IEOR 4650 Business Analytics Group Project

PREDICTING NEWS POPULARITY IN SOCIAL NETWORKS

Team 15
Pearly Ang ha2488
Darren Lu dl3228
Aries Li jl5239
Xiaosu Qi xq2183
Xavier Sallent xs2350

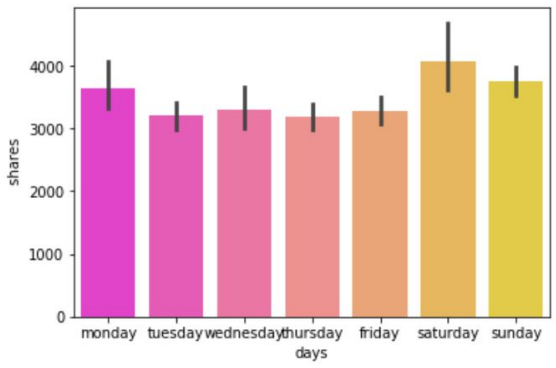
1. Abstract

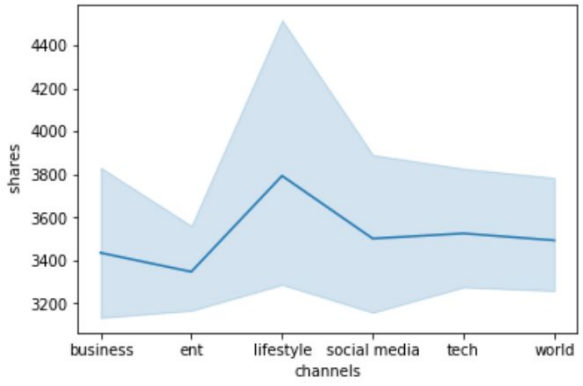
The project aims to extract value and opportunity from the publication of online news articles by maximizing the popularity of articles, measured by the number of shares on social media, to optimize the content creation process for delivering the most benefit at the least cost. Such a tool will help publishers and editors in maximizing the popularity of their articles and in seizing the best pricing opportunities for digital advertisement. The dataset[1] contains features describing the sentiment of the text and title after Natural Language Processing (NLP), the characteristics of the word features and the article, the dependent feature of number of article shares and other ancillary information variables for the articles on Mashable.com during 2013-2015.

2. Exploratory Data Analysis

In the correlation matrix[2], we have a preliminary look at the relationships that the various features have with each other. Ostensibly, the features with high negative and positive correlations tend to be variables with similar fields such as minimum, maximum, average words, negative words and sentiment polarity and scores by LDA topic. However, we also observe interesting relationships such as the relationship between keyword tokens the range for the number of shares of the article and, the absolute title subjectivity and sentiment polarity scores.

Further, we explored several hypotheses on the relationships between certain features. For instance, there might be a relationship between the day of the week and the number of shares for an article, or the type of category and the popularity of an article. As seen below, the proportion of shares increase over Friday and the weekend, implying that articles on 'off' days benefit from a sizably larger rate of interest, and we note that lifestyle articles garner most interest.





We explored other hypotheses as well, and the aggregated results has some clear implications on resource allocation and article formulation. From observing the data, we

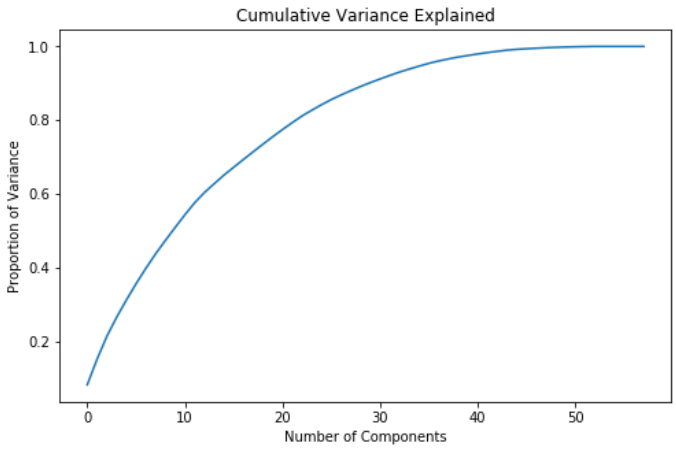
[1] Refer to Table 1 in Appendix A for the full table of features and their types.
[2] Refer to Chart 1 in Appendix A for the correlation matrix.

see clear relationships between individual factors and the popularity of an article, precipitating some preliminary business insights. In order to improve the popularity of the article, journalists for Mashable.com ought to increase the amount of multimedia attachments (e.g. images, videos), boost the number of clickable references to other articles or sources, expand the amount of subjectivity in title, content, and utilize more popular words. On the flipside, it might be a better idea to reduce the number of lengthy words, avoid multi-topic articles and leave out words with negative connotations.

However, to assess how a combination of these factors would affect the popularity of article, we will further explore methods for feature engineering and feature vector selection (PCA) as well as run machine learning models to improve our proposed recommendations for maximizing the impact of our articles.

3. Principal Component Analysis

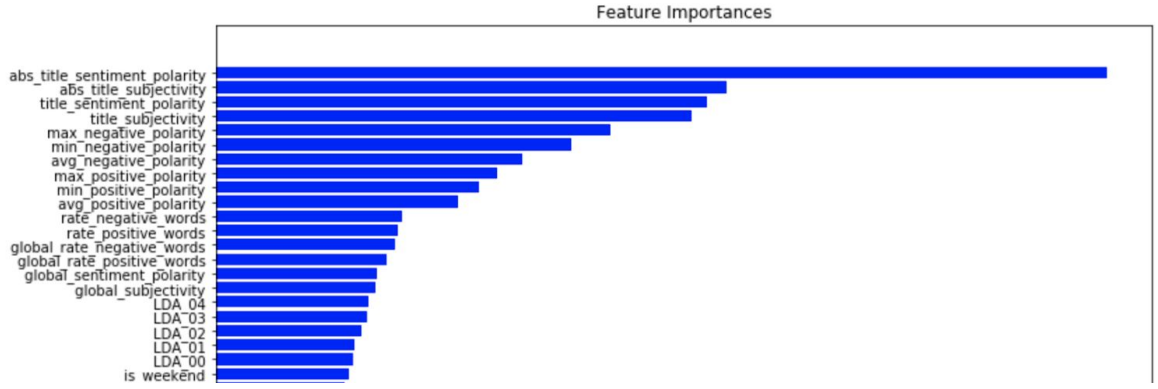
Our original dataset contains 59 features including the dependent variable (number of shares per article). Since this is a large number, we believe it might be useful to run a Principal Component Analysis to understand by how many variables could our data be explained. As a reminder, PCA analysis accounts only for the variance within the data, without taking into account what is the dependent variable. Below there is a plot with the cumulative variance explained when more principal components are added:



The plot shows an almost linear relationship between the number of components and the cumulative variance explained, which is not good news. This means that in order to achieve a 95% explained variance in the independent set, we still need around 35 variables, which is a significant reduction from our 58 initial ones but still has high dimensionality.

4. Feature Importance

By implementing classification tree and classification random forest, we can find the feature importances of our model. Both methods return the same rankings of feature importance. The top 10 are shown in the graph below.



The training accuracy of classification tree is 0.637 and the testing accuracy is 0.640. In the random forest, the accuracy of both training set and testing set are 0.661. By looking at the features names, we suspect that these variables are highly correlated. By looking at the pair correlations between variables, some of these variables do highly correlate with each other, such as 'abs_title_subjectivity', and 'title_subjectivity'. After deleting those highly correlated variables however, the performances of both algorithms got worse. The accuracy drops down to around 0.6, which is surprising. A possible explanation is that the original model actually did not overfit the data. Thus even though some of the features are highly correlated, deleting some of them could negatively impact our capability to explain the variance of the number of shares of articles.

By looking at the correlation between theses important features and the number of shares, we can synthesize some findings. Articles with high shares tend to have a title with strong sentiment words and people are more likely to share those with positive sentiments. A title with strong subjectivity is more attractive, and publishing an article on the weekend can help to increase the shares.

5. Machine Learning Models and Model Selection

Regression

In the regression models part, we try 7 different models using all features for the prediction in the future. We selected Linear Regression Model, Lasso and Ridge regressions for us to choose the best model from.

At the beginning, 75% data was used for training the model, and 25% was reserved for testing using Python libraries. We then built each model for our training data using the scikit-learn library.To compare the performance of each model, we conducted k-fold cross-validation for each model, with k = 10. Following that, we calculated the mean of

RMSE for each fold and obtained the final RMSE for the model comparison. For Lasso and Ridge penalties, we also determine the best parameter lambda for each of those two. In the end, we get the RMSE table below.

Algorithm	RMSE (After k fold cross-validation with k = 10)
Linear Regression	11760.99
Lasso	11656.98
Ridge	11024.96

After comparing the results above, we decided to use Ridge Regression as our final model, the RMSE we ran on the test dataset is 10659.10.

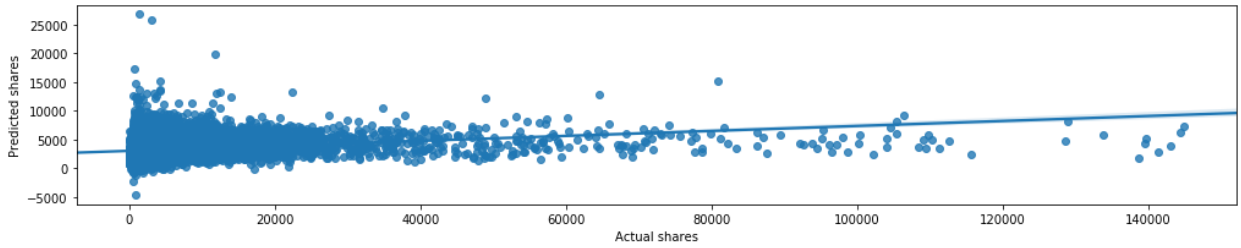


figure 1: Linear Regression

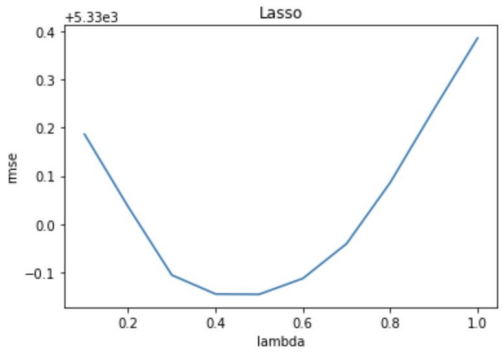


figure 2: Linear Regression

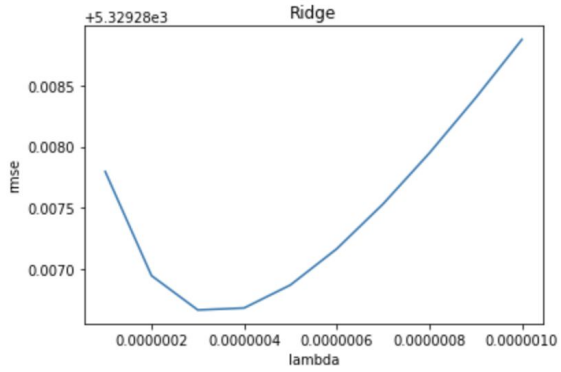


figure 3: Linear Regression

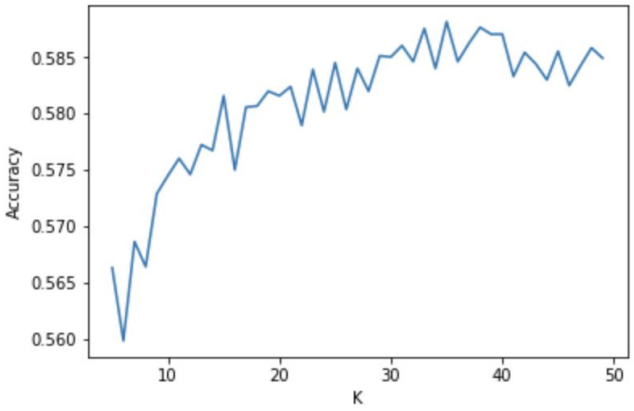
Classification

With classification models, we took a threshold of 1400 shares (where 1400 is the median of all the 36879 articles) as our criterion, which means we identify the news with number of shares greater than 1400 as 1 and others with shares less than 1400 as 0. We implemented four different algorithm for classification: logistic regression, decision tree, random forest and K-nearest neighbors.

The table of results below show the outcome of a 3-fold cross validation approach.

Algorithm	Accuracy
Logistic Regression	0.599
Decision Tree	0.637
Random Forest	0.661
K-Nearest Neighbour (k=30)	0.584

As seen in the graph, for KNN, the best K we found is 30, with an accuracy score about 0.584.



After tuning the hyper parameter using cross validation, we decided to use Random Forest as our final choice for our classification model, with a maximum depth of 8, 70 estimators, a minimum samples split of 8 and a minimum samples leaf of 4. The test accuracy is 0.661.

6. Business Insights / Conclusions

We have seen that our best models reach a maximum performance of 66.1% in new unseen data. These results are rather weak, taking into account that we have around a 50% of positive cases in the dataset.

However, we can still extract some business insights from the correlation matrix. When more complex techniques fail, as they have in our case, a rather simple look at correlations can be used to extract meaningful conclusions. We see there are some weak but still statistically relevant correlations between the number of shares and some of the features. In order to as to raise popularity, Mashable editors should maximize:

- News on LDA topic number 3. Scraping the web for their actual content and drawing a word cloud (refer to the figure on the right), we notice a common pattern that news about the Winter Olympics in Sochi (Russia) in 2014 attracted an above average interest in Mashable readers.

- The number of links in a page
- Readers are attracted to images; the more the better
- The more subjectivity in both the title and the content the better. This is understandable, we have already seen how scandals and “fake news” usually travel easier and faster through the internet.
- Put strong positive sentiment words in titles.
- Try to publish the articles on weekend if possible.

On the other hand, editors should minimize:

- News on LDA topic number 2. Looking at its word cloud, it suggests news about science are less shared
- Too many words in the title have a negative impact in the number of shares
- Categorizing news in the “World” section

7. Limitations

- News have been analyzed independently. Some news may rank high in our model but could be “eclipsed” by others published in the same day.
- Readers could face saturation, especially if a website only focuses on a specific topic which is supposed to attract more attention.
- More popular is not always better. In the long term, news agencies could incur in reputational costs if they only produce content to be shared rather than taking a more responsible and objective approach.
- More importantly, we note that the best performing model only yields an accuracy of about 70%, which is not as ideal as what we initially set out to do, taking into account that the positive / negative split is 50%. However, we found that similarly published research papers achieved similar accuracy (70%).

9. Appendix

Feature	Type (#)
Words	
Number of words in the title	number (1)
Number of words in the article	number (1)
Average word length	number (1)
Rate of non-stop words	ratio (1)
Rate of unique words	ratio (1)
Rate of unique non-stop words	ratio (1)
Links	
Number of links	number (1)
Number of Mashable article links	number (1)
Minimum, average and maximum number of shares of Mashable links	number (3)
Digital Media	
Number of images	number (1)
Number of videos	number (1)
Time	
Day of the week	nominal (1)
Published on a weekend?	bool (1)

Feature	Type (#)
Keywords	
Number of keywords	number (1)
Worst keyword (min./avg./max. shares)	number (3)
Average keyword (min./avg./max. shares)	number (3)
Best keyword (min./avg./max. shares)	number (3)
Article category (Mashable data channel)	nominal (1)
Natural Language Processing	
Closeness to top 5 LDA topics	ratio (5)
Title subjectivity	ratio (1)
Article text subjectivity score and its absolute difference to 0.5	ratio (2)
Title sentiment polarity	ratio (1)
Rate of positive and negative words	ratio (2)
Pos. words rate among non-neutral words	ratio (1)
Neg. words rate among non-neutral words	ratio (1)
Polarity of positive words (min./avg./max.)	ratio (3)
Polarity of negative words (min./avg./max.)	ratio (3)
Article text polarity score and its absolute difference to 0.5	ratio (2)

Target	Type (#)
Number of article Mashable shares	number (1)

Table 1: Features in the dataset by group. Source: Fernandes et. al.

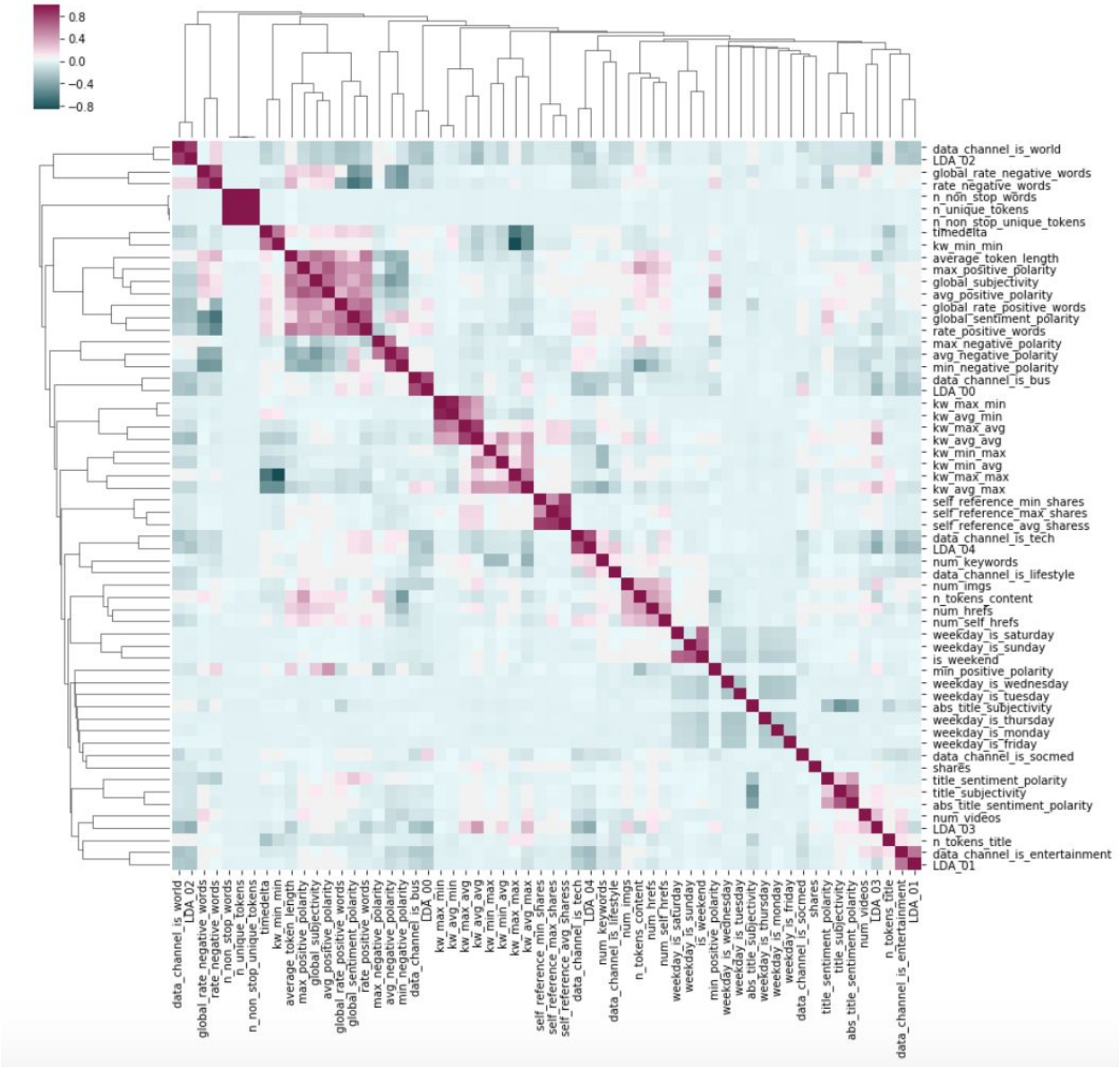


Chart 1: Correlation matrix of all of the feature variables.

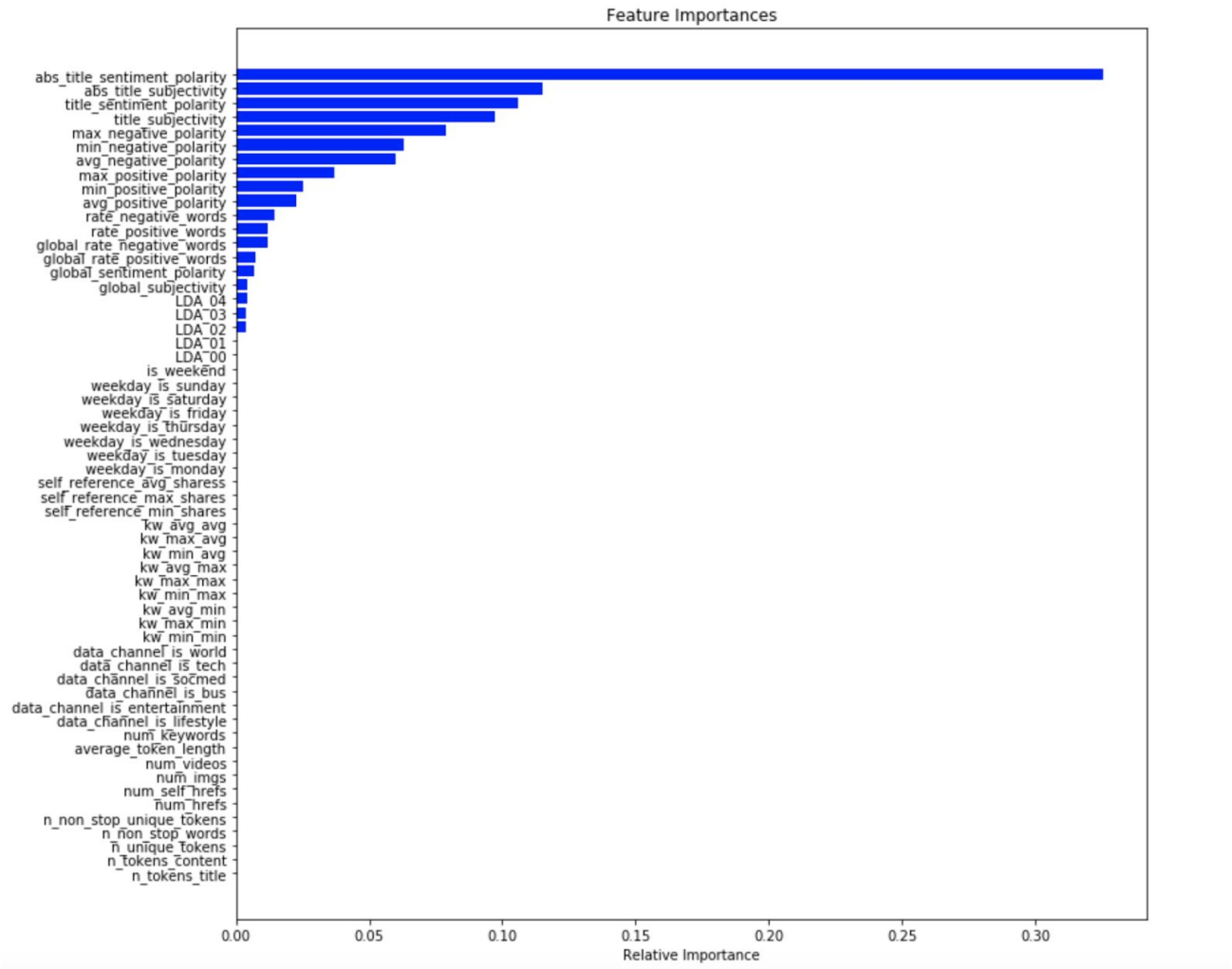


Chart 2: Feature Importance from Decision Tree

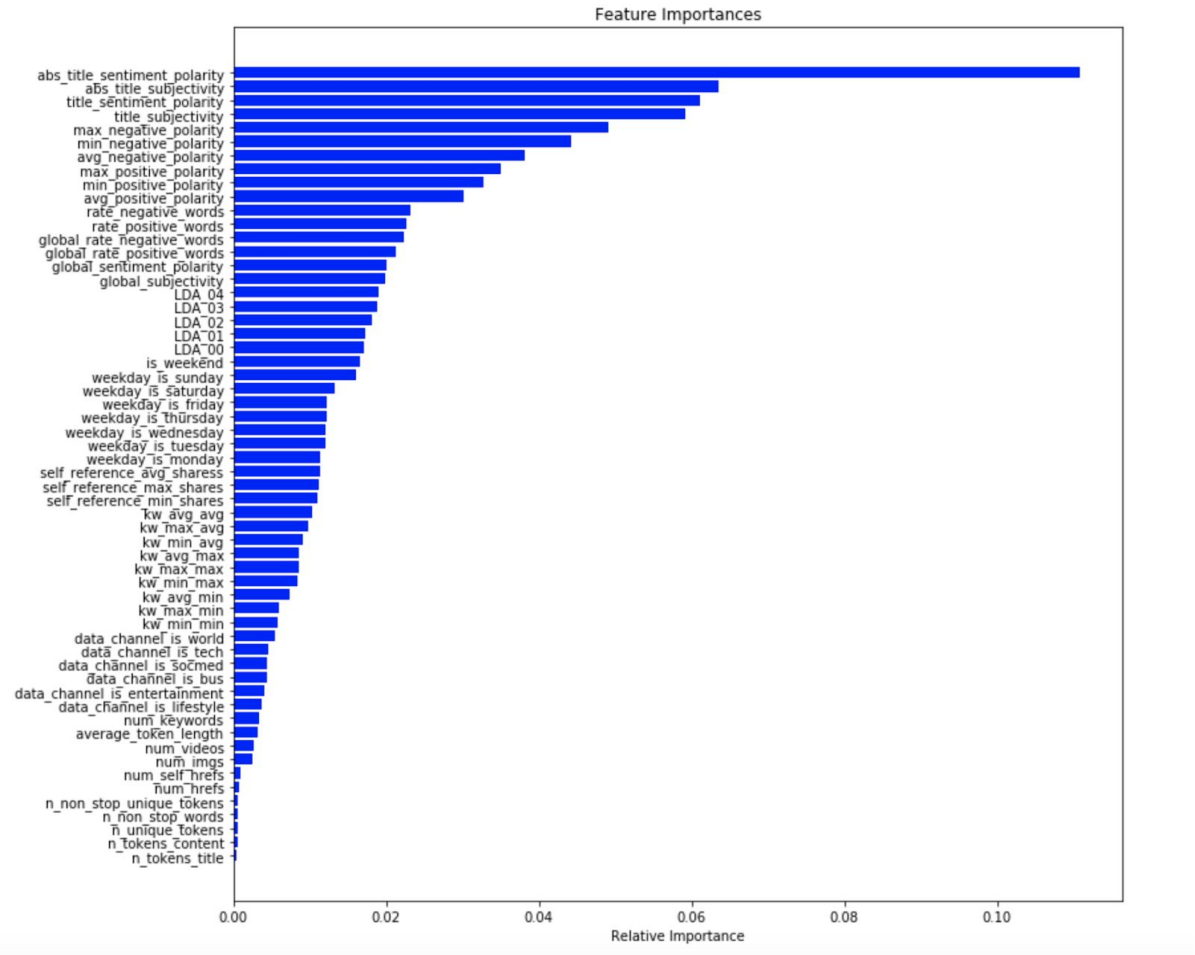


Chart 3: Feature Importance from Random Forest

Top Absolute Correlations		
kw_max_min	kw_avg_min	0.944639
n_unique_tokens	n_non_stop_unique_tokens	0.917913
self_reference_max_shares	self_reference_avg_sharess	0.850831
rate_positive_words	rate_negative_words	0.842706
self_reference_min_shares	self_reference_avg_sharess	0.842039
n_non_stop_words	average_token_length	0.828308
data_channel_is_world	LDA_02	0.811644
kw_max_avg	kw_avg_avg	0.805605
global_rate_negative_words	rate_negative_words	0.800993
data_channel_is_bus	LDA_00	0.768814
avg_negative_polarity	min_negative_polarity	0.752249
global_sentiment_polarity	rate_positive_words	0.747740
kw_min_min	kw_max_max	0.745415
global_sentiment_polarity	rate_negative_words	0.732173
title_subjectivity	abs_title_sentiment_polarity	0.725456
kw_max_max	kw_avg_max	0.716188
weekday_is_sunday	is_weekend	0.711219
data_channel_is_tech	LDA_04	0.702803
global_rate_negative_words	rate_positive_words	0.674179
weekday_is_saturday	is_weekend	0.653247

Chart 4: Highly Correlated Variables