Online News Popularity Analysis and Prediction

Group 16

Abhijit Kannepalli, Bhavesh Chatnani, Hemil Kothari,
Saurabh Saoji, Shrey Shah

Introduction

- With the advent of the internet and social media, online news has become integral to everyone’s life.

- The most significant increase in attention paid to a specific news piece is known as news popularity.

Literature Review

- Omar (2007) talks about the transition from conventional to online news and factors that contributed to online news becoming popular. Immediacy and polarity plays a key role for a news category to spread faster than others.

- Shirsat et al. (2017) estimated the polarity of words (positive, negative, neutral). The articles were divided into genre groups as positive, negative, or neutral.

- Deshpande (2017) studied the dimensionality reduction techniques, compared different models and identified the optimal model to predict web news popularity.

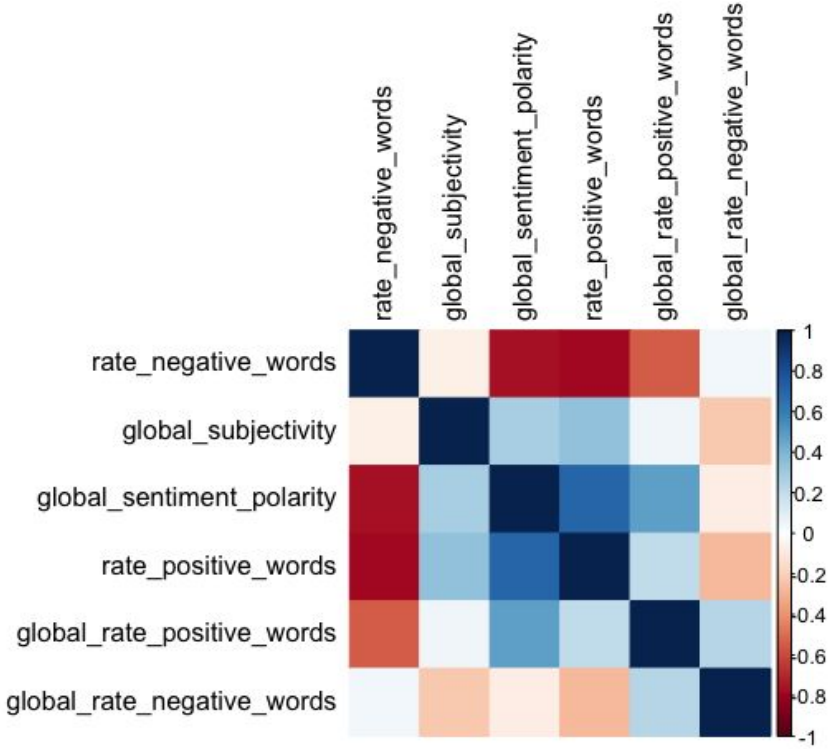
Research Question 1

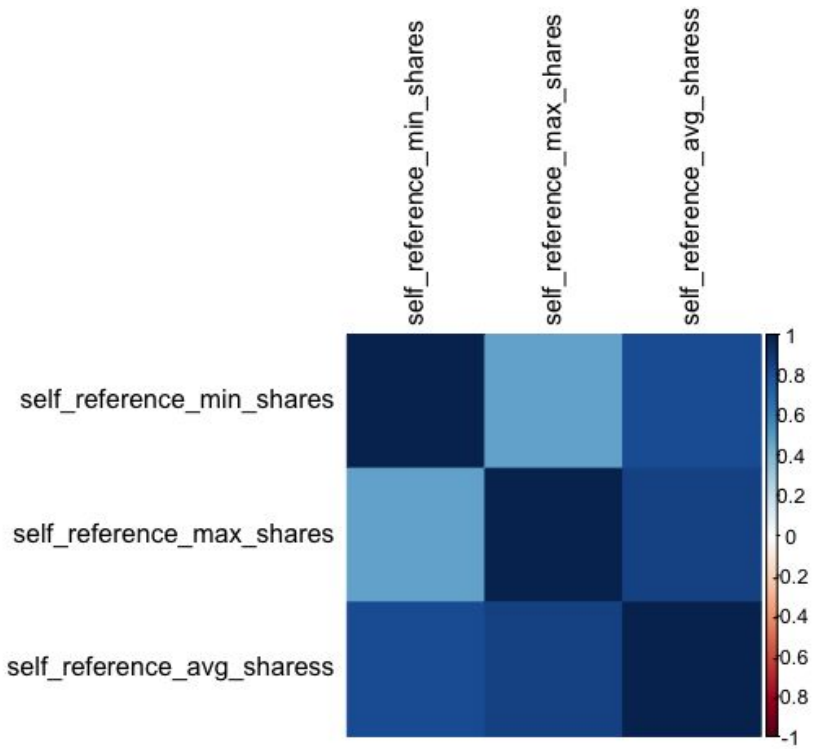
Which factors are critical in determining the popularity of online news in terms of its shares?

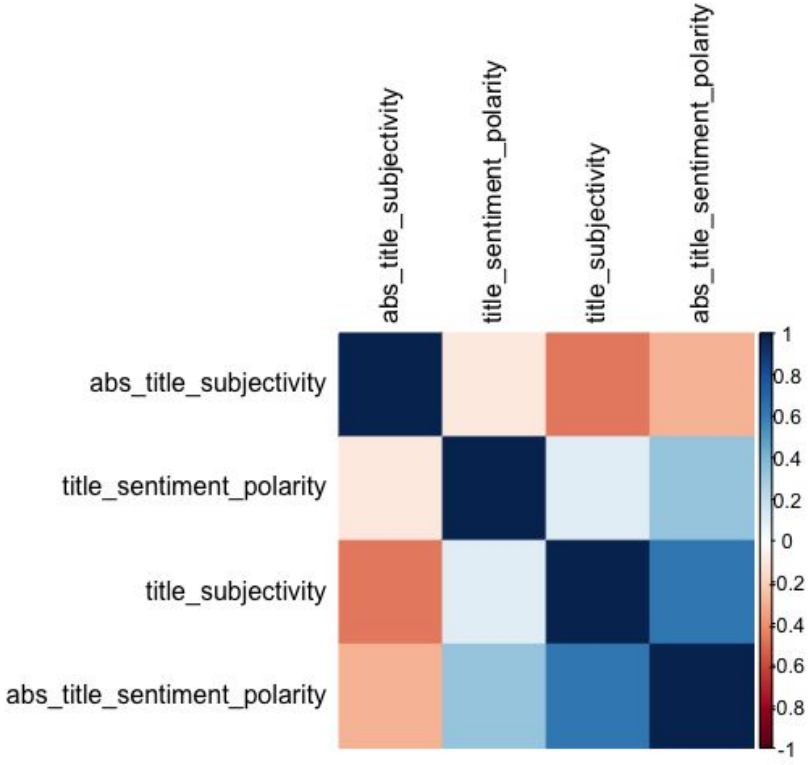
Application

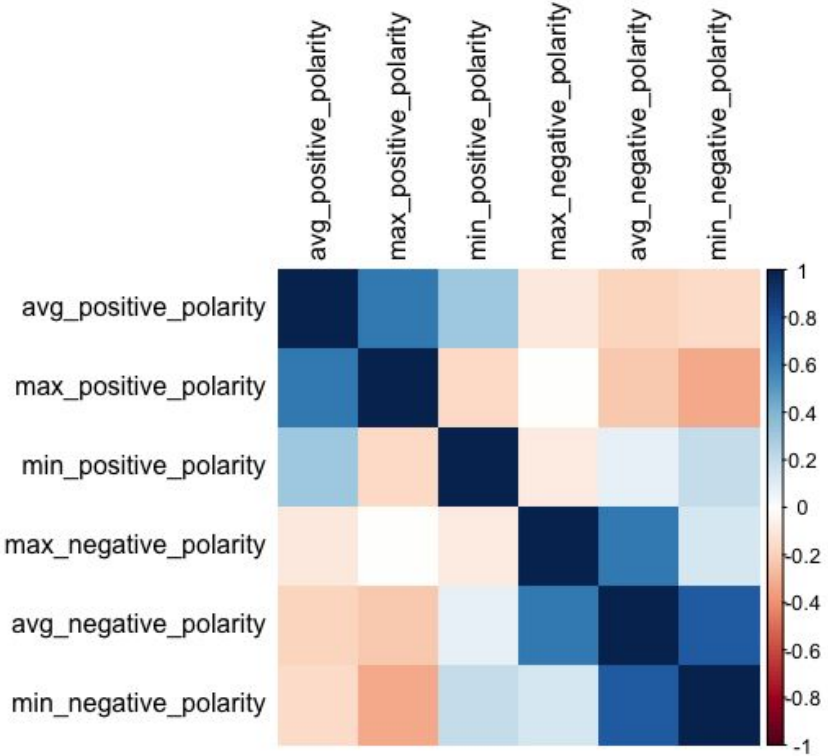
- Whether a news is popular or not is dependent on a wide array of features
- Analysing all of them together can be complex. What if we can identify the significant factors that should be considered for identifying popular news articles?
- The reduced feature set can be used in further analysis like:
  - predicting the shares of the news articles
  - Bloggers and social media content creators can understand what type of articles gain more popularity
- News/Text mining can be performed based on article content and self reference shares to other articles of the same type to find relationships between articles

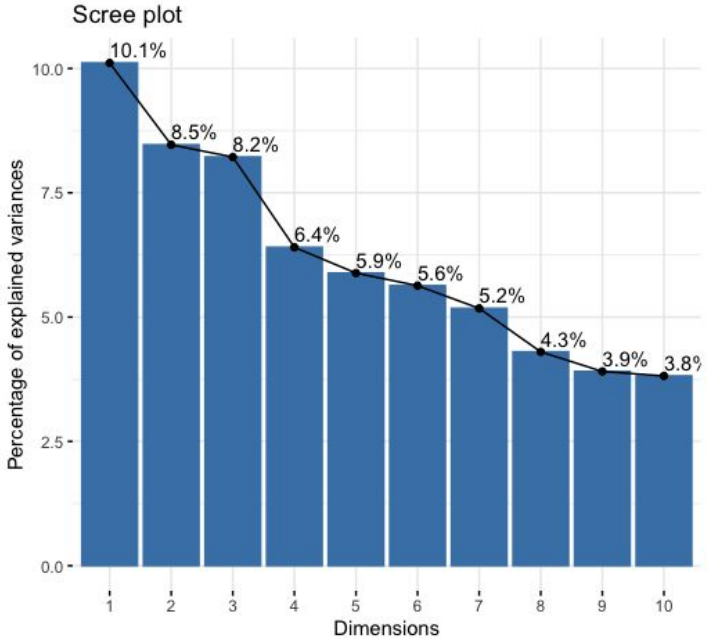
Quick Visualizations






abs_title_subjectivity
title_sentiment_polarity
title_subjectivity
abs_title_sentiment_polarity
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


avg_positive_polarity
max_positive_polarity
min_positive_polarity
max_negative_polarity
avg_negative_polarity
min_negative_polarity
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1



Number of components = 5

Analysis & Discussion

- Is it significant?

  KMO test has a value of 0.5

  Bartlett's Test: 12926832 (df = 44, p < 0.001)

  Cronbach's Alpha = 0.53

```
> # 4th PCA
> # Increasing the cutoff to remove crossloadings
> comps = print(pca_with_5_components_promax_4$loadings, cutoff=0.45, sort=T)

Loadings:
                           RC2    RC4    RC1    RC3    RC5
n_tokens_content            0.870
n_unique_tokens            -0.904
n_non_stop_unique_tokens   -0.846
num_hrefs                   0.649
num_imgs                    0.682
self_reference_min_shares          0.881
self_reference_max_shares          0.909
self_reference_avg_sharess         0.813
self_reference_avg_shares          0.885
kw_min_min                                -0.728
kw_min_max                                 0.583
kw_max_max                                 0.834
kw_avg_max                                 0.896
kw_min_avg                                 0.588
kw_avg_avg                                 0.588
global_sentiment_polarity                         0.882
global_rate_positive_words                        0.660
rate_positive_words                               0.926
rate_negative_words                              -0.935
n_non_stop_words                                         0.902
average_token_length                                     0.889

                 RC2   RC4   RC1   RC3   RC5
SS loadings    3.239 3.104 3.084 2.964 2.175
Proportion Var 0.154 0.148 0.147 0.141 0.104
Cumulative Var 0.154 0.302 0.449 0.590 0.694
```

Interpretation of Principal Components

- **PC2: Number of Tokens, Images and Links** – Number of unique, content and non-stop tokens, images and links present in articles
- **PC4: Self Reference Shares** – Shares of articles when it refers/links to other articles of similar type
- **PC1: Keyword Based Shares** – Shares of articles based on the best, worst and average keywords
- **PC3: Positive Content News** – Positive sentiment polarity, the proportion of positive and negative words
- **PC5: Non-Stop Words and Token Length** – Average length of tokens and the amount of non-stop words

Research Question 2

Predicting if an article is popular or not?

Application

- Predicting the popularity based on specific attributes of an article would help publishers improve those aspects for better reach.

- By analyzing the type of polarity in the popular articles, news agencies or content writers could mold their articles in that direction.

Method and Outcome

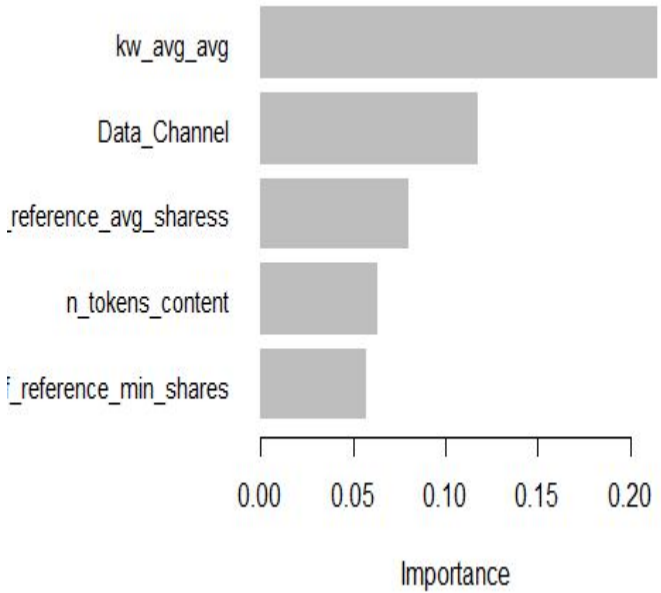
- XGBoost to predict the binary output variable - Popular / Not popular (Threshold: 2000)
- Summary of outcome:
  Accuracy: **71%**
  Sensitivity**: 73%**
  Specificity**: 63%**

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4197  350
         1 1569  586
```

Here ‘0’ represents the class of unpopular articles

Significant features in determining the popularity

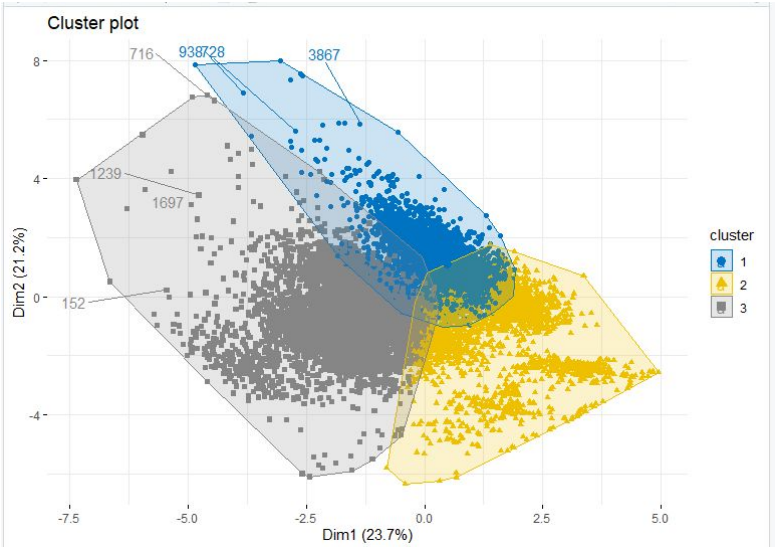


Research Question 3

Do authors prefer to write polarizing articles?

Methods and Outcome

- Used K-means clustering technique
- Articles were split into 3 clusters:
  - Positive
  - Negative
  - Neutral
- Around 78% percent of the articles were either positive or negative.




```
K-means clustering with 3 clusters of sizes 5931, 3061, 4908

Cluster means:
  avg_positive_polarity min_positive_polarity max_positive_polarity avg_negative_polarity min_negative_polarity
1             0.1128972          -0.18064712             0.2645179            -0.3098292           -0.3491159
2            -0.5575088           0.43637679            -0.9504595             1.0635771            1.1072897
3             0.2112758          -0.05385723             0.2731257            -0.2889186           -0.2687057
  max_negative_polarity title_subjectivity title_sentiment_polarity abs_title_subjectivity abs_title_sentiment_polarity
1            -0.1275734         -0.7807579              -0.1796215             0.70382189                  -0.5083924
2             0.4567408         -0.2578490              -0.1668563             0.07895857                  -0.3748514
3            -0.1306939          1.1043095               0.3211251            -0.89976769                   0.8481450
```

Research Question 4

Is the popularity of an article dependent on the data channel publishing it?

https://drive.google.com/file/d/1heCxqHvbr0oCvaUi0fmhmfrlVblOZpur/view

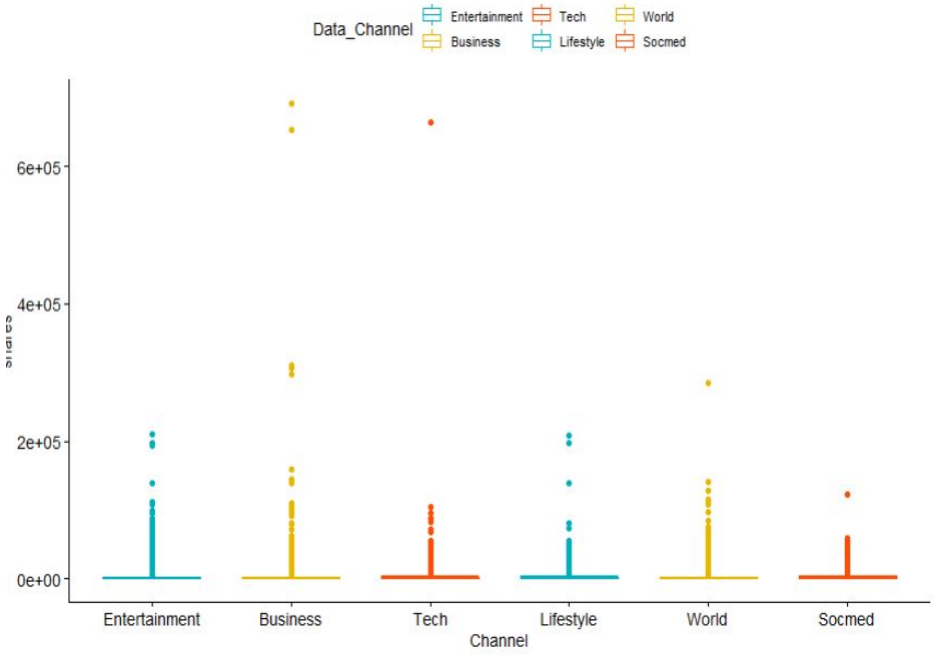
Application

- Understanding the significance of data channels can help us understand if the category of an article, ie, business, technology etc. is a contributing factor to the popularity

- This can also allow us to analyze the readers of the article as to which category of news are they are more likely to read and share.

Methods and Outcome

- Original data channel columns were pre-processed to form one single column forming categorical attributes
- Data is non-normal, hence Kruskal-Wallis test was used
- P-value < 0.001 suggests that null hypothesis is rejected, and data channel of article does not influence the popularity
- Thus we conclude that the data channel an article belongs to does not influence the popularity of the article




```
        Kruskal-Wallis rank sum test

data:  df2[, "shares"] by df2[, "Data_Channel"]
Kruskal-Wallis chi-squared = 2163.6, df = 5, p-value < 2.2e-16
```

Research Question 5

Is the popularity of an article dependent on the day of the week it is published on?

https://drive.google.com/file/d/1AR8mUgFZ-6Q3qJC7FPtXSPcPj2dOCirB/view

Applications

- Predicting the popularity based on the day of the article being published can be an important factor to consider

- This will help us understand if people specifically prefer to watch/read news on weekends (when one is relatively free) or weekdays.

Methods and Outcome

- Used Mann-Whitney U Test algorithm.
- P-value for all of the days: <0.05
- All p-values are less than 0.05, and hence there is no association between the day of week when article was published and the popularity of a news article.

```
        Wilcoxon rank sum test with continuity correction

data:  shares by weekday_is_monday
W = 113179968, p-value = 9.23e-05
alternative hypothesis: true location shift is not equal to 0
```

```
        Wilcoxon rank sum test with continuity correction

data:  shares by weekday_is_tuesday
W = 126429355, p-value = 2.996e-16
alternative hypothesis: true location shift is not equal to 0
```

```
        Wilcoxon rank sum test with continuity correction

data:  shares by weekday_is_wednesday
W = 128258150, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Similarly, for the rest of the days, the p-value is still below 0.05.

Limitations

- We can represent approximately same amount of information with less data. The trade-off here, is that it does not always guarantee the interpretability of the components.
- Popularity is subjective. In our research, we took a threshold of 2000 for the number of shares to be popular.
- Due to cluster overlap, sentiment in many articles is not clearly determined.
- This research has not achieved predictions based on the news article’s polarity.

Future Work

- In the future, we can focus on working on even larger datasets which can also include actual words present in the article, to help us improve the factors determining popularity.

- We would also like to try to understand why factors like data channel and time of article being published is not relevant when studying the data but are important when used to predict the popularity using our model.

Conclusion

- The major factors which can be used to determine online news popularity are based on token proportion, reference to similar articles, keyword based shares, the proportion of positive content and token lengths.
- The grouping of articles based on the positive, negative or neutral sentiments can be helpful in understanding the mindset of authors.
- The popularity prediction and the factors can be used to understand the article content for gaining maximum traction.
- It is observed that article popularity does not depend on the genre or when the article is read indicating.

References

Deshpande, D. (2017). Prediction & Evaluation of online news popularity using Machine Intelligence. *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. https://doi.org/10.1109/iccubea.2017.8463790

Omar, B. (2007). The Switch to Online Newspapers Could Immediacy Be a Factor?

Shirsat, V. S., Jagdale, R. S., & Deshmukh, S. N. (2017). Document level sentiment analysis from news articles. *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. https://doi.org/10.1109/iccubea.2017.8463638

THANK YOU