PREDICTING POPULARITY OF ONLINE NEWS ARTICLES

CSC424 - June 6, 2018

Akbar Aidarov
Michael Janke
Monica Stettler
Samiyah Reed
Siravich Khongrod

ABSTRACT
Online news has become increasingly popular in recent years. It is the preferred channel for many to receive their news content. The goal of our research was to build a model that would predict the number of shares of online news article using several exploratory and predictive modeling techniques. The methods utilized for this research include decision trees, random forest, principal component analysis (PCA), factor analysis (CFA), canonical correlation analysis (CCA), correspondence analysis (CA), and linear regression. Our research revealed that strong negative and positive words are the biggest influencers in whether an online news item get shared a lot.

INTRODUCTION

Online news has become increasingly popular in recent years. It is the preferred channel for many to receive their news content. Therefore, there has been much research in this area, trying to help authors and news media companies to create content that gets seen and shared. Popularity is often measured by considering the number of interactions in the Web and social networks (e.g., number of shares, likes and comments). The prediction such popularity metrics is valuable for authors, content providers, advertisers and even activists/politicians (e.g., to understand or influence public opinion) (Bandari et al., 2015). It is our goal to identify the key features that contribute to the popularity of a news item and then to predict the popularity of the item based on the number of shares it receives.

LITERATURE REVIEW

The source of our dataset was the UCI Machine Learning Repository. The original researchers Fernandez et al (2015) donated the dataset after publication of their work. The dataset summarizes a heterogeneous set of features about articles published by Mashable over a period of two years. It is a multivariate dataset with no multiple related tables. There are 39,797 items in the dataset with 61 total attributes, of which 46 are metric and 14 are categorical.

Some of the more interesting metric variables are: the number of words in the title and content, especially the number of non-stop words, the number images, videos, and links and the rates of negative and positive words.

The dataset has only a few categorical variables, all of which are interesting: type of the news channel where the article was posted (lifestyle, entertainment, business, soc/medicine, technology, world), day of the week article was published, and whether the article was published on a weekend. Our literature review found that the digital publishing industry measures engagement by the number of likes, shares and comments, according to Silver (2017).

The researchers, Fernandez et al (2015), who donated our dataset worked to create an optimization model to help authors evaluate their news items *before* publishing them. The model would work in two steps, first to predict whether it would become popular and then to optimize features that could be changed to enhance popularity. They chose a binary assessment of popularity - popular or unpopular. Tatar et al (2012) evaluated user comments as a predictor of

popularity. Bandari et al. (2015) had a similar goal to Fernandez et al. (2015) but applied different techniques. Our approach is to predict the number of times a news item gets shared.

METHODS

The dataset of 61 variables includes 46 numerical variables and 14 categorical variables. The non-metric are actually two categorical variables: data channel and day of week. The numeric variables fell into six main groups: words, links, digital media (images & videos), time, keywords and natural language processing.

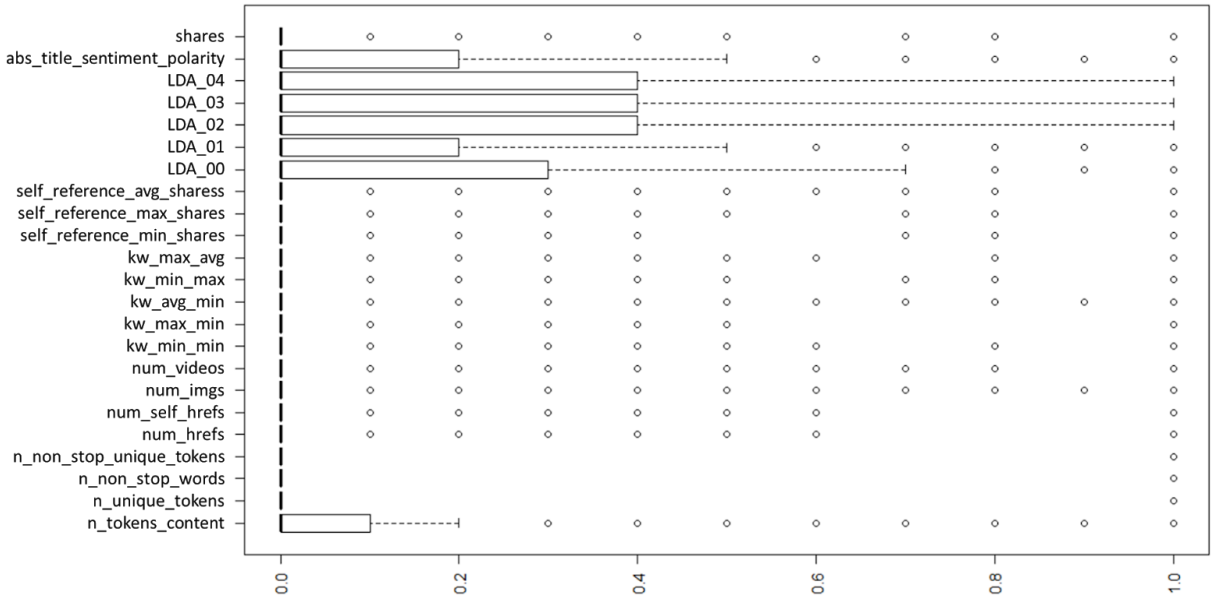
There are no missing values in the dataset, but the dataset did require a fair amount of preprocessing. Many of the features were skewed and thus were logarithmically transformed. There were many outliers that were replaced with the variable mean. Some outliers were removed. Variables needed to be scaled as some were ratios, while others were count values. Some were low count values, and some were extremely high-count values.

Using both the correlation matrix and Bartlett's test of Sphericity indicated that there was significant correlation among the independent variables. Therefore, principal component analysis and factor analysis were also used as pre-processing steps, which revealed interesting patterns and relationships within the data.

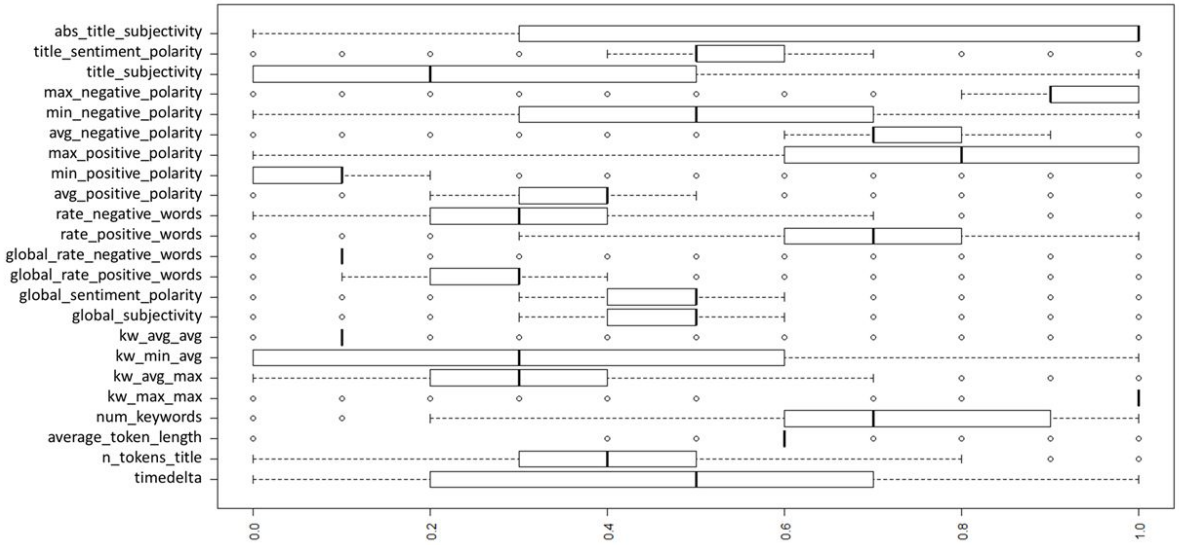
Most of the variables are word count. These are whole numbers, for instance, frequencies of a set of tokens, number of web elements such as videos, images, etc. Other variables are decimal numbers. They are normalized to a scale of zero to one such as similarity measure (LDA), positive/negative words.

The metric variables can be classified as highly skewed and not highly skewed. We defined a threshold by observing the median value of each variables but did not compare median with other statistics for effective visualization. The graphics show that most highly skewed variables contain a large number of outliers.

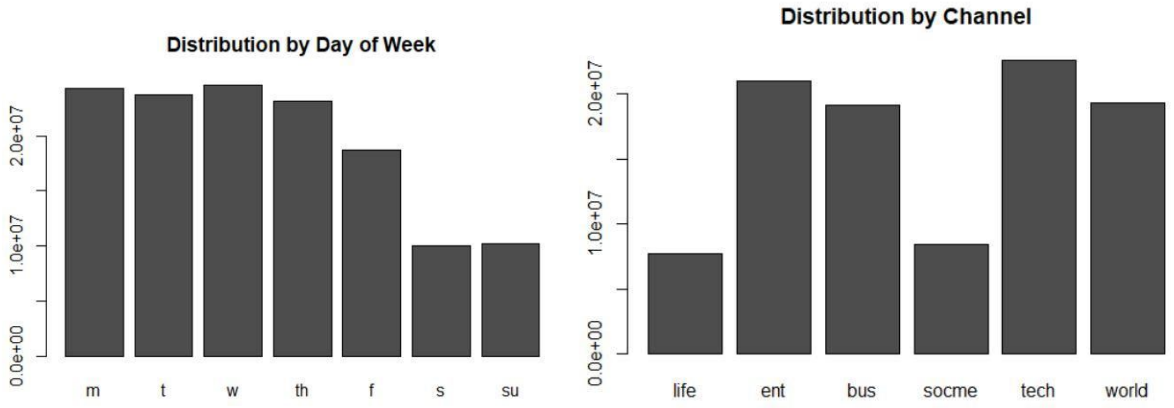
Highly skewed variables



Not Highly Skewed



The distribution of the non-metric variables is shown in the figures below.



Principal Component Analysis

As part of our preprocessing assessment of the dataset, we checked for correlations between the target variable and independent variables and also among the independent variables. The correlations with the target variable mean we have some predictors and the correlations among the independent variables mean we have to manage multicollinearity. PCA is a great tool to uncover latent meaning in our data as well as for managing the multicollinearity.

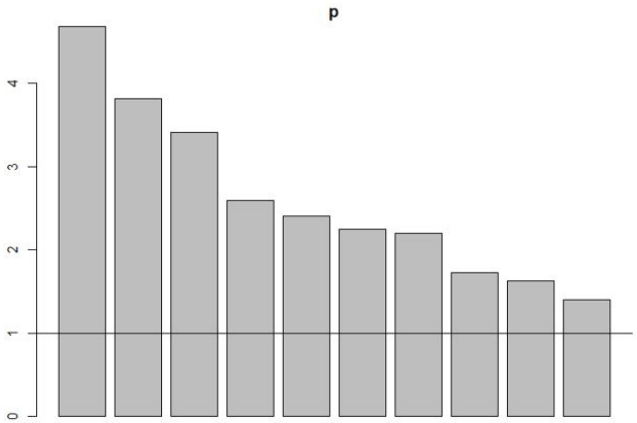
In addition, we ran the Bartlett's test of Sphericity which indicated that that there were significant correlations in the data sufficient to run the PCA.

Before running the PCA, we removed all the categorical variables. As mentioned before, the data was in different scales and so it was important to also scale and center the data. We also removed n_unique_tokent, n_non_stop_words and n_non_stop_unique_tokens as they had too few correlations for this analysis. We use the varimax rotation and cutoff the loadings at a 0.4 significance.

n_tokens_title	28	kw_max_max	30	global_rate_positive_words	33
n_tokens_content	36	kw_avg_max	36	global_rate_negative_words	33
n_unique_tokens	4	kw_min_avg	29	rate_positive_words	39
n_non_stop_words	6	kw_max_avg	34	rate_negative_words	32
n_non_stop_unique_tokens	4	kw_avg_avg	37	avg_positive_polarity	36
num_hrefs	36	self_reference_min_shares	23	min_positive_polarity	30
num_self_hrefs	31	self_reference_max_shares	33	max_positive_polarity	33
num_imgs	38	self_reference_avg_sharess	31	avg_negative_polarity	37
num_videos	34	LDA_00	32	min_negative_polarity	34
average_token_length	41	LDA_01	28	max_negative_polarity	32
num_keywords	33	LDA_02	40	title_subjectivity	32
kw_min_min	31	LDA_03	36	title_sentiment_polarity	29
kw_max_min	13	LDA_04	35	abs_title_subjectivity	21
kw_avg_min	22	global_subjectivity	37	abs_title_sentiment_polarity	32
kw_min_max	28	global_sentiment_polarity	33	shares	24

Table PCA 1 - Number of Significant Correlations per Variable

We considered the three main methods to determine the optimal number of components to use. The Kaiser method indicated that 10 PCs was optimal, however many of those did not add much to the discovery of latent meanings. The scree plot knee showed that 5 PCs would be best. In the end, we chose 6 components based on the percent variance explained. The 6 PCs accounted for 46% of the cumulative variance. Below are the PCs and their loadings with cutoff at .4, along with their descriptive names.



	RC1	RC2	RC3	RC4	RC6	RC5
SS loadings	4.435	3.557	3.156	3.117	2.454	2.443
Proportion Var	0.106	0.085	0.075	0.074	0.058	0.058
Cumulative Var	0.106	0.190	0.265	0.340	0.398	0.456

Figure PCA 1 - Scree Plot and Cumulative Proportion of Variance Explained

Each of the principal components we chose had a theme and meaning to them. Below are our six components with their descriptive names.

RC1 (Positive words)

average_token_length	0.712
global_subjectivity	0.818
global_rate_positive_words	0.598
rate_positive_words	0.626
avg_positive_polarity	0.789
max_positive_polarity	0.708

RC2 (Word Polarity & negativity)

global_sentiment_polarity	0.775
global_rate_negative_words	-0.769
rate_negative_words	-0.888
avg_negative_polarity	0.552
min_negative_polarity	0.631

RC3 (Fewest good keywords)

kw_max_min	0.811
kw_avg_min	0.799

kw_max_avg	0.840
kw_avg_avg	0.769

RC4 (Most good keywords)

kw_min_min	-0.618
kw_max_max	0.646
kw_avg_max	0.842
kw_min_avg	0.539
kw_min_max	0.442

RC6 (Links & images)

n_tokens_content	0.743
num_hrefs	0.605
num_self_hrefs	0.506
num_imgs	0.458

RC5 (Language subjectivity & sentiment – NLP)

self_reference_min_shares	-0.532
self_reference_max_shares	-0.539
self_reference_avg_sharess	-0.618
title_subjectivity	0.622
abs_title_subjectivity	-0.500
abs_title_sentiment_polarity	0.632

Factor Analysis

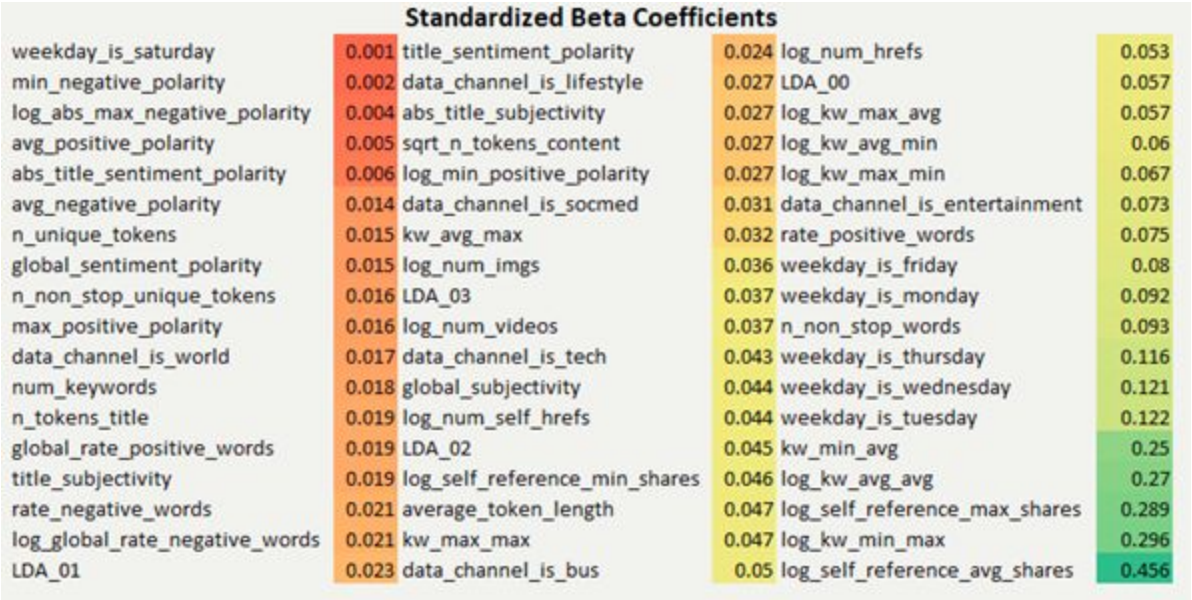
Analysis and Transformation of the Variables

Normal distribution was achieved through transformation of the variables. Most of the variables required log transformation with customized offsets to keep the values > 0 (absolute value in one case). Some required sqrt transformation. The results of this transformation gave us more or less tolerable skewness and kurtosis values.

Checking for Multicollinearity and Assessing Importance of Variables

To get VIF values and standardized parameter estimates, we need to fit a regression model on the transformed data assuming it meets the assumptions of normality. We take out the non-predictive variables and use log_shares as our dependent variable. We also take out the variables, which have undefined coefficients due to singularity.

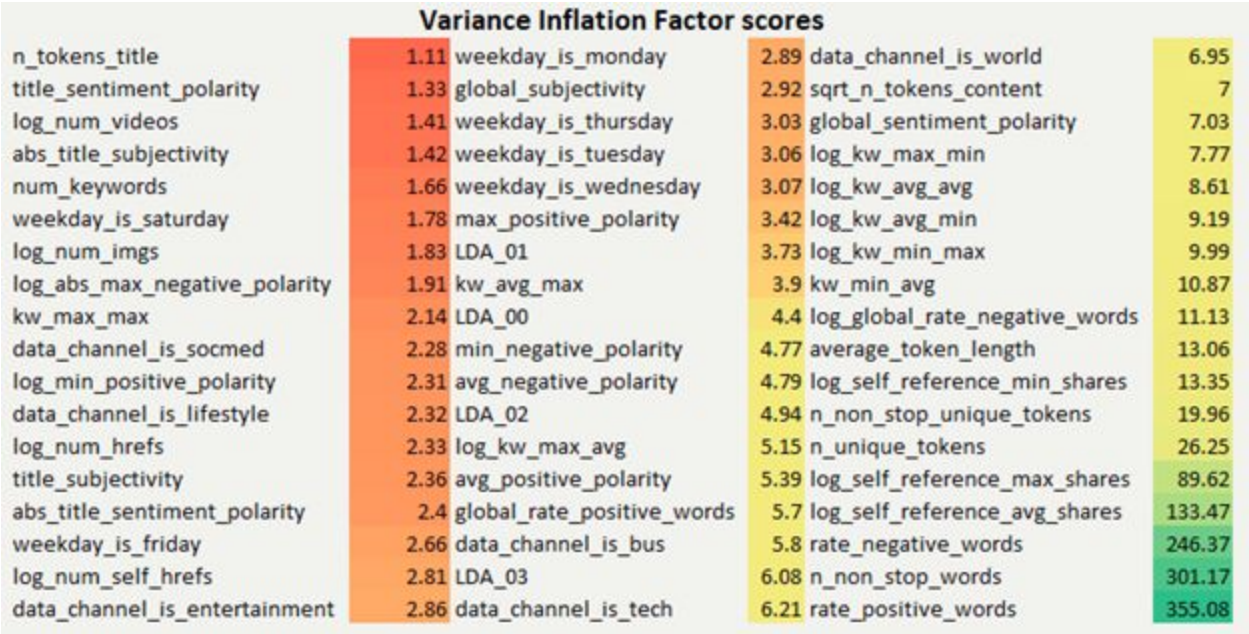
The results show that we have enough multicollinearity amongst variables to conduct factor analysis. The standardized beta coefficients give us a preliminary look into which variables could potentially be important factor constructs.



Standardized Beta Coefficients

weekday_is_saturday	0.001	title_sentiment_polarity	0.024	log_num_hrefs	0.053		
min_negative_polarity	0.002	data_channel_is_lifestyle	0.027	LDA_00	0.057		
log_abs_max_negative_polarity	0.004	abs_title_subjectivity	0.027	log_kw_max_avg	0.057		
avg_positive_polarity	0.005	sqrt_n_tokens_content	0.027	log_kw_avg_min	0.06		
abs_title_sentiment_polarity	0.006	log_min_positive_polarity	0.027	log_kw_max_min	0.067		
avg_negative_polarity	0.014	data_channel_is_socmed	0.031	data_channel_is_entertainment	0.073		
n_unique_tokens	0.015	kw_avg_max	0.032	rate_positive_words	0.075		
global_sentiment_polarity	0.015	log_num_imgs	0.036	weekday_is_friday	0.08		
n_non_stop_unique_tokens	0.016	LDA_03	0.037	weekday_is_monday	0.092		
max_positive_polarity	0.016	log_num_videos	0.037	n_non_stop_words	0.093		
data_channel_is_world	0.017	data_channel_is_tech	0.043	weekday_is_thursday	0.116		
num_keywords	0.018	global_subjectivity	0.044	weekday_is_wednesday	0.121		
n_tokens_title	0.019	log_num_self_hrefs	0.044	weekday_is_tuesday	0.122		
global_rate_positive_words	0.019	LDA_02	0.045	kw_min_avg	0.25		
title_subjectivity	0.019	log_self_reference_min_shares	0.046	log_kw_avg_avg	0.27		
rate_negative_words	0.021	average_token_length	0.047	log_self_reference_max_shares	0.289		
log_global_rate_negative_words	0.021	kw_max_max	0.047	log_kw_min_max	0.296		
LDA_01	0.023	data_channel_is_bus	0.05	log_self_reference_avg_shares	0.456		

Figure CFA 1 - Standardized Beta Coefficients



Variance Inflation Factor scores

n_tokens_title	1.11	weekday_is_monday	2.89	data_channel_is_world	6.95
title_sentiment_polarity	1.33	global_subjectivity	2.92	sqrt_n_tokens_content	7
log_num_videos	1.41	weekday_is_thursday	3.03	global_sentiment_polarity	7.03
abs_title_subjectivity	1.42	weekday_is_tuesday	3.06	log_kw_max_min	7.77
num_keywords	1.66	weekday_is_wednesday	3.07	log_kw_avg_avg	8.61
weekday_is_saturday	1.78	max_positive_polarity	3.42	log_kw_avg_min	9.19
log_num_imgs	1.83	LDA_01	3.73	log_kw_min_max	9.99
log_abs_max_negative_polarity	1.91	kw_avg_max	3.9	kw_min_avg	10.87
kw_max_max	2.14	LDA_00	4.4	log_global_rate_negative_words	11.13
data_channel_is_socmed	2.28	min_negative_polarity	4.77	average_token_length	13.06
log_min_positive_polarity	2.31	avg_negative_polarity	4.79	log_self_reference_min_shares	13.35
data_channel_is_lifestyle	2.32	LDA_02	4.94	n_non_stop_unique_tokens	19.96
log_num_hrefs	2.33	log_kw_max_avg	5.15	n_unique_tokens	26.25
title_subjectivity	2.36	avg_positive_polarity	5.39	log_self_reference_max_shares	89.62
abs_title_sentiment_polarity	2.4	global_rate_positive_words	5.7	log_self_reference_avg_shares	133.47
weekday_is_friday	2.66	data_channel_is_bus	5.8	rate_negative_words	246.37
log_num_self_hrefs	2.81	LDA_03	6.08	n_non_stop_words	301.17
data_channel_is_entertainment	2.86	data_channel_is_tech	6.21	rate_positive_words	355.08

Figure CFA 2 - VIF Scores

Exploring Correlations

The purpose of this analysis is to explore the covariance of the variables in the "Online Popularity" dataset and to identify interpretable meaning in the found components and factors.

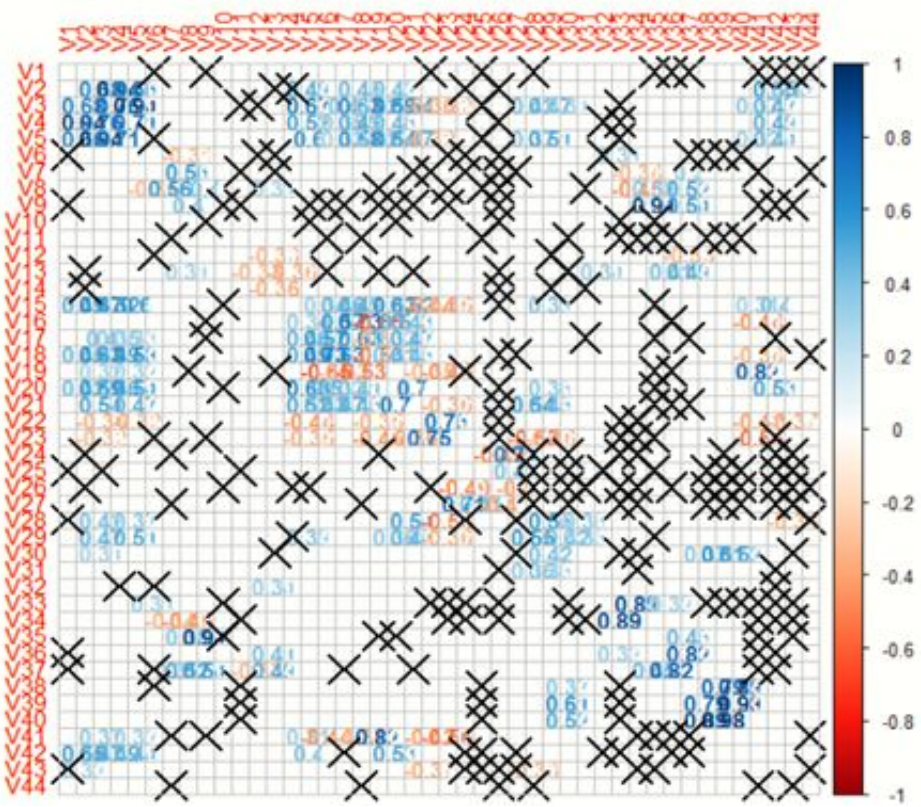


Figure CFA 3 - Strong and Significant Correlations (37 Variables)

We began by evaluating the correlations. Given that we have a total of 44 integer and ratio variables (for simplicity, they were renamed to be "V1" through "V44"), we removed the variables that showed zero significant and strong correlations with other variables ("V1", "V10", "V11", "V12", "V14", "V26", "V44"), which left us with 37 numeric and integer variables. The visualization above is a plot of the correlations that are strong ($r > 0.3$) and significant($p < 0.05$). Insignificant correlations are crossed out.

Bartlett's test of sphericity shows that there is correlation to exploit in the data (K-squared = 30500000, df = 36, p-value < 0.01). KMO measure of sample adequacy indicates stability in factor computations (Overall MSA = 0.72) (see Appendix CFA).

Scree Plot Analysis

We built scree plot to determine the number of factors to extract. It tells us that we could use up to 11 factors, according to Kaiser criterion and Parallel analysis. Acceleration analysis suggests that using only 1 factor is sufficient, whereas Optimal Coordinates analysis shows relevance of using 4 factors.

Taking into account that Factor 6 has only 2 loadings (ideally factors should have at least 3 loadings), we recommend proceeding with 5 factors as a compromise between Acceleration Factor and Kaiser criterion analyses.

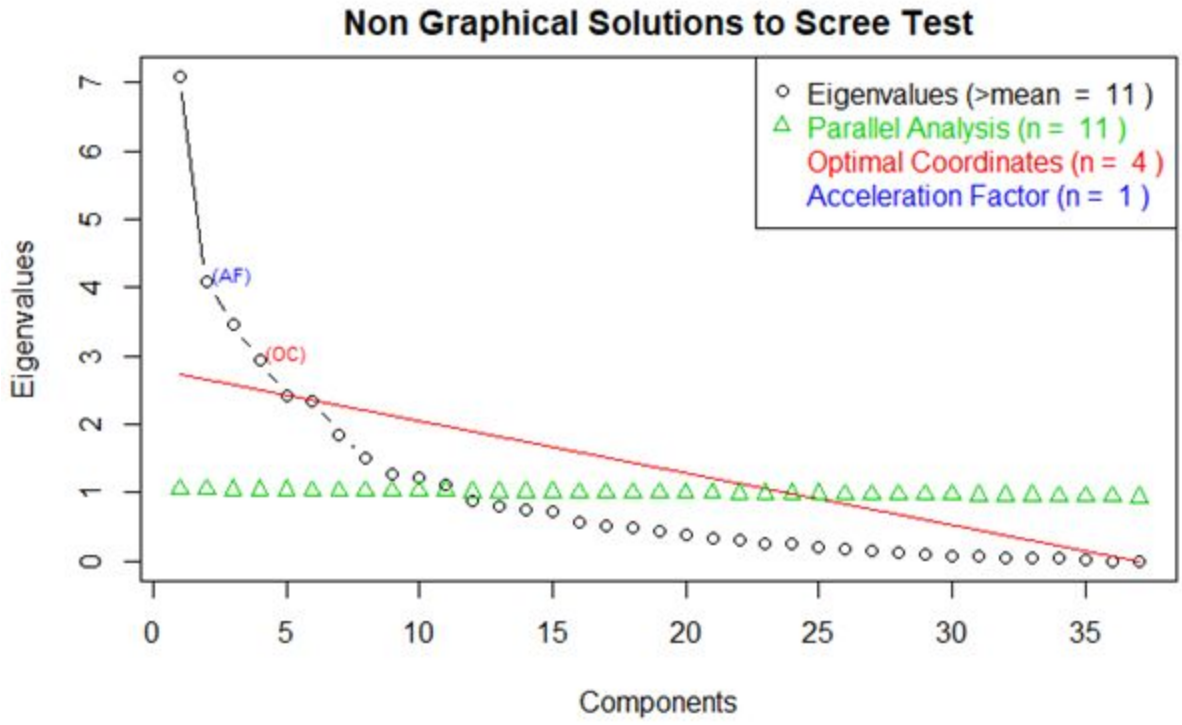


Figure CFA 4 - Scree Plot Analyses

As a result of the Factor Analysis, we were able to establish the following 5 factors and interpret their meaning based on their item loadings:

"Positivity of Non-stop Words" - Factor 1

n_non_stop_words	0.928
n_non_stop_unique_tokens	0.903
average_token_length	0.885
n_unique_tokens	0.875
global_subjectivity	0.642
avg_positive_polarity	0.569
log_min_positive_polarity	0.539

"Positivity of Content" - Factor 2

rate_negative_words	-0.948
global_sentiment_polarity	0.753
rate_positive_words	0.752
log_global_rate_negative_words	-0.748
global_rate_positive_words	0.467

"Popularity and Number of Links to Other Mashable Articles" - Factor 3

log_self_reference_avg_shares	0.963
log_self_reference_max_shares	0.943
log_self_reference_min_shares	0.842
log_num_self_hrefs	0.471

"Popularity of Keywords" - Factor 4

kw_min_avg	0.982
log_kw_min_max	0.927
log_kw_avg_avg	0.565
kw_avg_max	0.436

"Size of the Content" - Factor 5

sqrt_n_tokens_content	0.853
log_num_hrefs	0.478
max_positive_polarity	0.439
log_num_imgs	0.434
min_negative_polarity	-0.426

The uniqueness scores demonstrate that avg_negative_polarity, log_abs_max_negative_polarity, LDA_03, num_keywords, log_kw_max_avg, kw_max_max, title_sentiment_polarity, log_num_videos, log_kw_avg_min, log_kw_max_min, abs_title_sentiment_polarity, and title_subjectivity are too unique to be explained by covariance of the variables (all uniqueness values over 0.8). They are not shown in the output of the analysis because their absolute loadings are less than 0.4. The cumulative proportion of explained variance is 46.2%. As shown by the correlation plot below, the scores of the factors are independent from each other ($p < 0.05$).

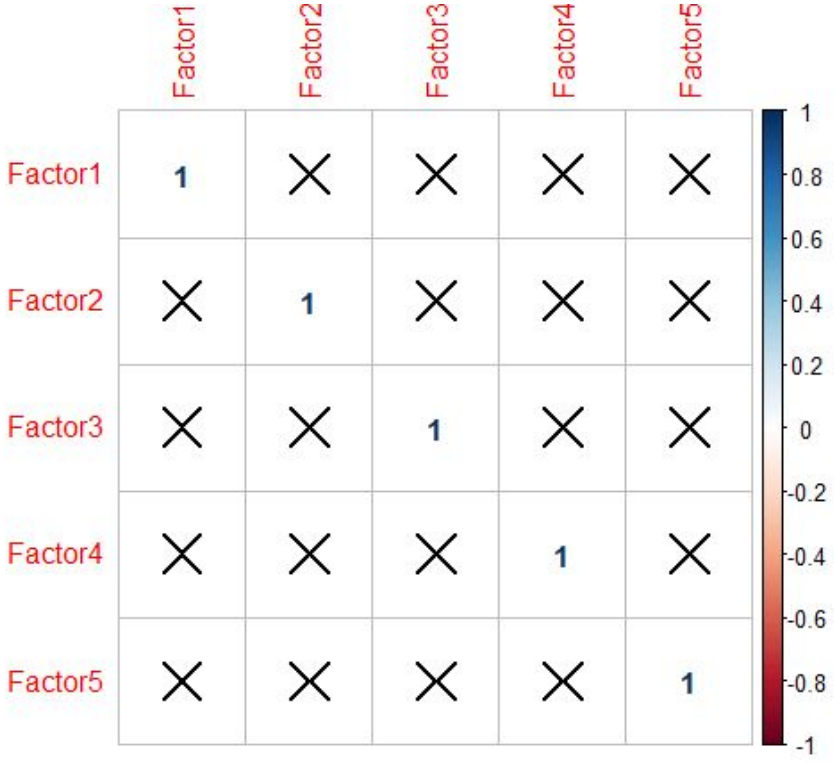


Figure CFA 5 - Correlation of Factor Scores

Canonical Correlation Analysis

Canonical correlation analysis was performed using the components from the PCA as the variates. The intent was to determine if there exists, any predictive relationships between any of the components. In most cases, based upon redundancies, there did not appear to be any such relationship. Table CCA 1 lists the aggregate redundancies for each of the variate pairs.

Component Pairs	X explained by Y	Y explained by X
Positive Words and Word Polarity & Negativity	**0.4893**	**0.6205**
Positive Words and Most Good Keywords	0.0070	0.0099
Positive Words and Fewest Good Keywords	0.0209	0.0177
Positive Words and Links & Images	0.0480	0.0944
Positive Words and Language Subjectivity & Sentiment	0.0207	0.0218
Word Polarity & Negativity and Most Good Keywords	0.0065	0.0087
Word Polarity & Negativity and Fewest Good Keywords	0.0093	0.0130
Word Polarity & Negativity and Links & Images	0.0455	0.1067
Word Polarity & Negativity and Language Subjectivity & Sentiment	0.0097	0.0166
Most Good Keywords and Fewest Good Keywords	0.2377	0.0773
Most Good Keywords and Links & Images	0.0134	0.0167
Most Good Keywords and Language Subjectivity & Sentiment	0.0051	0.0056
Fewest Good Keywords and Links & Images	0.0131	0.0101
Fewest Good Keywords and Language Subjectivity & Sentiment	0.0234	0.0205
Links & Images and Language Subjectivity & Sentiment	0.0274	0.0101

Table CCA 1 - Aggregate Redundancies

In one case, Positive Words and Word Polarity & Negativity, there appears to be a meaningful relationship with 49% of the Positive Words variate explained by the Word Polarity variate and 62% of the Word Polarity variate explained by the Positive Words variate. This pairing was the focus for further examination and referred to as PC1 and PC2 going forward. The relationship between Most Good Keywords and Fewest Good Keywords may warrant further analysis based upon the 0.24 value for Most Good Keywords explained by Fewest Good Keywords. However, this analysis was not performed as part of this research.

To test for significance of the variate pair, the Bartlett Chi-squared test was used. All five variates are significant. The adequacy coefficients are fairly low for each of the variates. But variate one may be adequately significant to justify further examination with values of 0.22 and 0.48. (Figures CCA 1 and CCA 2)

```
          rho^2      Chisq df    Pr(>X)
CV 1 9.9541e-01 2.8366e+05 30 < 2.2e-16 ***
CV 2 7.4610e-01 7.6651e+04 20 < 2.2e-16 ***
CV 3 3.4683e-01 2.3937e+04 12 < 2.2e-16 ***
CV 4 1.1328e-01 7.5580e+03  6 < 2.2e-16 ***
CV 5 7.3481e-02 2.9349e+03  2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure CCA 1 - Bartlett’s Chi-squared results for PC1 and PC2

```
       X Vars:
      CV 1       CV 2       CV 3       CV 4       CV 5
0.22251197 0.26938603 0.13624657 0.09291088 0.12273603

       Y Vars:
      CV 1       CV 2       CV 3       CV 4       CV 5
0.48133647 0.11061474 0.08167948 0.16338389 0.16298542
```

Figure CCA 2 - Canonical Variate Adequacies for PC1 and PC2

Based upon the canonical variate coefficients, in variate one, we see that rate_positive_words has a much larger impact within the component than do the other variables. And similarly, rate negative words has a greater impact than the others. (Figure CCA 3) Looking at the communalities, it appears that average_token_length does not belong which makes some sense considering the subject matter of the other variables (Figure CCA 4) Based upon this finding, average_token_length was removed from the component and the canonical correlation was rerun.

X Vars:

	CV 1	CV 2	CV 3	CV 4	CV 5
average_token_length	-0.0009907451	-0.0249384	0.09610013	-0.008122698	-0.03194641
global_subjectivity	0.0001932682	0.1819429	-0.02912299	-0.352219618	-1.06113701
global_rate_positive_words	0.0025578207	0.7652861	-0.70191931	0.658570774	0.27991853
avg_positive_polarity	0.0005659141	0.4726612	0.92681779	0.714331105	0.28620373
max_positive_polarity	-0.0006972624	0.1054196	-0.20233522	-1.209352414	0.36187372
rate_positive_words	0.9986487231	-0.4789058	0.35838228	-0.143230068	-0.10557149

Y Vars:

	CV 1	CV 2	CV 3	CV 4	CV 5
global_sentiment_polarity	6.872543e-03	1.09433086	1.2051509	-0.001202819	0.23953465
global_rate_negative_words	1.839427e-03	0.94428819	-1.0886455	0.662604982	0.05931234
rate_negative_words	-9.966843e-01	0.02296794	1.7738194	-0.087263907	0.22117317
avg_negative_polarity	-2.471012e-03	-0.25773637	-0.3766110	-0.398997923	1.32795323
min_negative_polarity	-2.212519e-05	-0.12675243	0.2656907	1.337152605	-0.68662787

Figure CCA 3 - Canonical Variate Coefficients for PC1 and PC2

X Vars:

average_token_length	global_subjectivity	global_rate_positive_words	avg_positive_polarity	max_positive_polarity	rate_positive_words
0.07013277	0.99950137	0.99557202	0.99759525	0.99994815	0.99999932

Y Vars:

global_sentiment_polarity	global_rate_negative_words	rate_negative_words	avg_negative_polarity	min_negative_polarity
1	1	1	1	1

Figure CCA 4 - Canonical Communalities for PC1 and PC2

After removing average_token_length, the communalities are now all 1. Also, the adequacy for positive words increased slightly from 0.22 to 0.26. (Figures CCA 5 and CCA 6)

X Vars:

global_subjectivity	global_rate_positive_words	avg_positive_polarity	max_positive_polarity	rate_positive_words
1	1	1	1	1

Y Vars:

global_sentiment_polarity	global_rate_negative_words	rate_negative_words	avg_negative_polarity	min_negative_polarity
1	1	1	1	1

Figure CCA 5 - Canonical Communalities After average_token_length Removed

X Vars:

CV 1	CV 2	CV 3	CV 4	CV 5
0.2663493	0.3154910	0.1593247	0.1114744	0.1473607

Y Vars:

CV 1	CV 2	CV 3	CV 4	CV 5
0.48136649	0.11049478	0.08198159	0.16320263	0.16295451

Figure CCA 6 - Canonical Variate Adequacies After average_token_length Removed

Taking another look at aggregate redundancies, the percentage of variance for positive words explained by Word Polarity & Negativity increased by about 9%, from 0.489 to 0.579. (Figure CCA 7)

```
Aggregate Redundancy Coefficients (Total Variance
Explained by All CVs, Across Sets):

        X | Y: 0.5785872
        Y | X: 0.6201762
```

Figure CCA 7 - Aggregate Redundancies After average_token_length Removed

Looking at the cross loadings, we see very strong correlation between positive and negative word rates and the opposing variate. There are also strong correlations with PC1 for global_sentiment_polarity and global_rate_negative_words. Variates one and two show the strongest associations. In CV1, rate_positive_words and rate_negative_words are exact opposites, suggesting one of the two isn't necessary. Furthermore, PC1 may sufficiently predict PC2, rendering the latter unnecessary. Figure CCA 8 shows the cross loadings.

	CV 1	CV 2	CV 3	CV 4	CV 5
global_subjectivity	0.11906779	0.490185055	0.0614672037	-0.0685226950	-2.117970e-01
global_rate_positive_words	0.53067980	0.574053197	-0.2879794803	0.0561708484	2.104003e-02
avg_positive_polarity	0.06382354	0.583388289	0.4258713703	-0.0119302233	2.792094e-02
max_positive_polarity	0.17422910	0.515828472	0.0759489793	-0.2348185711	8.948441e-02
rate_positive_words	0.99769784	-0.001628149	0.0005483883	-0.0002776258	4.058354e-05

	CV 1	CV 2	CV 3	CV 4	CV 5
global_sentiment_polarity	0.7803195	0.354460154	0.240119101	2.319070e-02	0.0589373468
global_rate_negative_words	-0.7570395	0.407341138	-0.251075863	4.543508e-02	0.0029002597
rate_negative_words	-0.9976893	0.003679837	0.001140723	-9.201818e-05	-0.0001568895
avg_negative_polarity	0.2599989	-0.221144321	0.039146723	1.326256e-01	0.2277757241
min_negative_polarity	0.3883106	-0.267337798	0.136670755	2.687693e-01	0.0666892315

Figure CCA 7 - Cross Loadings After average_token_length Removed

The helio plot gives a visual cue as to the strengths of the correlations between the variables in word polarity and the positive words variate. There are strong negative correlations for rate_negative_words and global_rate_negative_words as well as moderate to strong positive correlations for global_sentiment_polarity and min_negative_polarity. (Figure CCA 8)

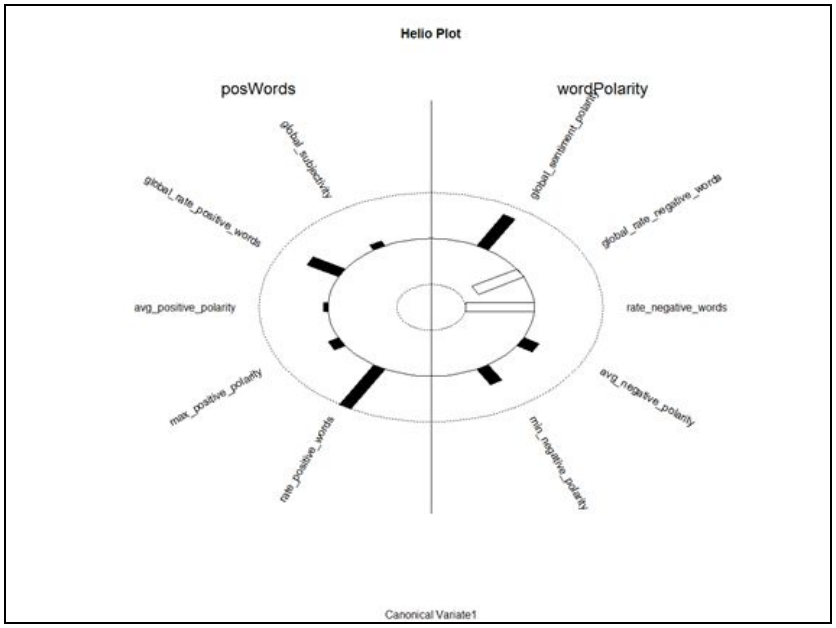


Figure CCA 8 - Helio Plot PC1 and PC2

Conclusions that have been drawn from the canonical correlation analysis are as follows. First, the variable average_token_length appears to be misplaced within the Positive Words variate and could be removed. Second, based upon the aggregate redundancies, there appears to be no predictive relationship between the principal components, other than Positive Words and Word Polarity & Negativity. Third, there is a moderate to strong correlation between Positive Words and Word Polarity & Negativity, suggesting one could be discarded. The redundancy coefficient is greater for Word Polarity & Negativity, so it is the likely candidate. Lastly, rate_positive_words and rate_negative_words are nearly perfectly inversely correlated suggesting that one or the other could be removed.

The canonical correlation results summarized here were obtained using R and the yacca package. See Appendix CCA for full results.

Correspondence Analysis

The opportunity to perform correspondence analysis presented itself as the dataset had 2 sets of categorical variables – Data Channel and Day of week (channel and day). Subjective scaling by visual inspection was used to determine the categorical 2 factors. Detected grouping patterns for channel and day. The dataset contained binary 1s and 0s for the values of each categorical variable. These columns were converted to 2 factors before performing correspondence analysis. The 2 factors were then used as a correlation matrix via a 2-way contingency table, with days as columns and channels as rows. Due to the matrix only 6 variables were allowed – Sunday and Weekend variables did not yield anything, so they were excluded from the correlation matrix. CA figures 1 and 2 show the correlation between the data channels and weekday.

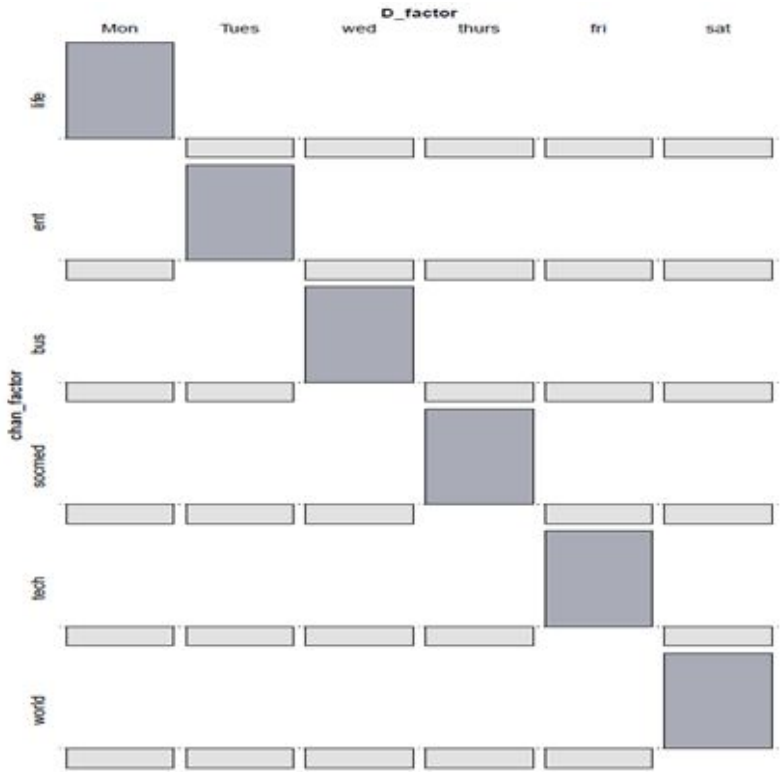


Figure CA 1 - Data Channel by Weekday (Association Plot)

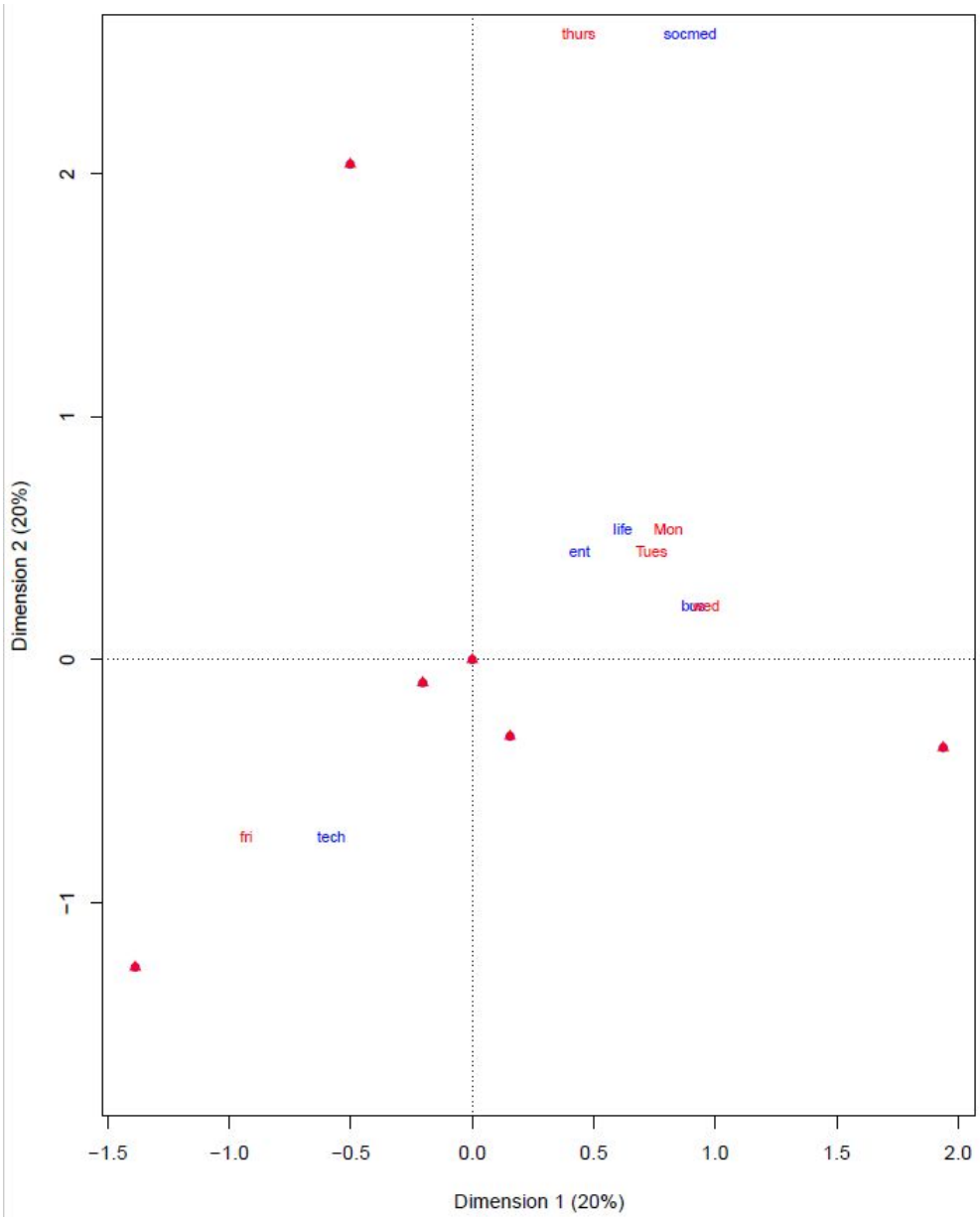


Figure CA 2 - Data Channel by Weekday (CA Plot)

Linear Regression

As our original goal was to predict the number of times an article was shared, we had anticipated that a linear regression would be a useful model. Our pre-work to this point was intended to help with the linear model building.

We ran countless combinations. We used the components from the PCA analysis. We ran with all six and then different combinations. We included the variables that had been removed as not correlated with the others. We dummy coded the categorical variables and gave those a run with the PCs. The models were either not valid or they were not predictive. In the end we used

the transformed dataset from the CFA for the linear regression model. The best model was fit by implementing stepwise model selection in R ("MASS" package). The final model ended up including 45 variables and excluding the following attributes: weekday_is_saturday, min_negative_polarity, rate_negative_words, avg_positive_polarity, n_unique_tokens, n_non_stop_unique_tokens, log_abs_max_negative_polarity, abs_title_sentiment_polarity, and global_sentiment_polarity. See Appendix LM for further details about the model.

Multiple R-squared of the chosen model is 14.6% (Adjusted R-squared = 14.5%). The model is robust with F-statistic of 147 on 45 and 38824 degrees of freedom at p-value < 0.001.

```
Residual standard error: 0.859 on 38824 degrees of freedom
Multiple R-squared:  0.146,     Adjusted R-squared:  0.145
F-statistic:  147 on 45 and 38824 DF,  p-value: <0.0000000000000002
```

Figure LM 1 - Model Statistics

The table of absolute standardized beta coefficients suggest that the information about the performance of references and keywords, as well as the days of the week, are the best predictors of number of shares.

Standardized Beta Coefficients

avg_negative_polarity	0.014	kw_avg_max	0.032	log_kw_max_avg	0.057
max_positive_polarity	0.017	LDA_03	0.037	log_kw_avg_min	0.061
num_keywords	0.018	log_num_imgs	0.037	log_kw_max_min	0.068
data_channel_is_world	0.018	log_num_videos	0.037	n_non_stop_words	0.072
n_tokens_title	0.019	global_subjectivity	0.042	data_channel_is_entertainment	0.073
LDA_01	0.023	average_token_length	0.043	weekday_is_friday	0.08
global_rate_positive_words	0.023	data_channel_is_tech	0.043	weekday_is_monday	0.092
title_subjectivity	0.023	rate_positive_words	0.043	weekday_is_thursday	0.117
title_sentiment_polarity	0.025	log_num_self_hrefs	0.044	weekday_is_wednesday	0.122
abs_title_subjectivity	0.026	LDA_02	0.046	weekday_is_tuesday	0.123
sqrt_n_tokens_content	0.026	log_self_reference_min_shares	0.046	kw_min_avg	0.25
log_global_rate_negative_words	0.026	kw_max_max	0.047	log_kw_avg_avg	0.271
log_min_positive_polarity	0.026	data_channel_is_bus	0.051	log_self_reference_max_shares	0.29
data_channel_is_lifestyle	0.027	log_num_hrefs	0.054	log_kw_min_max	0.295
data_channel_is_socmed	0.031	LDA_00	0.057	log_self_reference_avg_shares	0.457

Table LM 1 - Standardized Beta Coefficients (Linear Regression)

Next we assess the linear model for multicollinearity. The VIF scores suggest that there are a lot of correlations among the variables that are inflating the variance. However, using principal components in the regression did not perform well (R-Squared = 2.8%, see Appendix LM).

Variance Inflation Factor Scores					
n_tokens_title	1.1	max_positive_polarity	2.14	LDA_03	6.04
title_sentiment_polarity	1.14	data_channel_is_socmed	2.28	data_channel_is_tech	6.18
log_num_videos	1.39	log_num_hrefs	2.29	data_channel_is_world	6.88
title_subjectivity	1.41	data_channel_is_lifestyle	2.3	log_kw_max_min	7.76
abs_title_subjectivity	1.41	sqrt_n_tokens_content	2.62	log_kw_avg_avg	8.6
avg_negative_polarity	1.48	global_subjectivity	2.69	log_kw_avg_min	9.17
log_num_imgs	1.56	log_num_self_hrefs	2.8	log_kw_min_max	9.98
num_keywords	1.66	data_channel_is_entertainment	2.85	log_global_rate_negative_words	10.11
weekday_is_friday	1.82	LDA_01	3.71	kw_min_avg	10.86
log_min_positive_polarity	1.84	kw_avg_max	3.9	average_token_length	11.78
weekday_is_monday	1.93	LDA_00	4.39	log_self_reference_min_shares	13.34
weekday_is_thursday	2	LDA_02	4.93	rate_positive_words	19.6
weekday_is_tuesday	2.01	global_rate_positive_words	4.97	n_non_stop_words	25
weekday_is_wednesday	2.02	log_kw_max_avg	5.15	log_self_reference_max_shares	89.38
kw_max_max	2.14	data_channel_is_bus	5.78	log_self_reference_avg_shares	133.21

Table LM 2 - VIF Scores (Linear Regression)

Since our linear regression was disappointing, we considered alternatives. What we really cared about was predicting popularity, in order to help bloggers and journalists get their articles seen and shared. So, we decided to bin the shares and create a classification model that would predict popularity.

Random Forest

To predict the number of times the news gets shared using a classification algorithm, the number of shares was discretized into ten bins. The binning was done using equal frequency so that the classes are not imbalanced. Prior to that, min-max normalization was done on all variables except the number of shares, the number of shares was transformed using logarithm function and outliers that falls out of the range of 3 standard deviations are removed. Note that Random Forest is not prone to different scales. This step was done as an exploratory data analysis in order to see the performance of the classification algorithm with higher number of bins compared to Fernandez et al. (2015).

Shares Range (log scale)	[4.69,6.57)	[6.57,6.77)	[6.77,6.91)	[6.91,7.09)	[7.09,7.24)	[7.24,7.44)	[7.44,7.74)	[7.74,8.1)	[8.1,8.65)	[8.65,10.3]
Count	3903	3910	3343	3485	3758	4052	4778	3986	3883	3999

Table RF 1- Distribution of Discretized No of Shares

The performance of the classification of multiclass shares was not comparable to classification of binary class. The decision tree archived 16.71% accuracy and the Random Forest archived roughly 17% accuracy. Note that the number of trees for random forest was chosen as 50 as the error rate starts to stabilize when iterating this parameter. What was interesting about the result was that there are three series of peaks in the diagonal as shown in the confusion matrix below:

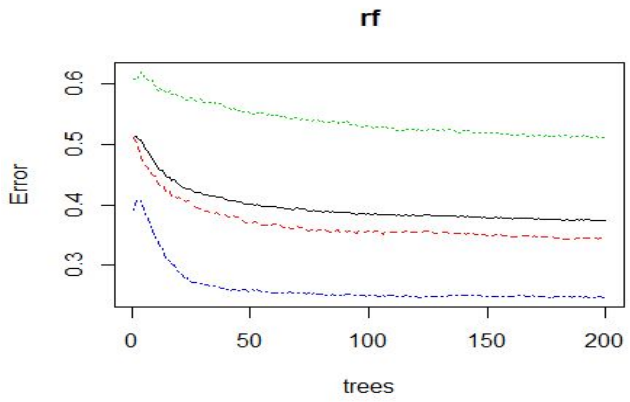


Figure RF 1 - Iterations of Trees for Random Forest

Shares	[4.69,6.57)	[6.57,6.77)	[6.77,6.91)	[6.91,7.09)	[7.09,7.24)	[7.24,7.44)	[7.44,7.74)	[7.74,8.1)	[8.1,8.65)	[8.65,10.3]
[4.69,6.57)	396	292	216	197	182	167	154	92	87	93
[6.57,6.77)	262	258	198	158	146	124	112	80	82	88
[6.77,6.91)	107	113	96	95	67	73	74	52	44	49
[6.91,7.09)	91	100	85	99	84	90	80	60	51	46
[7.09,7.24)	93	106	74	99	118	109	107	74	83	69
[7.24,7.44)	72	88	90	121	142	127	189	127	110	109
[7.44,7.74)	132	150	138	164	191	272	341	304	249	213
[7.74,8.1)	63	62	69	96	105	113	190	162	183	137
[8.1,8.65)	45	80	61	71	117	110	188	197	196	207
[8.65,10.3]	85	108	106	118	137	170	202	190	239	328

Table RF2 - Confusion Matrix of Random Forest (10 Class)

The first peak corresponds to the threshold in Fernandez et al. (2015); D1 = 1400 where the split between the binary class was defined. The second peak is roughly at 2300 shares. From this result, 3 gaussian distributions were inferred. So, a trinary class of target variable was hypothesized to be most appropriate.

To discretize the number of shares into three bins that corresponds to the peaks, K-means discretization was implemented. The class distribution was as imbalanced, so oversampling was used to evenly distribute them. This is a comparison of the original and the oversampled.

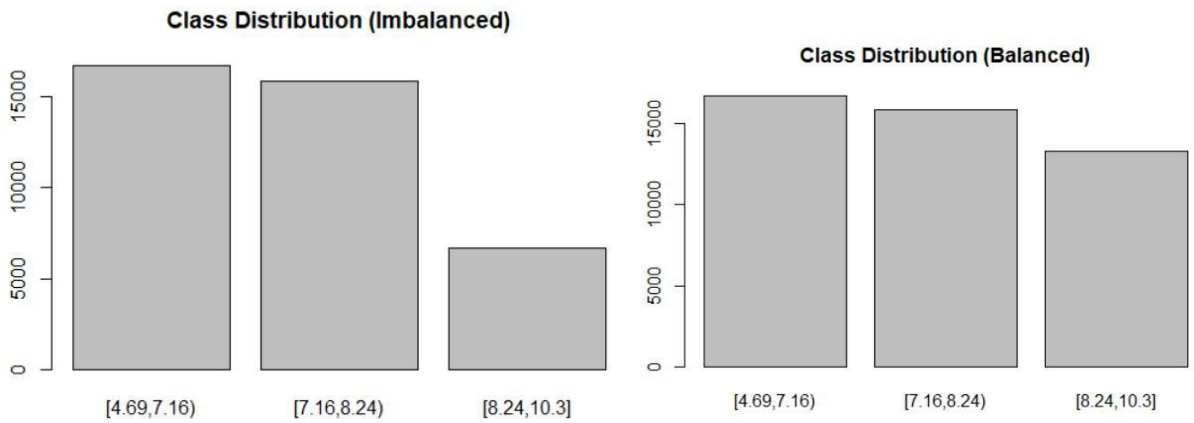


Figure RF 2 - Distribution of Imbalance vs Balanced Trinary Discretization

The result on the trinary classification was more sound and was comparable to Fernandez et al (2015) with the accuracy and precision of 0.62 and 0.62 for trinary class and 0.67 and 0.67 for binary class.

The most important features are listed by class. The frequent items are highlighted.

[4.69,7.16)		[7.16,8.24)		[8.24,10.3]	
kw_avg_avg	0.03169	is_weekend	0.00230	kw_avg_avg	0.135203
kw_max_avg	0.01717	LDA_04	0.00184	kw_max_avg	0.109074
LDA_02	0.01268	global_sentiment_polarity	0.00143	self_reference_min_shares	0.089487
kw_min_avg	0.01220	n_non_stop_unique_tokens	0.00132	self_reference_avg_sharess	0.089358
data_channel_world	0.01201	kw_min_min	0.00115	self_reference_max_shares	0.067520
self_reference_avg_sharess	0.01086	data_channel_is_socmed	0.00104	kw_min_avg	0.057282
self_reference_min_shares	0.01043	min_positive_polarity	0.00088	LDA_04	0.052864
data_channel_entertainment	0.00891	kw_max_max	0.00088	LDA_03	0.049175
is_weekend	0.00825	rate_negative_words	0.00078	LDA_02	0.047093
LDA_01	0.00727	kw_min_max	0.00075	kw_avg_max	0.044792

Table RF 3 - Important Features by Class

Although the attempt to discretize to a relatively high number of bins leading towards mimicking prediction of continuous variable was not successful, this experiment had led to a trinary class prediction which is different from previous work of binary class. The performance is comparable while the classification was able to provide more detailed information in terms of defining the second threshold of news popularity. Perhaps the news can be classified as unpopular, moderately popular and highly popular.

DISCUSSION AND RESULTS

The factor analysis demonstrated five latent predictors that could be investigated deeper to understand what makes news articles popular. Much like principal components, the factors can be described as polarity of the content (positive/negative), the size of the content, and the performance of the keywords and referenced articles.

While the items are providing operationalization, we suggest using summated scales as a way of using the determined latent factors in CFA in further research, because it will reduce measurement error, avoid literal interpretation of factors and make it less sensitive to new data.

These factors would be best tested using more elaborate techniques of Structural Equation Modeling.

There were a number of common threads that we could see throughout various analyses. Information from the PCA and factor analysis was used in the canonical analysis as well as in the decision tree. All these contributed to the selection of respective variables included in the final regression model. The patterns and messages we kept seeing over and over again were related to emotion. Strong positive and strong negative emotions conveyed in word choice were the most influencing factors determining whether an online news items will become popular or not.

The biggest limitation of our research became the lack of information about the variables in the dataset. Many of the attributes are products of feature engineering done by previous researchers. Operating with these attributes without being able to trace back the context and reason under which they were created turned out to be especially challenging.

Another limitation in our research is the definition of popularity across news items when it comes to classification. How many shares do exactly constitute a popular news article? How is popularity measured and what are the “thresholds” of popularity across different channels? How popularity is defined in relation to the time period within which it stays relevant? Should we consider popular and viral to be separate goals? We could address these questions only partially in our analyses and welcome feedback and ideas in further research.

An additional research question was posed: How are days and channels paired to determine the best days to publish channel-specific content? The digital publishing industry places heavy emphasis on online engagement. As discussed in a Forbes.com article by Curtis Silver, web analytics is a major factor in the industry. According to Silver (2017) refers to a study by Kya, a digital publisher firm, that researched the best days and times to post general content to increase page views. Within the context of our research, the articles in the Mashable dataset found the following results: Mon-Lifestyle, Tues-Entertainment, Wednesday-Business, Thursday-Social Media, Friday-Tech, Saturday-World. Future application of this result can be used for bloggers or those in the digital marketing industry to decide when to post subject-specific articles to increase shares and website visits.

CONCLUSION

The purpose of this paper was to evaluate the possibility of predicting popularity of news items by predicting the number of times they are shared online. To this end, almost 40 thousand news articles from Mashable.com were analyzed using the existing dataset donated to the Machine Learning Repository of the University of California - Irvine by Fernandes et al. (2015). Our work used a combination of several different statistical techniques including random forest classification, principal component analysis, exploratory factor analysis, canonical correlation analysis, correspondence analysis, as well as linear regression analysis to understand the key interactions and correlations among the variables that affect the popularity of a news item.

The main conclusion drawn from these techniques is that emotion is the key influence to an online news article being shared by many. Not surprisingly, news articles with very strong positive or negative messages get shared the most. Journalists and bloggers can learn from this

research to be more effective in their writing to motivate readers to share their articles with their networks. Such general conclusion may come as a result of each article being relevant and idiosyncratic to the context of the time period when it was posted. Further, authors need to be aware of the danger, that in an effort to increase a news item's popularity, emotional bias is introduced, clouding the facts.

The further use of the techniques we propose can also be complemented by a timelier measure of public perception and media intelligence techniques that could operationalize and account the user comments under the news items (Tatar et al., 2011).

Future work that needs to be done is to get a better understanding of the types of positive and negative words that have the most impact. In addition, study of the outliers would be worthwhile. While the median of our dataset was 1400 shares, there was a group of outliers in the hundreds of thousands that would be worth digging into.

References:

Fernandes, K., Vinagre, P., & Cortez, P. (2015) A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. *Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence*, September, Coimbra, Portugal.

Tatar, A., Antoniadis, P., De Amorim, M. D., & Fdida, S. (2014). From popularity prediction to ranking online news. *Social Network Analysis and Mining*, *4*(1), 174.

Bandari, R., Asur, S., & Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity. *ICWSM*, *12*, 26-33

Tatar, A., Leguay, J., Antoniadis, P., Limbourg, A., de Amorim, M. D., & Fdida, S. (2011, May). Predicting the popularity of online articles based on user comments. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics* (p. 67). ACM.

Wu, B., & Shen, H. (2015). Analyzing and predicting news popularity on Twitter. *International Journal of Information Management*, *35*(6), 702-711.

Leo, B. (October 01, 2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Silver, C. (2017). Study Reveals Best Time To Publish Online For Those Sweet, Sweet Page Views. *Forbes.com.* Retrieved from https://www.forbes.com/sites/curtissilver/2017/05/23/study-reveals-best-time-to-publish-online-for-those-sweet-sweet-page-views/#4abf86a3f124

Appendix CFA

Bartlett's Chi-Squared test and KMO factor adequacy test for CFA & PCA:

```
        Bartlett test of homogeneity of variances

data:  A_PopNumeric
Bartlett's K-squared = 30500000, df = 36, p-value <0.0000000000000002

Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = A_PopNumeric)
Overall MSA =  0.72
MSA for each item =
  V2   V3   V4   V5   V6   V7   V8   V9  V13  V15  V16  V17  V18  V19  V20  V21  V22  V23  V24
0.72 0.73 0.77 0.93 0.64 0.63 0.77 0.63 0.79 0.96 0.80 0.73 0.66 0.56 0.81 0.90 0.68 0.77 0.58
 V25  V27  V28  V29  V30  V31  V32  V33  V34  V35  V36  V37  V38  V39  V40  V41  V42  V43
0.73 0.57 0.76 0.91 0.66 0.74 0.61 0.52 0.55 0.64 0.66 0.72 0.60 0.59 0.58 0.77 0.86 0.59
```

Factor analysis with 5 factors enforced:

```
Loadings:
                               Factor1 Factor2 Factor3 Factor4 Factor5
n_unique_tokens                 0.875                           -0.471
n_non_stop_words                0.928
n_non_stop_unique_tokens        0.903
average_token_length            0.885
global_subjectivity             0.642
avg_positive_polarity           0.569
log_min_positive_polarity       0.539
global_sentiment_polarity               0.753
rate_positive_words             0.615   0.752
rate_negative_words                    -0.948
log_global_rate_negative_words         -0.748
log_self_reference_min_shares                   0.842
log_self_reference_max_shares                   0.943
log_self_reference_avg_shares                   0.963
kw_min_avg                                              0.982
log_kw_min_max                                          0.927
log_kw_avg_avg                                          0.565
sqrt_n_tokens_content                                           0.853
num_keywords
kw_max_max
kw_avg_max                                              0.436
LDA_03
global_rate_positive_words      0.405   0.467
max_positive_polarity           0.417                           0.439
avg_negative_polarity
min_negative_polarity                                           -0.426
title_subjectivity
title_sentiment_polarity
abs_title_sentiment_polarity
log_num_hrefs                                                   0.478
log_num_self_hrefs                              0.471
log_num_imgs                                                    0.434
log_num_videos
log_kw_max_min
log_kw_avg_min
log_kw_max_avg
log_abs_max_negative_polarity

                Factor1 Factor2 Factor3 Factor4 Factor5
SS loadings       5.734   3.131   3.003   2.766   2.473
Proportion Var    0.155   0.085   0.081   0.075   0.067
Cumulative Var    0.155   0.240   0.321   0.396   0.462
```

Appendix CFA (continued)

Uniqueness of variables:

n_unique_tokens	0.005000
n_non_stop_words	0.005000
rate_positive_words	0.005000
rate_negative_words	0.005000
log_self_reference_avg_shares	0.005000
kw_min_avg	0.010061
log_self_reference_max_shares	0.040997
n_non_stop_unique_tokens	0.086081
average_token_length	0.106507
log_kw_min_max	0.109171
log_self_reference_min_shares	0.221863
sqrt_n_tokens_content	0.227643
log_global_rate_negative_words	0.305260
global_sentiment_polarity	0.370039
global_subjectivity	0.533431
max_positive_polarity	0.598233
global_rate_positive_words	0.602346
avg_positive_polarity	0.632943
log_kw_avg_avg	0.644133
log_num_hrefs	0.646428
min_negative_polarity	0.653118
log_num_self_hrefs	0.664086
log_min_positive_polarity	0.665348
log_num_imgs	0.773210
kw_avg_max	0.779948
avg_negative_polarity	0.801601
log_abs_max_negative_polarity	0.861991
LDA_03	0.880387
num_keywords	0.910215
log_kw_max_avg	0.914129
kw_max_max	0.941212
title_sentiment_polarity	0.960307
log_num_videos	0.967969
log_kw_avg_min	0.977666
log_kw_max_min	0.994095
abs_title_sentiment_polarity	0.995620
title_subjectivity	0.996522	

Appendix CCA

Positive Words and Word Polarity & Negativity

All five of the canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.99541	283660	30	0.0000
CV2	0.74610	76651	20	0.0000
CV3	0.34683	23937	12	0.0000
CV4	0.11328	7558	6	0.0000
CV5	0.07348	2935	2	0.0000

The canonical correlations and shared variance for each variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Positive Words)	Adequacy Y-vars (Word Polarity)	Redundancy X by Y	Redundancy Y by X
CV1	0.9977	0.9954	0.2225	0.4813	**0.2215**	**0.4791**
CV2	0.8638	0.7461	0.2694	0.1106	**0.2010**	0.0825
CV3	0.5890	0.3468	0.1362	0.0817	0.0473	0.0283
CV4	0.3366	0.1133	0.0929	0.1634	0.0105	0.0185
CV5	0.2711	0.0735	0.1227	0.1630	0.0090	0.0120

Positive Words and Most Good Keywords

Three of the five canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0208	1432	30	0.0000
CV2	0.0136	623	20	0.0000
CV3	0.0024	97	12	0.0000
CV4	0.0001	6	6	0.4315
CV5	0.0000	2	2	0.4072

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Positive Words)	Adequacy Y-vars (Most Good Keywords)	Redundancy X by Y	Redundancy Y by X
CV1	0.1442	0.0208	0.1892	0.1713	0.0039	0.0036
CV2	0.1166	0.0136	0.1863	0.4427	0.0025	0.0060
CV3	0.0486	0.0024	0.2316	0.1150	0.0005	0.0003

Positive Words and Fewest Good Keywords

Two of the four canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0808	3421	24	0.0000
CV2	0.0045	182	15	0.0000
CV3	0.0002	8	8	0.4071
CV4	0.0000	1	3	0.7760

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Positive Words)	Adequacy Y-vars (Fewest Good Keywords)	Redundancy X by Y	Redundancy Y by X
CV1	0.2842	0.0808	0.2445	0.2152	0.0197	0.0174
CV2	0.0671	0.0045	0.2459	0.0490	0.0011	0.0002

Positive Words and Links & Images

All four of the canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.2105	14549	24	0.0000
CV2	0.1210	5460	15	0.0000
CV3	0.0102	499	8	0.0000
CV4	0.0027	105	3	0.0000

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Positive Words)	Adequacy Y-vars (Links & Images)	Redundancy X by Y	Redundancy Y by X
CV1	0.4588	0.2105	0.1187	0.3243	0.0250	0.0683
CV2	0.3479	0.1210	0.1649	0.1899	0.0200	0.0230
CV3	0.1010	0.0102	0.2709	0.2447	0.0028	0.0025
CV4	0.0522	0.0027	0.0943	0.2410	0.0003	0.0007

Positive Words and Language Subjectivity & Sentiment

Five of the six canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0531	3864	36	0.0000
CV2	0.0284	1767	25	0.0000
CV3	0.0146	661	16	0.0000
CV4	0.0022	94	9	0.0000
CV5	0.0002	10	4	0.0354
CV6	0.0001	3	1	0.1112

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Positive Words)	Adequacy Y-vars (Language Subjectivity & Sentiment)	Redundancy X by Y	Redundancy Y by X
CV1	0.2304	0.0531	0.2632	0.2970	0.0140	0.0158
CV2	0.1684	0.0284	0.1426	0.1733	0.0040	0.0049
CV3	0.1210	0.0146	0.1675	0.0599	0.0025	0.0009
CV4	0.0467	0.0022	0.0891	0.0944	0.0002	0.0002
CV5	0.0142	0.0002	0.1779	0.3331	0.0000	0.0001
CV6	0.008121875	0.0001	0.1596	0.0424	0.0000	0.0000

Word Polarity & Negativity and Most Good Keywords

Four of the five canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0249	1506	25	0.0000
CV2	0.0102	535	16	0.0000
CV3	0.0033	141	9	0.0000
CV4	0.0003	15	4	0.0048
CV5	0.0001	2	1	0.1577

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Word Polarity)	Adequacy Y-vars (Most Good Keywords)	Redundancy X by Y	Redundancy Y by X
CV1	0.1579	0.0249	0.0594	0.1459	0.0015	0.0036
CV2	0.1009	0.0102	0.4468	0.4464	0.0046	0.0045
CV3	0.0573	0.0033	0.1099	0.1536	0.0004	0.0005
CV4	0.0184	0.0003	0.1419	0.1433	0.0000	0.0000
CV5	0.0072	0.0001	0.2420	0.1107	0.0000	0.0000

Word Polarity & Negativity and Fewest Good Keywords

Three of the four canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0561	2395	20	0.0000

CV2	0.0041	174	12	0.0000
CV3	0.0003	14	6	0.0299
CV4	0.0000	1	2	0.5103

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Word Polarity)	Adequacy Y-vars (Fewest Good Keywords)	Redundancy X by Y	Redundancy Y by X
CV1	0.2369	0.0561	0.1414	0.2250	0.0079	0.0126
CV2	0.0644	0.0041	0.3156	0.0638	0.0013	0.0003
CV3	0.0181	0.0003	0.2847	0.2961	0.0001	0.0001
CV4	0.0059	0.0000	0.1606	0.4152	0.0000	0.0000

Word Polarity & Negativity and Links & Images

All four of the canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.3126	15987	20	0.0000
CV2	0.0299	1570	12	0.0000
CV3	0.0096	404	6	0.0000
CV4	0.0009	35	2	0.0000

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Word Polarity)	Adequacy Y-vars (Links & Images)	Redundancy X by Y	Redundancy Y by X
CV1	0.5591	0.3126	0.1231	0.3152	0.0385	0.0986
CV2	0.1728	0.0299	0.0901	0.1818	0.0027	0.0054
CV3	0.0978	0.0096	0.4427	0.2600	0.0025	0.0025
CV4	0.0302	0.0009	0.1319	0.2429	0.0001	0.0002

Word Polarity & Negativity and Language Subjectivity & Sentiment

All five of the canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0417	2653	30	0.0000
CV2	0.0180	1017	20	0.0000
CV3	0.0048	317	12	0.0000
CV4	0.0030	133	6	0.0000
CV5	0.0005	18	2	0.0001

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Word Polarity)	Adequacy Y-vars (Language Subjectivity & Sentiment)	Redundancy X by Y	Redundancy Y by X
CV1	0.2041	0.0417	0.1430	0.3244	0.0060	0.0135
CV2	0.1343	0.0180	0.0724	0.1292	0.0013	0.0023

CV3	0.0690	0.0048	0.3347	0.0936	0.0016	0.0004
CV4	0.0547	0.0030	0.2614	0.0524	0.0008	0.0002
CV5	0.0216	0.0005	0.1885	0.3524	0.0001	0.0002

Most Good Keywords and Fewest Good Keywords

All four of the canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.5620	5189	20	0.0000
CV2	0.0799	3443	12	0.0000
CV3	0.0049	240	6	0.0000
CV4	0.0014	52	2	0.0000

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Most Good Keywords)	Adequacy Y-vars (Fewest Good Keywords)	Redundancy X by Y	Redundancy Y by X
CV1	0.7497	0.5620	0.3777	0.1278	0.2123	0.0718
CV2	0.2827	0.0799	0.3092	0.0475	0.0247	0.0038
CV3	0.0699	0.0049	0.1073	0.1495	0.0005	0.0007
CV4	0.0368	0.0014	0.1475	0.6752	0.0002	0.0009

Most Good Keywords and Links & Images

All four of the canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0261	2281	20	0.0000
CV2	0.0214	1263	12	0.0000
CV3	0.0105	429	6	0.0000
CV4	0.0006	23	2	0.0000

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Most Good Keywords)	Adequacy Y-vars (Links & Images)	Redundancy X by Y	Redundancy Y by X
CV1	0.1616	0.0261	0.0971	0.3267	0.0025	0.0085
CV2	0.1464	0.0214	0.4180	0.2549	0.0090	0.0055
CV3	0.1025	0.0105	0.1821	0.2427	0.0019	0.0025
CV4	0.0246	0.0006	0.0488	0.1757	0.0000	0.0001

Most Good Keywords and Language Subjectivity & Sentiment

Two of the five canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0148	668	30	0.0000
CV2	0.0019	93	20	0.0000
CV3	0.0003	20	12	0.0663
CV4	0.0002	7	6	0.3232
CV5	0.0000	1	2	0.6639

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Most Good Keywords)	Adequacy Y-vars (Language Subjectivity & Sentiment)	Redundancy X by Y	Redundancy Y by X
CV1	0.1218	0.0148	0.3051	0.3490	0.0045	0.0052
CV2	0.0436	0.0019	0.2687	0.1956	0.0005	0.0004

Fewest Good Keywords and Links & Images

Three of the four canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0520	2310	16	0.0000
CV2	0.0062	258	9	0.0000
CV3	0.0004	18	4	0.0013
CV4	0.0000	1	1	0.2518

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Fewest Good Keywords)	Adequacy Y-vars (Links & Images)	Redundancy X by Y	Redundancy Y by X
CV1	0.2280	0.0520	0.2451	0.1621	0.0127	0.0084
CV2	0.0789	0.0062	0.0396	0.2585	0.0002	0.0016
CV3	0.0208	0.0004	0.2331	0.1936	0.0001	0.0001

Fewest Good Keywords and Language Subjectivity & Sentiment

Three of the four canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0563	2573	24	0.0000
CV2	0.0083	346	15	0.0000
CV3	0.0005	26	8	0.0010
CV4	0.0001	6	3	0.1325

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Fewest Good Keywords)	Adequacy Y-vars (Title Subjectivity & Sentiment)	Redundancy X by Y	Redundancy Y by X
CV1	0.2372	0.0563	0.4069	0.3333	0.0229	0.0188
CV2	0.0910	0.0083	0.0434	0.2056	0.0004	0.0017
CV3	0.0231	0.0005	0.1116	0.1111	0.0001	0.0001

Links & Images and Language Subjectivity & Sentiment

All four of the canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.0589	3557	24	0.0000
CV2	0.0236	1223	15	0.0000

CV3	0.0066	303	8	0.0000
CV4	0.0013	49	3	0.0000

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Links and Images)	Adequacy Y-vars (Language Subjectivity & Sentiment)	Redundancy X by Y	Redundancy Y by X
CV1	0.2427	0.0589	0.3395	0.0519	0.0200	0.0031
CV2	0.1537	0.0236	0.2488	0.2120	0.0059	0.0050
CV3	0.0813	0.0066	0.1879	0.2431	0.0012	0.0016
CV4	0.0356	0.0013	0.2239	0.3815	0.0003	0.0005

Positive Words and Word Polarity & Negativity Redux

All five of the canonical variates were deemed significant based on the Bartlett's Chi-Squared Test.

Bartlett's Chi-Squared Test:

	Canonical R^2	Chi-Square	df	P-value
CV1	0.9954	283410	25	0.0000
CV2	0.7457	76402	16	0.0000
CV3	0.3438	23753	9	0.0000
CV4	0.1133	7555	4	0.0000
CV5	0.0734	2932	1	0.0000

The canonical correlations, shared variance, and redundancy for each significant variate:

	Canonical Correlation	Shared Variance	Adequacy X-vars (Positive Words)	Adequacy Y-vars (Word Polarity & Negativity)	Redundancy X by Y	Redundancy Y by X
CV1	0.9977	0.9954	0.2663	0.4814	0.2651	0.4792
CV2	0.8635	0.7457	0.3155	0.1105	0.2352	0.0824
CV3	0.5863	0.3438	0.1593	0.0820	0.0548	0.0282
CV4	0.3366	0.1133	0.1115	0.1632	0.0126	0.0185
CV5	0.2709	0.0734	0.1474	0.1630	0.0108	0.0120

Appendix LM

Stepwise selection - linear regression model output:

```
    Min      1Q  Median      3Q     Max
-8.075  -0.543  -0.153   0.383   5.907

Coefficients:
                                Estimate    Std. Error t value             Pr(>|t|)
(Intercept)                   2.7467074304  0.2575351915   10.67 < 0.0000000000000002 ***
n_tokens_title                0.0083946509  0.0021674077    3.87              0.00011 ***
n_non_stop_words             -0.3913060243  0.1278603927   -3.06              0.00221 **
average_token_length         -0.0474946864  0.0176790578   -2.69              0.00722 **
num_keywords                  0.0091028182  0.0029769278    3.06              0.00223 **
data_channel_is_lifestyle    -0.1139627625  0.0295693481   -3.85              0.00012 ***
data_channel_is_entertainment -0.1778391573 0.0191993918   -9.26 < 0.0000000000000002 ***
data_channel_is_bus          -0.1291757542  0.0286405126   -4.51    0.000006494194888 ***
data_channel_is_socmed        0.1228323432  0.0279606772    4.39    0.000011207274748 ***
data_channel_is_tech          0.1018704786  0.0277808878    3.67              0.00025 ***
data_channel_is_world        -0.0402235936  0.0282242492   -1.43              0.15412
kw_max_max                   -0.0000002057  0.0000000298   -6.90    0.000000000005240 ***
kw_avg_max                   -0.0000002232  0.0000000645   -3.46              0.00054 ***
kw_min_avg                    0.0002045445  0.0000126441   16.18 < 0.0000000000000002 ***
weekday_is_monday            -0.2299606197  0.0162489310  -14.15 < 0.0000000000000002 ***
weekday_is_tuesday           -0.2924644427  0.0158592660  -18.44 < 0.0000000000000002 ***
weekday_is_wednesday         -0.2897694075  0.0158586183  -18.27 < 0.0000000000000002 ***
weekday_is_thursday          -0.2796916560  0.0159050514  -17.59 < 0.0000000000000002 ***
weekday_is_friday            -0.2122593817  0.0167863106  -12.64 < 0.0000000000000002 ***
LDA_00                        0.1998553446  0.0345936307    5.78    0.000000007651952 ***
LDA_01                       -0.0969653233  0.0381456275   -2.54              0.01103 *
LDA_02                       -0.1514346829  0.0346315693   -4.37    0.000012301834235 ***
LDA_03                       -0.1156073970  0.0361160321   -3.20              0.00137 **
global_subjectivity           0.3315340008  0.0611540136    5.42    0.000000059520616 ***
global_rate_positive_words   -1.2402371933  0.5576060635   -2.22              0.02614 *
rate_positive_words           0.2100862541  0.1014358431    2.07              0.03835 *
max_positive_polarity        -0.0648431994  0.0256803117   -2.53              0.01157 *
avg_negative_polarity        -0.0989866537  0.0414891315   -2.39              0.01704 *
title_subjectivity            0.0659667468  0.0159748759    4.13    0.000036446166976 ***
title_sentiment_polarity      0.0860420313  0.0175630700    4.90    0.000000966974133 ***
abs_title_subjectivity        0.1289642780  0.0273902096    4.71    0.000002505205558 ***
sqrt_n_tokens_content         0.0026367985  0.0007764311    3.40              0.00068 ***
log_num_hrefs                 0.0615246841  0.0081389591    7.56    0.000000000000041 ***
log_num_self_hrefs           -0.0590062441  0.0105266949   -5.61    0.000000020919291 ***
log_num_imgs                  0.0355538036  0.0055758928    6.38    0.000000000183401 ***
log_num_videos                0.0504685036  0.0075116578    6.72    0.000000000018589 ***
log_kw_max_min               -0.0777914976  0.0149788479   -5.19    0.000000207501997 ***
log_kw_avg_min                0.0804541072  0.0186946785    4.30    0.000016846850161 ***
log_kw_min_max               -0.0807562200  0.0040511497  -19.93 < 0.0000000000000002 ***
log_kw_max_avg               -0.1213025987  0.0224782057   -5.40    0.000000068368207 ***
log_kw_avg_avg                0.7473933422  0.0379852607   19.68 < 0.0000000000000002 ***
log_self_reference_min_shares -0.0286929558 0.0106360987   -2.70              0.00699 **
log_self_reference_max_shares -0.1696480345 0.0259471297   -6.54    0.000000000063026 ***
log_self_reference_avg_shares  0.2994815800 0.0355110051    8.43 < 0.0000000000000002 ***
log_global_rate_negative_words 0.0721657524 0.0420377465    1.72              0.08604 .
log_min_positive_polarity    -0.0494539186  0.0120277662   -4.11    0.000039365546351 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.859 on 38824 degrees of freedom
Multiple R-squared:  0.146,     Adjusted R-squared:  0.145
F-statistic:  147 on 45 and 38824 DF,  p-value: <0.0000000000000002
```

Appendix LM (continued)

Linear regression model using PCA components - output:

```
Coefficients:
               Estimate Std. Error t value             Pr(>|t|)
(Intercept)      3281.5       39.9   82.34 < 0.0000000000000002 ***
p2$scores[, 1]    268.0       39.9    6.72        0.000000000018 ***
p2$scores[, 3]   1120.8       39.9   28.12 < 0.0000000000000002 ***
p2$scores[, 4]    693.3       39.9   17.40 < 0.0000000000000002 ***
p2$scores[, 6]    110.3       39.9    2.77                0.0056 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7940 on 39639 degrees of freedom
Multiple R-squared:  0.0281,     Adjusted R-squared:  0.028
F-statistic:  287 on 4 and 39639 DF,  p-value: <0.0000000000000002
```