An Analysis of Sharing in Social Media: What do we choose to share?

Introduction

Along with the rise of Internet, social media has been growing rapidly over the last decade and has become an important part of our daily life. Sharing on social media a is very powerful means of information dissemination. To better understand the behavior of sharing, we would like to know the relationship between the number of shares in social networks and several predictors. We are given a sample of 513 articles collected over two years from Mashable, a digital media website. They have several hypothesis regarding this underlying relationship. One is that the length of the article is negatively associated with the probability of sharing. We believe, at the beginning of the study, that the relationship between the type of website channel and the number of shares changes depending on the number of links in the article. Another one is that the more negative an article is, the more likely it gets shared. An interaction between the number of images in the article and the day it was published is also assumed.

Exploratory Data Analysis

We have a randomly selected sample of size 513. All observations in this sample contain one response variable “Shares” (the number of shares in social networks) and predictors: “Title” (the number of words in the title), “Content” (the number of words in the article content), “Links” (the number of links in the article), “Images” (the number of images in the article), “Videos” (the number of videos in the article), “Positive” (rate of positive words in the article content), “Negative” (rate of negative words in the article content), “Channel” (the type of website: Business, Entertainment, Lifestyle, Other, SocialMedia, Tech, World), “DayPublished” (the day of the week the article was originally published (Monday, ..., Sunday)).

The distribution of the response variable “Shares” is strongly right-skewed, with most of data being small along with very few very large outliers over 50000. Its mean of “Shares” is 3451 and median is 1500.

Among the 9 predictor variables, “Title”, whose distribution looks fairly normal and unimodal, is the only one that does not have noticeable skewness. The mean value of “Title” is 10.47 and median is 10. No outlier either. “Content”, “Images”, “Videos” and “Links” are strongly right-skewed. “Positive” and “Negative” are slightly right-skewed. All the essential statistics of these predictor variables are listed down below in Table 1 and Table 2. For individual univariate distributions of both the response variable and the predictors, see Figure 1.

Figure 1

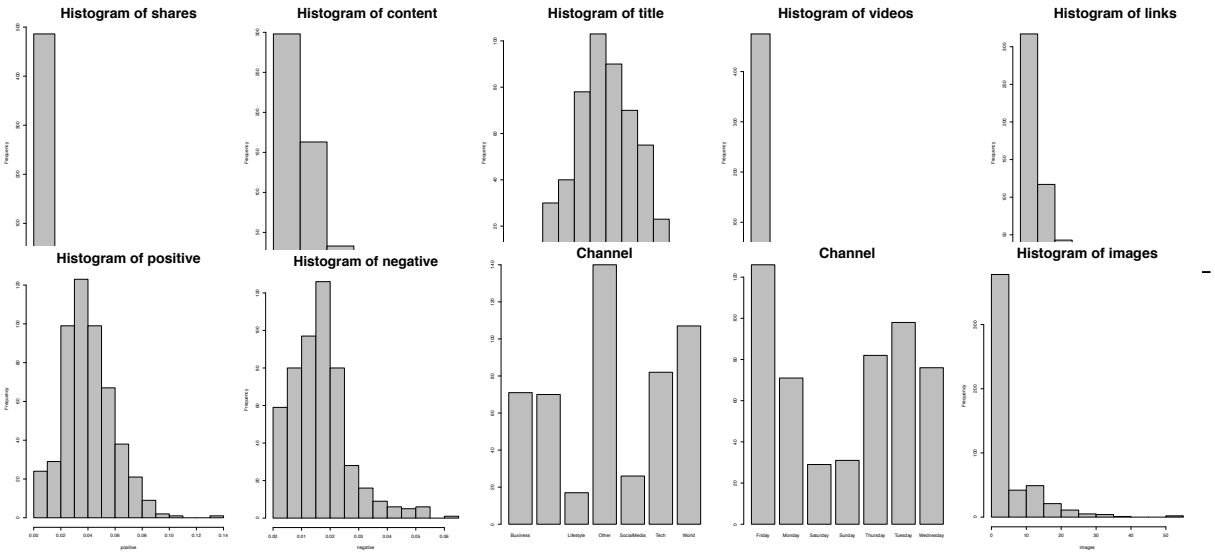


Table 1

	Mean	Median	SD	1st Quartile	3rd Quartile
Shares	3451	1500	7498.713	997	2900
Title	10.47	10.00	2.068986	9.00	12.00
Content	525.2	423.0	418.9672	243.0	680.0
Links	11.97	8.00	12.28331	4.00	15.00
Images	4.719	1.00	7.374415	1.00	7.00
Videos	1.62	0.00	4.634045	0.00	1.00
Positive	0.04011	0.03876	0.01873021	0.02809	0.05096
Negative	0.016260	0.015720	0.01033533	0.009579	0.021050

Table 2

Channel	Day Published
Business: 71	Sunday: 31
Entertainment: 70	Monday: 71
Lifestyle: 17	Tuesday: 98
Social Media: 26	Wednesday: 76
Tech: 82	Thursday: 82
World: 107	Friday: 126
Other: 140	Saturday: 29

We further studies the bivariate characteristics of this sample data. First, we test the correlation between the response variable and continuous quantitative predictor variables using the cor() function. See Figure 2. It appears that expect for Links, which has a correlation value of 0.165, the continuous predictors are not strongly correlated with the response variable. Since we discovered earlier that most predictors showed a strong right-skewness, this lack of correlation could be due to the outliers that also caused the skewness.

Then, we take a look at the correlation between the response variable and categorical predictor variables, with the help of conditional box plot. Initially, we plotted the two conditional box plots using the whole data set. However, because of the larger outliers, the actual box plot itself is fairly small and indiscernible. We hence adjust the data set, excluding the shares greater than 6000 and only focus on the smaller side. The range of the box plot itself is now way larger than before, which is desirable. As you can see from Figure 3, there is not really a strong linear relationship between shares and either channel and daypublished. The distribution of shares does not differ significantly over different days or channels, meaning that shares is not strongly affected by either channel or daypublished. Therefore, neither of these two can be ordered categorical variables.

When it comes to the individual groups, the group Social Media under the variable Channel shoes a much wider range and a higher mean than others. However, the size of Social Media is 26 and is significantly smaller than other groups so this abnormality could be the result of small category size. Same for Saturday and Sunday in Daypublished. Sunday and Saturday have the similar pattern, showing higher means and slight wider range. But, both of them are smaller comparing to other groups.

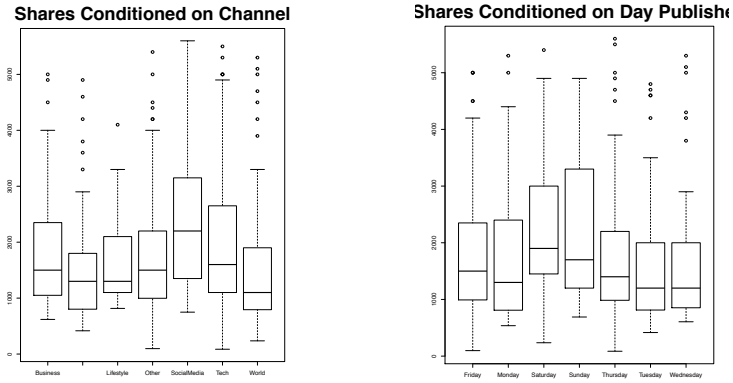
Table 3

Correlation between the response variable and continuous predictor variables

Shares	Content: -0.004896723	Links: 0.16461489	Images: 0.01626444
	Videos: -0.01356831	Positive: 0.06429951	Negative: -0.044801678
	Title: 0.028470082		

Figure 2

Correlation between the response variable and predictor categorical variables



Modeling The Data

We first examine the categorical variable “Day Published”, which has the potential to be included in this model as a ordered categorical variable. The way to determine whether or not we should do it this way is to see if one level change in the “Day Published” leads to same amount of change in the response variable “Shares”. However, as shown in Figure 2, day published is not very suitable for being treated as a ordered categorical variable. We should also consider the possibility of combining groups because of the small sizes of Saturday and Sunday, which could potentially increase the instability of the model.

Table 4

	Sunday	Monday	Tuesday	Wednesday	Thursday	Saturday	Friday
Coefficient	-243.0	-454.1	117.2	-1076.2	179.3	-953.7	
SE	1509.5	1117.3	1014.1	1093.5	1068.3	1550.7	Reference

p-value	-0.161	0.2991	0.116	-0.984	0.168	-0.615	Group
Size	31	71	98	76	82	29	

We then take a look at another categorical variable “Channel”. “Channel” is obviously not ordered, so we would like to know if it is better to just collapse multiple levels to one to enhance stability and interpretability, since the group size of Lifestyle(17) and Social Media(26) are significantly smaller than that of other groups. The downside of this choice, however, is to lose information and to average out the effects of each level. Choosing other as reference group (since the size of Other is the greatest(140), setting Other as the reference group increases stability). Combine everything other than “Other” and rerun lm() function. Coefficient = 1515.6. SE = 740.9. p-value = 0.0413. It does indeed reflects the individual groups but having a group of Other and non-Other is not significant in a sense that even the conclusion is statistically significant, it is not useful. So we decide to leave as it is.

Table 5

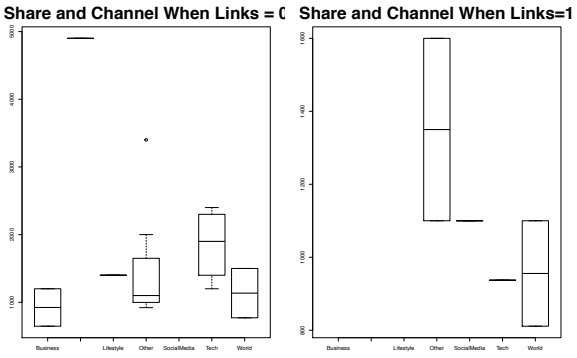
	Business	Social Media	Entertainment	Lifestyle	Tech	World	Other
Coefficient	-1743.2	-1664.2	-1344.7	-1038.8	-686.5	-2151.5	
SE	1092.4	1601.2	1097.6	1925.8	1042.7	962.8	
p-value	0.1112	0.2991	0.2211	0.5898	0.5106	0.0259	Reference Group
Size	71	26	70	17	82	107	

Diagnostics

As stated above in the Introduction section, we have several hypothesis prior to the study. The first is whether or not the longer the article the less likelihood of sharing. According to the initial model, that appear to be the case. The coefficient of content is -0.7457 with p-value equal to 0.4417. The second hypothesis is that more negative articles will receive more shares. From the model, we have that predictor negative has a very large negative coefficient, indicating that one adjusted unit increase in negative, holding other predictors constant, will result in a huge decrease in shares. Therefore, we believe that the more negative the article is, the less shares it gets, which proves the hypothesis wrong. The variable positive also supports this conclusion, since the more positive words an article contains, naturally speaking, the less negative it is. The coefficient of positive is fairly large and positive. Hence, the more positive (the less negative) an article is, the more likely it will be shared.

The other one is that the relationship between channel and shares depends on the number of links. From our test, that appears to be the case. The distribution of channel vs shares does change quite a lot under different link values, meaning that the relationship of channel vs shares depends on link values. Figure 3.

Figure 3



After running the diagnostics of our initial model, we see a couple features that are worth concerning. The QQ-plot is quite right-skewed on the top. The residual vs fitted plot shows a strong negative linear pattern, indicating inconsistency in error variance and the variance of the model. The Scale-Location plot looks positively linear, too.Also, boxcox shows need of adjustment because the lambda is close to 0 with1 being desirable. We first try log transformation of y. The boxcox appeared to be worse than that of the original data. So we then try share^(-1/4). Based on the diagnostics of the post-transformation model, we believe that this first try is indeed successful. QQ-plot is no longer skewed. Boxcox shows lambda is very close to the ideal value one. Residual vs Fitted plot is random, not showing any pattern. Scale-Location plot is also random. The new transformed model pass all the diagnostics.

Figure 4

Original data diagnostics. Pre-transformation.

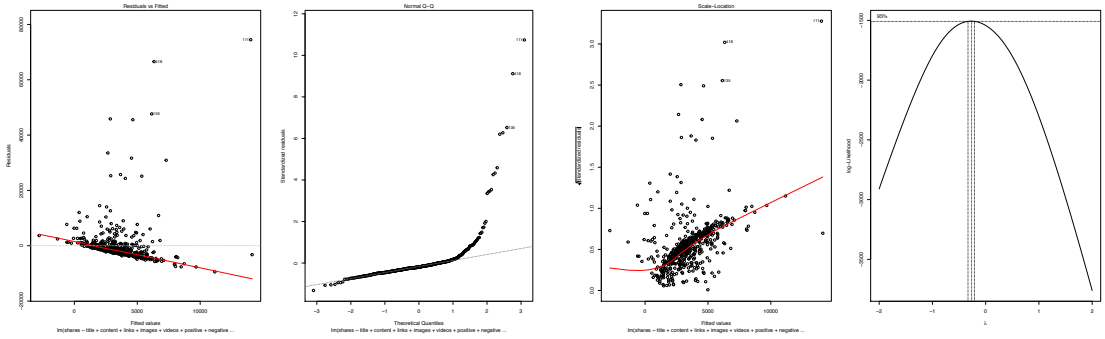
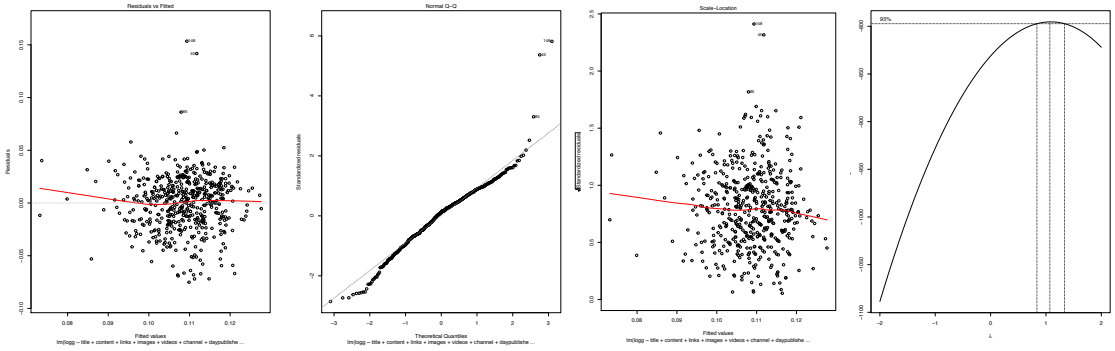


Figure 5
Post-transformation diagnostics.



Model Inference & Results

A summary of the new model is attached down below. p-value of the new model is 0.002816, low enough to ensure the statistical significancy of the model. R-square is 0.07759. Adjusted R-square is 0.04204.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.597e-01	9.516e-03	16.781	<2e-16	***
title	-6.134e-04	7.277e-04	-0.843	0.3997	
content	-6.266e-06	4.234e-06	-1.480	0.1395	
links	-3.434e-04	1.372e-04	-2.503	0.0126	*
images	7.332e-05	2.248e-04	0.326	0.7445	
videos	1.290e-04	3.459e-04	0.373	0.7094	
positive	-2.002e-02	8.274e-02	-0.242	0.8090	
negative	2.838e-01	1.487e-01	1.908	0.0569	.
d1	-2.705e-05	5.085e-03	-0.005	0.9958	
d2	4.942e-03	4.611e-03	1.072	0.2844	
d3	3.938e-03	4.954e-03	0.795	0.4271	
d4	3.425e-03	4.866e-03	0.704	0.4819	
d6	-3.977e-03	6.771e-03	-0.587	0.5572	
d7	-1.224e-02	6.755e-03	-1.811	0.0707	.
c1	5.671e-03	5.448e-03	1.041	0.2984	
c2	1.193e-02	5.145e-03	2.318	0.0209	*
c3	4.835e-03	8.762e-03	0.552	0.5814	
c5	-5.579e-03	7.284e-03	-0.766	0.4441	
c6	2.416e-03	5.093e-03	0.474	0.6354	
c7	1.171e-02	4.926e-03	2.378	0.0178	*

	Coefficient
title	-0.0006134
content	-6.266E-06

	Coefficient
links	-3.434E-04
images	7.332E-05
videos	1.290E-04
positive	-2.002E-02
negative	2.838E-01
Monday	-2.705E-05
Tuesday	4.942E-03
Wednesday	3.938E-03
Thursday	3.425E-03
Saturday	-3.977E-03
Sunday	-1.224E-02
Bussiness	5.671E-03
Entertainment	1.193E-02
Social Media	-5.579E-03
Tech	2.416E-03
World	1.171E-02
Lifestyle	4.835E-03

Discussion & Result

Our final model satisfies all criteria. We believe that this model is sufficient to predict the number of shares in social networks using the provided predictors. Among the 9 predictors, we favor NEGATIVE to be especially useful in predicting the number of shares in the social network. There is a non-zero negative linear relationship between these two and p-value is small.

Although the new model pass all the diagnostics and is indeed statistically significant, there are still some improvements yet to be done. During out univariate analysis, we discovered a couple of outliers. But we did not identify and exclude those outliers. Next, our R-square value is lower than the adjusted R-square value, which means there might be variables that are not containing enough regression information, hence not useful, existing in our model. Our next step would be identifying these insignificant variables and get rid of them to make out model stabler. Also the R-square is not large, leaving more than 90 percent of the information in the error. It should also be improved.

R Appendix

```
da<-read.table("da2.txt")
> shares<-da[,1]
> title<-da[,2]
> content<-da[,3]
> links<-da[,4]
> images<-da[,5]
> videos<-da[,6]
> channel<-da[,7]
> daypublished<-da[,8]
> positive<-da[,9]
> negative<-da[,10]


##Univariate
>hist(shares,cex.main=3,col=8)
> hist(content,cex.main=3,col=8)
> hist(title,cex.main=3,col=8)
> hist(videos,cex.main=3,col=8)
> hist(links,cex.main=3,col=8)
> hist(positive,cex.main=3,col=8)
> hist(negative,cex.main=3,col=8)
> hist(channel,cex.main=3,col=8)
> plot(channel,cex.main=3,col=8)
> plot(channel,cex.main=3,col=8,main="Channel")
> plot(daypublished,cex.main=3,col=8,main="Channel")
> hist(images,cex.main=3,col=8)

##Bivariate
> pairs(cor(var))
> cor(pairs(var))
Error in cor(pairs(var)) : supply both 'x' and 'y' or a matrix-like 'x'
> plot(cor(var))
> boxplot(shares~channel)
> Channel<-channel[shares<30000]
> boxplot(shares~Channel)
Error in model.frame.default(formula = shares ~ Channel) :
  variable lengths differ (found for 'Channel')
> Shares<-shares[shares<30000]
> boxplot(Shares~Channel)
> Shares<-shares[shares<15000]
> Channel<-channel[shares<15000]
> boxplot(Shares~Channel)
> Channel<-channel[shares<10000]
> Shares<-shares[shares<10000]
> boxplot(Shares~Channel)
> Shares<-shares[shares<6000]
> Channel<-channel[shares<6000]
> boxplot(Shares~Channel)
```

```
> boxplot(Shares~Channel,main="Shares Conditioned on Channel")
> boxplot(Shares~Channel,main="Shares Conditioned on Channel",cex.main=3)
> daypublished<-daypublished[shares<6000]
> boxplot(Shares~daypublished,main="Shares Conditioned on Day Published”,cex.main=3)

##Building Model
> c1<-ifelse(channel=="Business",1,0)
> c2<-ifelse(channel=="Entertainment",1,0)
> c3<-ifelse(channel=="Lifestyle",1,0)
> c4<-ifelse(channel=="Other",1,0)
> c5<-ifelse(channel=="SocialMedia",1,0)
> c6<-ifelse(channel=="Tech",1,0)
> c7<-ifelse(channel=="World",1,0)
> summary(lm(shares~c1+c2+c3+c4+c5+c6+c7))
> d1<-ifelse(daypublished=="Monday",1,0)
> d2<-ifelse(daypublished=="Tuesday",1,0)
> d3<-ifelse(daypublished=="Wednesday",1,0)
> d4<-ifelse(daypublished=="Thursday",1,0)
> d5<-ifelse(daypublished=="Friday",1,0)
> d6<-ifelse(daypublished=="Saturday",1,0)
> d7<-ifelse(daypublished=="Sunday",1,0)
> summary(lm(shares~d1+d2+d3+d4+d7+d6+d5))

##Model&Diagnostics
> notother<-ifelse(channel=="",1,0)
> line<-
lm(shares~title+content+links+images+videos+positive+negative+d1+d2+d3+d4+d6+d7+c1+c2
+c3+c5+c6+c7)
> summary(line)
> Shares1<-Shares[links==0]
> Channel0<-Channel[links==0]
> Shares0<-Shares[links==0]
> boxplot(Shares0~Channe0)
Error in eval(expr, envir, enclos) : object 'Channe0' not found
> boxplot(Shares0~Channel0)
> boxplot(Shares0~Channel0,main="Share and Channel When Links = 0")
> boxplot(Shares0~Channel0,main="Share and Channel When Links = 0", cex.main=3)
> boxplot(Shares0~Channel0,main="Share and Channel When Links = 0")
> Channel1<-Channel[links==1]
> Shares1<-Shares[links==1]
> boxplot(Shares1~Channel1,main="Share and Channel When Links=1", cex.main=3)
> lm(Share1~Channel1)
Error in eval(expr, envir, enclos) : object 'Share1' not found
> lm(Shares1~Channel1)
trans<-lm(logg~title+content+links+images+videos+channel+daypublished+positive+negative)
> boxcox(trans)
> boxcox(trans)
> plot(title)
> boxcox(trans)
```

```
> lm(shares~title+content+links+images+videos+channel+daypublished+positive+negative)
boxcox(line)
> logg<-shares^0.2
> trans<-lm(logg~title+content+links+images+videos+channel+daypublished+positive+negative)
> boxcox(trans)
> logg<-shares^(-0.2)
> trans<-lm(logg~title+content+links+images+videos+channel+daypublished+positive+negative)
> boxcox(trans)
> logg<-shares^(-0.1)
> trans<-lm(logg~title+content+links+images+videos+channel+daypublished+positive+negative)
> boxcox(trans)
> logg<-shares^(-0.3)
> boxcox(trans)
> trans<-lm(logg~title+content+links+images+videos+channel+daypublished+positive+negative)
> boxcox(trans)
> logg<-shares^(-0.25)
> boxcox(trans)
> plot(trans)
Hit <Return> to see next plot: trans
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
>
> max(shares)
```