Caption Analysis for PSI

Prepared by:
Huixin Yang
Shiqi Wang
Ningxin Zhang

Dec 13, 2022

Contents

Background

In this digital age, the impact of social media is beyond anyone's imagination. According to Statista, the average daily social media usage of users worldwide has incremented to 147 minutes in 2022. As a result, companies would seek every possibility to promote their brand image and expand their network influence on different social media platforms. Our client Partnerships for Strategic Impact (PSI) is a non-profit consulting firm that aims to support small to medium sized nonprofits to develop their impact stories. The goal of our project is to help our client investigate what characteristics of a caption text could increase the engagement of their social media posts, so that PSI could refer to our findings when they attempt to post more engaged texts and attract more audience either for themselves or on their clients' behalf.

1 Research Question

The question we seek to answer is: As for companies similar to PSI, with the control of number of followers, what factors of caption have the strongest association with the engagement of their social media posts? If strong correlations exist, are they positive or negative? For the purpose of this analysis, we evaluate the popularity of a social media post by its number of likes and comments. We would like to find out what kind of caption of a social media post will potentially attract more likes and comments, and are there certain tones or features that will perform the best, specifically for social media accounts that are similar with PSI's account.

2 Data

2.1 Data Collection

In order to obtain relatively up to date and complete data, the data was scraped from three social media platforms of interest. For each of the platforms, we used an open sourced package from GitHub. Data were derived from the most recent post from PSI designated companies till October 19th, 2022. For LinkedIn, there are 149 observations from 5 Companies; for Facebook, there are 150 observations from 2 companies; for Twitter, there are 105 observations from 8 companies. Raw data collected from target posts include the following information: Name of the company, Text of the caption, Likes, Comments, and Followers.

2.2 Data Cleaning

From the raw data, cleaning process started from three aspects that we believe may influence the audience response on social media. Firstly, sentiment analysis is conducted to compute sentiment of the caption since audience may have a preference to respond to more positive or more negative contend. Secondly, special characters that include question mark, hashtag, and emoji are counted for each caption. Questions in text may provoke audience to answer as response, while hashtags can increase the exposure of the post. Emoji is a trending internet language that would influence the sentimental effect.Thirdly, we counted the word count of the post to see whether audience have a preference over longer or shorter post; we also calculated the average length of the words used in sentence after removing all the common stop words to investigate whether audience prefers easier or more difficult words. The response variable is response, computed by adding the number of likes and comments that a post received.

2.3 Data

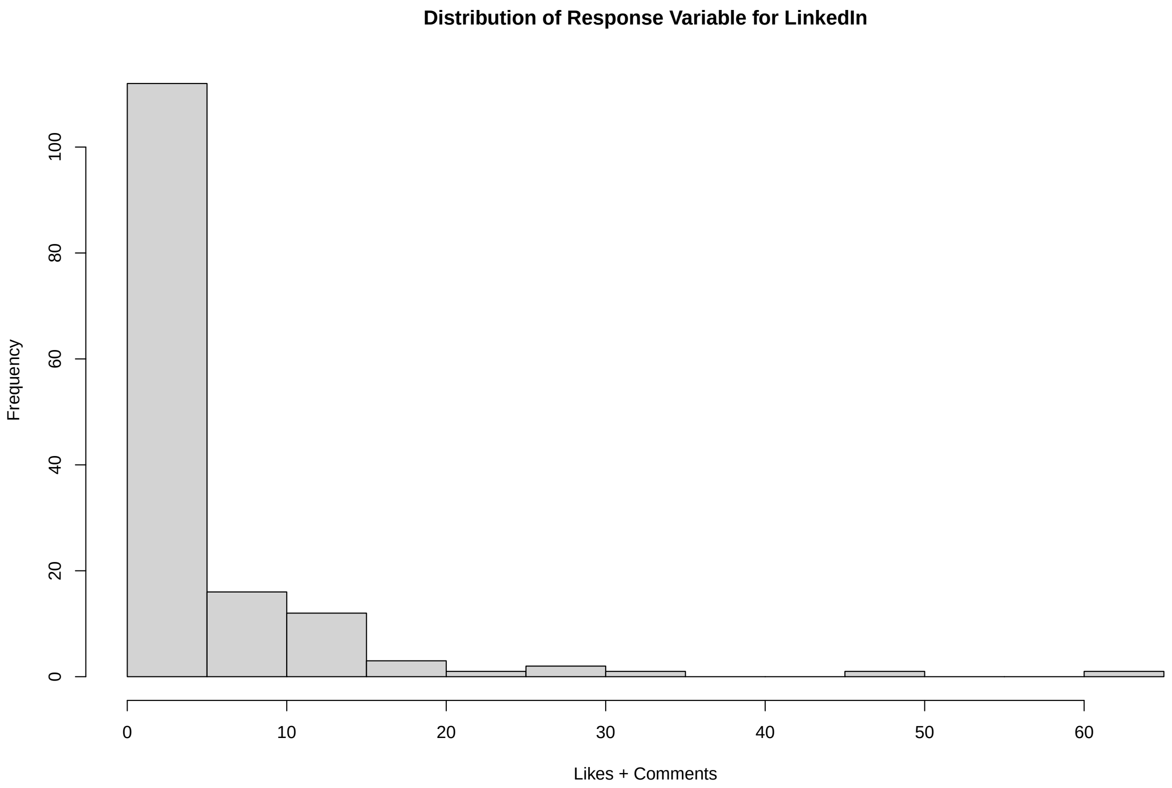
Interested predictors and response variable:

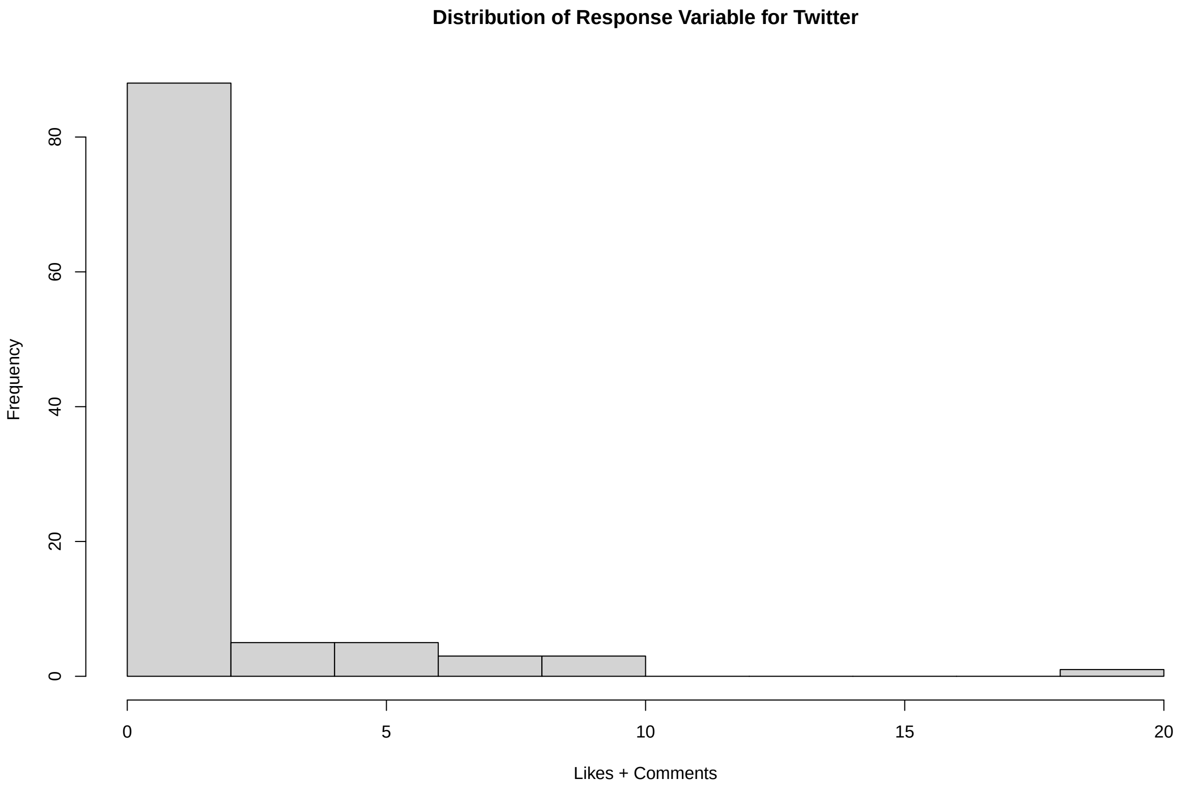
Predictors	Description
SentimentGI	The sentiment of the post text; Quantitative; Ranging from -1 to 1
Hashtag	The number of hashtags included in the post; Quantitative
Question mark	The number of question marks included in the post; Quantitative
Emoji	The number of emojis included in the post; Quantitative
Average length	The average length (number of letters) of the words in the post; Quantitative
Word count	The number of words in the post; Quantitative
Followers*	The number of followers that the account has; Control variable, not of interest
Response	Response variable. The responses that a post receives; Comments + Likes; Quantitative

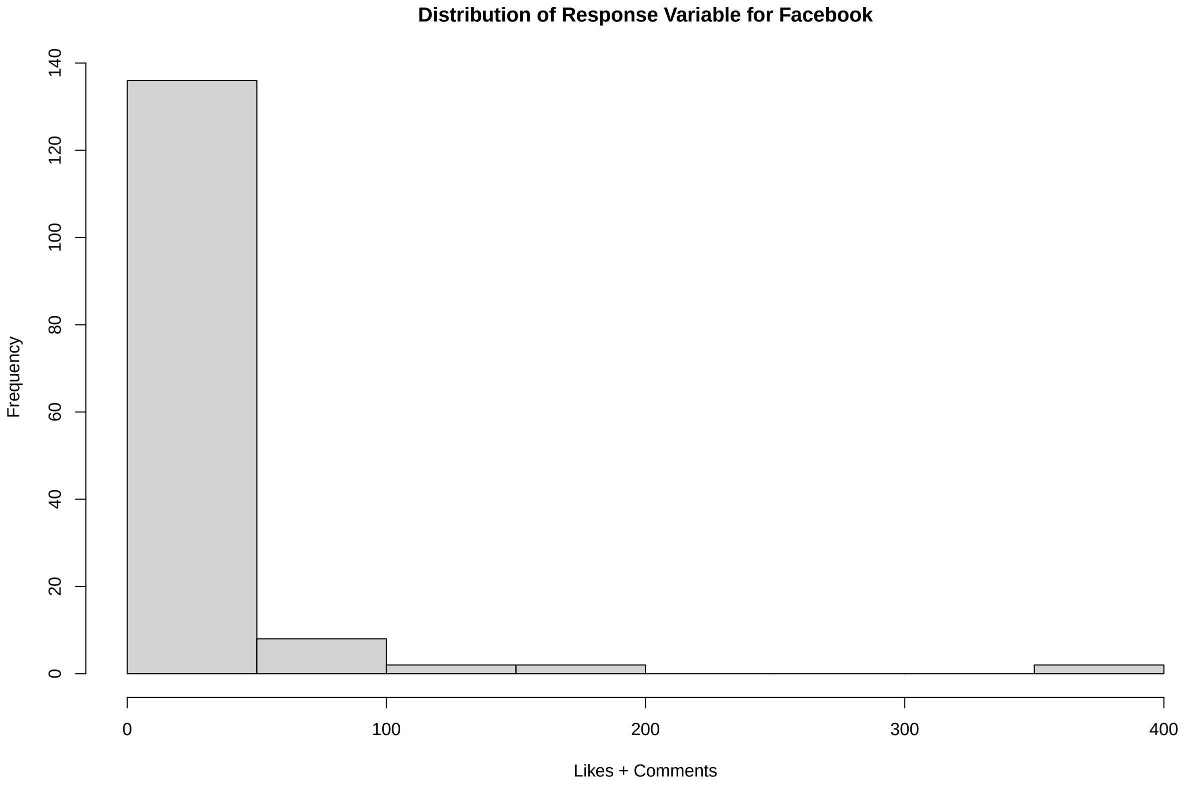
3 Exploratory Data Analysis

3.1 Log Transformation of the Response Variable

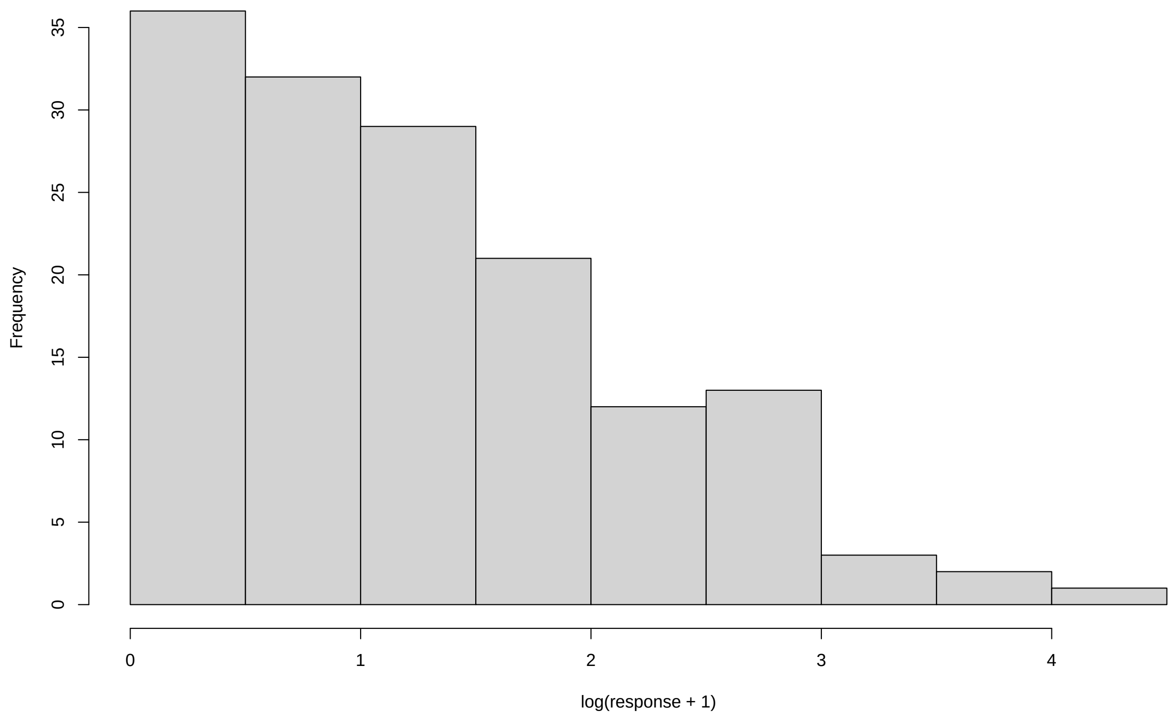
While we were collecting the data, we noticed that there were quite a few zero values. Therefore, we plotted histograms to check the distribution of our response variable (Likes + Comments) for all three platforms. Due to the small-scale nature of the companies we focused on, these accounts had limited likes and comments on their posts, which leads to the heavily right-skewed distribution of our response variables. We realized that we had to correct the distribution by performing a log transformation. However, because of the presence of zero values, log(0) would give us an error in R. So, we came up with a new method, which was log(y+1) that allowed us to perform log transformation on zero values. We also conducted research to validate our approach and found a blog that also acknowledged this method. "log(x+1) transformation is often used for transforming data that are right-skewed, but also include zero values." In addition, we computed the value of log(y) and log(y+1), making sure that these two values do not differ a lot.

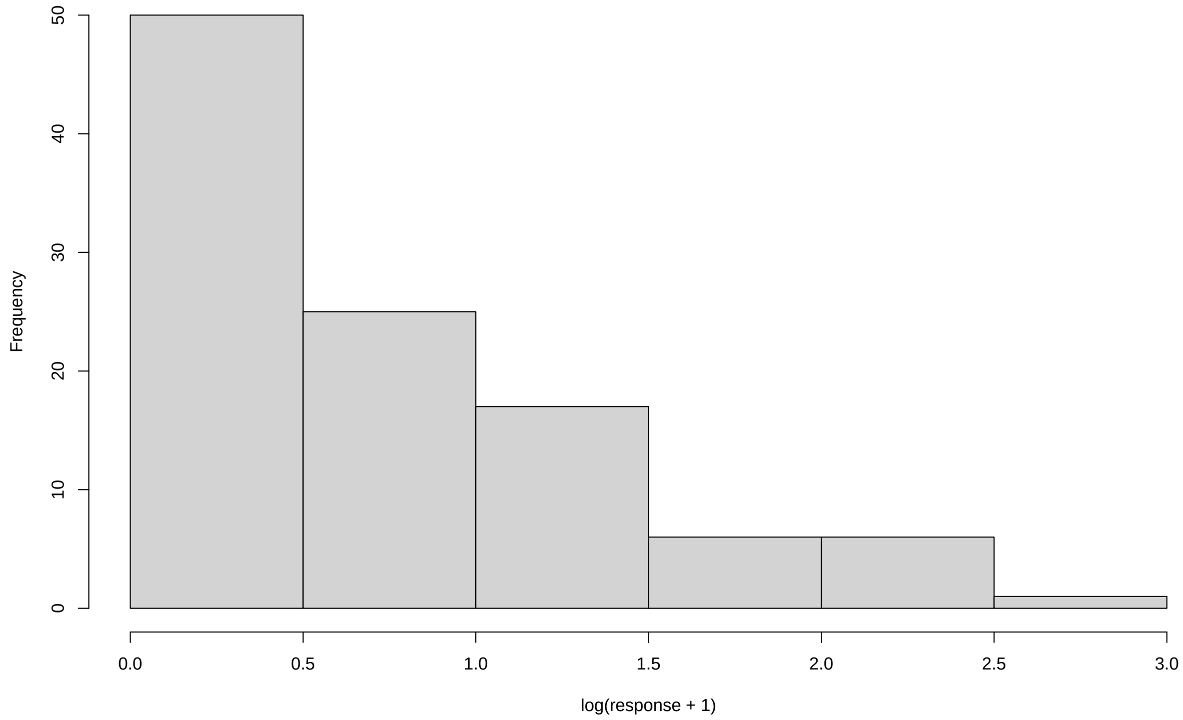

Distribution of Response Variable for LinkedIn

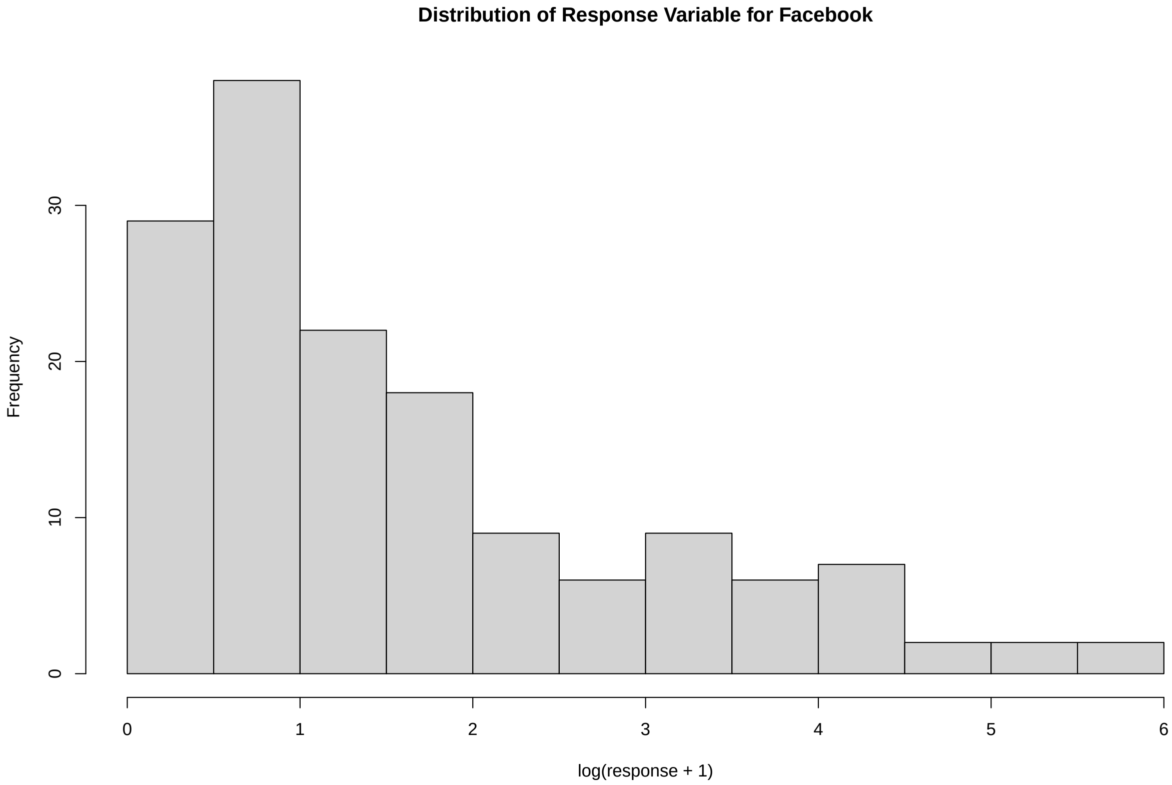

Distribution of Response Variable for Twitter



After we have conducted the log(y+1) transformation on the response variable for all three platforms, we compared the distribution of the values again using histograms below. It is quite obvious that the distribution for each platform has become a lot smoother and significantly less skewed. So, we believed that the log(y+1) transformation was effective and had alleviated our problem to a certain extent. But due to the zero values in our dataset, it is impossible to transform the response variable to a normal distribution.


Distribution of Response Variable for LinkedIn


Distribution of Response Variable for Twitter

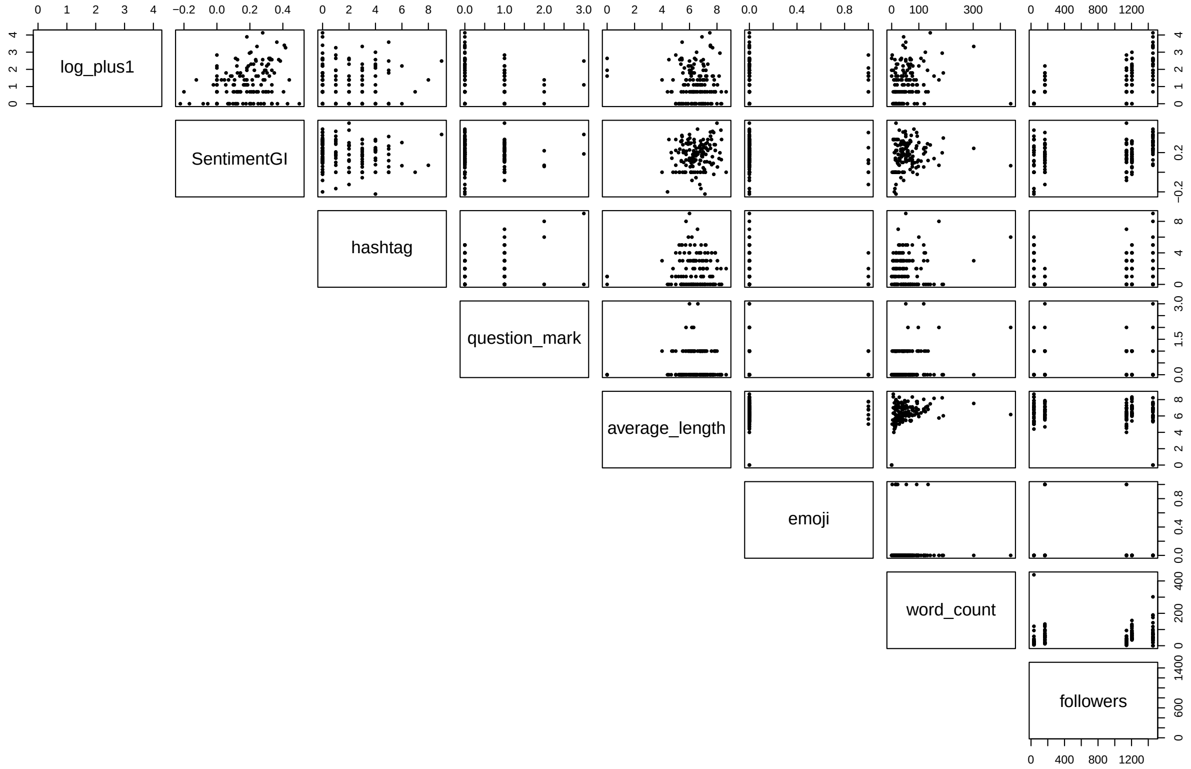


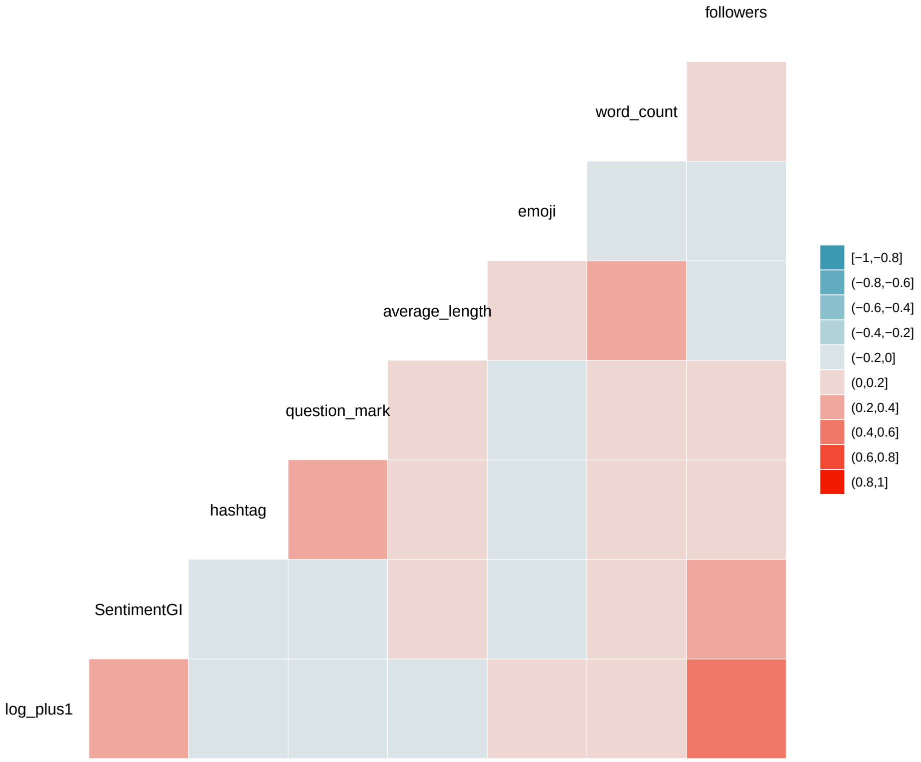
3.2 Exploratory Data Analysis

In order to explore the relationship between our variables of interest, we chose to perform scatterplot matrices and correlation heatmaps. Scatterplot matrices allow us to scrutinize the distribution between two variables, checking for multicollinearity and the trends between the response and predictors. On the other hand, correlation heatmaps can visually showcase the association between two variables, which is more straightforward and easier to read.

3.2.1 LinkedIn

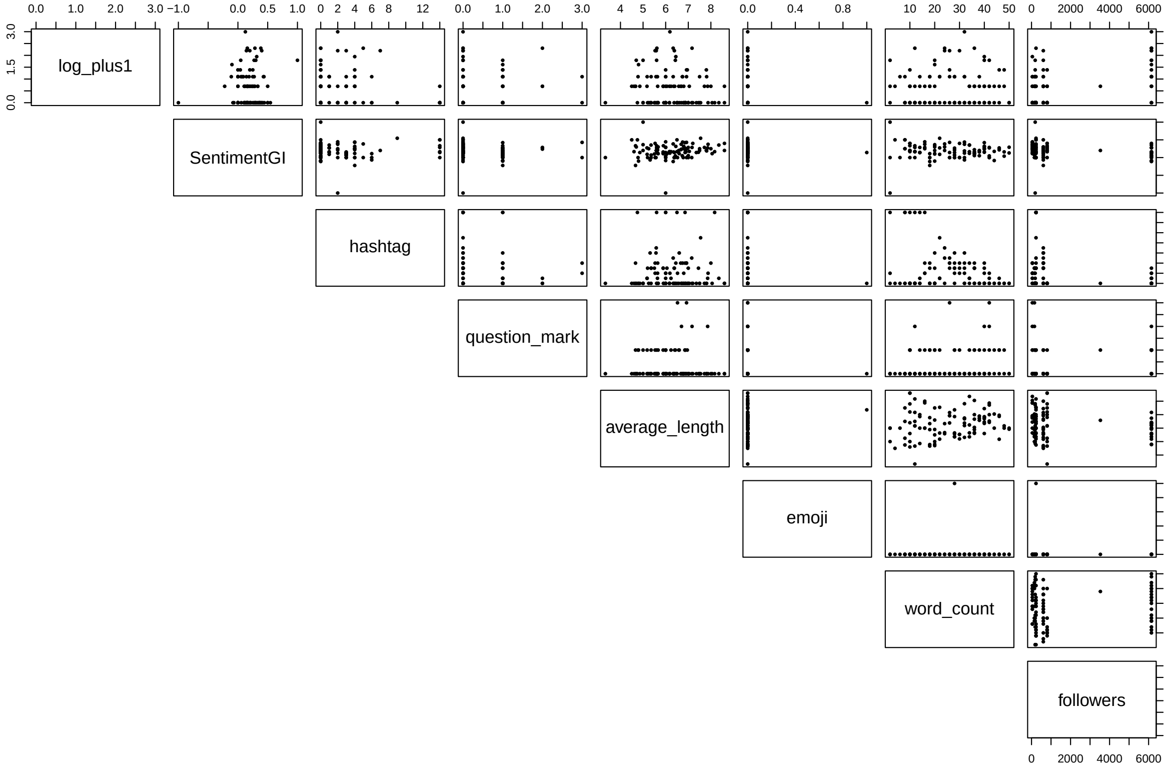
In the scatterplot matrix, we noticed that companies rarely used emojis in their LinkedIn posts. Even when they do include emojis in their captions, they would just use one. Moreover, their captions were mostly between 0 to 200 words. Some captions' sentiment scores were between -0.2 to 0, which indicated a negative tone of the caption. In the correlation matrix, we noticed that there is a relatively strong correlation between followers and the response variable (between 0.4 to 0.6). Sentiment score and word_count also has a slightly positive correlation with the response variable. No multicollinearity issue was detected from the two graphs.

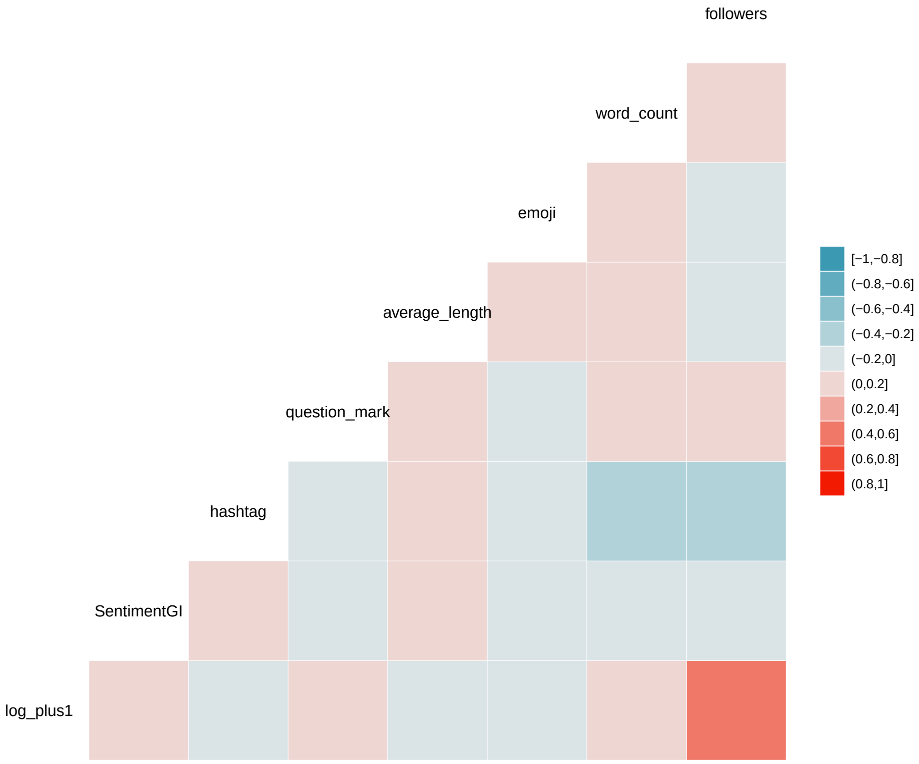




3.2.2 Twitter

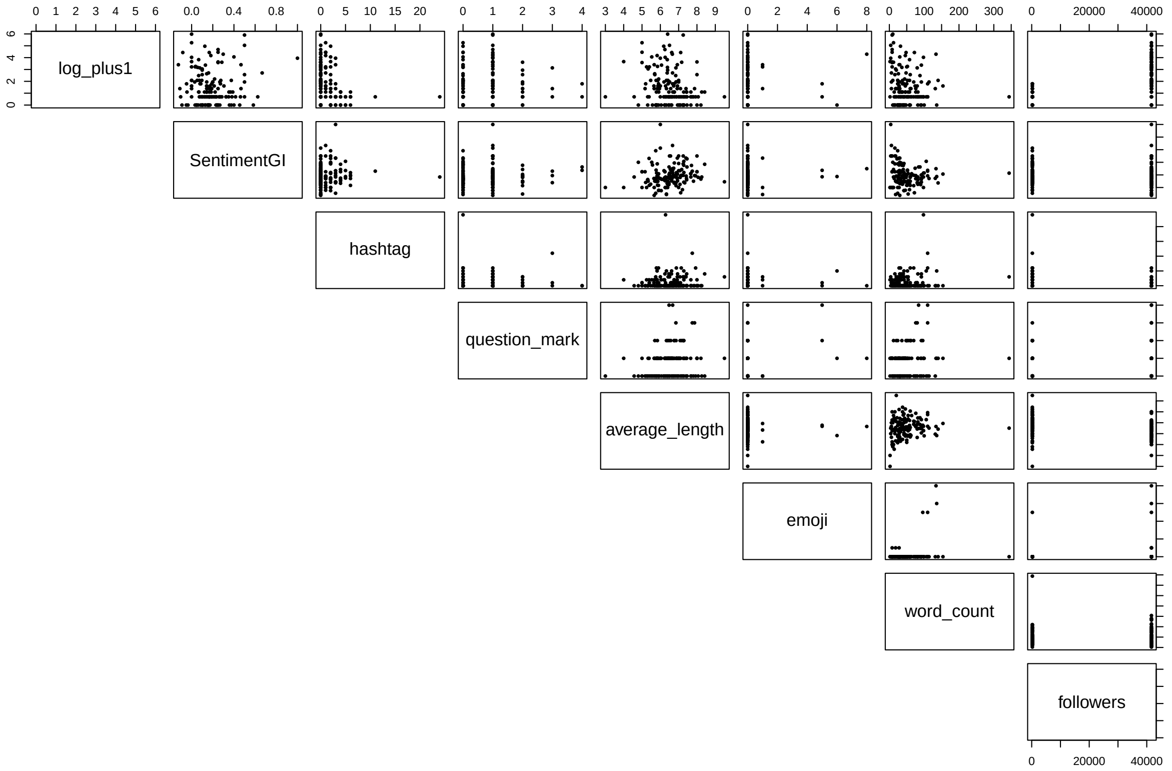
In the scatterplot matrix, we noticed that companies almost never used emojis in their Twitter posts. Moreover, their captions were mostly between 10 to 40 words, which were significantly less than LinkedIn posts. There seemed to be one notable outlier that had a sentiment score of -1.0, indicating a strong negative polarity of the post. In the correlation matrix, we noticed that there is a relatively strong correlation between followers and the response variable (between 0.4 to 0.6), same as LinkedIn. There were no more notable correlations between two variables. No multicollinearity issue was detected from the two graphs.

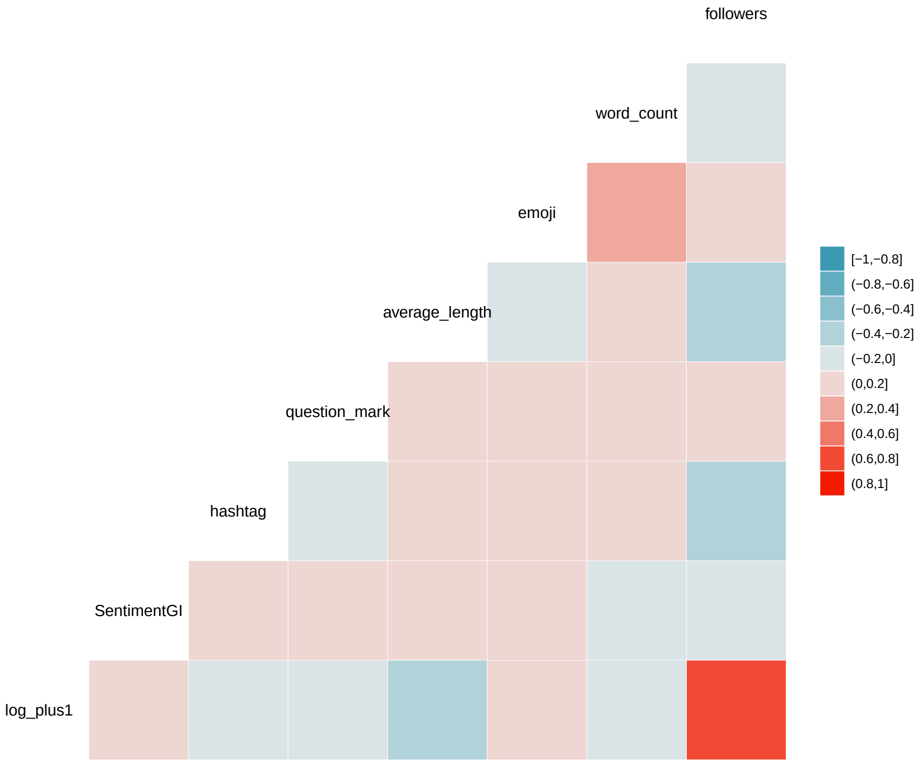




3.2.3 Facebook

In the scatterplot matrix, we noticed that companies rarely used emojis in their Facebook posts, but there were some outliers that used more than 5 emojis in a single post (the maximum number of emojis used was 8). Moreover, their captions were mostly between 0 to 200 words, which were similar to LinkedIn posts. For Facebook posts, the majority of the captions had a positive polarity score and there seemed to be one outlier that had a sentiment score of 1.0. In the correlation matrix, we noticed that there is a strong correlation between followers and the response variable (between 0.6 to 0.8). Average_length had a slightly negative correlation with the response variable. No multicollinearity issue was detected from the two graphs.


log_plus1
SentimentGI
hashtag
question_mark
average_length
emoji
word_count
followers
0 1 2 3 4 5 6
0.0 0.4 0.8
0 5 10 15 20
0 1 2 3 4
3 4 5 6 7 8 9
0 2 4 6 8
0 100 200 300
0 20000 40000


followers
word_count
emoji
average_length
question_mark
hashtag
SentimentGI
log_plus1
[−1,−0.8]
(−0.8,−0.6]
(−0.6,−0.4]
(−0.4,−0.2]
(−0.2,0]
(0,0.2]
(0.2,0.4]
(0.4,0.6]
(0.6,0.8]
(0.8,1]

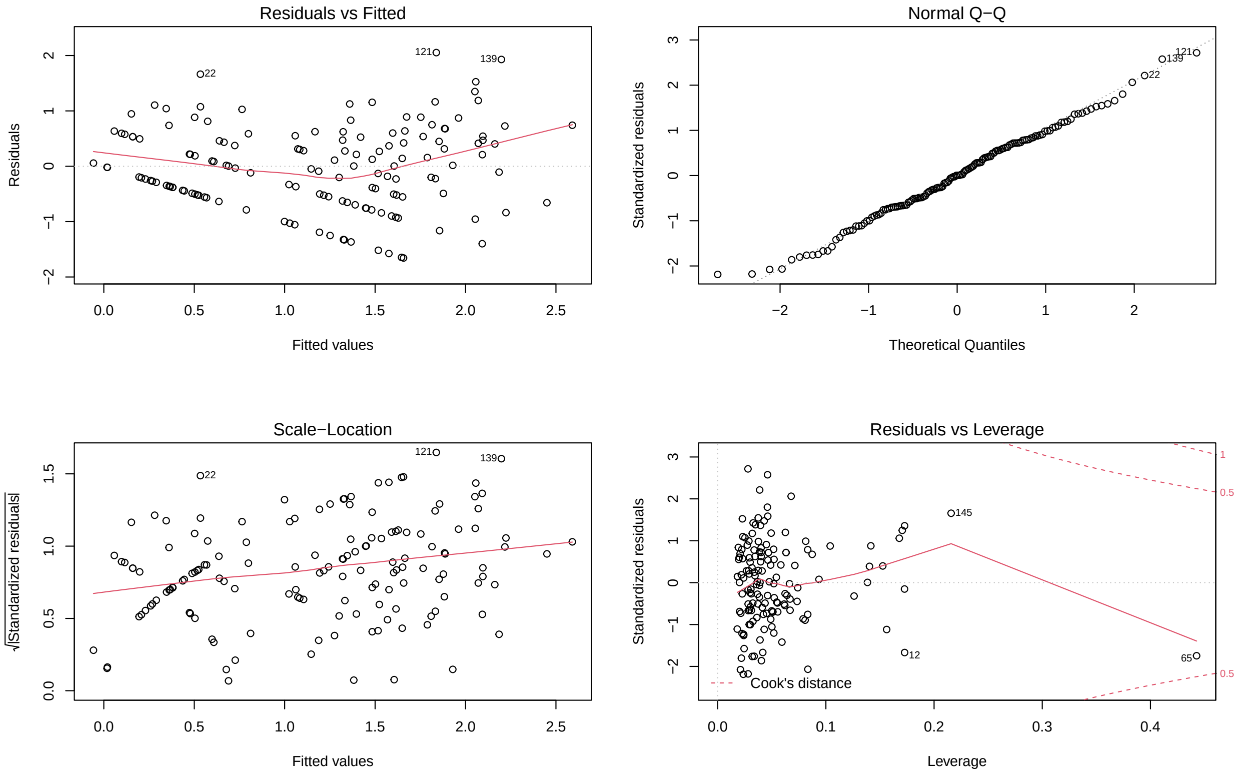
4 Models

Due to the natural differences of user volume, usage scenarios, platform tonality and advertising delivery among Linkedin, Twitter and Facebook, it is unconvincing to come up with one general result for the three platforms. Thus, for each kind of model below, we made three models for each platform separately, and interpreted the result separately.

4.1 LinkedIn

4.1.1 Linear Regression Model

We start with building a linear regression model for the numbers of likes and comments, with each social media post representing one observation. In addition to these key variables of interest we mentioned above, we added the number of followers in the model as control variables. Before interpreting the model, we first check if all the assumptions for linear regression were met. Linear regression models assume that the errors are independent and identically distributed. Breaking down the assumptions, for each fixed value, we want the error to have mean 0, constant variance, and normal distribution. In the Residual vs. Fitted plots and Scale-Location plots for each platform, we identified moderate curvature for the gray line along the x-axis, meaning we did not fully meet the assumption that the mean of error should be 0, even though the residuals seem to have decent spread in the plot. Fortunately, in the Normal Q-Q plots, the standardized residuals fall along 45 degrees, which represents that the residuals are normally distributed. In the Residuals vs. Leverage plots, all of our data points fall inside of Cook's Distance, thus none of our observations are influential and should be dropped.



After checking the assumption, we performed a T-test for each predictor and F-test for the whole model, and interpreted the finding for Linkedin. The whole model has F-statistic 14.8 ($p < .001$), meaning that our model is reasonable and useful. We identified the significant variables of interest as SentimentGI (p = .0343), Question_mark (p = .040), Emoji (p = .002) and Word_count (p = .024). The adjusted R-squared we got is 0.3998.

```
##
## Call:
## lm(formula = log(likes + comments + 1) ~ SentimentGI + hashtag +
##     question_mark + average_length + emoji + followers + word_count,
##     data = linkedin)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65726 -0.51372  0.00378  0.53597  2.05347
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.3338731  0.4783368   0.698  0.48636
## SentimentGI     1.0437475  0.4881051   2.138  0.03425 *
## hashtag        -0.0072577  0.0344716  -0.211  0.83356
## question_mark  -0.2247075  0.1084438  -2.072  0.04012 *
## average_length -0.0293008  0.0719516  -0.407  0.68447
## emoji           0.9098001  0.2852374   3.190  0.00176 **
## followers       0.0009567  0.0001154   8.293 8.84e-14 ***
## word_count      0.0028220  0.0012367   2.282  0.02402 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7667 on 138 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.4288, Adjusted R-squared:  0.3998
## F-statistic:  14.8 on 7 and 138 DF,  p-value: 2.638e-14
```

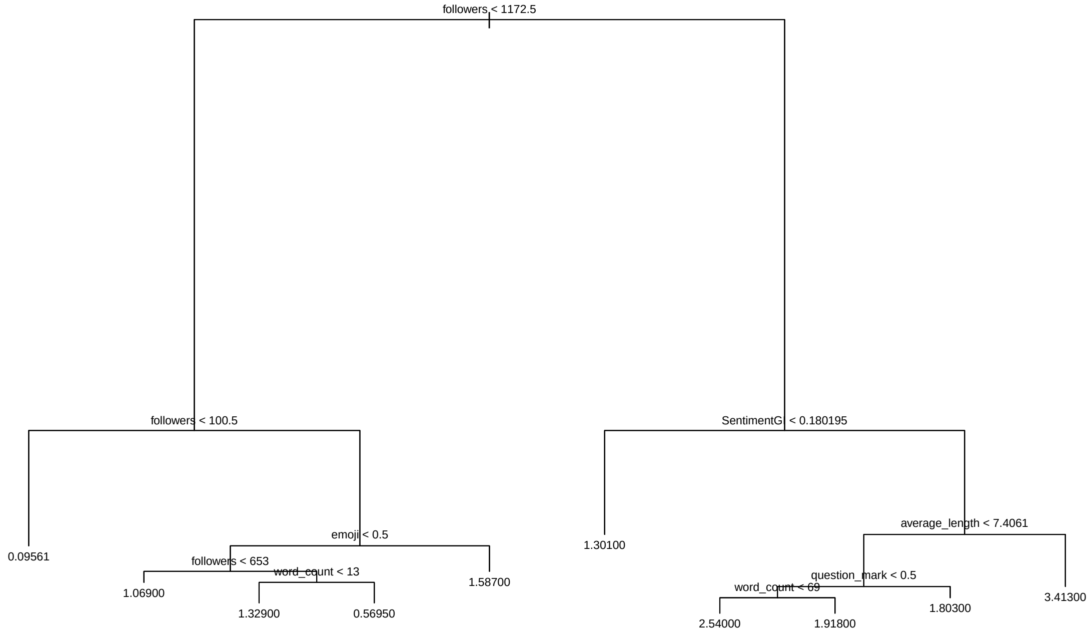
To improve our model, we dropped some insignificant predictors with relatively high p-value, and we are left with the reduced model below.

```
##
## Call:
## lm(formula = log(likes + comments + 1) ~ SentimentGI + question_mark +
##     emoji + word_count + followers, data = linkedin)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6302 -0.5241 -0.0087  0.5201  2.0494
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.1393971  0.1423708   0.979  0.32921
## SentimentGI    1.0311503  0.4754367   2.169  0.03178 *
## question_mark -0.2270002  0.1038036  -2.187  0.03042 *
## emoji          0.9101241  0.2816855   3.231  0.00154 **
## word_count     0.0027064  0.0012012   2.253  0.02580 *
## followers      0.0009591  0.0001136   8.443 3.52e-14 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7618 on 140 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.428,  Adjusted R-squared:  0.4075
## F-statistic: 20.95 on 5 and 140 DF,  p-value: 1.362e-15
```

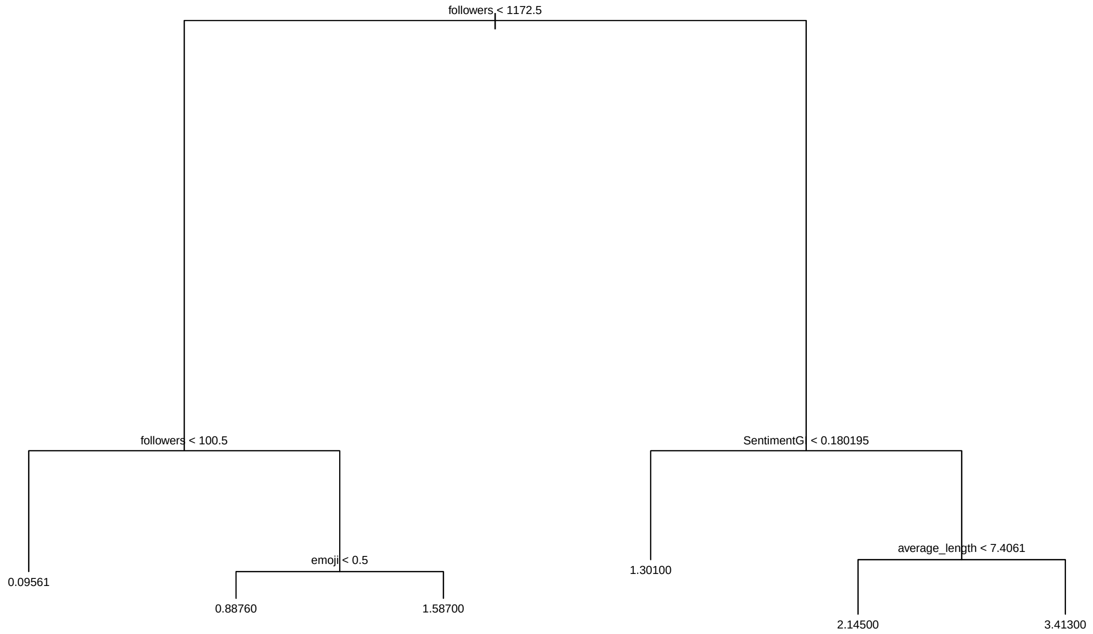
The likelihood ratio test gave us a p-value 0.897, thus we failed to reject this reduced model. Our improved model obtained an adjusted R-square 0.4075 (originally 0.3998), which is aligned to the result of the likelihood ratio test. From this model, we came to the conclusion that a post that has a caption with positive sentiment, less questions, more emojis, more word counts will potentially receive better engagement.

4.1.2 Regression Tree

A recursive binary tree model was built to analyze more specifically the significance and influence of important factors. The initial model has 10 terminal nodes, and predictors that are used include followers, emoji, word_count, SentimentGI, average_length, and question_mark. The biggest split is the control variable, follower. SentimentGI is also an outstanding predictor Excessive splits and number of terminal nodes made the results difficult to interpret, thus, pruning is necessary.



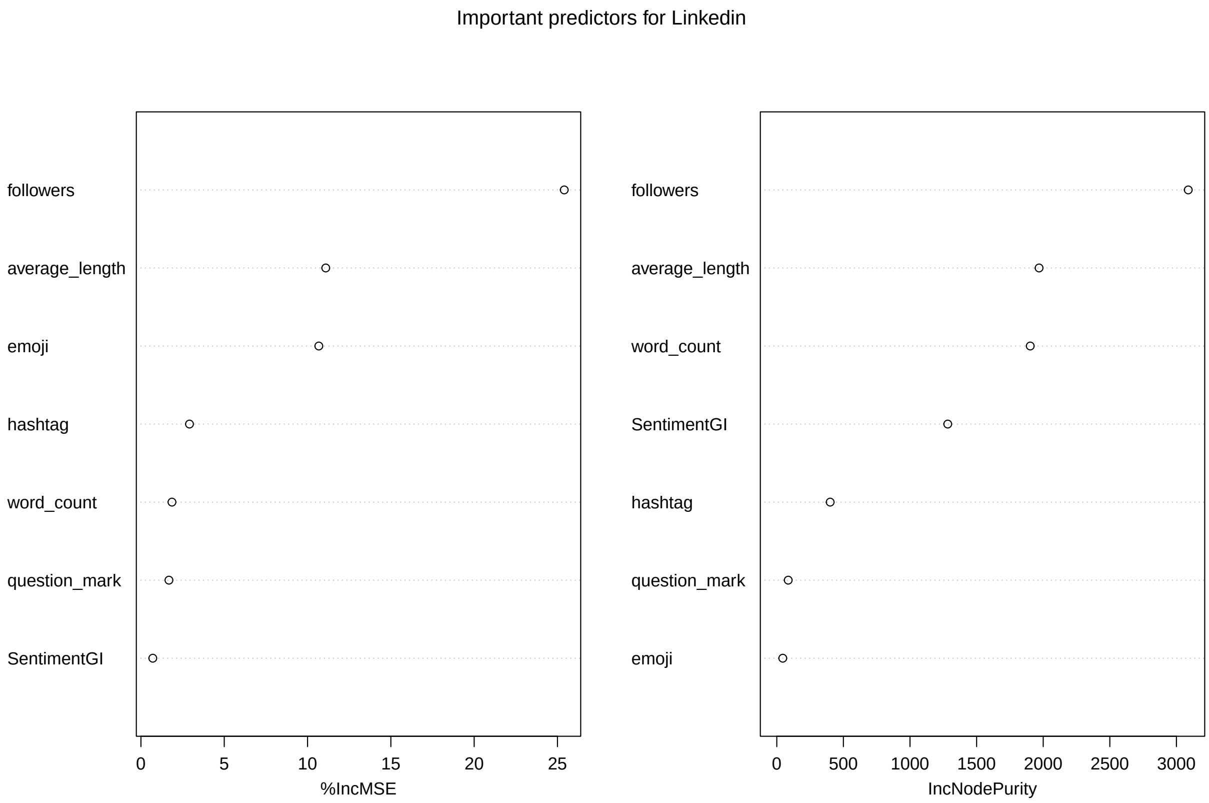
To determine the best number of terminal nodes, 10-fold Cross validation was performed, and the result shows that a model 6 terminal nodes has the lowest deviation. The recursive binary tree is pruned based on the result. Predictors used in the pruned model include followers, emoji, SentimentGI, and average_length. The residual mean deviance is 0.3432 with a total deviation of 48.04. An example interpretation is that, For accounts with more than 1173 followers, post with slightly positive sentiment (greater than 0.18) and average word length greater than 7 letters are predicted to receive 3 to 4 likes and comments.


followers < 1172.5
followers < 100.5
SentimentGI < 0.180195
0.09561
emoji < 0.5
1.30100
average_length < 7.4061
0.88760
1.58700
2.14500
3.41300

4.1.3 Random Forest

To improve the accuracy for regression trees, we then conduct a random forest model. In this case, the model has 500 trees and considers 4 variables at each split, which means it looks at 4 different combinations of variables to determine the best way to split the data at each step. The mean of squared residuals is a measure of how well the model is able to predict the outcome. A lower value indicates a better fit. In this case, the mean of squared residuals is 0.50, which suggests that the model is able to make relatively accurate predictions. The percentage of variance explained by the model indicates how much of the variation in the data can be explained by the model. In this case, the model explains 4.55% of the variation in the data. This may be considered a relatively low amount of variance explained, but it could still be useful for making predictions depending on the specific goals and context of the model.

```
##                   %IncMSE IncNodePurity
## SentimentGI     0.7110132    1283.30781
## hashtag         2.9140219     400.79725
## question_mark   1.6812499      85.74043
## average_length 11.0923112    1969.07745
## emoji          10.6755369      44.92066
## followers      25.4059958    3088.76190
## word_count      1.8625787    1902.64327
```

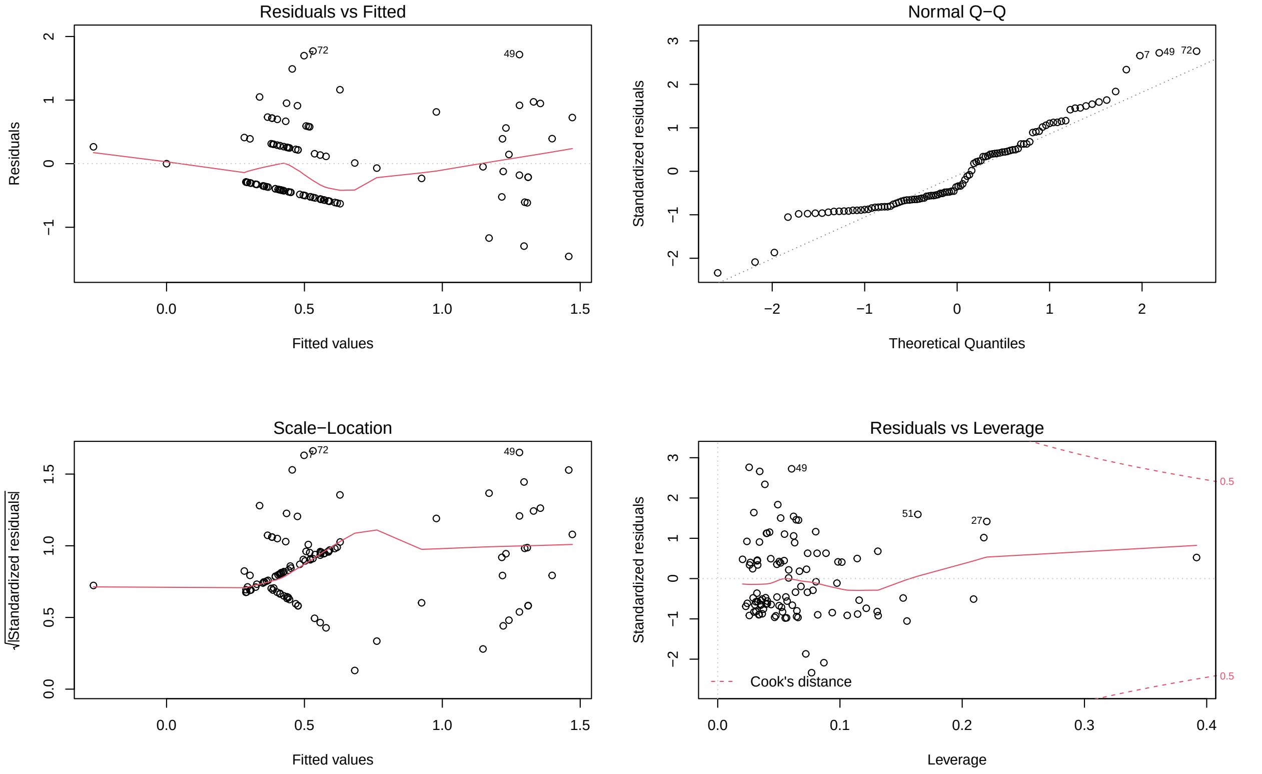


The model evaluates two different measures for each predictor: Mean Decrease Accuracy (%IncMSE) and Mean Decrease Gini (IncNodePurity). %IncMSE shows how much the model's accuracy decreases if we leave out a particular variable, while IncNodePurity is a measure of variable importance based on the Gini impurity index used for calculating the splits in trees. When looking at %IncMSE, we notice that the most influential predictor is the hashtag variable. When looking at IncNodePurity, the most influential predictors are SentimentGI and average_length. This suggests that these variables are the most important for making accurate predictions using this model.

4.2 Twitter

4.2.1 Linear Regression Model

As we mentioned before, for each fixed value, we want the error to have mean 0, constant variance, and normal distribution. In the Residual vs. Fitted plots and Scale-Location plots for each platform, we identified apparent curvature for the gray line along the x-axis, meaning we did not meet the assumption that the mean of error should be 0. This was caused by unbalanced distribution in our data set, and thus led to two obvious clusters in the residual plots. Even though we already transformed our response value in order to solve the right-skewed distribution problem, we still failed to meet part of the assumption. Fortunately, in the Normal Q-Q plots, the standardized residuals fall along 45 degrees, which represents that the residuals are normally distributed. In the Residuals vs. Leverage plots, all of our data points fall inside of Cook's Distance, thus none of our observations are influential and should be dropped.



After checking the assumption, we performed a T-test for each predictor and F-test for the whole model, and interpreted the finding for Linkedin. The whole model has F-statistic 4.584 (p < .001), meaning that our model is reasonable and useful. We identified the only significant variable of interest as SentimentGI (p = .061). The adjusted R-squared we got is 0.1943.

```
##
## Call:
## lm(formula = log(likes + comments + 1) ~ SentimentGI + hashtag +
##     question_mark + average_length + emoji + followers + word_count,
##     data = twitter)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.4585 -0.4509 -0.2127  0.3136  1.7716
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.602e-01  4.287e-01   1.307   0.1944
## SentimentGI     5.965e-01  3.141e-01   1.899   0.0605 .
## hashtag        -4.646e-03  1.843e-02  -0.252   0.8015
## question_mark   1.673e-03  1.039e-01   0.016   0.9872
## average_length -4.118e-02  6.621e-02  -0.622   0.5354
## emoji          -3.663e-01  6.601e-01  -0.555   0.5802
## followers       1.499e-04  2.996e-05   5.002 2.52e-06 ***
## word_count     -3.848e-04  5.288e-03  -0.073   0.9421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6498 on 97 degrees of freedom
## Multiple R-squared:  0.2486, Adjusted R-squared:  0.1943
## F-statistic: 4.584 on 7 and 97 DF,  p-value: 0.0001852
```

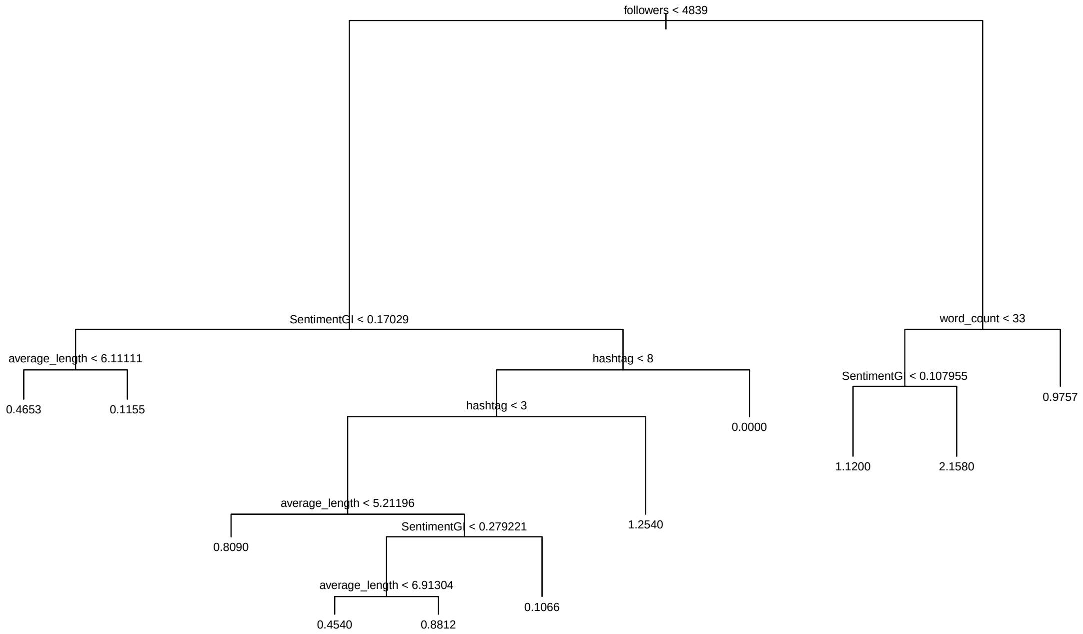
To improve our model, we dropped some insignificant predictors with relatively high p-value, and we are left with the reduced model below.

```
##
## Call:
## lm(formula = log(likes + comments + 1) ~ SentimentGI + average_length +
##     emoji + followers, data = twitter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4595 -0.4692 -0.2118  0.3192  1.7613
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.406e-01  4.135e-01   1.307   0.1940
## SentimentGI     5.959e-01  3.089e-01   1.929   0.0566 .
## average_length -4.162e-02  6.393e-02  -0.651   0.5165
## emoji          -3.546e-01  6.479e-01  -0.547   0.5854
## followers       1.515e-04  2.807e-05   5.397 4.57e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6402 on 100 degrees of freedom
## Multiple R-squared:  0.2481, Adjusted R-squared:  0.218
## F-statistic: 8.247 on 4 and 100 DF,  p-value: 8.636e-06
```

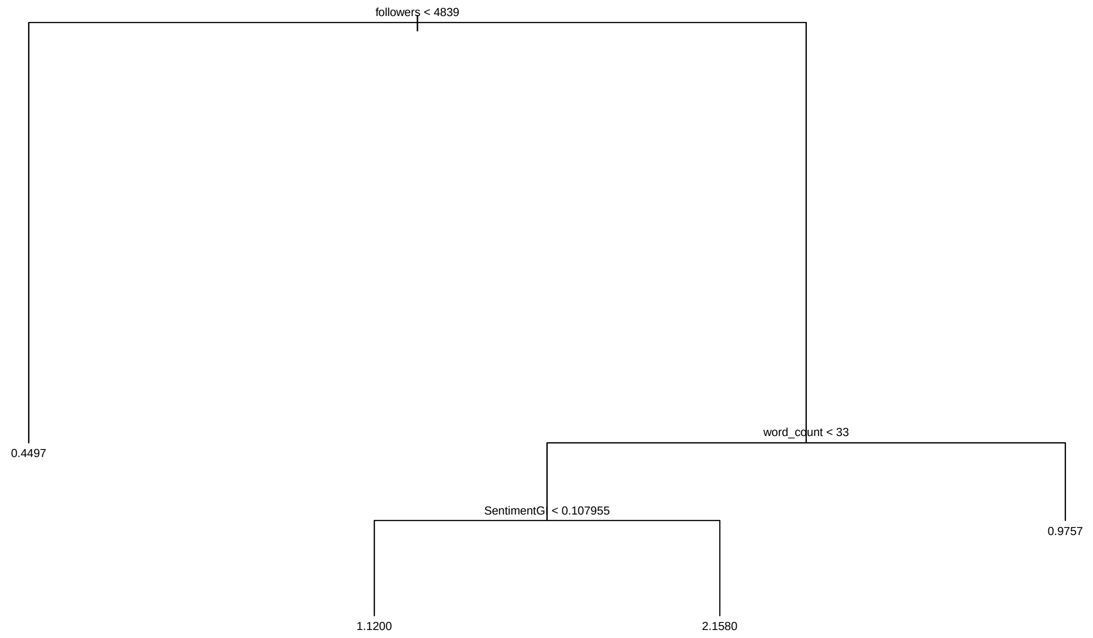
The likelihood ratio test gave us a p-value 0.9953, thus we fail to reject this reduced model. Our improved model obtained an adjusted R-square 0.218 (originally 0.1943), which is aligned to the result of the likelihood ratio test. From this model, we came to the conclusion that a post that has a caption with positive sentiment will potentially receive better engagement.

4.2.2 Regression Tree

The initial model has 10 terminal nodes, and predictors that are used include followers, SentimentGI, average_length, hashtag, and word_count. The biggest split is the control variable, follower. SentimentGI and word_count are also outstanding predictors. Excessive splits and number of terminal nodes made the results difficult to interpret, thus, pruning is necessary.



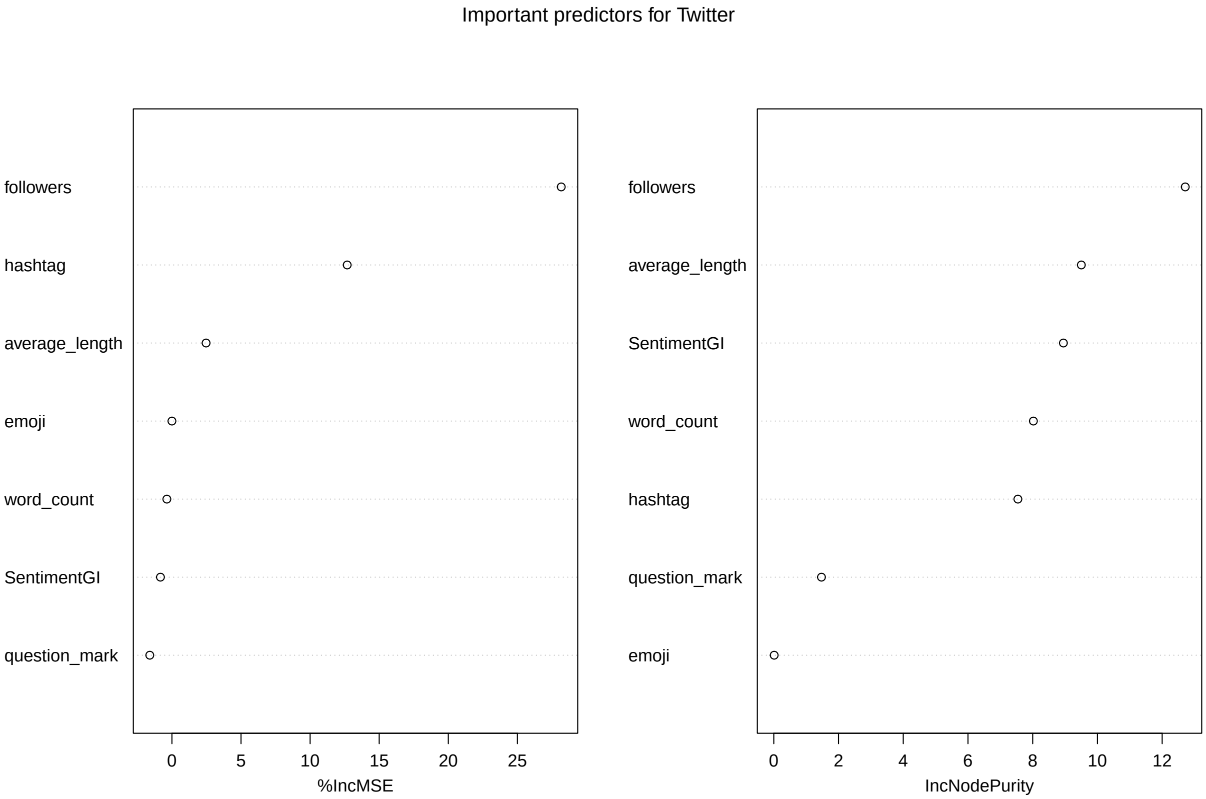
To determine the best number of terminal nodes, 10-fold Cross validation was performed, and the result shows that a model 4 terminal nodes has the lowest deviation. The recursive binary tree is pruned based on the result. Predictors used in the pruned model include followers, word_count, SentimentGI. The residual mean deviance is 0.3733 with a total deviation of 37.7. An example interpretation is that, For accounts with more than 4839 followers, post with less than 33 words and a slightly positive sentiment (greater than 0.1) are predicted to receive 2 likes and comments.



4.2.3 Random Forest

Similarly, this random forest model has 500 trees and considers 4 variables at each split, which means it looks at 4 different combinations of variables to determine the best way to split the data at each step.The mean of squared residuals is 0.48, which suggests that the model is able to make relatively accurate predictions. Also, the model explains 7.16% of the variation in the data. This may be considered a relatively low amount of variance explained, but it could still be useful for making predictions depending on the specific goals and context of the model.

```
##                    %IncMSE IncNodePurity
## SentimentGI    -0.8300165    8.94815009
## hashtag        12.6829963    7.54175824
## question_mark  -1.6018343    1.47333950
## average_length  2.4727565    9.50260007
## emoji           0.0000000    0.01214289
## followers      28.1725346   12.71304096
## word_count     -0.3668024    8.02073872
```

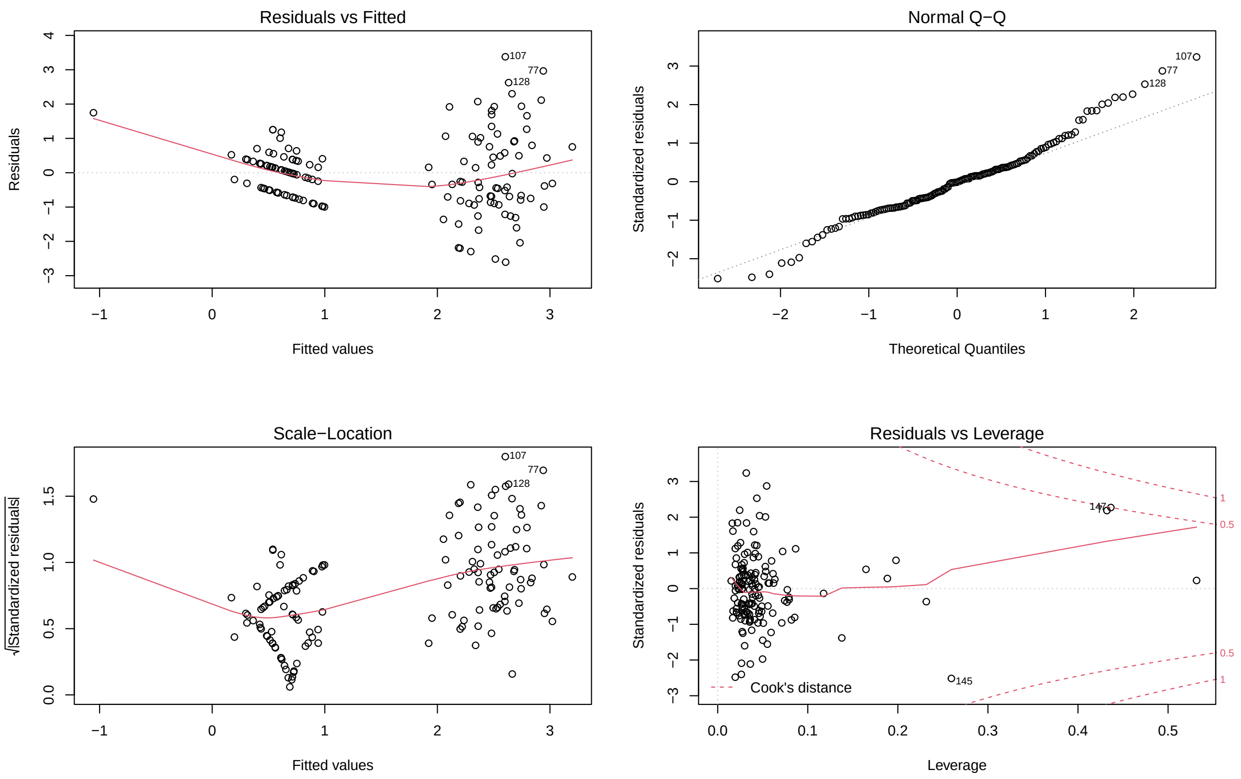


When looking at %IncMSE, we notice that the most influential predictors are the hashtag and average_length. When looking at IncNodePurity, the most influential predictors are SentimentGI and average_length. This suggests that these variables are the most important for making accurate predictions using this model.

4.3 Facebook

4.3.1 Linear Regression Model

Similar with Twitter, in the Residual vs. Fitted plots and Scale-Location plots for each platform, we identified apparent curvature for the gray line along the x-axis, meaning we did not meet the assumption that the mean of error should be 0. This was again caused by unbalanced distribution in our data set, and thus led to two obvious clusters in the residual plots. Fortunately, in the Normal Q-Q plots, the standardized residuals fall along 45 degrees, which represents that the residuals are normally distributed. In the Residuals vs. Leverage plots, all of our data points fall inside of Cook's Distance, thus none of our observations are influential and should be dropped.



After checking the assumption, we performed a T-test for each predictor and F-test for the whole model, and interpreted the finding for Linkedin. The whole model has F-statistic 18.45 (p < .001), meaning that our model is reasonable and useful. We identified the only significant variable of interest as Word_count (p = .018). The adjusted R-squared we got is 0.4522.

```
##
## Call:
## lm(formula = log(likes + comments + 1) ~ SentimentGI + hashtag +
##     question_mark + average_length + emoji + followers + word_count,
##     data = facebook)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6064 -0.6889 -0.0135  0.4867  3.3782
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.541e-01  7.828e-01   0.580   0.5628    
## SentimentGI     5.614e-01  5.454e-01   1.029   0.3051    
## hashtag         7.350e-03  3.491e-02   0.211   0.8335    
## question_mark  -1.084e-01  1.123e-01  -0.965   0.3360    
## average_length  5.355e-02  1.157e-01   0.463   0.6440    
## emoji           4.826e-02  9.287e-02   0.520   0.6041    
## followers       4.732e-05  4.739e-06   9.985   <2e-16 ***
## word_count     -5.552e-03  2.326e-03  -2.387   0.0183 *  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.061 on 141 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4781, Adjusted R-squared:  0.4522 
## F-statistic: 18.45 on 7 and 141 DF,  p-value: < 2.2e-16
```

To improve our model, we dropped some insignificant predictors with relatively high p-value, and we are left with the reduced model below.

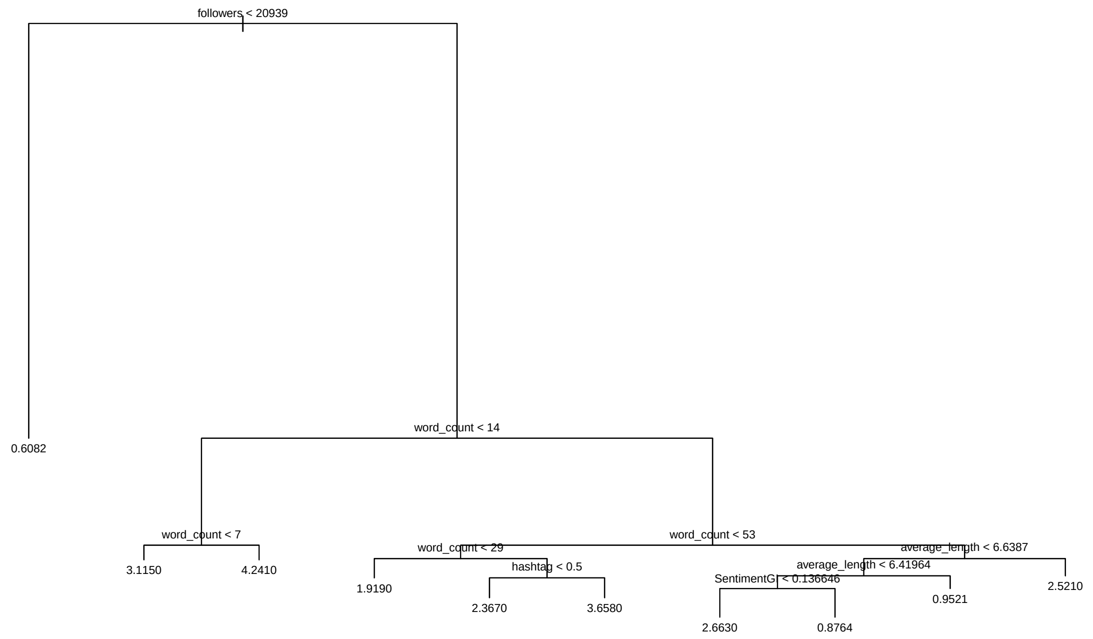
```
## 
## Call:
## lm(formula = log(likes + comments + 1) ~ word_count + followers, 
##     data = facebook)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.6400 -0.6816  0.0016  0.5024  3.2518 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  8.796e-01  1.616e-01   5.442 2.15e-07 ***
## word_count  -5.602e-03  2.117e-03  -2.646  0.00904 ** 
## followers    4.578e-05  4.134e-06  11.073  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
```

```
## Residual standard error: 1.048 on 147 degrees of freedom
## Multiple R-squared:  0.4703, Adjusted R-squared:  0.4631
## F-statistic: 65.25 on 2 and 147 DF,  p-value: < 2.2e-16
```

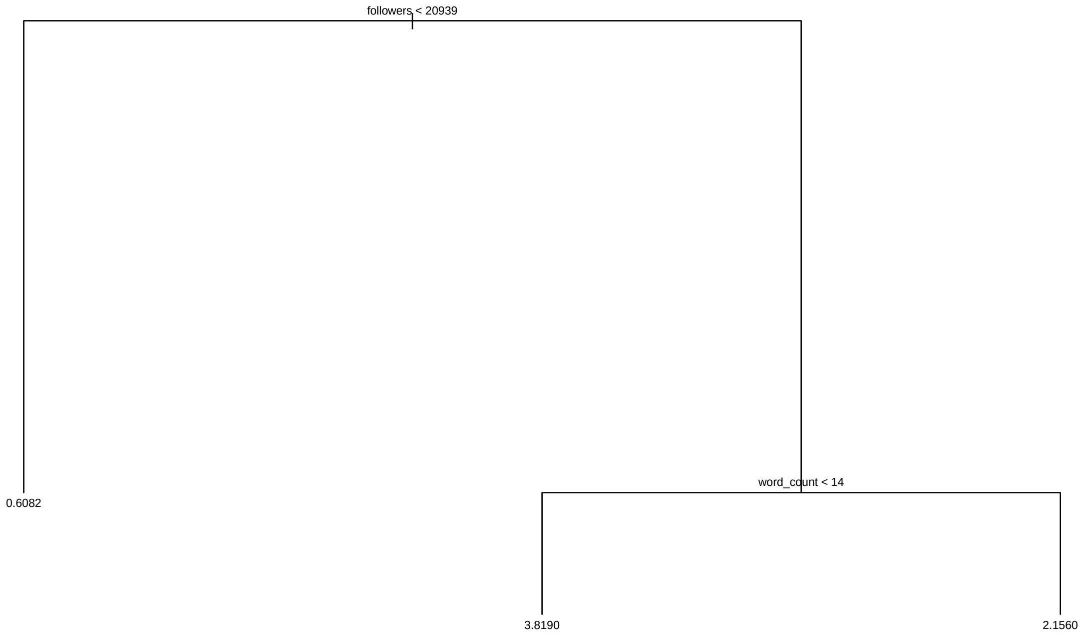
The likelihood ratio test gave us a p-value 0.768, thus we fail to reject this reduced model. Our improved model obtained an adjusted R-square 0.4618 (originally 0.4522), which is aligned to the result of the likelihood ratio test. From this model, we came to the conclusion that a post that has a caption with more word counts will potentially receive better engagement.

4.3.2 Regression Tree

The initial model has 10 terminal nodes, and predictors that are used include followers, word_count, hashtag, average_length, SentimentGI. The biggest split is the control variable, follower. Word_count is also an outstanding predictor. Excessive splits and number of terminal nodes made the results difficult to interpret, thus, pruning is necessary.



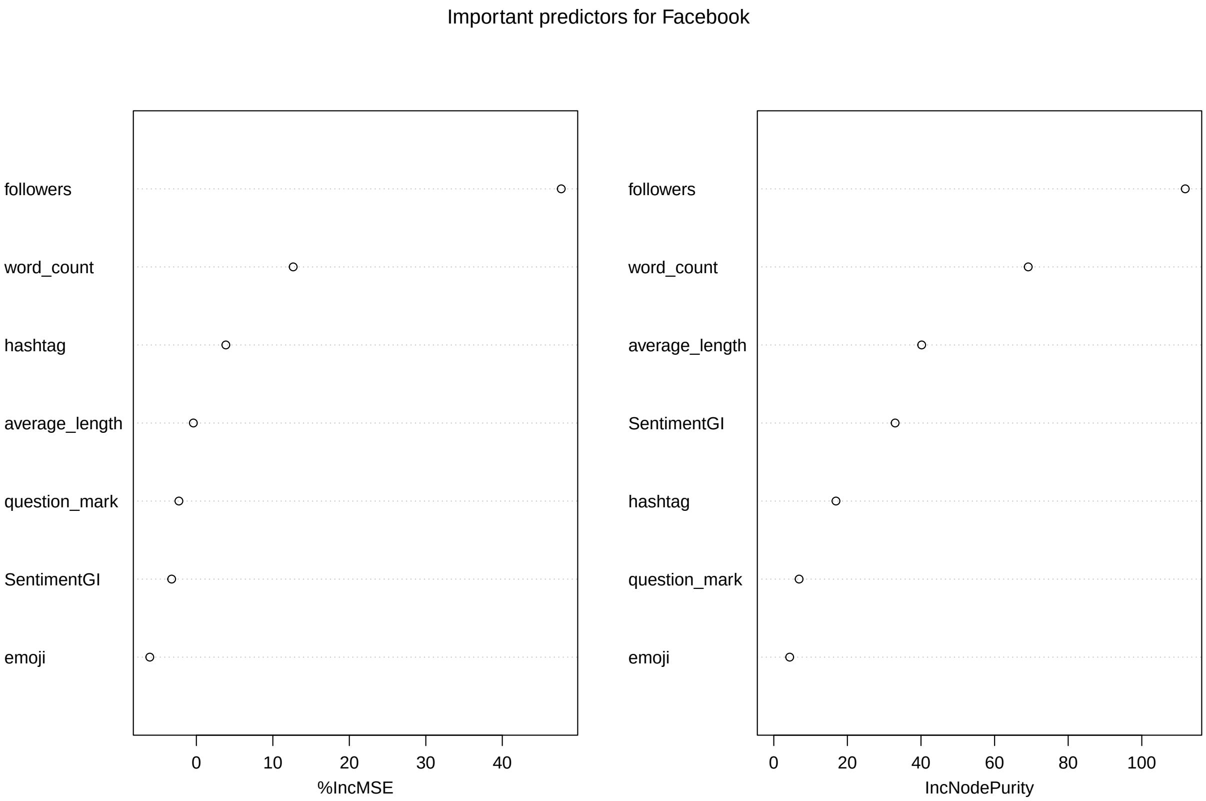
To determine the best number of terminal nodes, 10-fold Cross validation was performed, and the result shows that a model 3 terminal nodes has the lowest deviation. The recursive binary tree is pruned based on the result. Predictors used in the pruned model include followers, word_count, SentimentGI. The residual mean deviance is 0.92 with a total deviation of 134.3, which is extremely high. An example interpretation is that, For accounts with more than 20939 followers, post with less than 14 words are predicted to receive 3.8 likes and comments.



4.3.3 Random Forest

As mentioned before, this random forest model has 500 trees and considers 4 variables at each split, which means it looks at 4 different combinations of variables to determine the best way to split the data at each step. The mean of squared residuals is 0.49, which suggests that the model is able to make relatively accurate predictions. Also, the model explains 6.14% of the variation in the data. This may be considered a relatively low amount of variance explained, but it could still be useful for making predictions depending on the specific goals and context of the model.

```
##                   %IncMSE IncNodePurity
## SentimentGI    -3.2302973     32.974022
## hashtag          3.8540018     16.894614
## question_mark  -2.2824049      6.879562
## average_length -0.3916373     40.188000
## emoji          -6.0896475      4.321289
## followers      47.7056637    111.810992
## word_count     12.6597889     69.140714
```



When looking at %IncMSE, we notice that the most influential predictors are the hashtag and average_length. When looking at IncNodePurity, the most influential predictors is word count. This suggests that this variable is the most important for making accurate predictions using this model.

5 Limitation and Improvement

In the data collection process, one limitation that emerges is that the response variable is extremely right skewed. According to the histogram, most of the observations are in the first bin, thus, transformation is necessary to perform further linear regression models. However, there are many 0 values in the response variable, making it impossible to simply apply log transformation. Therefore, we decided to apply log(y+1) transformation to eliminate the 0 values before transformation. After the transformation, 0 observations become 0 value again, elevating the validity of the transformation.

Another limitation that strongly affected the power of the machine learning models that we built is that the amount data we collected is insufficient. There are only less than 150 observations for each of the three platforms, which severely increases the variance. Also, small sample size results in inability to split training set and sample set. The best method to solve this limitation is to collect more data. However, due to the limited ability of the open-source GitHub packages, collecting more post are currently unavailable because of social media platform restriction. Another possible improvement is to apply Ridge Regression to reduce variance when sample size is small.

6 Recommendations and Conclusion

After we had successfully built three models for each of the three platforms, we wanted to summarize our findings and provide recommendations to our client for individual platforms.

6.1 LinkedIn

For LinkedIn, all three models agreed that sentiment score and emoji were significant predictors of the response variable. In addition, the regression tree and random forests suggested that average_length was also a useful predictor. To sum up, we would recommend our client to write in a more positive tone in their LinkedIn posts, use more emojis and more complicated words as well. In particular, using longer and more sophisticated words can make our client's LinkedIn posts seem more formal and professional. But of course, the key to including longer and more complicated words in a LinkedIn post is to maintain a good balance. We advise our client to use them sparingly and wisely, and make sure that those words will add value and authenticity to the post, shaping the image of the company.

6.2 Twitter

For Twitter, all three models agreed that sentiment score was a significant predictor of the response variable. There was no other overlap between model results. Thus, we would recommend our client to write in a more positive tone in their tweets. For example, instead of using words like "problem," "difficult," or "negative," try switching to words like "opportunity," "challenge," or "solution." This will help to frame the situation in a more positive light. Also, we wanted to emphasize the importance of ending the post on a positive note. This can be as simple as thanking your readers for their time, or expressing your hope that the information you have shared will be helpful to them. This will leave a positive impression and encourage the audience to engage with your post.

6.3 Facebook

For Facebook, the regression tree, linear regression, and random forest all suggested that word_count was a significant predictor. Noticeably, it had a negative association with the response variable. As a result, we would recommend our client to write shorter captions on Facebook, be concise and avoid using tedious long sentences in their expressions. More specifically, we advise our client to choose words carefully and only include information that is relevant to the gist of the Facebook post. This means avoiding unnecessary words or details and being straight to the point. For instance, our client can use bullet points or lists to organize all the information and make it easy for the followers to comprehend. When editing the post, it is best to read aloud and identify words or phrases that can be removed without losing the meaning of the message.

References

n.d.a. https://github.com/shaikhsajid1111/twitter-scraper-selenium.

n.d.b. https://github.com/kevinzg/facebook-scraper.

n.d.c. https://advertools.readthedocs.io/en/master/.

n.d.d. https://github.com/daniyalrmb/LinkedIn-Scraping-Sentiment.

n.d.e. https://www.nltk.org/howto.html.

n.d.f. https://cran.r-project.org/web/packages/SentimentAnalysis/SentimentAnalysis.pdf.