Uncovering the Relationship Between Online News Characteristics and Popularity

Abstract

Given online news' tremendous popularity and its potential societal impacts, this study seeks to understand which characteristics within the text of a news article corresponds to a higher number of shares. Using a data set obtained from the UC Irvine Machine Learning Repository that contains information on 39,644 articles published on Mashable, we utilized 58 predictive attributes to determine which had the greatest impact on the variable of interest. AIC and BIC stepwise regression were used to determine the best multiple linear regression model, and a regression tree was created to provide extra information about which variables are most useful. This final model suggests that day of the week, category, subjectivity, and amount of positive words are key characteristics of online news articles.

Background and Significance

The popularity and reach of online news in recent years has sparked numerous conversations about its political, social, and economic effects. In the United States alone, 86% of people often or sometimes receive their news online via smartphone, computer, or tablet, and 52% prefer to receive their news through digital platforms (Shearer). However, the rise in popularity of reading news on social media or other online publications has led to an alarming increase in misinformation, with “23% [of Americans] saying they had shared fabricated political stories themselves – sometimes by mistake and sometimes intentionally” (Anderson and Rainie). The danger of online news is that there are so many platforms which make it impossible to fact check every publication. Therefore, if these providers of misinformation can figure out the factors within their articles associated with a rise in shares, it could have very dangerous consequences.

Another impact of the rise in online news includes an increase in online advertising. In 2018, digital advertising revenue reached $111 billion, in 2020 it reached $152 billion, and in coming years it is expected to grow at an even higher rate (“Digital News Fact Sheet”). As so many individuals prefer to receive their news online and this growing audience can have so much influence over impacts such as misinformation and advertising, our research question aims to answer what characteristics of online news lead to greater popularity. Our conclusions may help address issues of misinformation if fake news articles were targeted whose attributes suggest it may reach a wide audience. Further, our conclusions may help online advertisers identify which articles to focus their ads on to reach a larger audience. Overall, because online news reaches such a wide audience, understanding what makes certain articles more popular than others is essential to understanding this rising industry.

Methods

To answer the research question, the dataset “Online News Popularity” from the UC Irvine Machine Learning Repository was utilized (Fernandez). This data set contains 61 characteristics of 39,644 Mashable articles obtained between 2013 and 2015. These characteristics and their descriptions can be found in table 1, but some examples include the type of channel the article was shared in, the day of the week it was published, and the rate of positive words as well as the number of shares which is the dependent variable of interest used as a determinant of popularity. The data set did not contain any missing values, so no data imputation was necessary. The URL variable was removed since given our methods of analysis, it would not provide any useful insights. Other variables were removed that were part of sets of dummy variables to allow us to perform VIF. Prior to running the regressions, we checked for multicollinearity using VIF stepwise variable selection. From this we decided to remove 5 variables - n_unique_tokens, n_non_stop_words, self_reference_avg_sharess, rate_positive_words, and kw_max_min.

Following this, we checked the assumptions of multiple linear regression (linearity, independence, constant variance, and normal distribution) using the diagnostic plots from running a linear regression on all remaining predictors in the data set, as shown in figure 1. Looking at the Residuals vs. Fitted Values plot, we discovered the data did not follow the constant variance and linearity assumptions since the values are not centered around zero and are not evenly distributed across the line. The data also does not follow the normal distribution assumption. We see this from looking at the Normal Q-Q plot. The data should fall across the line; however, we see the tails are skewed away from the line. Because of these violations of the assumptions, it was decided that the best course of action was to log the shares variable as well as removing any outliers shown in the diagnostic plots. It was decided to remove outliers since, in the context of our data set, it did not seem practical to determine characteristics for an article that had extreme numbers of shares since those characteristics behind that are most likely also due to outside factors besides the textual components of the article itself. We then ran the full regression again, which resulted in the diagnostic plots shown in figure 2. We now see the assumptions fully met with the variances more evenly distributed and centered around zero and the observations following the line in the Normal Q-Q plot.

Multiple linear regression was deemed the best method of analysis since this can provide the variables that are associated with the greatest percentage point increase in the number of shares. A regression tree will also be utilized to support any results found by the multiple linear regression. Given that there are so many predictor variables, PCA was initially considered because it could help to reduce dimensionality; however, the goal of this study was to interpret the output, which would not have been as easily done if PCA were involved. Interpretability was extremely important given that these results could actually be applied by companies to increase engagement with articles. To find the regression model that best fits the data, AIC and BIC stepwise regressions were performed. The number of variables, R-squared, adjusted R-squared, and CV-scores were then compared to determine which model better fit the data. Once the best linear regression model was determined, diagnostic plots were once again utilized to remove outliers. A regression tree was then created and pruned to determine if any of the significant characteristics found in the regression model are especially important in their association with increasing the number of shares.

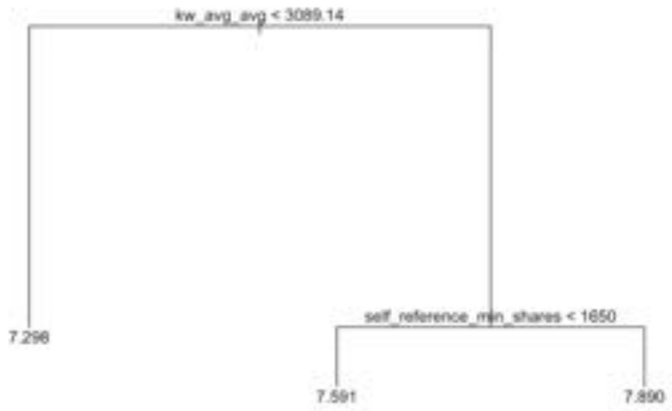
Results

The AIC stepwise regression resulted in a model consisting of 38 variables, an R-squared value of 0.1233, an adjusted R-squared score of 0.1255, and a CV-score of 0.780, as shown in table 3. The BIC stepwise regression resulted in a model consisting of 29 variables, an R-squared value of 0.1226, an adjusted R-squared score of 0.1219, and a CV-score of 0.759. Given that these measures of fit are extremely similar between the two models and that the BIC model contains far fewer predictors, it was decided that the BIC model was more parsimonious and the better fit. This is primarily due to the number of predictors - by using fewer variables, it is less likely that the model will overfit the data.

Our results are primarily focused around the BIC multiple linear regression model and what this model suggests about the most popular attributes of online news. All of the predictors for our final model are significant at the 99.9% significance level. These predictors and their associated values are listed in table 2. The model output demonstrates that the attributes of popular articles include day of the week published, channel published in, global and title subjectivity, and polarity of content. In particular, our results show that using Monday as a baseline, publishing on Saturday and Sunday will lead to a roughly 23% increase in shares while publishing on Tuesday leads to a 7% decrease. Further, with the Business channel as the baseline, the Social Media channel is most popular leading to a 29% increase in engagement. On the other hand, the Entertainment channel is least engaging, and is associated with a 18% decrease. By far the most important attribute was global subjectivity, which led to a 53% increase. Title subjectivity was also relatively significant with a 15% increase in shares. Finally, the minimum polarity of positive words is associated with a 26% decrease in engagement.

A regression tree was also created to confirm results from the multiple linear regression or suggest other possible important predictors for online news' success. After pruning the tree, the ideal tree size included 3 nodes, shown in figure 3. The splits in the tree included the average number of shares of the keywords prior to the date of publication and the minimum number of shares of the referenced articles within Mashable. For the keywords, if the average number of shares was less than roughly 3,000, then the article would expect to receive about 1,400 ($e^{7.298}$) shares. If the article's keywords had more than 3,000 shares and the article's referenced articles had less than 1,650 shares, then the article would expect roughly 1,980 ($e^{7.591}$) shares. If the article's referenced articles had more than 1,650 shares, then the article would expect roughly 2,670 ($e^{7.890}$) shares. This indicates that keywords and the popularity of referenced articles are important factors in the popularity of other articles. Both of these variables are also included in the multiple linear regression model which adds further support to their usefulness as predictors of article popularity, especially given that the R-squared values of the two models are not extremely different (0.1226 vs. 0.1175).

Figure 3. Pruned regression tree



Discussion and Conclusions

Overall, the significant attributes as identified by our multiple regression analysis are relatively intuitive. Our finding that weekend publications lead to increased engagement makes sense, as readers have more time on weekends to set aside for reading and sharing online news. Additionally, the importance of global subjectivity suggests that readers engage more often with sensationalized and opinionated news. An interesting implication from our analysis is that articles with greater minimum positive polarity, or more negative content, reach a wider audience. This suggests that negative news is more popular than positive news, which follows the negatively skewed content of news channels and sites. Yet, the most important attributes of an article might be those that were included in both the regression tree and the multiple linear regression - the popularity of an article's keywords and any of its referenced articles. This makes sense given that if keywords and other articles are being heavily shared, then the topic of the article in question must also be very relevant.

The conclusions of our analysis have several limitations, stemming largely from the specificity of our data set. This can be seen by the very low R-squared scores of both the linear regression and regression tree models. The attributes included in this data set might be too specific to the actual content of the article and might not pick up on other societal trends. Additionally, Because our data set is based on Mashables data, our results cannot be generalized across articles from other social media sites. Further, the Mashables platform is primarily published in English, thus the results are more representative of an English speaking audience. For future research, it may be interesting to focus on multiple social media platforms across different languages to create more generalizable results.

References

Anderson, J., & Rainie, L. (2017, October 19). The Future of Truth and Misinformation Online. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/

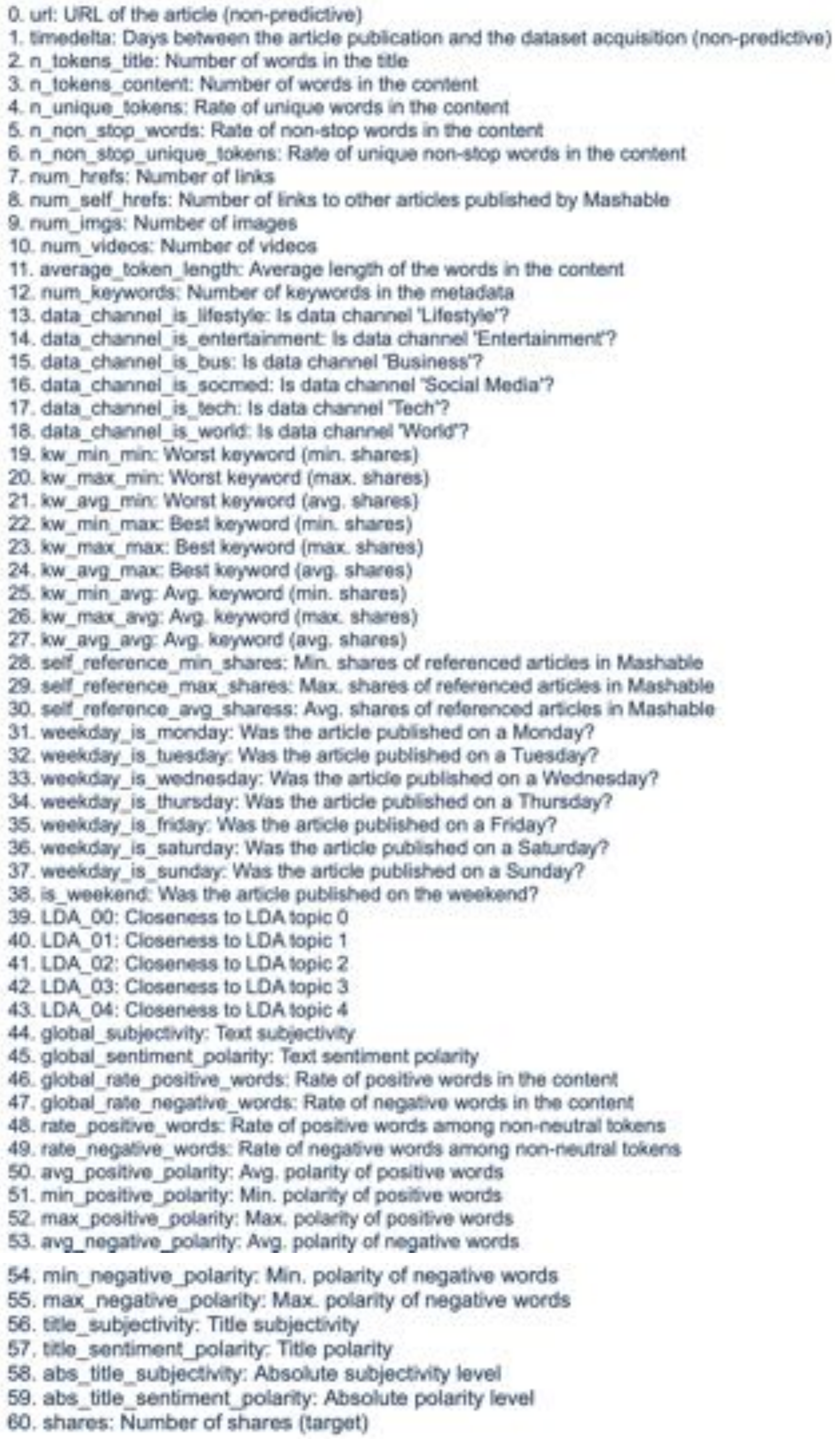
Digital News Fact Sheet. (n.d.). *Pew Research Center's Journalism Project*. Retrieved May 15, 2022, from https://www.pewresearch.org/journalism/fact-sheet/digital-news/

Fernandes, K. (n.d.). *UCI Machine Learning Repository: Online News Popularity Data Set*. Retrieved May 12, 2022, from https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity#

Shearer, E. (n.d.). More than eight-in-ten Americans get news from digital devices. *Pew Research Center*. Retrieved May 15, 2022, from https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/

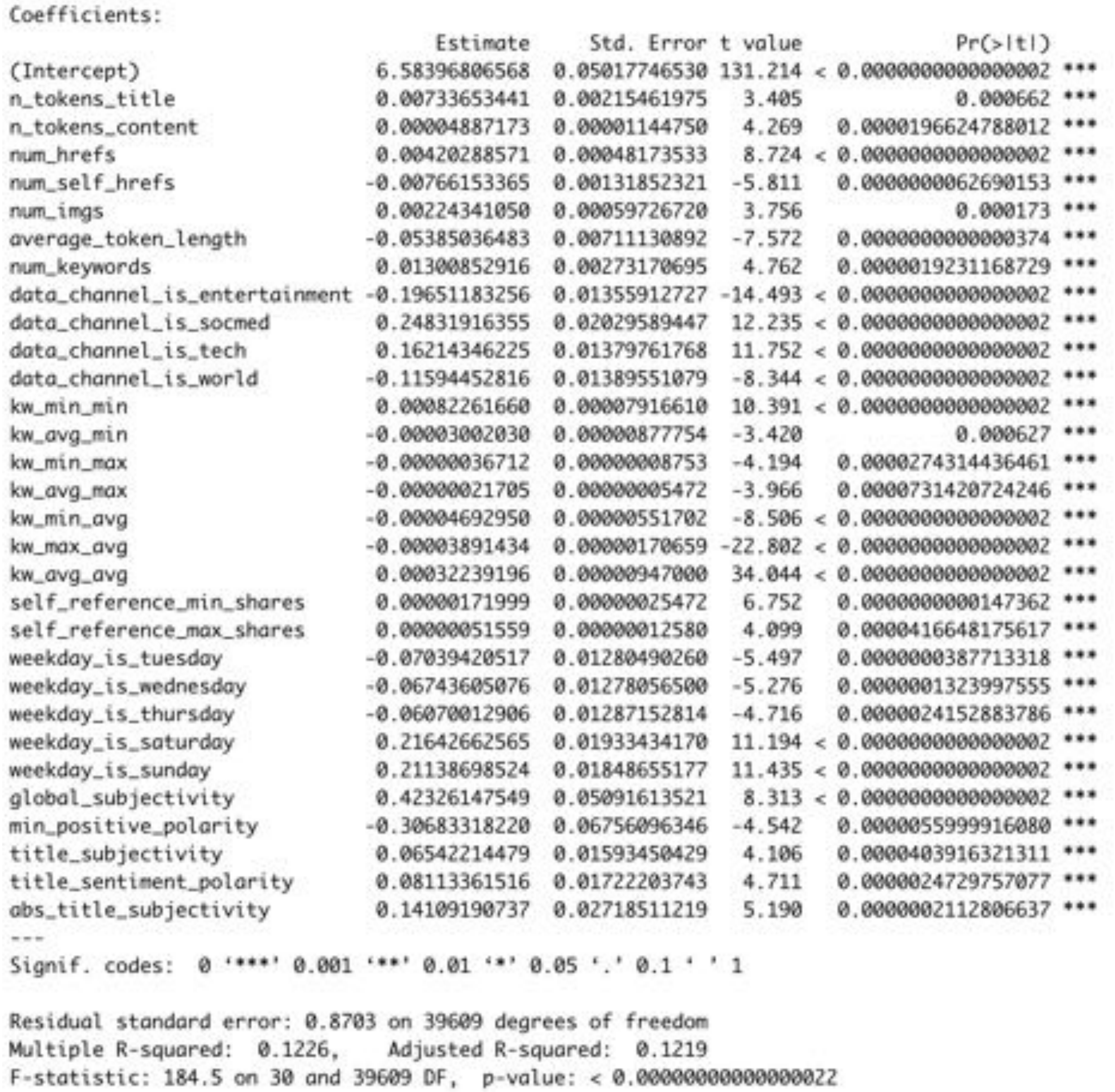
Appendices

Table 1. All predictors included in original data set



0. url: URL of the article (non-predictive)
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
2. n_tokens_title: Number of words in the title
3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
14. data_channel_is_entertainment: Is data channel 'Entertainment'?
15. data_channel_is_bus: Is data channel 'Business'?
16. data_channel_is_socmed: Is data channel 'Social Media'?
17. data_channel_is_tech: Is data channel 'Tech'?
18. data_channel_is_world: Is data channel 'World'?
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)
21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)
23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
25. kw_min_avg: Avg. keyword (min. shares)
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: Avg. keyword (avg. shares)
28. self_reference_min_shares: Min. shares of referenced articles in Mashable
29. self_reference_max_shares: Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
31. weekday_is_monday: Was the article published on a Monday?
32. weekday_is_tuesday: Was the article published on a Tuesday?
33. weekday_is_wednesday: Was the article published on a Wednesday?
34. weekday_is_thursday: Was the article published on a Thursday?
35. weekday_is_friday: Was the article published on a Friday?
36. weekday_is_saturday: Was the article published on a Saturday?
37. weekday_is_sunday: Was the article published on a Sunday?
38. is_weekend: Was the article published on the weekend?
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2
42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity
46. global_rate_positive_words: Rate of positive words in the content
47. global_rate_negative_words: Rate of negative words in the content
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words
51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: Number of shares (target)

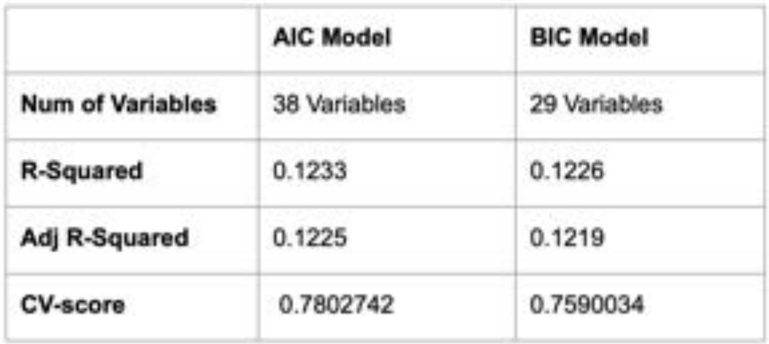
Table 2. Output of final regression model



```
Coefficients:
                                    Estimate    Std. Error t value             Pr(>|t|)    
(Intercept)                    6.58396806568 0.05017746530 131.214 < 0.0000000000000002 ***
n_tokens_title                 0.00733653441 0.00215461975   3.405             0.000662 ***
n_tokens_content               0.00004887173 0.00001144750   4.269   0.0000196624788012 ***
num_hrefs                      0.00420288571 0.00048173533   8.724 < 0.0000000000000002 ***
num_self_hrefs                -0.00766153365 0.00131852321  -5.811   0.0000000062690153 ***
num_imgs                       0.00224341050 0.00059726720   3.756             0.000173 ***
average_token_length          -0.05385036483 0.00711130892  -7.572   0.0000000000000374 ***
num_keywords                   0.01300852916 0.00273170695   4.762   0.0000019231168729 ***
data_channel_is_entertainment -0.19651183256 0.01355912727 -14.493 < 0.0000000000000002 ***
data_channel_is_socmed         0.24831916355 0.02029589447  12.235 < 0.0000000000000002 ***
data_channel_is_tech           0.16214346225 0.01379761768  11.752 < 0.0000000000000002 ***
data_channel_is_world         -0.11594452816 0.01389551079  -8.344 < 0.0000000000000002 ***
kw_min_min                     0.00082261660 0.00007916610  10.391 < 0.0000000000000002 ***
kw_avg_min                    -0.00003002030 0.00000877754  -3.420             0.000627 ***
kw_min_max                    -0.00000036712 0.00000008753  -4.194   0.0000274314436461 ***
kw_avg_max                    -0.00000021705 0.00000005472  -3.966   0.0000731420724246 ***
kw_min_avg                    -0.00004692950 0.00000551702  -8.506 < 0.0000000000000002 ***
kw_max_avg                    -0.00003891434 0.00000170659 -22.802 < 0.0000000000000002 ***
kw_avg_avg                     0.00032239196 0.00000947000  34.044 < 0.0000000000000002 ***
self_reference_min_shares      0.00000171999 0.00000025472   6.752   0.0000000000147362 ***
self_reference_max_shares      0.00000051559 0.00000012580   4.099   0.0000416648175617 ***
weekday_is_tuesday            -0.07039420517 0.01280490260  -5.497   0.0000000387713318 ***
weekday_is_wednesday          -0.06743605076 0.01278056500  -5.276   0.0000001323997555 ***
weekday_is_thursday           -0.06070012906 0.01287152814  -4.716   0.0000024152883786 ***
weekday_is_saturday            0.21642662565 0.01933434170  11.194 < 0.0000000000000002 ***
weekday_is_sunday              0.21138698524 0.01848655177  11.435 < 0.0000000000000002 ***
global_subjectivity            0.42326147549 0.05091613521   8.313 < 0.0000000000000002 ***
min_positive_polarity         -0.30683318220 0.06756096346  -4.542   0.0000055999916080 ***
title_subjectivity             0.06542214479 0.01593450429   4.106   0.0000403916321311 ***
title_sentiment_polarity       0.08113361516 0.01722203743   4.711   0.0000024729757077 ***
abs_title_subjectivity         0.14109190737 0.02718511219   5.190   0.0000002112806637 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8703 on 39609 degrees of freedom
Multiple R-squared:  0.1226,	Adjusted R-squared:  0.1219 
F-statistic: 184.5 on 30 and 39609 DF,  p-value: < 0.00000000000000022
```

Table 3. Comparison of evaluation metrics between AIC and BIC models



	AIC Model	**BIC Model**
Num of Variables	38 Variables	29 Variables
R-Squared	0.1233	0.1226
Adj R-Squared	0.1225	0.1219
CV-score	0.7802742	0.7590034

Figure 1. Diagnostic plots of original data

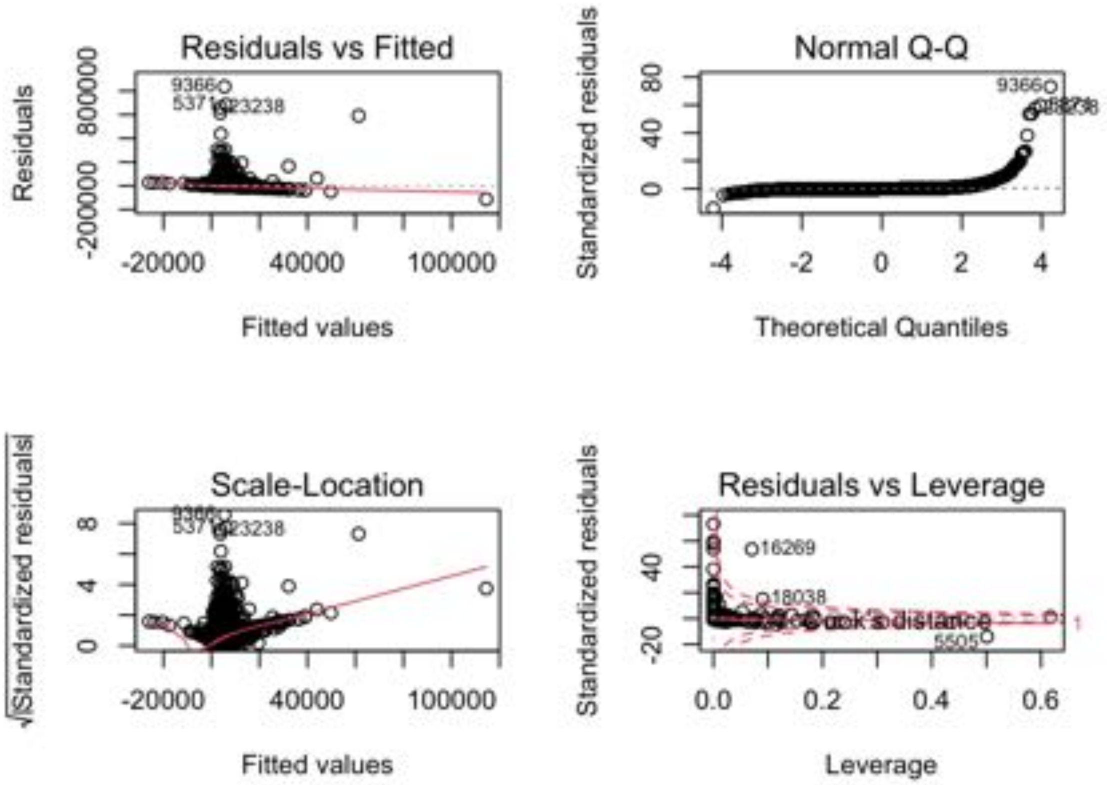
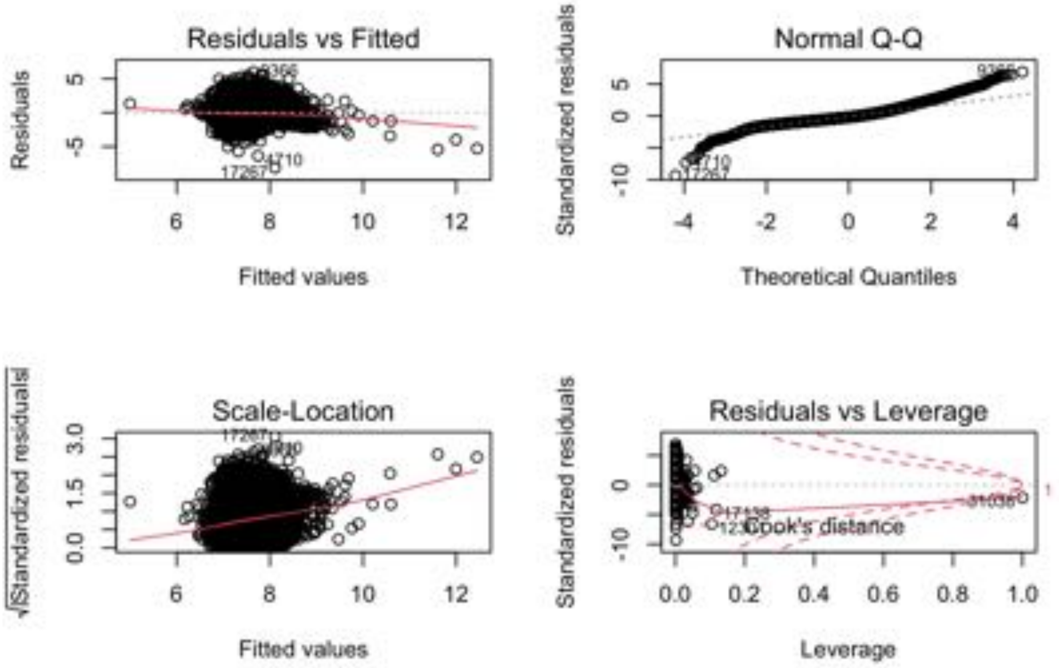


Figure 2. Diagnostic plots of data after logging shares variable


Residuals vs Fitted
Residuals
5
-5
9366
4710
17267
6
8
10
12
Fitted values
Normal Q-Q
Standardized residuals
5
0
-10
9366
4710
17267
-4
-2
0
2
4
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
3.0
1.5
0.0
17267
4710
6
8
10
12
Fitted values
Residuals vs Leverage
Standardized residuals
0
-10
17138
Cook's distance
1
0.0
0.2
0.4
0.6
0.8
1.0
Leverage