Online News Popularity

Jiali Huang

Professor Alexander Petersen

PSTAT 126

Winter 2017

Abstract:

When browsing social media sites, there are often articles that more than one of your friends are sharing. As people share it amongst the friends, a ripple effect leads to some articles to go viral. This study examines a few attributes of articles, such as number of images and videos, to find predictors that affect the number of times an article is shared. The data used has 39644 observations, and is extracted from “Mashables.com”. The regression models derived from the data shows that articles with a larger proportion of positive words and articles that were released during the weekend have the strongest positive effect on shares, while a longer average word length tends to make an article less sharable.

Problem and Motivation:

Online articles are becoming more and more prominent, as they have already overshadowed tradition news outlets, and are continuing to expand. However, due to oversaturation of the medium, many articles only get a modest amount of views due to other media forms, such as videos, becoming more and more popular. There are very few articles that go viral, but those select few that do go viral, get a substantially larger number of shares and views than most others articles. This is a direct result of how easily information can get spread to the masses through the internet. The audience for online media is still growing as internet connections continue to expand across the globe, allowing for some articles to reach much larger audiences than traditional news outlets such as newspapers and magazines. More importantly, the demographic for viewers of online articles is disproportionally skewed towards teens and young adults, a prime target for many companies to run advertisement campaigns on. If there are ways to make articles more attractive for sharing and consumption, websites may find themselves having more web traffic, and thus more advertising revenue. It should be noted that the results found in this study is specific to Mashable articles, and its applicability to other online articles may be limited.

Questions of Interest:

Are there specific attributes that help articles reach a broader audience by increasing shareability, and are there some attributes that hinder an articles’ ability to go viral? Does the weekend affect how many articles are shared? Is there a relationship between the weekend and other predictors?

Data:

The data is a set of web analytics of “Mashables.com”, titled “Online News Popularity Data Set”, found on UCI machine learning repository website. The source of the data is from University of Porto, Portugal.

Response Variable = Shares
The number of times the article is shared onto social media. This is likely related to the number of views of an article (not part of the data), and whether the article goes viral.

Predictors =
n_tokens_content: the number of words in the article
num_hrefs: the number of URL links on the article
num_imgs: the number of images in the article
num_videos: the number of videos in the article
is_weekend: a dummy variable for if the article was published over the weekend
global_rate_postive_words: percent rate of positive words,
global_rate_negative_words: percent rate of negative words
n_unique_tokens: rate of unique words in the article

Note: There are interactions between is_weekend and all other predictors.

Regression Methods:

First, a multilinear regression was done on base predictors and interactions with is_weekend onto the response variable, shares. The second model is a log transformed version of the first model, using log(shares) as the response variable instead. The final model is attained by using backwards selection from the second model, reducing the number of predictors in the regression. The final model will show us which predictors have effects on the number of shares of an article based on the AIC criteria.

Regression Analysis with Diagnostics:

First, start with the full untransformed model, with shares as the response. Then diagnostic tests were done on the first regression to check linear regression model assumptions. To check normality of residuals, we use a normal QQ plot.* The plot for our first model shows that the normality assumption fails. There is an increase for values in standardized residuals on the right. The Pearson residuals vs fitted values plot shows signs that support the linearity assumption, however there are a few outliers, such as observations 321 and 2843. The scale location plot shows that there seems to be non-constant variance, as points to the right are much closer to the fitted line, however, the lack of data for higher fitted values make it hard to determine. The outliers found in the data, having several magnitudes more views than the majority of other articles, giving the model a very low R-squared value at .005288. This means .5288% of the variability in the observed values for shared is explained by the predictors. Taking this into account, we transform the model so that we have log(shares) to better reflect the range of values for shares due to virality.

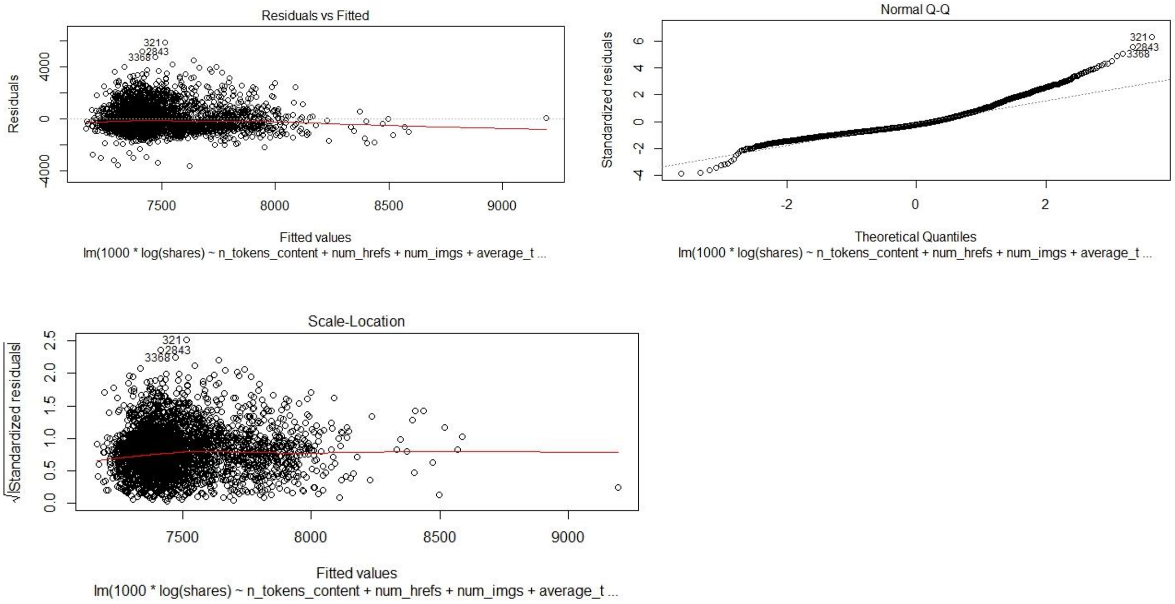
A Box-Cox plot in the appendix shows a lambda value that is sufficiently close to 0, and is one of the reasons why the log transformation was chosen. The normal QQ plot shows a weaker skew than before, and the residual vs fitted plot shows a better linear relationship, as well

*Model 1 and Model 2 Diagnostics can be found in the appendix.

as normality. The scale location seems to also have constant variance, and is more apparent than what is seen in the untransformed model. The R-squared value is .03804, a substantial improvement over the old one, but still a small value. Only 3.804% of the variability observed in log(shares) is explained by the predictors.

The final step is to simplify the model based on an information criteria (AIC). Using the backwards selection method, starting with the full model, we eliminated the predictors: is_week*average_token_length, is_weekend*num_hrefs, is_weekend*num_videos, and is_weekend*global_rate_positive_words. Running the same diagnostics shows that the removal of these predictors did not have a significant effect on the diagnostic tests or the R-squared, at .03796. Therefore, the final model shows that the predictors explain 3.796% of the variability in the observed value for log(shares).

Model 3 Diagnostics



To see whether the interaction terms are necessary in the final model, we do a partial f-test with the null and alternative hypotheses:

H_0: is_weekend interaction coefficients = 0

H_A: is_weekend interaction coefficients $\neq$ 0

From the ANOVA table of the full and partial model, we get an F value equal to 5.934, which leads to a very small p-value $\approx$ 0. The partial f-test suggests that we can reject the null hypothesis that the is_weekend interaction terms have no effect on shares, and we use the full model in the final analysis.

Interpretation:

The final model suggests that a higher rate of positive words in the article has the strongest positive effect on the number of shares, with its coefficient at 2940. Dividing this by 1000, we get 2.94. A one unit increase in the rate of positive words would lead to a 294% increase in the number of shares. Since this is a rate with upper bound 1, a .01 unit increase in the rate of positive words would lead to a 2.94% increase in in shares. The next largest positive coefficient is for the interaction between the weekend dummy variable and the number of unique words in the article. A 1 unit increase in the rate of unique words in the article over the weekend, leads to a 42.59% increase in shares. Similarly, since this is expressed as a rate, a .01 unit increase in the rate of unique words would lead to a .4259% increase in shares of an article over the weekend. On the other end, the interaction between the weekend dummy variable and the rate of negative words has the strongest negative coefficient. A .01 unit increase in the rate of negative words during the weekend leads to a 2.142% decrease in the number of shares compared to a similar article released during the weekdays. The average word length also seems to have a negative effect on how many shares an article receives. With each unit increase in the average word length leading to a 20.87% decrease in the number of shares of an article. This suggests that the higher the reading level of an article, the less likely it will get shared. Other variables had very small effects on the number of shares, but some of these small effects are more surprising than others. One of them being the weekend dummy, showing that, holding all else constant, an article released over the weekend only has 7.967% more shares than articles released during the weekday. This number is less than what is expected, since people have much more time over the weekends to read articles.

Final Conclusions:

Although the diagnostic tests had shown that linear model assumptions held for the final model, the predictors chosen were not good at explaining the variability in shares. There are some predictors that had statistically significant results, and strong effects on number of shares, however the explanatory power of the regression is lacking, having a large SSR. With an R-squared of only .003796, only 3.796% of the variability in the observed values in shares is explained by the predictors. These results are expected as the content likely has the greatest effect on an article. However, a surprising finding is how little effect images and videos had on the number of shares of an article, knowing the ease of consumption of the medium. Although this study has found that these predictors could not explain the number of shares of an article, future studies that contain analytics of the content of an article may lead to a much better analysis of how articles are shared. It could also be simply the fact that there is no recipe that creates viral articles, and the occur completely at random.

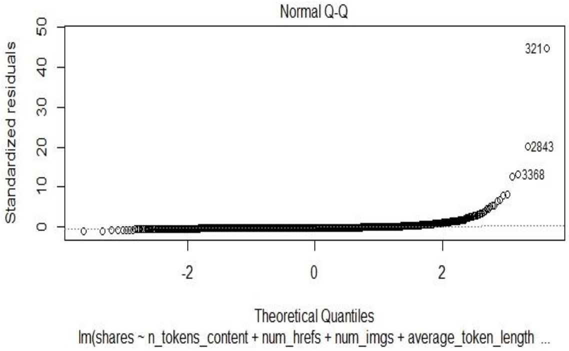
Appendix

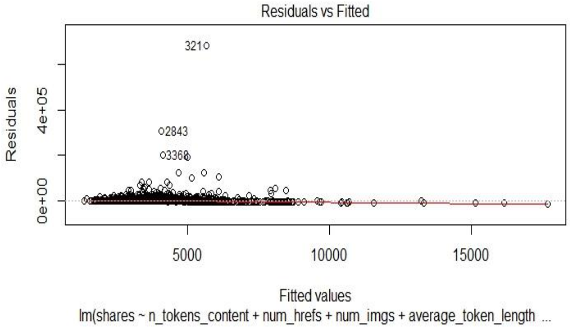
Code

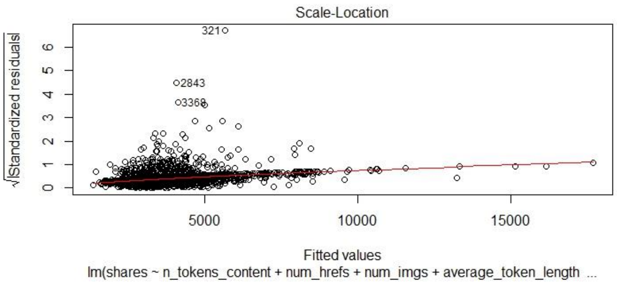
```
library(alr4)
onlineNewsData <- read.csv("C:/Users/Jiali/OneDrive/Documents/UCSB/PSTAT 
126/Project/OnlineNewsPopularity/OnlineNewsPopularity1.csv")

#Model 1

lm1 <- lm(shares ~ is_weekend*n_tokens_content + 
          is_weekend*num_hrefs + is_weekend*num_imgs + is_weekend*average_token_length + 
          is_weekend*num_videos + is_weekend*global_rate_positive_words + 
is_weekend*global_rate_negative_words + 
          is_weekend*n_unique_tokens, onlineNewsData)
summary(lm1)
avPlots(lm1)
pairs(onlineNewsData[c("n_tokens_content","num_hrefs","num_imgs","average_token_length", 
"num_videos","is_weekend",
                   "global_rate_positive_words", "global_rate_negative_words")])
plot(lm1, which = 1)
plot(lm1, which = 2)
plot(lm1, which = 3)
boxCox(lm1)

#Model 2: Log Transformation

lm2 <- lm(1000*log(shares) ~ is_weekend*n_tokens_content + 
          is_weekend*num_hrefs + is_weekend*num_imgs + is_weekend*average_token_length + 
          is_weekend*num_videos + is_weekend*global_rate_positive_words + 
is_weekend*global_rate_negative_words + 
          is_weekend*n_unique_tokens, onlineNewsData)
summary(lm2)
avPlots(lm2)
plot(lm2, which = 1)
plot(lm2, which = 2)
plot(lm2, which = 3)

#Model 3: Backward Selection

full = ~ is_weekend*n_tokens_content + 
  is_weekend*num_hrefs + is_weekend*num_imgs + is_weekend*average_token_length +
  is_weekend*num_videos + is_weekend*global_rate_positive_words +
is_weekend*global_rate_negative_words +
  is_weekend*n_unique_tokens
m0 = lm(1000*log(shares) ~ is_weekend, onlineNewsData)
m1 = update(m0, full)
lm3 = step(m1, scope = c(lower = ~ is_weekend), direction = "backward")

summary(lm3)
avPlots(lm3)
plot(lm3, which = 1)
plot(lm3, which = 2)
plot(lm3, which = 3)

#Partial F-test

partial <- lm(1000*log(shares) ~ is_weekend + n_tokens_content + num_hrefs + num_imgs +
average_token_length +
  num_videos + global_rate_positive_words + global_rate_negative_words + n_unique_tokens,
onlineNewsData)

anova(partial, lm3)
```

Additional Plots

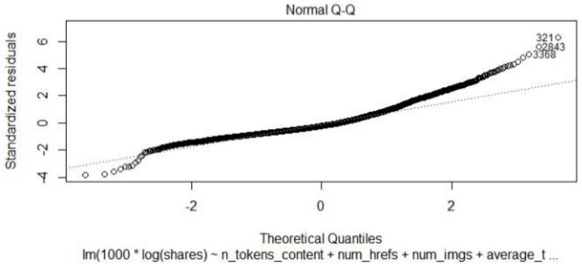
Model 1 Diagnostics

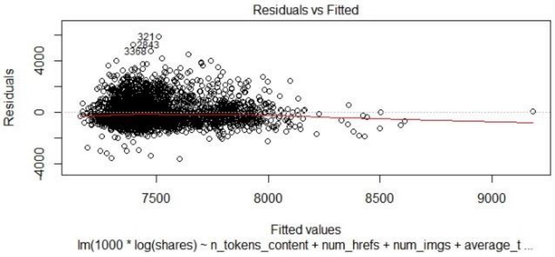


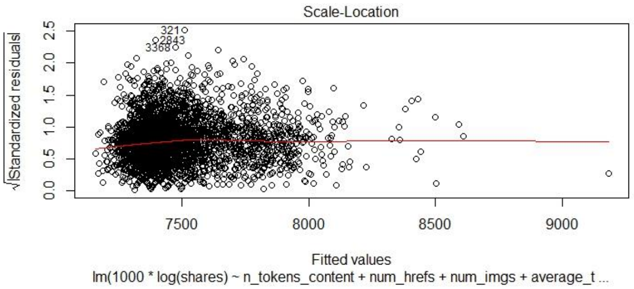




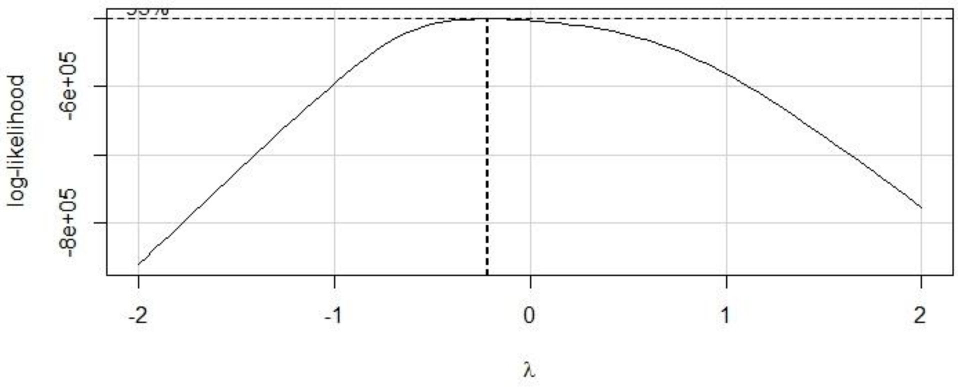
Model 2 Diagnostics







Box Cox Plot for Model 1


log-likelihood
-8e+05
-6e+05
λ