While these present limitations, they also reflect the complexities of real-world digital media. Recognizing these issues is key to adjusting the models and maintaining the IID assumption.

Hetero-Skedasticity: The Breusch-Pagan test revealed strong evidence of heteroskedasticity with a highly significant p-value ($< 2.2\text{e-}16$), indicating a violation of the constant variance assumption. To address this, we corrected for heteroskedasticity in our exploration by using robust standard errors, allowing for more reliable inference on the model's coefficients.

No Perfect Collinearity: We can examine the relationships among the numeric descriptor variables shown in the pairs plot (Fig 1 in Appendix). Specifically the plot displays the correlations between the descriptors: -0.042 (title tokens vs. sentiment), 0.026 (title tokens vs. digital media), and -0.048 (sentiment vs. digital media). These values are all very close to zero, indicating very weak linear associations between these variables. The corresponding scatter plots visually confirm this lack of strong linear patterns. Based on this evidence, the assumption of no perfect collinearity is met for this set of descriptors, and high multicollinearity does not appear to be a concern.

Linear Conditional Expectation: Evaluating the linear conditional expectation assumption requires examining the Residuals vs Fitted plots (Fig 2). The first plot (pre-transformation) shows clear heteroscedasticity, with residuals fanning out and this violates the constant variance assumption. A log transform to shares was chosen as the underlying distribution was right skewed from the analysis on multicollinearity. In contrast, the second plot (post transformation) shows a marked improvement: residuals are more evenly spread around zero, with reduced variance and less strong outliers. While the scatter still suggests slight non-linearity, the transformation significantly improved model assumptions, even if it may still contain minor violations.

Normal Distribution of Residuals: The initial analysis of the residuals shows a highly right-skewed distribution, clearly seen in the first histogram in Fig 3 where the frequency is heavily concentrated at low residual values with a long tail extending to the right. The second histogram, showing the residuals from the full model after transformation, demonstrates

a substantial improvement. The transformed residuals become more symmetrical and take on a bell-shaped distribution centered around zero, closely approximating a normal distribution. While this log transformation significantly fixed the original skewness and resulted in a distribution far more suitable for analysis, the improved result is still not perfectly normal.

Model results and Interpretation

From Table 1 in the appendix, the final robust linear regression model described article popularity using the log transform of shares based on timing, multimedia count, title length, sentiment, and content channel. Using robust standard errors to improve reliability, the analysis found that it is statistically significant that articles published on weekends had more shares.(coeff = 0.142, $p < 0.01$) In a practical sense, this means weekend articles has about 39% more shares than comparable weekday articles, holding other factors constant. Applying similar interpretations to other significant factors, including more images or videos (coeff = 0.002, $p < 0.01$) also had a significant but smaller positive relationship being roughly 5% more shares for every 10 items.

In contrast, most specific content channels underperformed relative to the baseline (non-category). For instance, the "World" (coeff = -0.229, $p < 0.01$) and "Business" (coeff = -0.162, $p < 0.01$) has substantially fewer shares (around 41% and 31% less, respectively), while "Social Media" (coeff = 0.015, $p > 0.1$) showed no significant difference. Notably, after accounting for these other factors, title length (coeff = -0.003, $p > 0.1$) and overall article sentiment (coeff = 0.012, $p > 0.1$) were not statistically significant. Meaning, we cannot confidently conclude these variables have a genuine effect on article log-shares. Their apparent effects could be due to random variation in the data.

Overall Effect

Table 2 ANOVA compares the reduced model to the full model. The extremely small p-value ($< 2.2e\text{-}16$) rejects the null hypothesis which tells us that adding these extra descriptors significantly improves how well the model explains article popularity.

The most actionable finding is the weekend advantage, suggesting that scheduling important articles for weekends could have more shares by a substantial margin (around 39%). Adding extra images or videos can help but with a smaller benefit of roughly 5% more shares for every 10 items. The analysis also shows that some content channels, like "World" or "Business," naturally receive lower engagement, which can inform topic choices.

Interestingly, this model result suggests that title length or overall sentiment might be less important than timing, adding media, and the choice of topic once those are considered. However, a key takeaway is the model's limited descriptive power (R^2 6.3%) for using only this information to build a growth strategy. While these factors provide measurable leverage, they are not guaranteeing success, as most of what drives shares likely comes from other factors the model didn't capture; elements like specific topic appeal, promotion, and inherent quality.

Appendix - Tables and Figures

Fig 1: Numeric Variable Pairs Plot

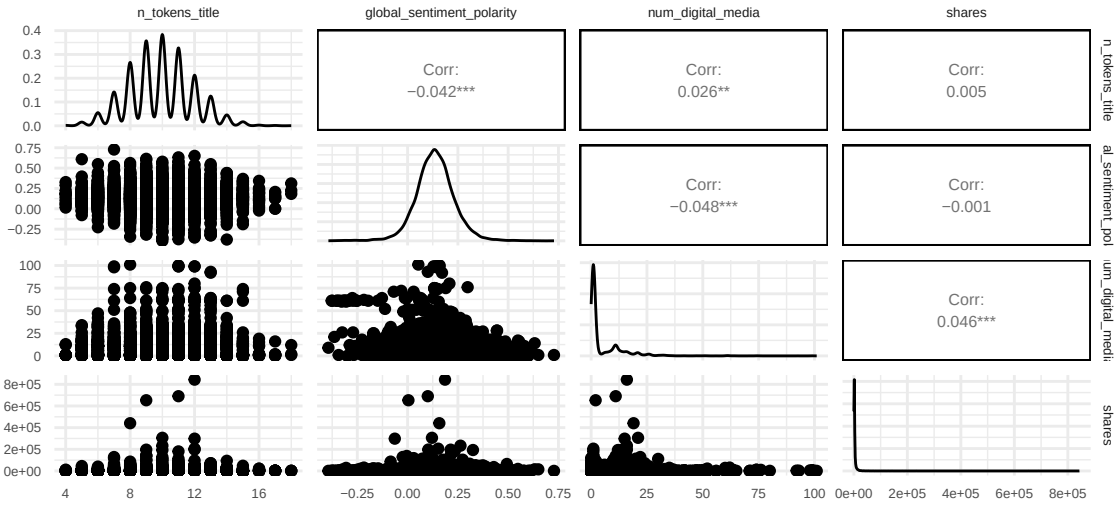

Fig 2: Residuals vs Fitted Values of Full−model

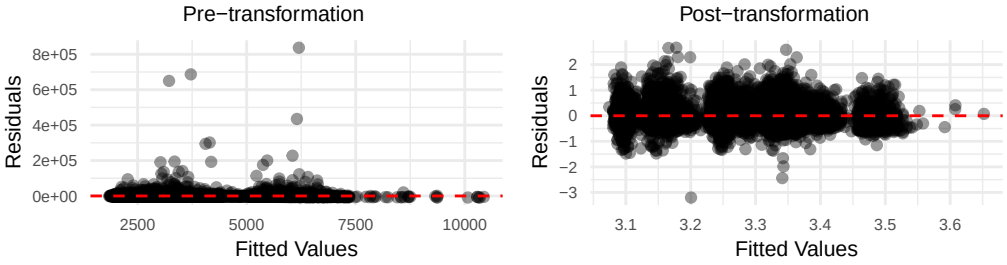

Fig 3: Residuals Histograms of Full−model

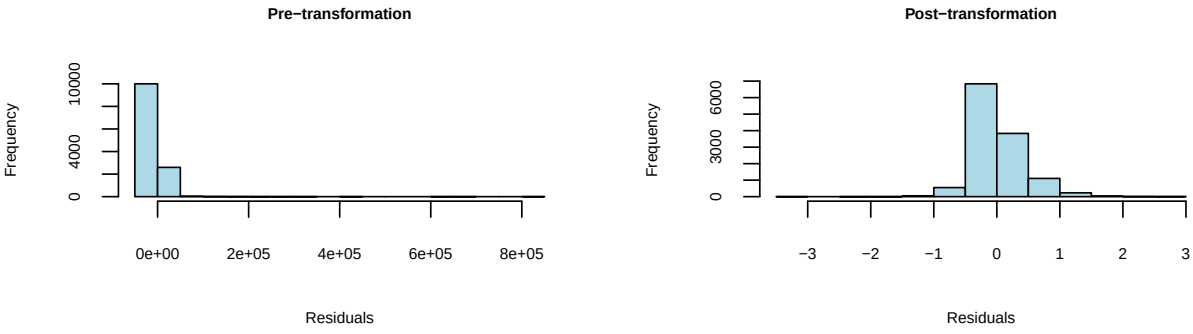

Table 1: Full-model Linear Regression Results with Robust SE

	Dependent variable:
	Shares (Robust SE)
is_weekend	0.142^{***}
	(0.010)
n_tokens_title	-0.003
	(0.002)
global_sentiment_polarity	0.012
	(0.040)
num_digital_media	0.002^{***}
	(0.0005)
data_channelBusiness	-0.162^{***}
	(0.014)
data_channelEntertainment	-0.182^{***}
	(0.014)
data_channelLifestyle	-0.073^{***}
	(0.018)
data_channelSocial Media	0.015
	(0.016)
data_channelTech	-0.081^{***}
	(0.013)
data_channelWorld	-0.229^{***}
	(0.015)
Constant	3.349^{***}
	(0.022)
Observations	12,652
R^2	0.063
Adjusted R^2	0.062
Residual Std. Error	0.401 (df = 12641)
F Statistic	84.353^{***} (df = 10; 12641)
Note:	*p<0.1; **p<0.05; ***p<0.01
	Standard errors are heteroskedasticity-robust.

Table 2: Anova test Reduced Model vs Full Model

Statistic	N	Mean	St. Dev.	Min	Max
Res.Df	2	12,645.500	6.364	12,641	12,650
RSS	2	2,081.678	68.595	2,033.174	2,130.182
Df	1	9.000		9	9
Sum of Sq	1	97.008		97.008	97.008
F	1	67.015		67.015	67.015
Pr(>F)	1	0.000		0	0