CIS 9660
Data Mining for Business Analytics
Final Project Report
Virality of Online Content

Group 3
Hyejin Ryoo, Guillermo Restrepo, Angus Lee, Pooja Shree Rajesh Babu,
Mariia Mohyla, Praveen Vungarala, Ali Hassan

Introduction and background

The rise of social media and the vast adoption of technologies has introduced a new way of social interaction. Sharing content has become a crucial part of our everyday life. According to a study, 59% of people share content frequently (J.Berger, K.L.Milkman, "What makes online content viral?", p.1). Hence, a new information-sharing environment is transforming channels of communication with customers. Understanding factors that influence content virality is essential for building effective marketing campaigns that resonate with the audience. The following analysis will rely on a psychological approach to understand what motivates consumers to share. A lot of studies have examined why certain pieces are more viral than others. One of the most popular research conducted by J. Berger and K.L.Milkman studied 7000 New York Times articles. Their research suggests that positive content is more viral than negative (J.Berger, K.L.Milkman, "What makes online content viral?"). In our analysis, we will accept or reject this statement by examining the relationship between rate of positive and negative words in the content and number of shares it received. In addition, as the study concludes, virality is partially driven by the activation of emotions, meaning that virality also depends on sentiment strength evoked by the words. Hence, we investigate the role of positive and negative sentiment strength in social transmission. Although we do not have enough data to analyze certain emotions, we consider the average positive and average negative polarity of the words as the extent of positive or negative sentiment expressed by the article to study the association with shares. Finally, we use all the factors to predict the number of shares the content gains using classification and tree-based methods.

Motivation of the research

Our research aims to determine how positive and negative content influence virality and what other factors are important to consider for the content to get noticed. We believe that by examining the relationship of important factors that shape content virality we can provide some business value. For example, we can help businesses appeal to their customers effectively through online marketing campaigns and increase customer reach. In addition, the classification methods used in the research help determine which campaigns will succeed in the marketplace, providing some cost-saving value.

Dataset description and variable introduction

Mashable is a large independent online news site founded in 2005. The dataset contains 61 attributes of about 40000 articles published by Mashable over a period of two years. The attributes were collected and preprocessed for learning purposes on May 31st, 2015 by K. Fernandes, P. Vinagre, and P. Cortez (K. Fernandes, P. Vinagre and P. Cortez. "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News."). We manually selected 16 predictors to test our hypotheses.

Our primary predictors are:

Rate of positive words and rate of negative words. We select these two predictors to indicate the degree to which the article is positive or negative. J. Berger and K.L. Milkman suggested in their research that if a percentage of negative words and positive words are included in the model as separate predictors, both of them will have a positive influence on virality (J.Berger, K.L.Milkman, "What makes online content viral?", p.5). In our analysis, we will examine the effect of positive and negative content on virality and compare how consistent our findings are with their results.

Average positive polarity and average negative polarity. Knowing the polarity helps us determine the extent of emotions evoked by the words in the article. If negative polarity is close to -1 this indicates activation of anxiety, fear, and other strong negative emotions. On the other hand, positive polarity closer to 1 indicates very strong positive emotions. Previous studies conducted by J. Berger and K.L. Milkman claimed that activation of positive and some negative emotions, like anger and anxiety, evoked by reading the content has an important positive effect on virality (J.Berger, K.L.Milkman, "What makes online content viral?", p.6-7). Although we do not have enough data to analyze emotions, we will examine the overall strength of positive and negative sentiments on the virality of the content.

Our control variables are:

Number of words in the title. Concise titles can attract audience attention and play a significant role in determining virality.

Number of words in the content. Longer articles might be more compelling for users, so it's important to control that variable.

Number of unique words in the content. Obscure words might be difficult to read and can negatively affect popularity. However, the diversity of words can be entertaining for users.

Number of images and videos. We assume that content that has more images can be perceived as more appealing.

Was the article published on the weekend? Articles published on weekends can get more popularity than articles published on weekdays.

'Entertainment', 'Lifestyle', 'Business', 'Social Media', 'Tech', 'World' **channels?** We assume that the category of the article plays an important role in determining virality.

Data summary statistics

Our dataset consists of 16 predictors (four primary predictors as shown in Table 1), 9 numerical predictors, and 7 categorical predictors. There are 39644 observations. The number of shares shows the highest mean value and standard deviation.

Table 1 Summary Statistics

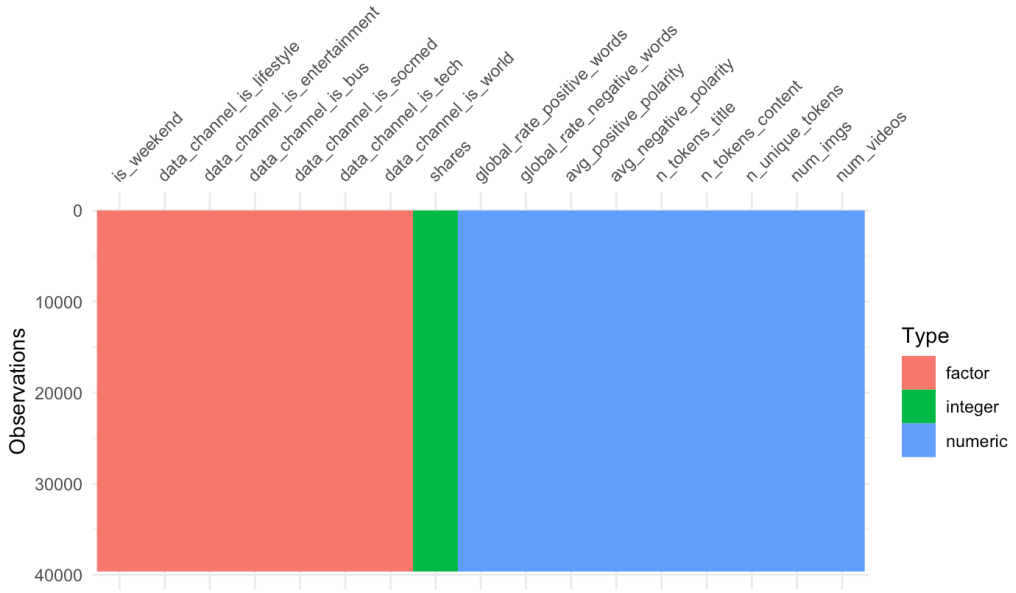
summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
Shares	39,644	3,395.380	11,626.950	1	843,300
Primary Predictor Variables					
Rate of positive words in the content	39,644	0.040	0.017	0.000	0.155
Rate of negative words in the content	39,644	0.017	0.011	0.000	0.185
Average polarity of positive words	39,644	0.354	0.105	0.000	1.000
Average polarity of negative words	39,644	-0.260	0.128	-1.000	0.000
Other Control Variables					
Number of words in the title	39,644	10.399	2.114	2	23
Number of words in the article's main text	39,644	546.515	471.108	0	8,474
Rate of unique words in the content	39,644	0.548	3.521	0.000	701.000
Number of images in the article	39,644	4.544	8.309	0	128
Number of videos in the article	39,644	1.250	4.108	0	91
Was the article published on the weekend?	39,644	0.131	0.337	0	1
Is data channel 'Lifestyle'?	39,644	0.053	0.224	0	1
Is data channel 'Entertainment'?	39,644	0.178	0.383	0	1
Is data channel 'Business'?	39,644	0.158	0.365	0	1
Is data channel 'Social Media'?	39,644	0.059	0.235	0	1
Is data channel 'Tech'?	39,644	0.185	0.389	0	1
Is data channel 'World'?	39,644	0.213	0.409	0	1

Missing Values

We plot missing values in each column and check data types. Based on the graph we didn't detect any missing values. There are three types of data: factor, integer, and numeric.

Graph 1 Plot of missing values



Correlation Matrix

The most significant correlations in the matrix are between the rate of positive words in the content and the average polarity of positive words (0.331), the rate of negative words in the content and the average polarity of negative words (-0.352), the number of average polarity of positive words and the number of words in article's main text (0.135). This suggests that the more positive words there are in an article, the more positive the average polarity of those words will be, and the more negative words there are in an article, the more negative the average polarity of those words will be. It also suggests the longer the article's main text will be, the more positive the average polarity of those words in the article.

There are also some weaker correlations like the positive correlation between the rate of negative words in the content and the number of images in the article (0.025). and a negative correlation between the rate of positive words in the content and the number of words in the title (-0.065). This suggests that articles with more negative words in the content are more likely to have images, and articles with more positive words in the content are less likely to have longer titles.

Overall, we can observe that there is no excessive correlation, which is beneficial for our research because it reduces multicollinearity issues.

Table 2 Correlation matrix

Correlation Matrix

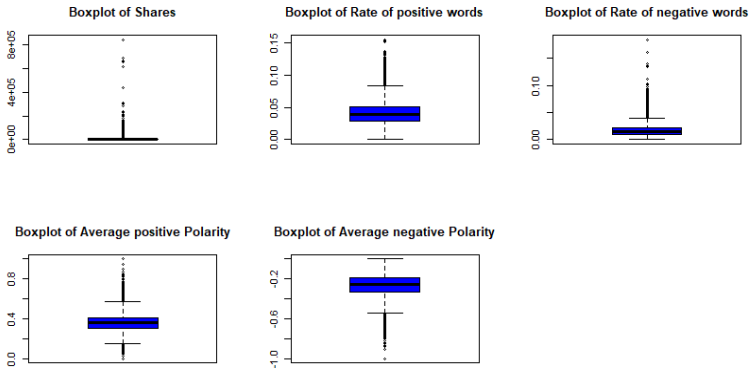
	Shares	Rate of positive words in the content	Rate of negative words in the content	Average polarity of positive words	Average polarity of negative words	Number of words in the title	Number of words in article's main text	Rate of unique words in the content	Number of images in the article	Number of videos in the article
Shares	1	0.001	0.007	0.012	-0.032	0.009	0.002	0.001	0.039	0.024
Rate of positive words in the content	0.001	1	0.107	0.331	-0.132	-0.065	0.134	0.00001	-0.042	0.072
Rate of negative words in the content	0.007	0.107	1	0.193	-0.352	0.016	0.125	-0.001	0.025	0.179
Average polarity of positive words	0.012	0.331	0.193	1	-0.276	-0.050	0.135	-0.0005	0.096	0.097
Average polarity of negative words	-0.032	-0.132	-0.352	-0.276	1	-0.017	-0.130	0.001	-0.072	-0.116
Number of words in the title	0.009	-0.065	0.016	-0.050	-0.017	1	0.018	-0.005	-0.009	0.051
Number of words in article's main text	0.002	0.134	0.125	0.135	-0.130	0.018	1	-0.005	0.343	0.104
Rate of unique words in the content	0.001	0.00001	-0.001	-0.0005	0.001	-0.005	-0.005	1	0.019	-0.001
Number of images in the article	0.039	-0.042	0.025	0.096	-0.072	-0.009	0.343	0.019	1	-0.067
Number of videos in the article	0.024	0.072	0.179	0.097	-0.116	0.051	0.104	-0.001	-0.067	1

Boxplots

By examining these boxplots, we can gain insights into the spread, central tendency, and potential outliers of these variables within the dataset. Boxplots were generated to analyze key emotional variables in our study, including shares, the rate of positive words, the rate of negative words, average positive polarity, and average negative polarity.

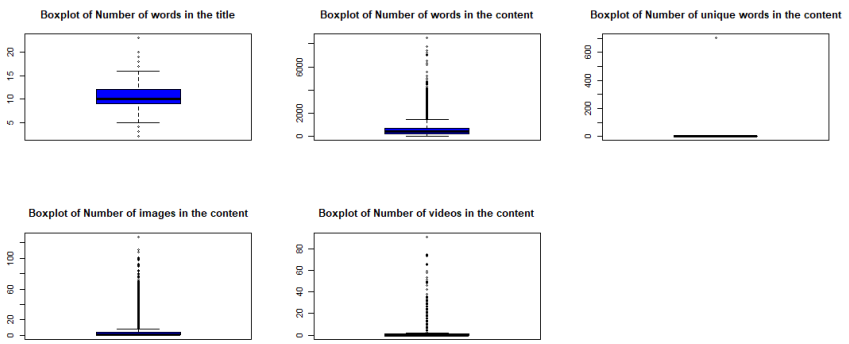
Boxplots for the rate of positive and negative words show the distribution and variations in word proportions in news articles. These plots indicate that positive words are more prevalent than negative words. The boxplots of average positive and negative polarity reveal the overall emotional tone of the articles, with a higher prevalence of positive sentiment.

Graph 2 Boxplots of numerical variables



The boxplots provide insights into the distribution and variability of variables related to the news articles. The "Number of words in the title" and "Number of words in the content" variables show typical distributions, with some outliers in the latter. The "Number of unique words in the content" variable also exhibits outliers. On the other hand, the "Number of images in the content" and "Number of videos in the content" variables have a majority of articles with a low count, with a few articles having a higher count. Overall, the boxplots help us understand the range and patterns of these variables in the dataset

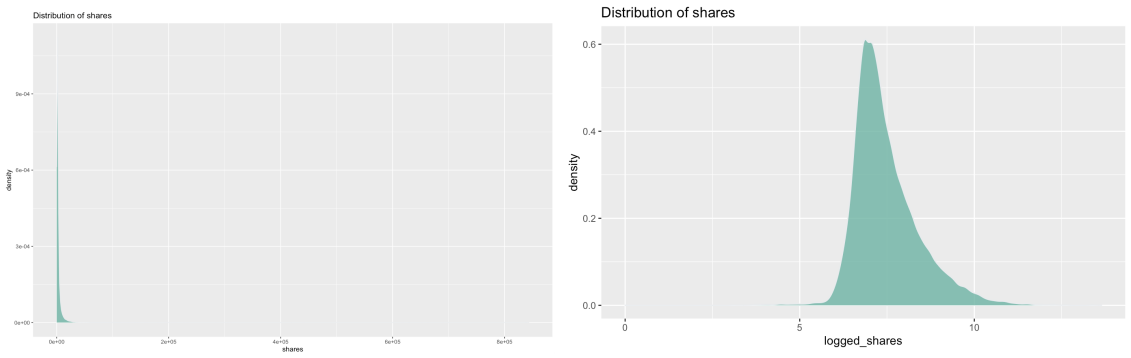
Graph 2 Boxplots of numerical variables



Transformation of dependent variable

As we plot the distribution of shares, we may notice that the variable is positively skewed and it might distort our results. To reduce the skewness of a variable we perform log transformation.

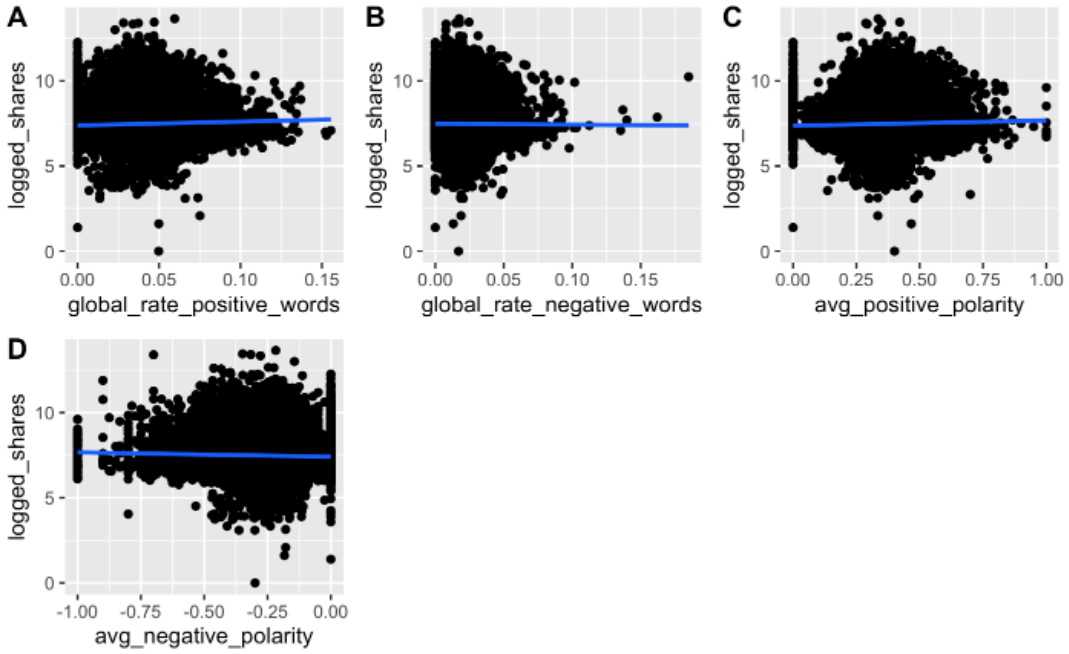
Graph 3 Distribution of dependent variable



Scatter plots

The scatter plots indicate the relationships between primary predictor variables and the logged number of shares of online news articles using a scatter plot. Although the relationship between the logged number of shares and our primary predictors based on our multi-linear regression was not particularly strong, the scatter plot allowed us to identify potential trends in the data. For example, the rate of positive words has a positive relationship with the number of shares. Also, we might observe that increasing the average negative polarity of the content might lead to a higher number of shares. Similarly, average positive polarity has a positive association with the number of shares.

Graph 4 Relationship between primary predictors and log shares



Data mining method description

In the first part of our research, we examine important factors that influence content virality. Firstly, we are looking at how the rate of positive and negative words affect virality. Secondly, we test how positive or negative sentiment affects the virality of the content. Consequently, we make the following hypotheses:

1. H1: Rate of positive words in the content positively affects the number of shares.
2. H2: Rate of negative words in the content positively affects the number of shares.
3. H3: Average positive polarity positively affects the number of shares.
4. H4: Average negative polarity positively affects the number of shares.

Our hypotheses are based on the existing theories, researched by J.Berger, and K.L.Milkman, in "What makes online content viral?". They studied 7000 New York Times articles to find how positive and negative content affect virality and the role of emotional activation in content transmission. Their results demonstrate that both positive and negative content positively affect content virality, although more positive content is more viral. On top of that, they suggest that the physiological arousal evoked by negative and positive emotions has a positive and significant influence on virality. For example, articles that evoke anxiety and anger have higher chances of getting viral, than those that evoke lower physiological arousal (J.Berger, K.L.Milkman, "What makes online content viral?", p.8).

For the purposes of our analysis, we run multi-linear regression to assess the relationship between our dependent variable and primary predictors while controlling other important content characteristics.

Table 3 Multi-linear regression summary

Multi-linear regression

	Dependent variable:
	Logged shares
Rate of positive words in the content	0.526*(0.297)
Rate of negative words in the content	-1.127**(0.459)
Average polarity of positive words	0.124**(0.053)
Average polarity of negative words	-0.277***(0.040)
Number of words in the title	0.004*(0.002)
Rate of unique words in the content	-0.342***(0.046)
Number of words in article's main text	0.00001(0.00001)
Number of images in the article	0.005***(0.001)
Number of videos in the article	0.005***(0.001)
Was the article published on the weekend?	0.283***(0.013)
Is data channel 'Lifestyle'?	-0.203***(0.023)
Is data channel 'Entertainment'?	-0.503***(0.016)
Is data channel 'Business'?	-0.344***(0.017)
Is data channel 'Social Media'?	-0.010(0.023)
Is data channel 'Tech'?	-0.195***(0.017)
Is data channel 'World'?	-0.568***(0.017)
Constant	7.742***(0.035)
Observations	39,643
R^2	0.074
Adjusted R^2	0.073
Residual Std. Error	0.896 (df = 39626)
F Statistic	196.994*** (df = 16; 39626)
Significance levels	*p**p***p<0.01

When holding other variables constant in the model we notice that the rate of positive words is positively associated with shares at 10% level. This means that positive content positively influences popularity. Our findings are consistent with J.Berger's and K.L. Milkman's results, therefore there is significant evidence to accept our first hypothesis that the rate of positive words in the content positively affects the number of shares.

Additionally, when holding other variables constant in the model, according to our model, the rate of negative words in the content is significant at 5% level and negatively associated with the number of shares. Since there is no second research that supports our claim we cannot reject our second hypothesis that the rate of negative words in the content positively affects the number of shares. As per J.Berger's, and K.L. Milkman's studies, the rate of negative words is positively associated with the virality of the content.

Similarly, when holding other variables constant in the model, we see that on average positive polarity of the words is positively associated with the number of shares at 5% level of significance, meaning that on average if we increase the strength of positive sentiment in the content, the number of shares will increase. In other words, positive sentiment positively affects the virality of the content. Our results are consistent with J.Berger's and K.L. Milkman's findings, therefore we have enough evidence to accept our third hypothesis that average positive polarity positively affects the number of shares.

In addition, when holding other variables constant in the model, average negative polarity is a very strong predictor at 0.1% level. Since the range of average negative polarity is from 0 to -1, increasing its value by 1 unit means we decrease the negative sentiment. In other words, articles with strong negative sentiment are more likely to become viral than articles with less negative sentiment. Consequently, we can conclude that the strength of negative emotions represented as average negative polarity has a significant positive effect on the virality of the content, and our fourth hypothesis is supported.
Furthermore, some content characteristics significantly negatively impact the number of shares an article receives. For example, the number of unique words in the content has a negative and significant association with the number of shares, while the number of images and videos have positive associations holding other variables constant in the model. As a result, we can conclude that people prefer to share engaging content with images and videos and less diversified vocabulary.

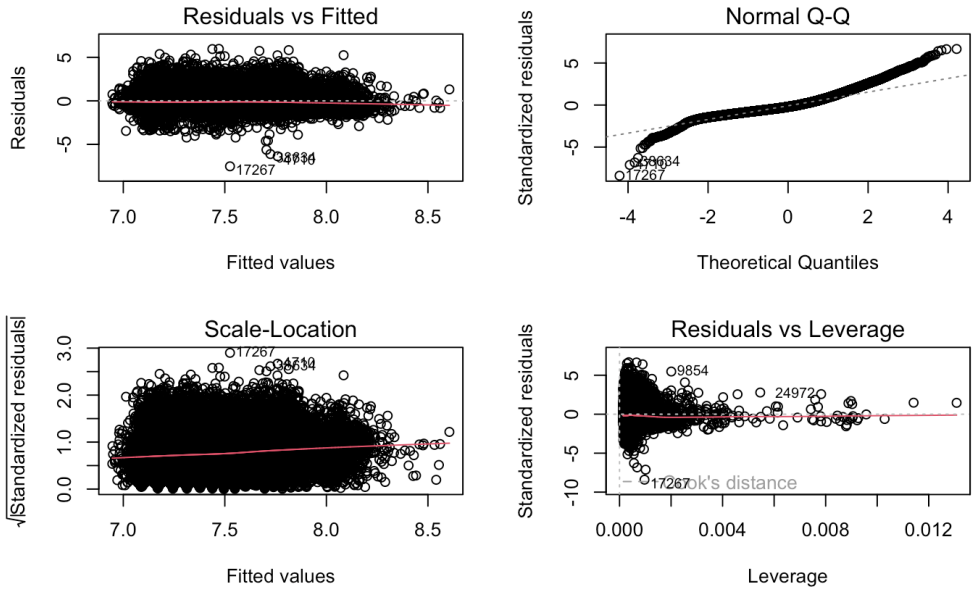
On top of that, the lifestyle category of articles on average receives fewer shares than the non-lifestyle category, while other variables are constant in the model. The same holds for entertainment, business, technology, and world categories.

It is important to mention that articles published on weekends receive more shares than the ones published on weekdays.

The adjusted R-squared value of 0.074 indicates that the model explains only a small amount of the variance in the number of shares an article receives. This means that there could be other variables beyond the ones included in the model that influence the number of shares.

The p-value of the model is very small indicating that the model is statistically significant.

Graph 5 Diagnostic Plots



These models demonstrate the diagnostic plots, to evaluate the assumptions, visually inspect our statistical model and detect any potential problems with the model fitting.

Residual vs Fitted.

The residual vs. fitted plot is shown in Figure 1. It displays the residuals on the vertical axis and the fitted values on the horizontal axis. The plot shows a random scatter of points with no discernible pattern, indicating that the residuals are independent of the fitted values. This suggests that the model is a good fit for the data and meets the assumption of linearity between the predictor and response variables.

Normal Q-Q

The Normal Q-Q plot is shown in Figure 2. The X and Y axis shows the theoretical quantiles and standardized residuals. As normal Q-Q is sensitive to sample size, both of the tails in the plot are a little deviated from the diagonal line. Since we have a large amount of data it is not concerning to us.

Scale- Location

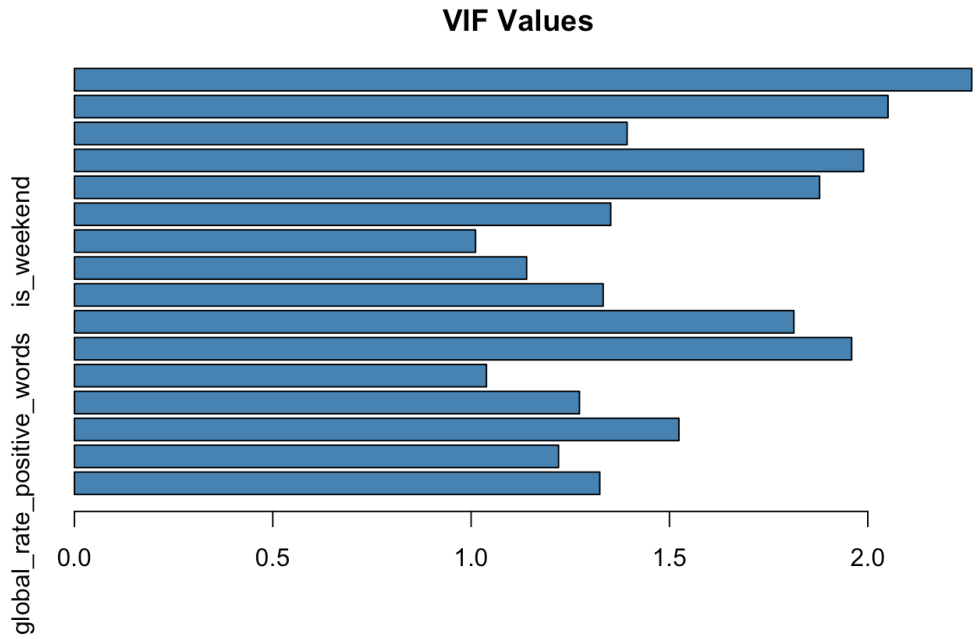
Figure 3 displays the scale-location plot. The square root of the absolute residuals (standardized residuals) is plotted against the fitted values (predicted values) from the model. The residuals from a statistical model are homoscedastic. It is evident that the horizontal line of points is observed, it indicates that the residuals have constant variance, and hence, the model fits the data well.

Residuals Vs. Leverage

The Figure 4 displays the residuals Vs leverage plot. This plot presents a comparison between the standardized residuals and the leverage values, which measure the distance of the predictor variables from their means. The plot shows that the leverage values are mostly close to zero and there are no extreme outliers or high leverage values. There was an influential point in the original

dataset which was beyond the cook's distance and it was discovered to be influencing the regression results. Therefore, the point was eliminated from the dataset. This suggests that the model provides a good fit to the data.

Graph 6 Collinearity



The VIF plot doesn’t show any collinearity between variables. Everything looks between 0 and 5.

Classification Methods

As part of the classification methods, we aim to predict the popularity of the content using various techniques like the Simple tree method, Logistic regression, and Random Forest. In order to convert the regression problem into a classification problem we have transformed our numerical dependent variable into categorical. In other words, observations that are larger than the median value of log shares are considered “popular”, the rest of the observations, below the median value, are considered not popular.

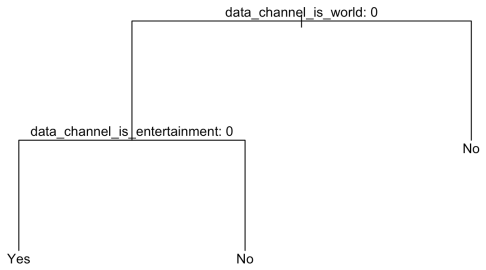
Single tree classification

We started our classification analysis with a single tree model to see as this is one of the simpler methods yet performs well on classification problems.

Table 4 Single tree summary

Residual mean deviance	1.339
Misclassification error rate	0.3962

Graph 7 Single tree method



The above tree is a binary classification problem that determines the popularity of the content. As it can be noted from the tree above, the predictors that are highly influential in determining popularity are whether the data channel is world and whether the data channel is entertainment. The residual mean deviance is 1.339. The misclassification error rate is close to 40% suggesting that the model predicts accurately in approximately 60% of the cases. If the data channel is not world and is not entertainment, then our tree predicts that the article is popular.

Logistic Regression

We have fitted a Logistic Regression model to predict the popularity of the article which has been defined as the number of shares higher than 1400 (exponential value of the median of log shares value of 7.244) as popular and not popular otherwise. The following table shows the logistic regression results of the predictors which helps us understand the significance in determining the popularity.

Table 5 Logistic regression summary

Logistic regression	
	Dependent variable:
	popular
Rate of positive words in the content	2.271**(0.979)
Rate of negative words in the content	-2.792*(1.519)
Average polarity of positive words	0.400**(0.175)
Average polarity of negative words	-0.452***(0.131)
Number of words in the title	-0.015**(0.007)
Rate of unique words in the content	-0.938***(0.152)
Number of words in the article's main text	0.0001**(0.00004)
Number of images in the article	0.007***(0.002)
Number of videos in the article	0.004(0.004)
Was the article published on the weekend?	0.785***(0.046)
Is data channel 'Lifestyle'?	-0.184**(0.076)
Is data channel 'Entertainment'?	-0.953***(0.053)
Is data channel 'Business'?	-0.447***(0.056)
Is data channel 'Social Media'?	0.459***(0.078)
Is data channel 'Tech'?	-0.029(0.054)
Is data channel 'World'?	-1.073***(0.054)
Constant	0.604***(0.114)
Observations	19,821
Log Likelihood	-12,921.790
Akaike Inf. Crit.	25,877.570
Significance levels	*p**p***p<0.01

The coefficients of the model reveal that several features are statistically significant predictors of an article's success at 5% level of significance. For example, an increase in the global rate of positive words in an article is positively associated with the popularity of the content (one unit increase in rate of positive words is associated with 2.271 increase in log odds, when other predictors are constant in the model). On top of that, the average negative polarity and average positive polarity of the article are also significant predictors of popularity, articles with a higher negative or positive polarity are more likely to be successful. These results reinforce our findings in multi-linear regression. Similarly, the number of images positively and significantly affects the number of shares. Publishing content on weekends has a significant positive effect on popularity. Articles published on weekends are much more likely to be popular than those published on weekdays. On the other hand, an increase in the rate of unique words in the article and number of words in the title have negative effects on the popularity of the content.

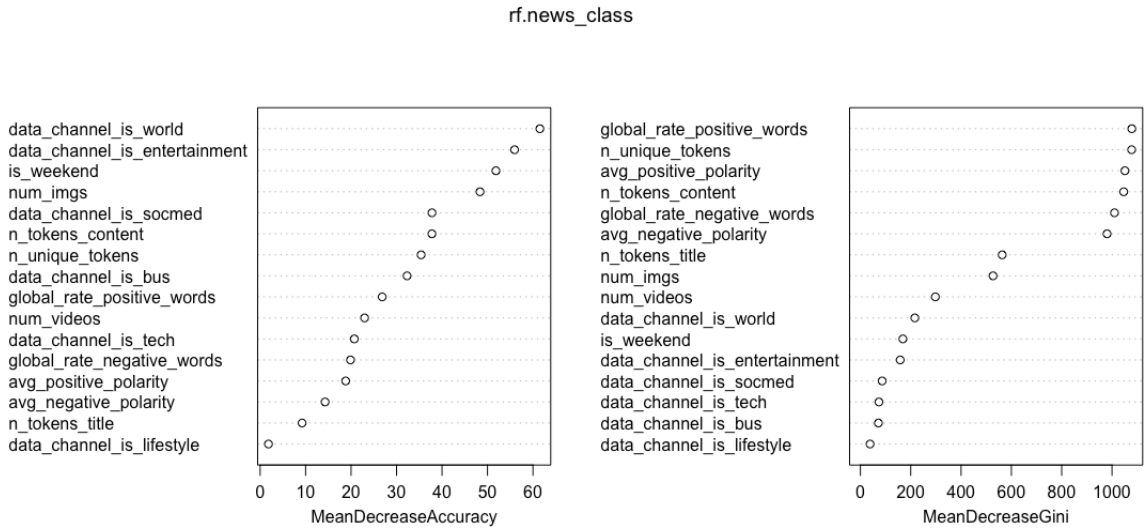
The subject of articles also seems to have a significant impact on popularity. For example, articles on social media are more likely to be successful, compared to non-social media channels, but it appears that articles on all the other subjects including world, business, entertainment, and lifestyle are more likely to be unsuccessful due to their negative but significant coefficients.

The findings suggest that articles with high negative polarity, high positive polarity, rate of positive words and images are most likely to be popular. Publishers may also want to consider publishing content on weekends to maximize the chances of them being popular. Finally, the model's results suggest that different categories of articles may have different popularity patterns, and publishers may want to take these patterns into account when designing and promoting content.

Random Forest

To improve the accuracy and robustness of our results, we have also tried Random Forest method.

Graph 8 Random Forest importance plot



Looking at the mean decrease in accuracy values, we can see that the category of the article is the most important predictor, with world and entertainment categories having the highest impact. The

second most important feature is whether the article was published on a weekend, followed by the number of images.

Now if we look at the mean decrease in Gini values, they suggest that the number of unique words in an article is the most important feature in reducing impurity in decision trees. This is followed by the rate of positive words, number of words in the article, average positive polarity, and rate of negative words and average negative polarity.

It can be deduced that the category of the article, if it is either world or entertainment, is the most important predictor of the outcome variable, followed by whether the article was published on a weekend and how many images it has. In view of these findings, publishers can publish more content on weekends and include images to make their content more entertaining for users as it is an integral factor in determining virality.

Table 6 Comparing metrics comparing

Models	Accuracy	Precision	Recall
Single Tree	0.601	0.574	0.714
Logistic Regression	0.629	0.616	0.645
Random Forest	0.635	0.629	0.622

As per the analysis based on the training and test datasets, it appears that the Random Forest model has the highest accuracy and precision rates while the Single Tree has the highest recall rate. The Random Forest is more appropriate for our analysis because if we predict a piece of content to be popular and if it's not it could cost dearly for the business if they allocate their promotional budgets based on these metrics.

Conclusions

Regression:
After fitting the regression model, we can draw a few important conclusions. Firstly, the strength of negative and positive sentiments positively influences virality. For example, articles with strong negative sentiments are more likely to be shared than articles with more neutral negative sentiments. Consequently, we suggest using strong negative or positive words in the content that evoke high sentiments to increase the virality of the content.
Secondly, positive content is more viral than negative content, according to our findings in multi-linear regression, logistic regression, and popular research “What makes online news viral?” (J.Berger, K.L.Milkman, “What makes online content viral?”, p.10).
We can conclude that people are motivated to share positive content. For example, they feel more valuable and involved with others in that way. To sum up, our analysis supports the findings of the popular research “What makes online news viral?”, that positive content has a high chance of getting viral and that the rise of positive or negative sentiments in the content is beneficial for increasing virality.

Classification:
As part of the classification methods, we took a multi-modal approach where we applied Single Tree, Logistic Regression and Random Forest techniques to understand whether the content is popular, which is defined as a binary variable based on the number of shares. Out of the three models, the results obtained from the Random Forest method stand out when compared across the metrics(i.e., accuracy, precision and recall).
The results from this analysis are very useful in multiple business use cases. For example, we can provide some confidence while choosing an article for its predicted popularity over others.

Practical implications

The rise of electronic circulation of content is ever-increasing across the globe with the adoption of the internet and mobile technology. Today, readers are finding most of the content online through news feeds, social media pages, and other networking platforms/apps. The current research looks at the data through various lenses and attempts to understand the relationship between virality/popularity and content characteristics. The findings from our analysis suggest that the rate of positive/negative words along with sentiments the content evokes does significantly impact virality/popularity. While these are the primary predictors, there are also a few control variables that seem to be significant in impacting virality/popularity. It has to be noted that though these findings are specific to Mashable dataset, they are in relation to the literature on factors influencing virality including positive/negative content, and emotions (J.Berger, K.L.Milkman, “What makes online content viral?”, p.10). Our analysis proves that it makes strong business sense for content providers to look at various features that affect the virality of the content and design the content accordingly.

References

1. Berger, J., & Milkman, K.L. (2011). What makes online content viral? *Journal of Marketing Research*, DOI: 10.1509/jmr.10.0353.

2. Fernandes, K., Vinagre, P., & Cortez, P. (2015). A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. In F. Pereira, P. Machado, E. Costa, & A. Cardoso (Eds.), *Progress in Artificial Intelligence*. EPIA 2015. Lecture Notes in Computer Science (Vol. 9273). Springer, Cham. https://doi.org/10.1007/978-3-319-23485-4_53.