DEC520Q-102-teamB27-Final Project

Ying Guo, Teddy Hector, Zhuoying Lin, Anirudh Reddy, Xiaoshi Zhu

8/13/2019

Table of Contents

Business Understanding

In the growing digital world, it is becoming increasingly important for news outlets to understand the nature, behavior, and interests of their readers in order to maintain the traffic and viewership. The goal of running a blog or a news website is to maximize the number of viewers and shares, which in turn will increase the revenue of the website. In order to increase the popularity of the website, writers need to mold their articles around the topics that interest viewers. The website also needs to know about the behavior patterns of its readers so it can release articles during certain periods to increase that article's views and shares. Businesses can also use this information to determine the best ways to increase clicks or traffic for their marketing website, and therefore boost their exposure to the public and increase their brand awareness.

Data Understanding

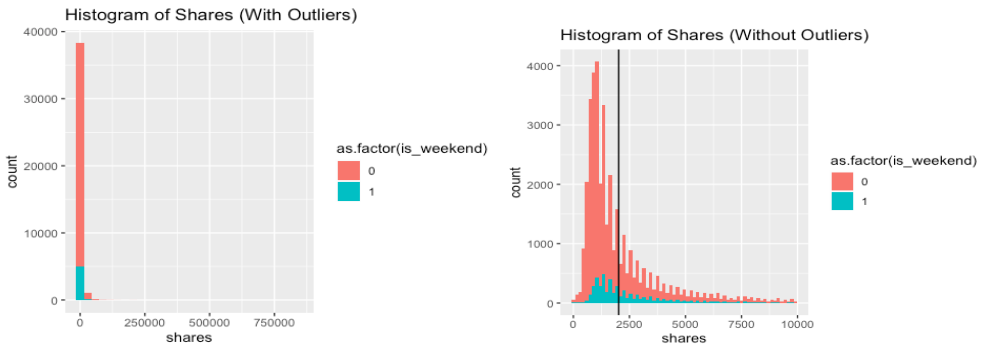
The dataset we obtained is Online News Popularity from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity). The data that has been

collected from the previous visits of the users to different articles on a blog website mashable.com.

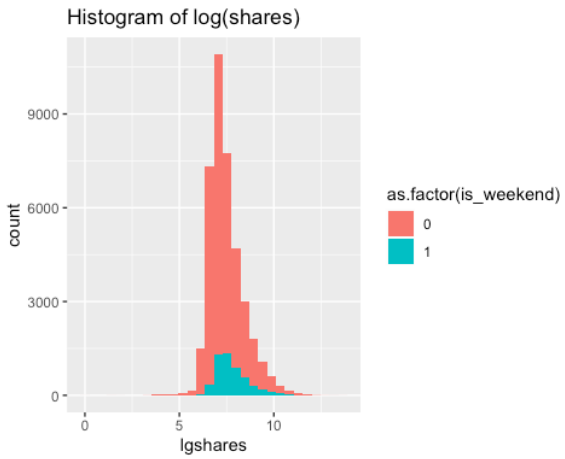
The dataset contains 61 variables, including the variable of interest, the number of shares, and two non-predictive variables, the URL of the article and time delta between the publish of the article and the archive of the data. The rest of 58 variables measure the article from the title, words choice, time the article was posted, etc., which could potentially be explanatory variables for number of shares.

Exploratory Data Analysis

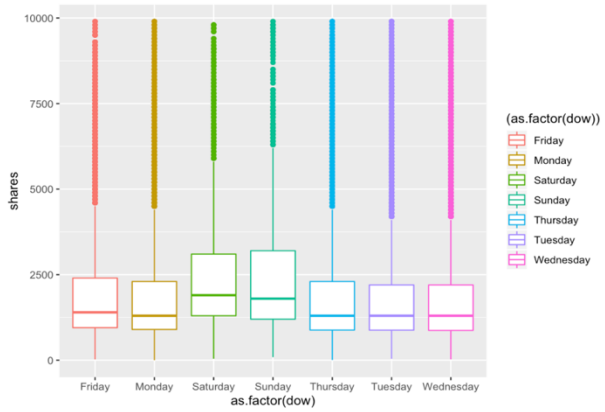
First, we examined our dependent variable, shares. From the histogram, we can see there are articles that have significantly higher shares than most of the other articles. We further investigated and concluded that shares fall mostly at the range between 0 to 10,000 shares. Observations with shares larger than 10,000 compose of 5% of the whole data set. The mean of shares is 2022 with the range of 0 and 10,000, but it is 3395 for the whole data set because of the significant effect of shares larger than 10,000.



Because of the outliers, we took the log of shares in order to bring the data closer to each other. After getting log(shares), we can tell from the histogram that our data is more concentrated and somewhat normally distributed.

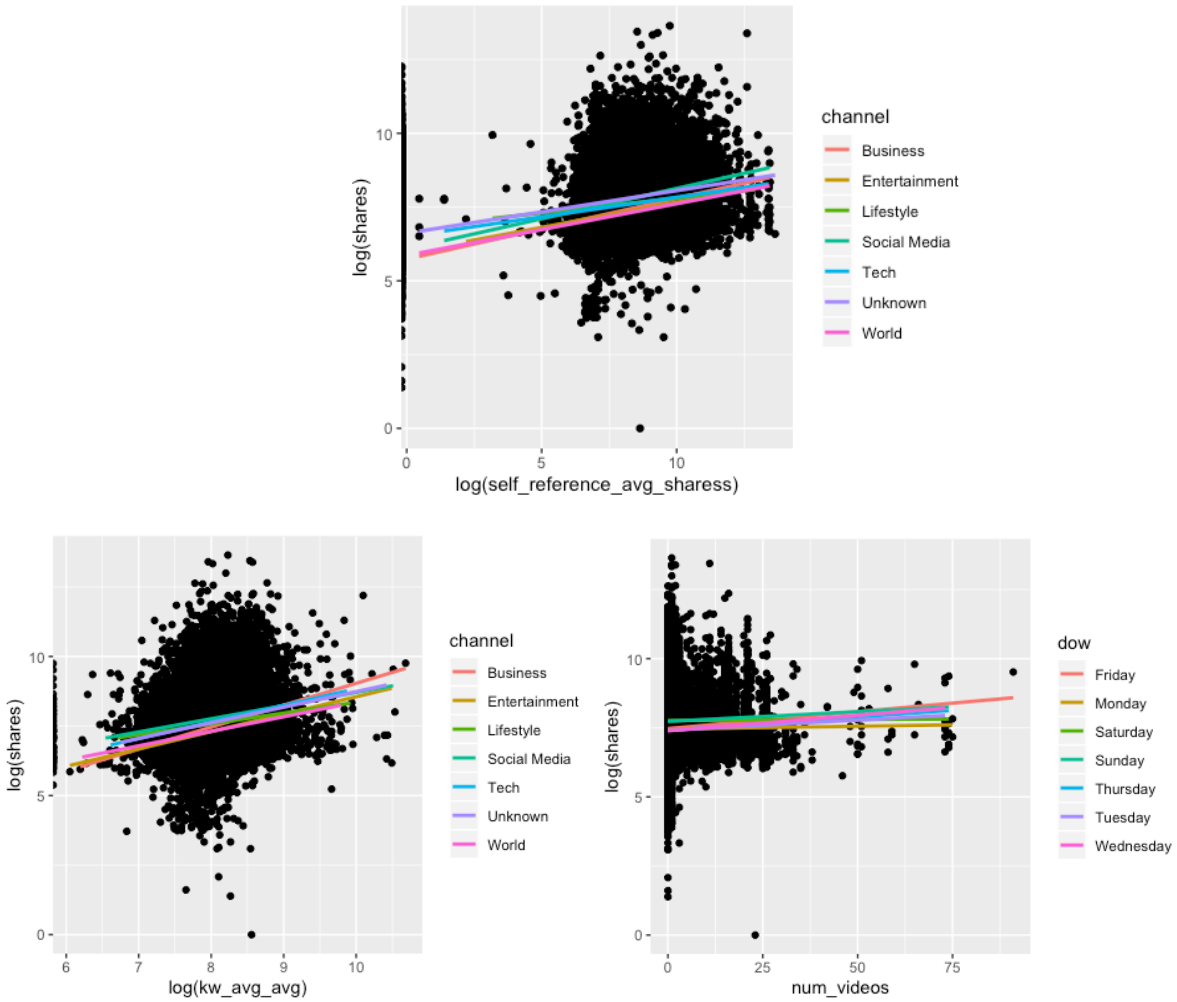


Next, we noticed that our data contained articles published Monday through Sunday. We wondered if the day of the week has an impact on shares that articles get. After filtering out outliers, we found that articles published on Saturdays and Sundays have higher shares on average than those on weekdays. As a result, we believe that day of week might be a good categorical variable that we can run interaction terms within our following analyses.



Interaction Exploration

Below, we investigated some interaction terms with categorical variables **dow** (day of the week when articles are published) and **channel** (channel in which articles belong).



From our interaction graphs, we found that categorical variables **channel** and **dow** do not seem to have much effect on distinguishing the effect of independent variables on **log(shares)**, given that the slopes are similar and close to each other. This seems contradictory to our previous observation that **dow** influences shares. However, even if their effects on **log(shares)** are small in value, it does not mean they are not statistically significant. So, we will further examine their effects later in **Modeling** section.

Data Preparation

We investigated each independent variable in detail and researched its relevance to the popularity of a news blog before cleaning our data. Our first step in cleaning the data is to remove non-predictive variables, **url** and **timedelta**. After this, we combined certain columns

together to get rid of redundant dummy variables. We noticed that the original data set contained many dummy variable columns, such as **weekday_is_monday**. These variables give a value of 1 for articles published on Monday and a value of 0 otherwise. We combined such variables into one column called **dow** to indicate the day of the week an article was published. We used the same logic to create a **channel** column, which indicates the genre of the article. We then deleted the old dummy variable columns.

Secondly, we decided to see if there are any outliers. When we checked the summary of each variable (48 variables currently), we noticed that **average_token_length** had 1100+ records with value of 0, which implied that those articles have no words at all. Moreover, these records have 0 for a number of other columns such as number of words in the title. These are suspicious instances and it is not reasonable to fix them as they have too many columns with wrong value, so we decided to remove them. We then noticed records with **n_unique_tokens**, **n_non_stop_words** and **n_non_stop_unique_tokens** greater than 1, which is impossible since these variables are rates. Therefore, we removed these records as well. Furthermore, the minimum number of **kw_min_min** and **kw_min_avg** were -1, which is not rational since these are number of shares. However, since 20,000+ records observed this issue, we can't simply remove them. We decided to keep them and investigate them later after we explored our dataset in more detail.

Third, we decided to see if we have any collinearity problems. We examined the correlation coefficient among all pairs of variables and found that there is high correlation (close to 1) among **n_unique_tokens**, **n_non_stop_words** and **n_non_stop_unique_tokens**, so we kept only **n_unique_tokens**. We also noticed that **kw_avg_avg** has the strongest correlation with **shares**, so we decided that **kw_avg_avg** must be included in our model.

Finally, we looked through the remaining data (46 variables currently) and decided to get rid of the max & min values of some variables because they cannot represent the whole data set. Moreover, there are repetitive or similar variables that roughly explain the same thing, such as **global_rate_negative_words** and **global_sentiment_polarity**, both of which are measures of how polarized the content is. So, we decided to keep only 1 variable for each property. (See Appendix Table 4 for details on what we kept, eliminate and reasons).

Modeling

Now we have cleaned the data based on intuition, correlation analysis, outlier analysis, and dummy variables analysis. This cleaner dataset will prevent us from the effect of outliers and irrelevant variables. We then ran both forward and backward AIC (Appendix Table 5) to determine the combination of variables that will best predict the number of shares.

Model 1: Basic Model

We started from a basic model using number of shares with all the other variables we kept (Summary See Appendix Table 1):

$$
\begin{aligned}
&shares \\
&= kw\_avg\_avg + channel + self\_reference\_avg\_sharess + num\_hrefs \\
&+ global\_subjectivity + n\_tokens\_title + num\_keywords + average\_token\_length \\
&+ global\_sentiment\_polarity + dow + num\_imgs
\end{aligned}
$$

In this model, we noticed that average shares of average keywords (kw_avg_avg) was significant, which aligned with our observation from correlation analysis. In this case, we can interpret that all else hold constant, 1 share average increase of the average shares of average keywords will lead to increase of 0.61 shares. A blogger can use this information to include appropriate keywords to boost shares. All the other variables that were significant in this model were quite self-explanatory and were easy to obtain, such as number of linke (num_hrefs), so the blogger

can use this model to find a baseline to estimate the share level of their articles. However, **channel** and **dow** were not significant, which was out of our expectation. We would like to further investigate these in our next model. Lastly, there are some insignificant variables in this model. If we want to use this model, we need to remove these variables.

Model 2: log(share) Model

According to **Exploratory Data Analysis**, it's better to conduct a log transformation to the **shares** in order to mitigate the effects of the outliers. Therefore, we ran second model selection and regression using **log(shares)** (See Appendix Table2).

$$\log(shares) = channel + kw\_avg\_avg + dow + num\_hrefs + global\_subjectivity + self\_reference\_avg\_shares + kw\_avg\_min + kw\_avg\_max + n\_unique\_tokens + title\_sentiment\_polarity + average\_token\_length + num\_keywords + title\_subjectivity + num\_imgs + global\_sentiment\_polarity + num\_videos$$

More variables were involved in this model, and along with previously included variables, this model better explained our business interest from a more thorough aspect. For example, average shares of best keywords (kw_avg_max) and average shares of worst keywords (kw_avg_min) were included in this model. Along with kw_avg_avg, the blogger can better understand which keywords to include and which keywords to avoid. More importantly, **channel** and **dow** become significant. The business now can better understand what its audiences like. For example, on average, a tech article has 0.119 less shares than a business article.

Model 3: Interaction Model

As mentioned previously, we found **kw_avg_avg** is an interesting variable to interact with **channel**. As a result, we ran a model to interact **kw_avg_avg** with **channel (Appendix Table 3).**

$$\log(shares) = channel + kw\_avg\_avg + channel * kw\_avg\_avg + dow + num\_hrefs + global\_subjectivity + self\_reference\_avg\_sharess + kw\_avg\_min + kw\_avg\_max + n\_unique\_tokens + title\_sentiment\_polarity + average\_token\_length +$$

$num\_keywords + title\_subjectivity + num\_imgs + global\_sentiment\_polarity + num\_videos$

In this model, the interaction between **channel** and **kw_avg_avg** is significant. Now we can interpret how different channels impact the relationship between **kw_avg_avg** and **shares**. For example, the slope of **kw_avg_avg** and **shares** on tech is 0.00024 and on entertainment is 0.000176, so tech has more impact on the relationship between kw_avg_avg and shares. Authors can use this information to better structure their topic and keywords to maximize shares.

As day of week (dow) is also a categorical variable, we interact it with several variables that interest us such as **global_subjectivity** and **global_sentiment_polarity**. Unfortunately, the interaction is not significant. Therefore, we decided not to interact **dow** with any other variables.

Model Comparison and Selection

MODEL	PROS	CONS
Model 1	• Simple model • More Interpretable as dependent variable is shares	• Includes the outliers which effects the fit • Has insignificant variables
Model 2	• Has better R squared • Account for outliers by using log • Model aligns with intuition	• Does not take the interactions between various variables into account
Model 3	• Take the interactions between channel and average shares of average keywords into account to better assist business decision	• Can be complicated to interpret

Finally, we decided to pick model 3 as our final model because:

1. Model 3 takes into consideration of outliers by using log(shares)
2. Model 3 uses interactions between channel and average shares of average key words which aligns with intuition and our graphical analysis
3. Model 3 has better r squared

Evaluation

The main purpose of our project is to solve real business problems in the media industry. We want to help media companies to understand their readers' interests and predict the shares of

their articles by using our model. Therefore, it is important to compare different models and choose the one we are most confident about.

The basic model does not provide as much significant variables as **log(share)** model. There're 11 variables in the basic model and only 5 of them are very significant. Some variables that are not significant in basic model such as **global_sentiment_polarity num_imgs** and **channelSocialMedia** become significant in the log(share)model.

Among 16 variables in the **log(share)** model, most of them are significant at 1% level and they help improve the goodness of fit of our model. As an evidence of our conclusion, the R-squared of the **log(share)** model is also higher than the basic model.

Next in our third model, we added interaction term **channel* kw_avg_avg**. By using this model, we can see if the effects of average share of average keyword on shares would vary according to different channels. Below is R-square and AIC for our three models:

Model	R-square Value	AIC
Basic Model	0.019	719128.9
Log(share) Model	0.108	-10298.7
Interaction Model	0.110	N/A

We can tell that our interaction model has the highest R-square value, which means it explains the most variations in y among the three models. According to our final model, to increase the shares of an article, one should release it on weekends, add more links, keywords, images and videos into the article, which is logical.

Deployment

There are many variables in our model that are statistically significant. However, to apply our results to solve real business problems, companies or websites should focus on certain practical aspects.

For example, our results tell that social media channel has most popularity, everything else equal. This implies that when advertising, websites should publish more articles in social media genre so that most general public could understand and become interested. Additionally, the time of publishing articles does matter. Companies or authors are likely to get more attention if they publish their articles on weekend because there will be more people free at home reading them. Moreover, adding features like images (num_imgs) and links to other articles (num_hrefs) can also bring more content to the article and therefore lead to greater popularity.

However, one should also be cautious when interpreting and deploying our result. For instance, our result indicates articles become more popular as titles of articles become sentimentally polarized. Nonetheless, this does not mean that authors should use the most agitative language just to catch people's eyes and increase popularity. This would lead to ethical and legal issues that could potentially be harmful to the website in the end.

It is also important to keep in mind that this dataset is taken only from one specific media website that does not necessarily represent other websites. There might be different characteristics of readers and value position for other websites. It is risky to adopt the exact same conclusion from this model without considering these new characteristics. For example, some websites are more oriented towards elite engineers who love to read complicated tech articles. Although our result suggests that the technology channel is not significant to number of shares, tech articles might be the only type of article such group of audience would feel like sharing. Therefore, business value will be generated out of our results only if authors and companies know about their individual circumstances and targets.

Appendix

Table 1: Regression Results from Model 1

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.24E+02	1.22E+03	0.51	0.60977	
kw_avg_avg	6.06E-01	5.18E-02	11.686	< 2e-16	***
channelEntertainment	-5.45E+02	2.08E+02	-2.618	0.00884	**
channelLifestyle	-1.04E+02	3.00E+02	-0.346	0.72938	
channelSocial Media	1.57E+02	2.82E+02	0.556	0.57829	
channelTech	-1.35E+02	2.06E+02	-0.657	0.51118	
channelUnknown	1.34E+03	2.38E+02	5.634	1.77E-08	***
channelWorld	-5.90E+02	2.02E+02	-2.924	0.00346	**
self_reference_avg_sharess	1.94E-02	2.43E-03	7.999	1.29E-15	***
num_hrefs	2.79E+01	5.74E+00	4.871	1.12E-06	***
global_subjectivity	3.25E+03	7.39E+02	4.396	1.11E-05	***
n_tokens_title	6.41E+01	2.83E+01	2.26	0.02382	*
num_keywords	7.44E+01	3.21E+01	2.316	0.02058	*
average_token_length	-4.86E+02	2.24E+02	-2.169	0.03007	*
global_sentiment_polarity	-1.28E+03	6.54E+02	-1.951	0.05104	.
dowMonday	4.79E+02	2.10E+02	2.276	0.02287	*
dowSaturday	6.12E+02	2.83E+02	2.166	0.03033	*
dowSunday	2.47E+02	2.72E+02	0.909	0.36318	
dowThursday	-2.16E+01	2.06E+02	-0.105	0.91654	
dowTuesday	-2.33E+01	2.06E+02	-0.113	0.90987	
dowWednesday	1.60E+02	2.05E+02	0.779	0.43611	
num_imgs	1.16E+01	7.78E+00	1.489	0.13642	

*** indicates significant at 0.1% level, ** 1%, * 5%, and "." for 10%.

Table 2: Regression Results from Model 2

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.30E+00	9.04E-02	80.707	< 2e-16	***
channelEntertainment	-2.09E-01	1.65E-02	-12.707	< 2e-16	***
channelLifestyle	-4.76E-03	2.34E-02	-0.204	0.83858	
channelSocial Media	2.54E-01	2.19E-02	11.598	< 2e-16	***
channelTech	1.19E-01	1.59E-02	7.454	9.27E-14	***
channelUnknown	1.10E-01	1.92E-02	5.761	8.43E-09	***
channelWorld	-1.95E-01	1.56E-02	-12.554	< 2e-16	***
kw_avg_avg	1.38E-04	4.99E-06	27.621	< 2e-16	***
dowMonday	-1.19E-02	1.60E-02	-0.742	0.45797	
dowSaturday	2.16E-01	2.15E-02	10.046	< 2e-16	***
dowSunday	2.10E-01	2.07E-02	10.14	< 2e-16	***

dowThursday	-6.59E-02	1.57E-02	-4.194	2.74E-05	***
dowTuesday	-7.47E-02	1.57E-02	-4.775	1.81E-06	***
dowWednesday	-6.98E-02	1.56E-02	-4.462	8.13E-06	***
num_hrefs	4.06E-03	4.61E-04	8.822	< 2e-16	***
global_subjectivity	4.98E-01	5.71E-02	8.721	< 2e-16	***
self_reference_avg_shares	1.74E-06	1.85E-07	9.408	< 2e-16	***
kw_avg_min	-7.99E-05	8.49E-06	-9.406	< 2e-16	***
kw_avg_max	-3.05E-07	4.49E-08	-6.803	1.04E-11	***
n_unique_tokens	-2.14E-01	5.22E-02	-4.088	4.35E-05	***
title_sentiment_polarity	7.13E-02	1.80E-02	3.962	7.43E-05	***
average_token_length	-7.85E-02	1.73E-02	-4.535	5.79E-06	***
num_keywords	1.04E-02	2.62E-03	3.965	7.35E-05	***
title_subjectivity	3.88E-02	1.45E-02	2.675	0.00748	**
num_imgs	1.87E-03	6.43E-04	2.905	0.00368	**
global_sentiment_polarity	-1.15E-01	5.15E-02	-2.225	0.0261	*
num_videos	2.29E-03	1.14E-03	2.011	0.04433	*

*** indicates significant at 0.1% level, ** 1%, * 5%, and "." for 10%.

Table 3: Regression Results from Model 3

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.30E+00	9.31E-02	78.418	< 2e-16	***
channelEntertainment	-3.43E-01	4.14E-02	-8.276	< 2e-16	***
channelLifestyle	1.53E-01	5.79E-02	2.637	0.00837	**
channelSocial Media	3.62E-01	5.45E-02	6.651	2.96E-11	***
channelTech	-5.98E-02	4.75E-02	-1.261	0.207414	
channelUnknown	2.35E-01	4.54E-02	5.17	2.35E-07	***
channelWorld	-2.72E-01	3.99E-02	-6.826	8.89E-12	***
kw_avg_avg	1.36E-04	8.27E-06	16.465	< 2e-16	***
dowMonday	-1.02E-02	1.60E-02	-0.634	0.526209	
dowSaturday	2.17E-01	2.15E-02	10.069	< 2e-16	***
dowSunday	2.10E-01	2.07E-02	10.131	< 2e-16	***
dowThursday	-6.58E-02	1.57E-02	-4.191	2.79E-05	***
dowTuesday	-7.41E-02	1.56E-02	-4.736	2.19E-06	***
dowWednesday	-6.79E-02	1.56E-02	-4.342	1.42E-05	***
num_hrefs	4.11E-03	4.60E-04	8.934	< 2e-16	***
global_subjectivity	5.04E-01	5.70E-02	8.847	< 2e-16	***
self_reference_avg_sharess	1.77E-06	1.86E-07	9.545	< 2e-16	***
kw_avg_min	-8.10E-05	8.57E-06	-9.456	< 2e-16	***
kw_avg_max	-3.21E-07	4.50E-08	-7.145	9.16E-13	***
n_unique_tokens	-2.19E-01	5.22E-02	-4.185	2.86E-05	***

title_sentiment_polarity	7.07E-02	1.80E-02	3.933	8.39E-05	***
average_token_length	-7.77E-02	1.73E-02	-4.485	7.30E-06	***
num_keywords	9.83E-03	2.62E-03	3.747	0.000179	***
title_subjectivity	3.85E-02	1.45E-02	2.656	0.007913	**
num_imgs	2.04E-03	6.44E-04	3.174	0.001502	**
global_sentiment_polarity	-1.18E-01	5.15E-02	-2.285	0.022289	*
num_videos	2.14E-03	1.14E-03	1.882	0.05985	.
channelEntertainment:kw_avg_avg	4.20E-05	1.24E-05	3.39	0.000699	***
channelLifestyle:kw_avg_avg	-4.65E-05	1.61E-05	-2.892	0.003832	**
channelSocial Media:kw_avg_avg	-3.40E-05	1.58E-05	-2.153	0.031293	*
channelTech:kw_avg_avg	6.43E-05	1.59E-05	4.046	5.21E-05	***
channelUnknown:kw_avg_avg	-2.77E-05	1.10E-05	-2.509	0.012097	*
channelWorld:kw_avg_avg	2.98E-05	1.38E-05	2.155	0.031181	*

*** indicates significant at 0.1% level, ** 1%, * 5%, and "." for 10%.

Table 4: Cleaning the Data

Dropped	Kept	Reasons
kw_min_min kw_max_min kw_min_max kw_max_max kw_min_avg kw_max_avg	kw_avg_avg	Extremes cannot represent the whole data set
min_positive_polarity max_positive_polarity min_negative_polarity max_negative_polarity avg_positive_polarity avg_negative_polarity	global_sentiment_polarity title_sentiment_polarity	Extremes cannot represent the whole data set Similar to global_sentiment_polarity and title_sentiment_polarity
num_self_hrefs self_reference_min_shares self_reference_max_shares	num_hrefs	Similar to num_hrefs
global_rate_positive_words global_rate_negative_words rate_positive_words rate_negative_words	global_subjectivity	Similar to global_subjectivity
abs_title_subjectivity abs_title_sentiment_polarity	Title_sentiment_polarity	Similar to sentiment polarity

Table 5: AIC Results

AIC	Basic Model	Log Model	Interaction Model
Backward Selection	719129.1	-10298.72	NA
Forward Selection	719128.9	-10298.72	NA
Stepwise Selection	719128.9	-10298.72	NA