Cyrus Benjamin C. Canape
Personal Project 7/2/2025

Online News Popularity Analysis (Mashable)

https://www.kaggle.com/datasets/deepakshende/onlinenewspopularity/data

I. Project Background

Mashable Inc. is a digital media website described as a "one-stop shop" for social media. As of November 2015, it has over 6,000,000 Twitter followers and over 3,200,000 fans on Facebook. The purpose of this project is to identify the different significant features for our target variable, the number of shares. By doing so, we can identify which set of features is significant and optimize the article to maximize the number of shares.

II. Methodology

A. Dataset

The dataset is publicly available on Kaggle with over 39,644 unique values and 61 columns. Some examples of the columns are the *URL, timedelta, n_tokens_title, etc.* These columns have their corresponding description, which can be seen in Kaggle.

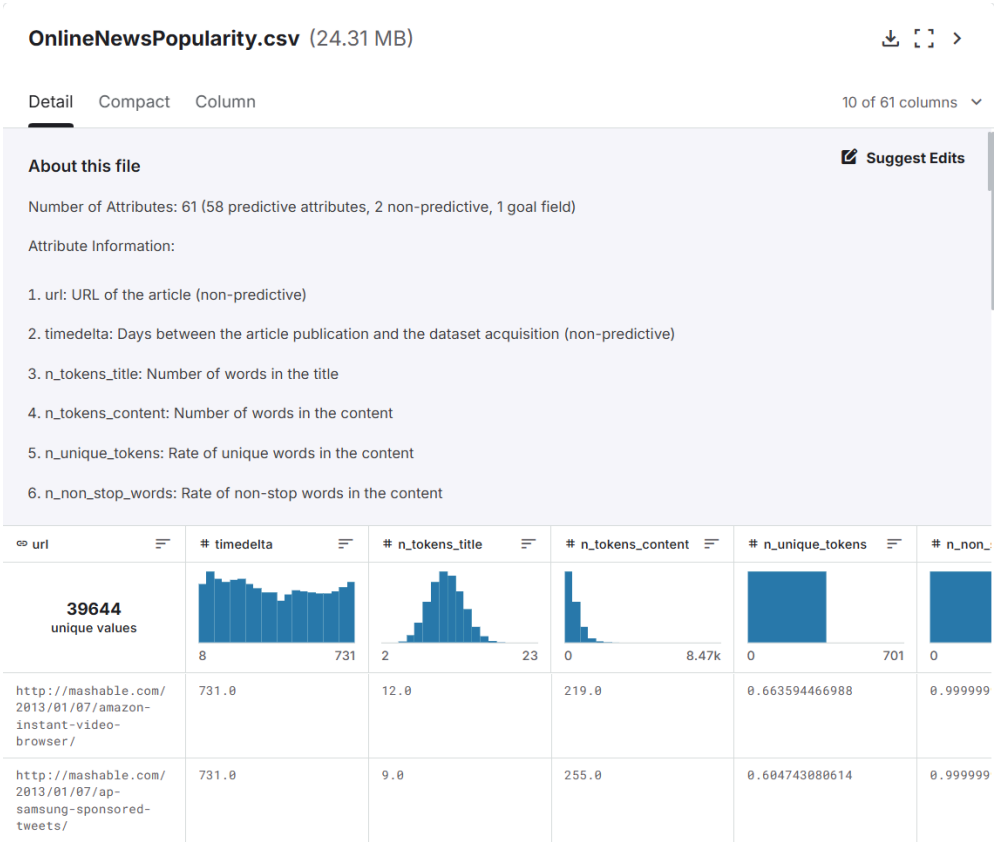

Figure 2.A.1. Kaggle Available Dataset

B. *Data Preprocessing*

The dataset is preprocessed before using it for EDA and analysis. This process can be seen under the folder *data > processed.*

The extra spaces from the column names were stripped. Unnecessary columns such as the *URL* and *timedelta* were dropped. Duplicates were also dropped, and missing values were checked. A new column was also added, named *log_shares,* which compressed the values of the data and made the distribution more normal. This preprocessed data is then saved in a new *.csv* file named *cleaned_data.csv*.

Data Preprocessing

```
[11]
```

```
import pandas as pd
import numpy as np


df = pd.read_csv("../raw/OnlineNewsPopularity.csv")

# Strip extra spaces from column names
df.columns = df.columns.str.strip()

# View the first 5 rows
print(df.head())
```

✓ 0.3s

```
                                                 url  timedelta  \
0  http://mashable.com/2013/01/07/amazon-instant-...      731.0
1  http://mashable.com/2013/01/07/ap-samsung-spon...      731.0
2  http://mashable.com/2013/01/07/apple-40-billio...      731.0
3  http://mashable.com/2013/01/07/astronaut-notre...      731.0
4   http://mashable.com/2013/01/07/att-u-verse-apps/      731.0

   n_tokens_title  n_tokens_content  n_unique_tokens  n_non_stop_words  \
0            12.0             219.0         0.663594               1.0
1             9.0             255.0         0.604743               1.0
2             9.0             211.0         0.575130               1.0
3             9.0             531.0         0.503788               1.0
4            13.0            1072.0         0.415646               1.0
```

Figure 2.B.1. Data Preprocessing

C. *Process*

After data preprocessing, 9 different sets of analysis was performed: *EDA, Feature Importance, Channel Popularity, Content Structure and Shares, Day of Week Trends, Keyword Significance, Positive and Negative Word Usage, Sentiment and Subjectivity Analysis, and Topic Modeling LDA Impact.*

1. *Exploratory Data Analysis*

The distribution of the shares was first bar plotted for visual representation. It showed a highly skewed distribution due to its large values.

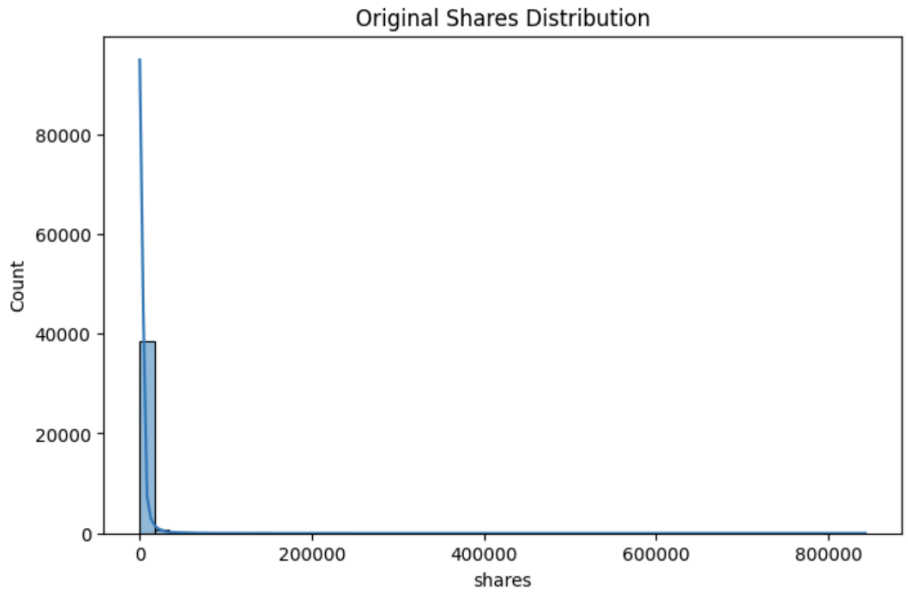

Figure 2.C.1 Original Share Distribution

However, by transforming the shares into a log share, it compressed the data which shows a more normal distribution.

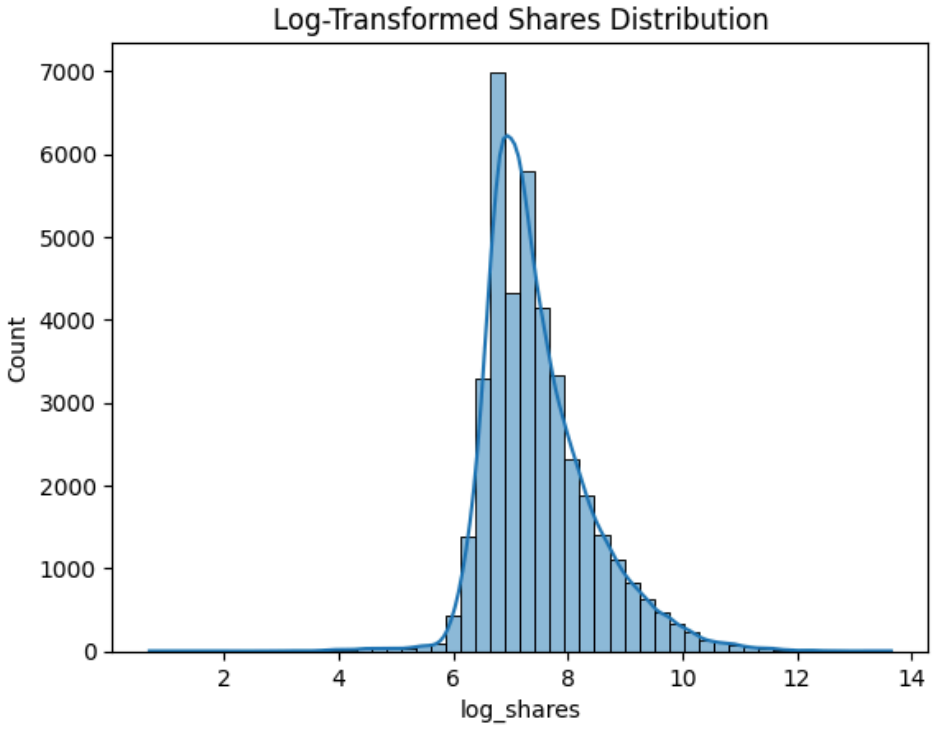

Figure 2.C.2 Log-Transformed Share Distribution

A correlation heatmap is also performed to visualize the significant features for the *log_share*.

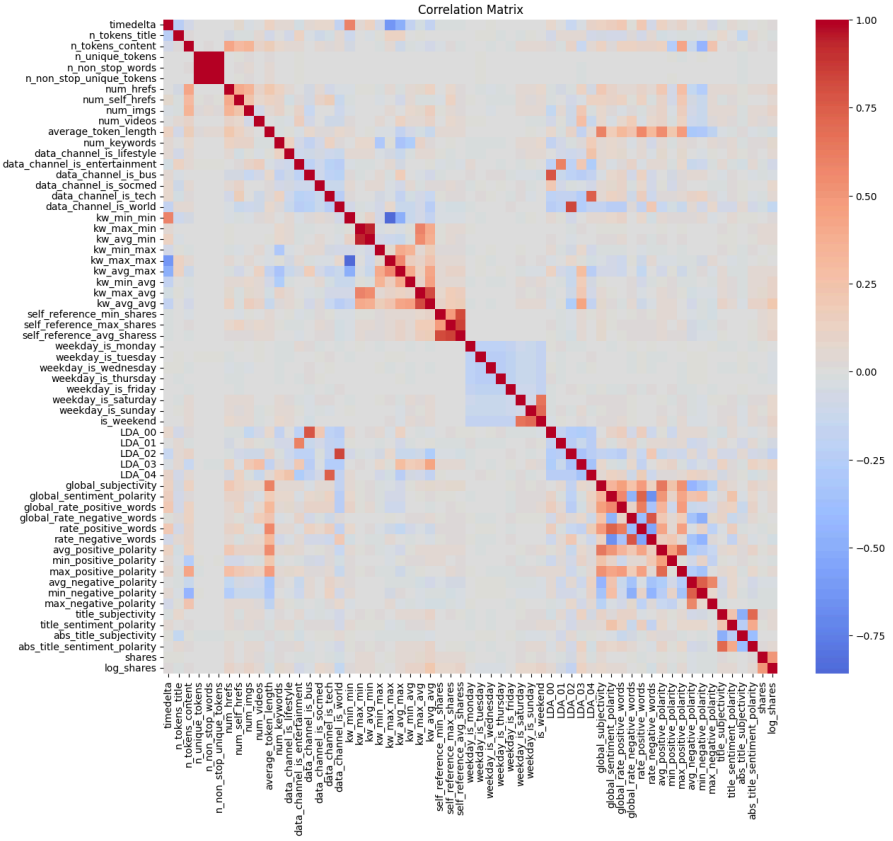

Figure 2.C.3 Features Correlation Heatmap

Other analyses, such as pairwise feature relationships, category distribution, feature and target relationships, and outlier check, were also done in the EDA analysis.

2. *Feature Importance Analysis*

A Spearman correlation with *log_shares* was executed to check the top 10 most and least correlated features with *log_shares*.

```
Top Spearman Correlations with log_shares:
shares                          1.000000
log_shares                      1.000000
kw_avg_avg                      0.255622
kw_max_avg                      0.223291
self_reference_avg_sharess      0.192174
self_reference_min_shares       0.181517
self_reference_max_shares       0.168725
is_weekend                      0.151718
data_channel_is_socmed          0.113572
global_subjectivity             0.113548
Name: log_shares, dtype: float64
```

Figure 2.C.4 Top 10 Spearman Correlation with log_shares

```
Least Spearman Correlations with log_shares:
n_unique_tokens                 -0.044842
min_positive_polarity           -0.045270
weekday_is_wednesday            -0.048131
average_token_length            -0.057335
LDA_01                          -0.068760
n_non_stop_unique_tokens        -0.070706
rate_negative_words             -0.070865
data_channel_is_entertainment   -0.114691
LDA_02                          -0.157179
data_channel_is_world           -0.168441
Name: log_shares, dtype: float64
```

Figure 2.C.4 Top 10 Least Spearman Correlation with log_shares

Random forest and XGBoost were also used to figure out the significant features. Both models have two different outputs.

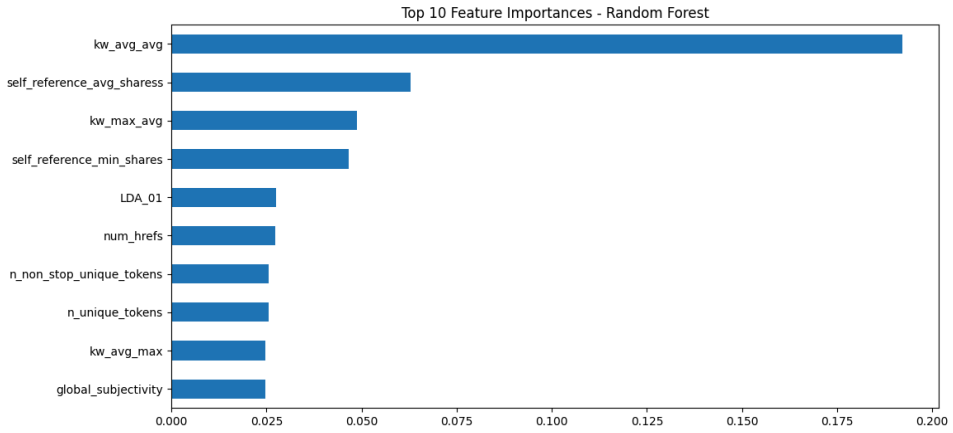

Figure 2.C.5 Top 10 Feature Importance - Random Forest

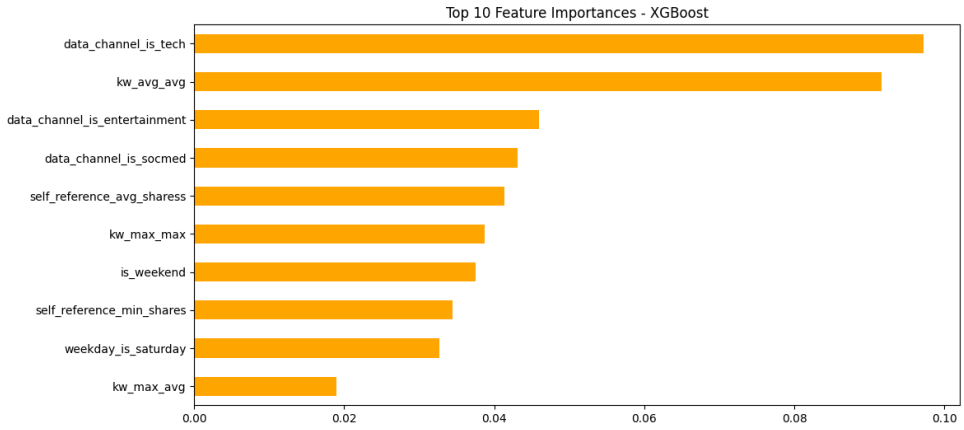

Figure 2.C.6 Top 10 Feature Importance - XGBoost

3. *Channel Popularity Analysis*

A channel popularity analysis was conducted to figure out which channel is the most shared.

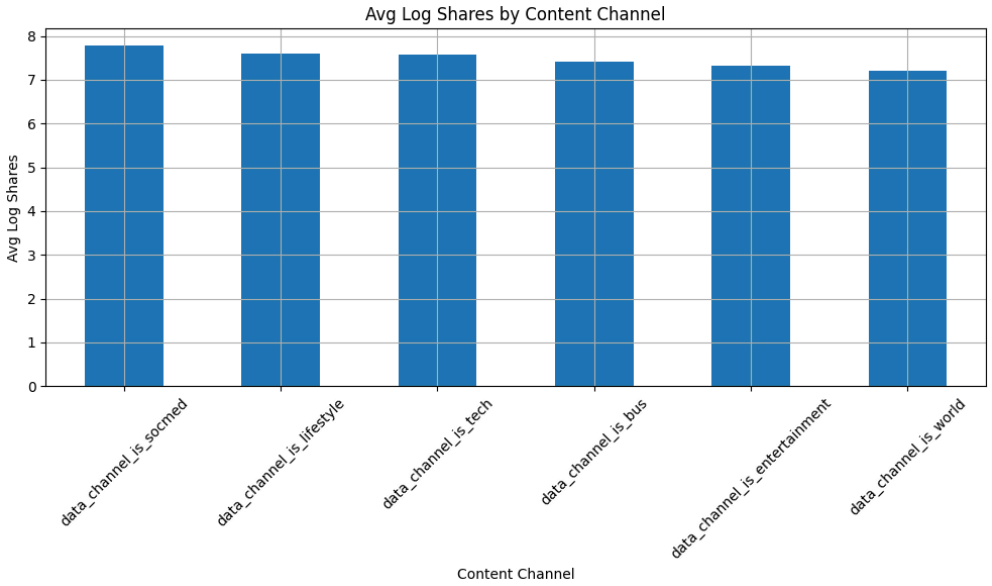

Figure 2.C.7 Content Channel Bar Chart

```
Average Log Shares by Channel:
data_channel_is_socmed: 7.78
data_channel_is_lifestyle: 7.61
data_channel_is_tech: 7.58
data_channel_is_bus: 7.41
data_channel_is_entertainment: 7.31
data_channel_is_world: 7.20
```

Figure 2.C.8 Content Channel Values

4. *Content Structure and Shares Analysis*

The content structure of the articles was also analyzed, including the length of the title, length of content, number of visuals, both images and videos, and the number of links.

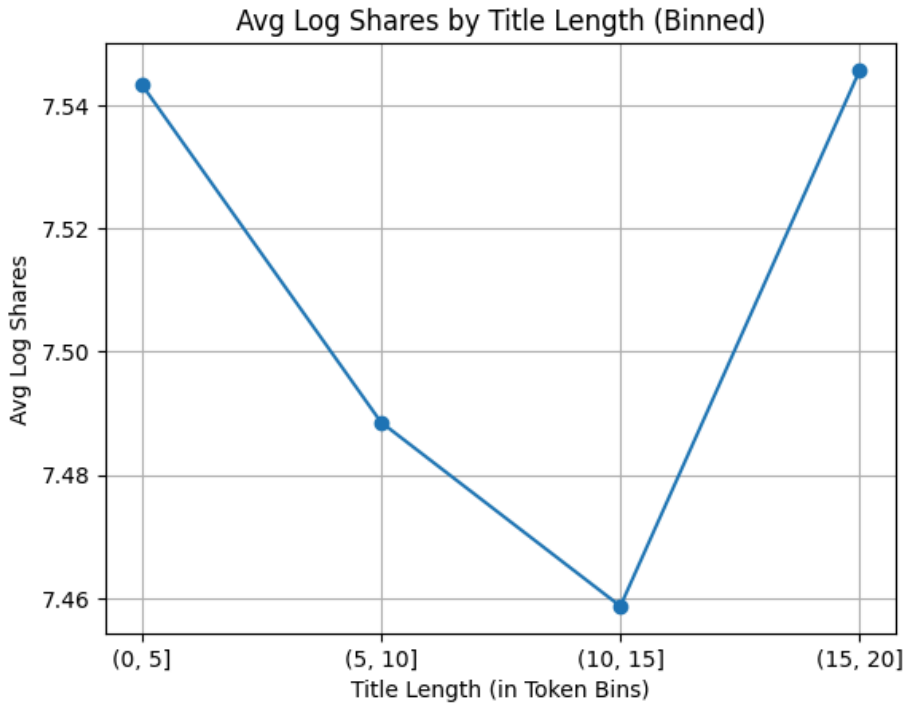

Figure 2.C.9 Average log_shares by Title Length

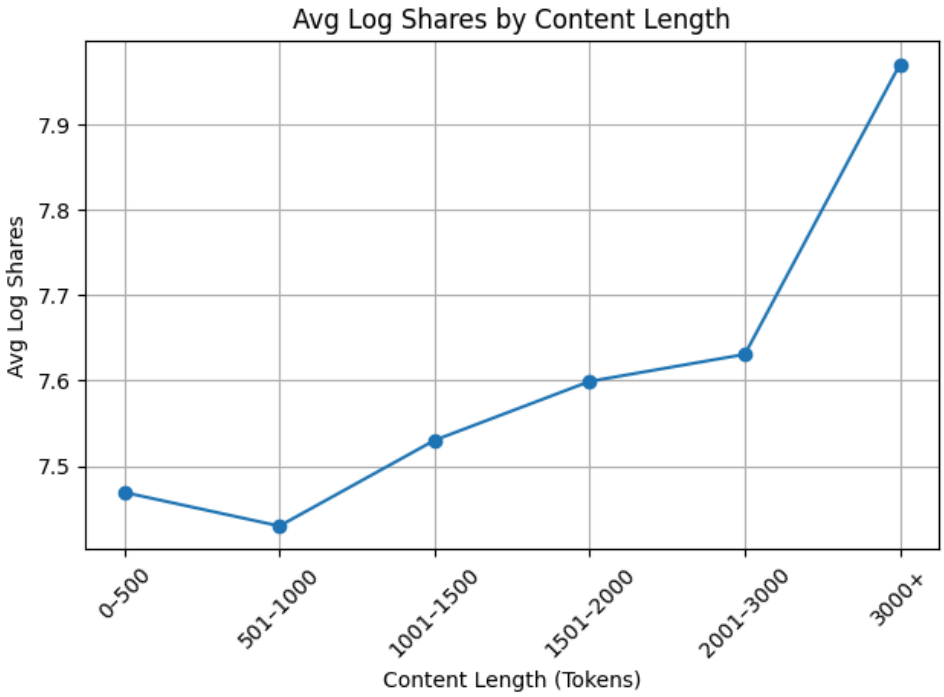

Figure 2.C.10 Average log_shares by Content Length

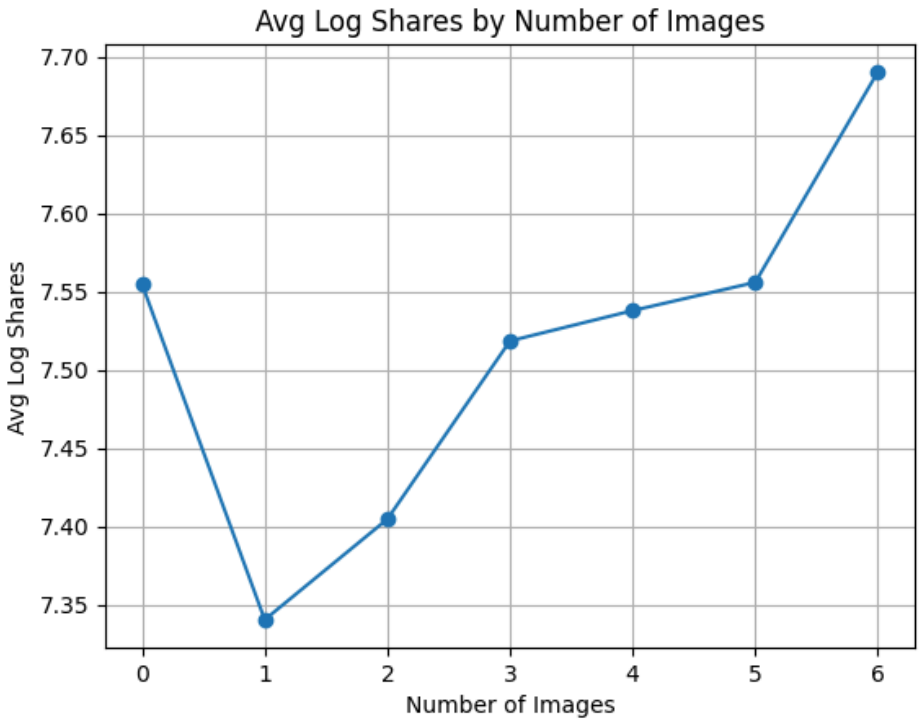

Figure 2.C.11 Average log_shares by Number of Images

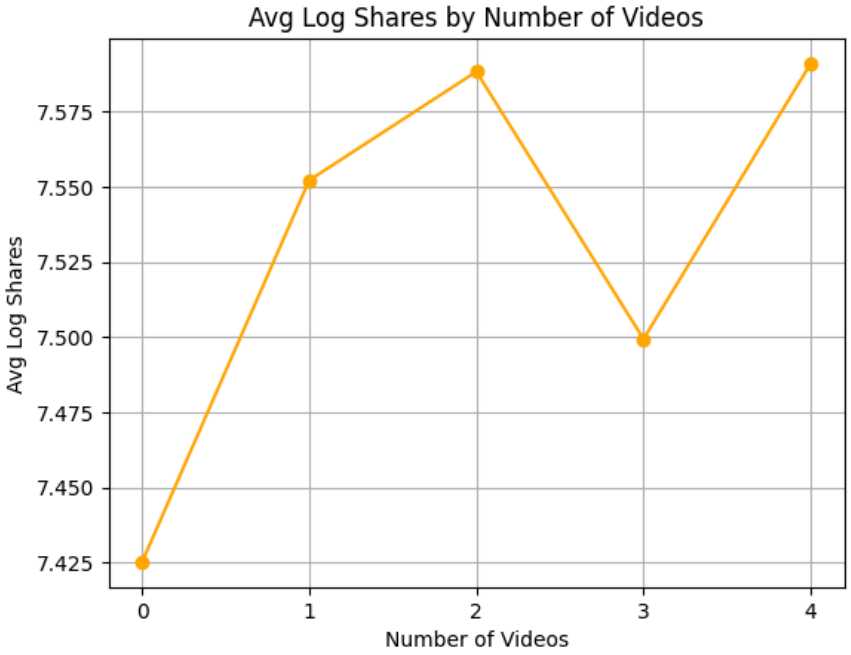

Figure 2.C.12 Average log_shares by Number of Videos

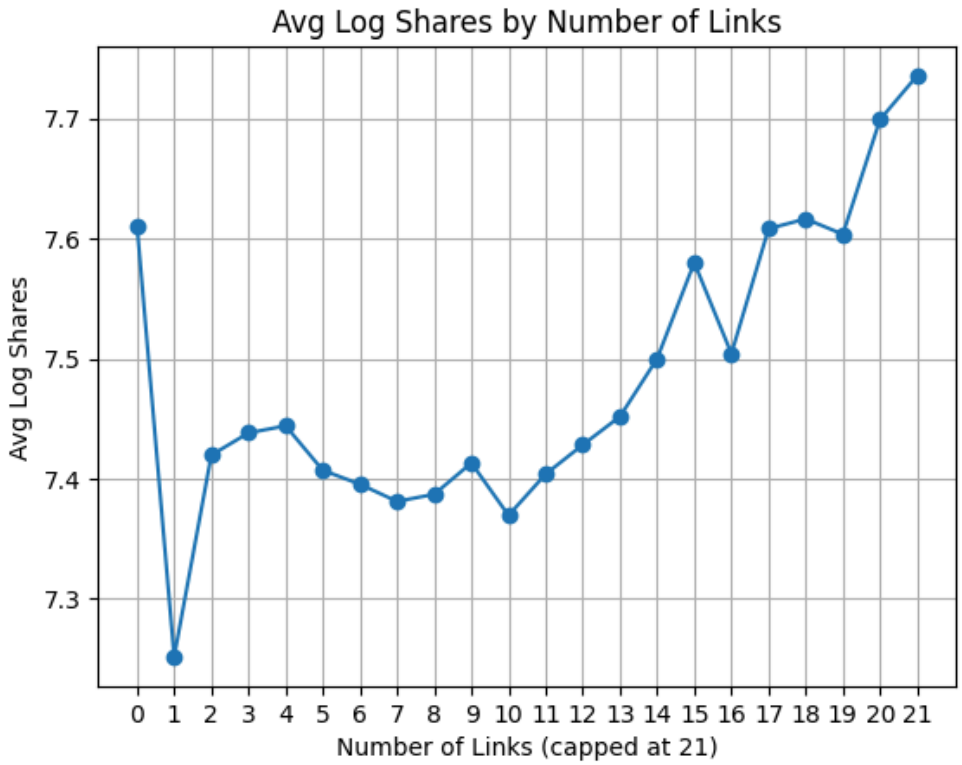

Figure 2.C.13 Average log_shares by Number of Links

5. *Day of Week Trends Analysis*

A day of week trends analysis was conducted to figure out which day of the week has the most log shares.

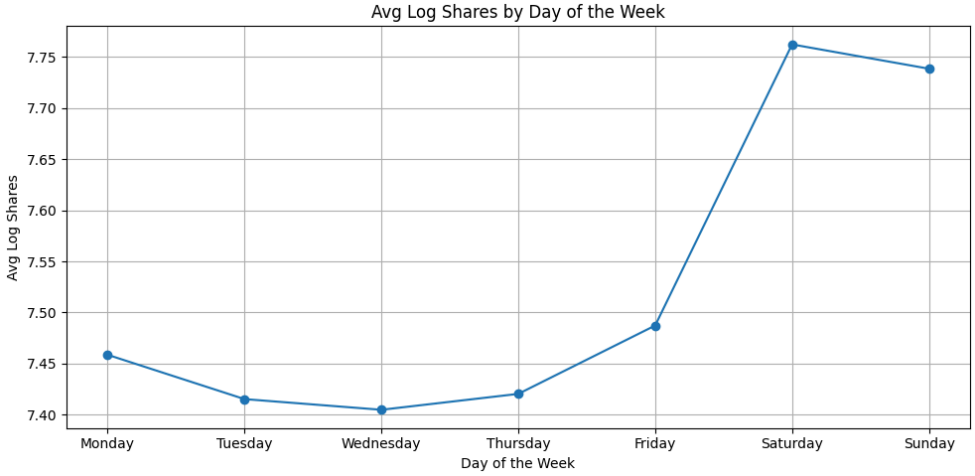

Figure 2.C.14 Average log_shares by Day of the Week

6. *Keyword Significance Analysis*

This analysis was conducted to figure out whether the usage of keywords is significant to a content's shareability. Self-reference within

the article is also analyzed in relation to its significance to the content's shareability.

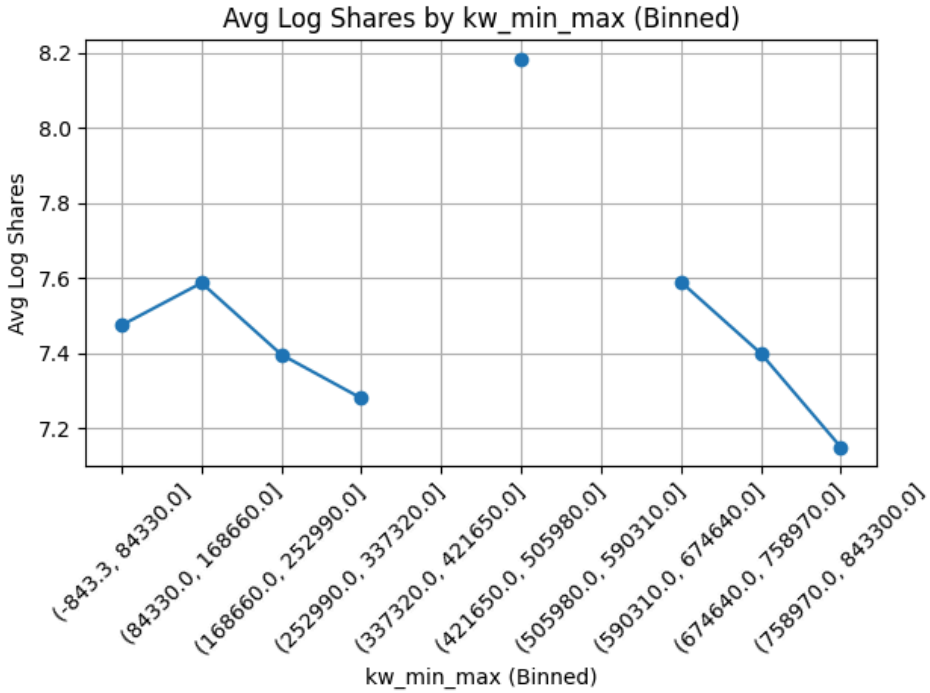

Figure 2.C.15 kw_min_max vs log_shares

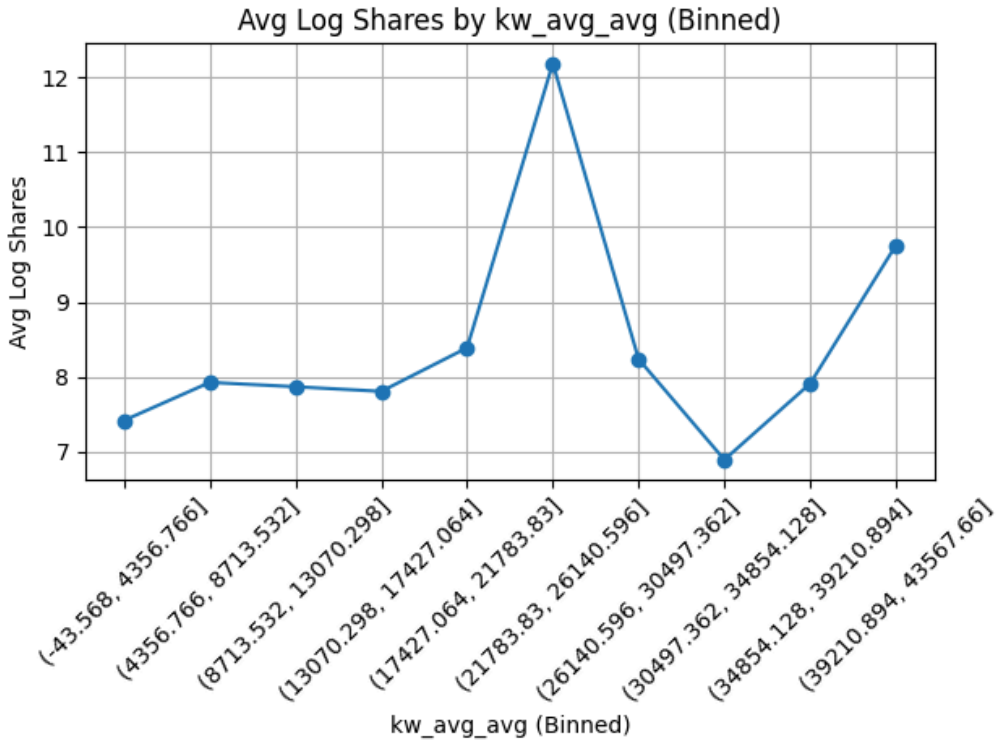

Figure 2.C.15 kw_avg_avg vs log_shares

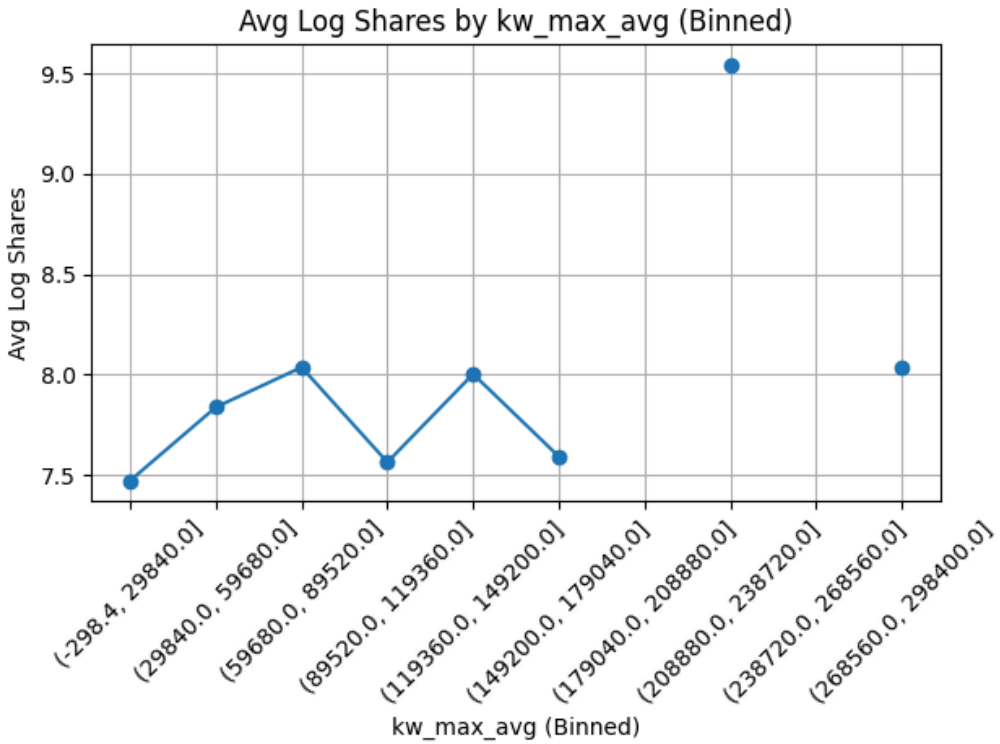

Figure 2.C.15 kw_max_avg vs log_shares

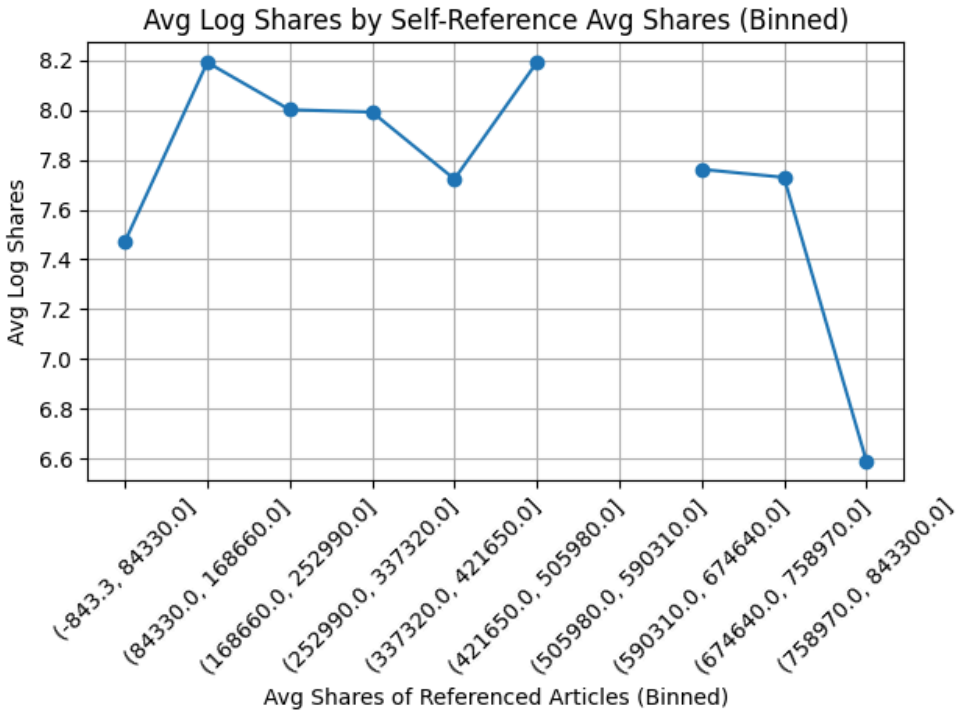

Figure 2.C.16 Average log_shares by Self-reference Average Shares

7. *Positive and Negative Word Usage Analysis*

This analysis was conducted to figure out the difference between the usage of positive and negative words.

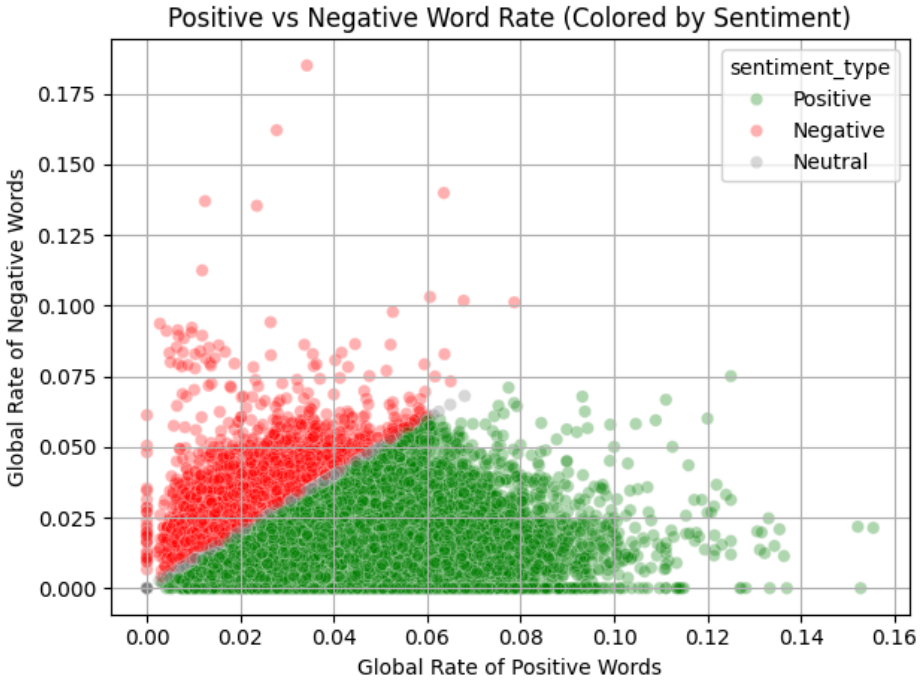

Figure 2.C.17 Positive vs Negative Word Rate

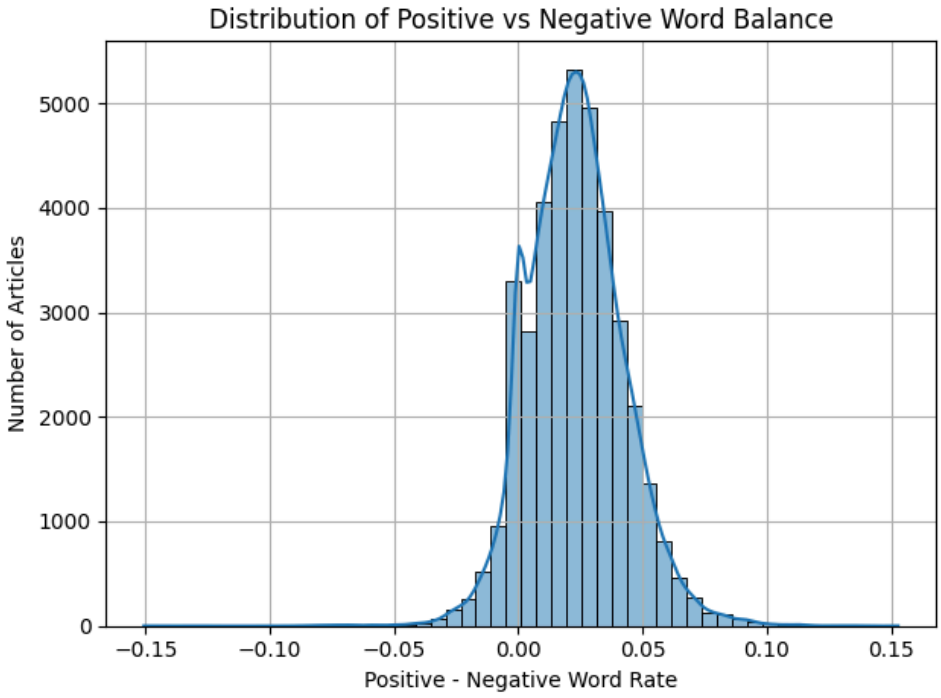

Figure 2.C.18 Positive-Negative Word Rate

8. *Sentiment and Subjectivity Analysis*

This analysis is conducted to check whether the sentiment and subjectivity of the article are significant to the content's shareability.

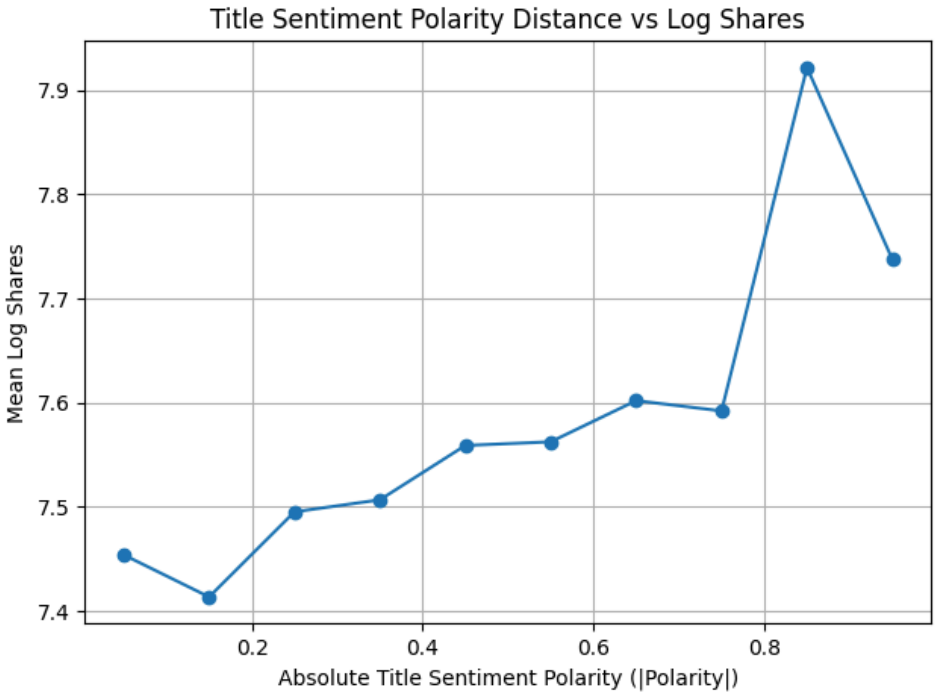

Figure 2.C.19 Absolute Title Sentiment Polarity

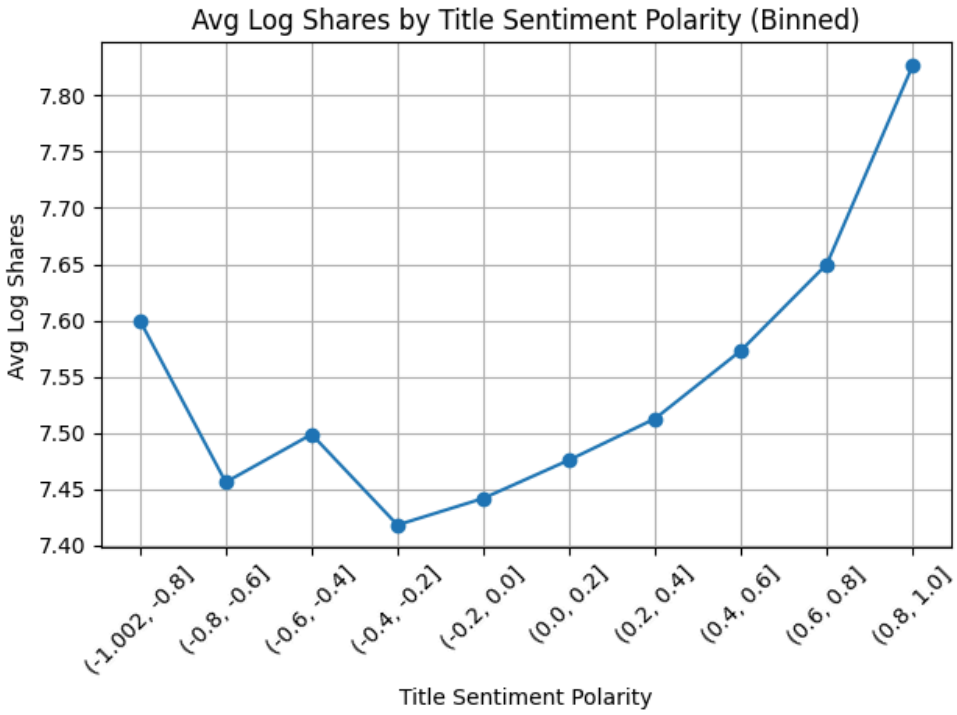

Figure 2.C.20 Title Sentiment Polarity

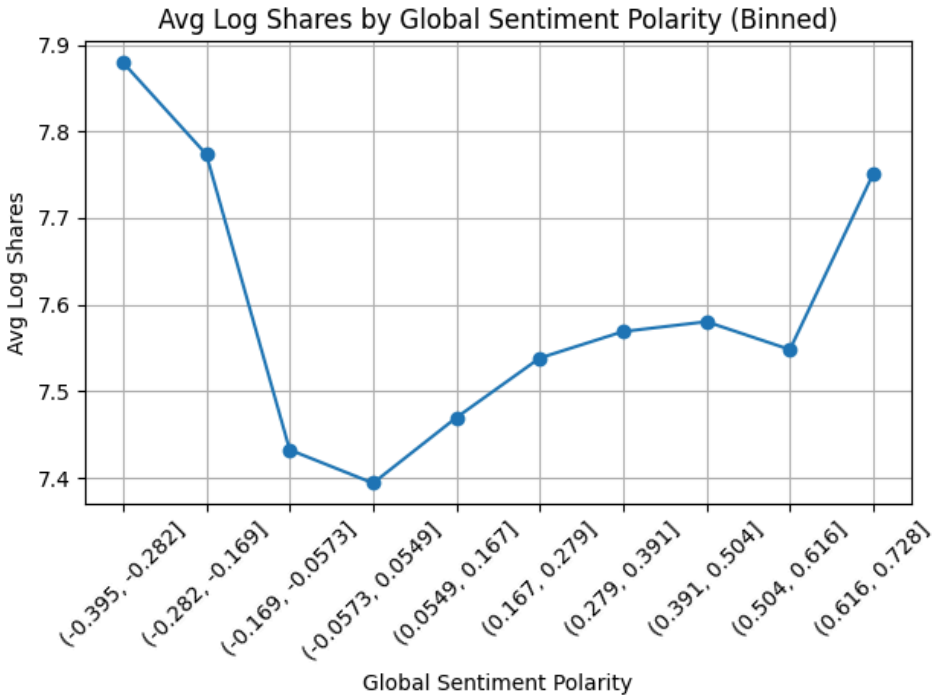

Figure 2.C.21 Global Sentiment Polarity

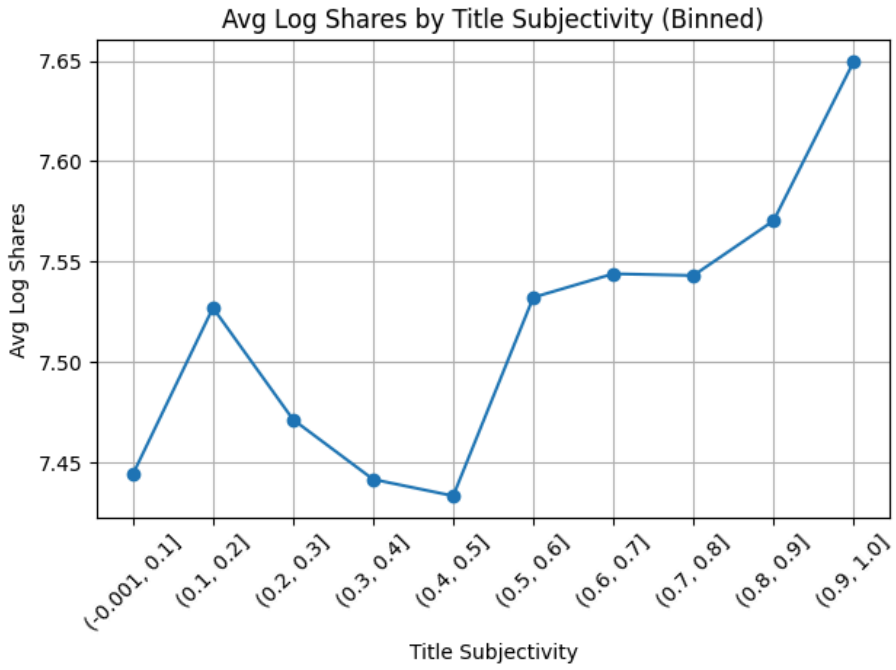

Figure 2.C.22 Title Subjectivity

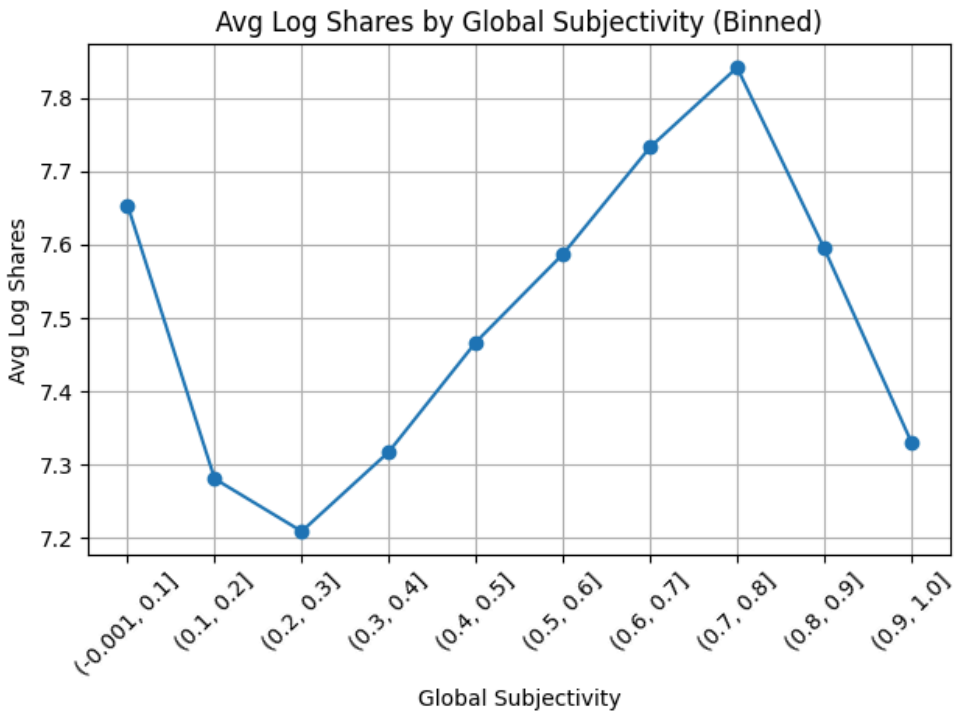

Figure 2.C.22 Global Subjectivity

9. *Topic Modeling LDA Impact Analysis*

This analysis measures the Latent Dirichlet Allocation, which identifies hidden topics in text data and assigns probability scores of each topic to each document.

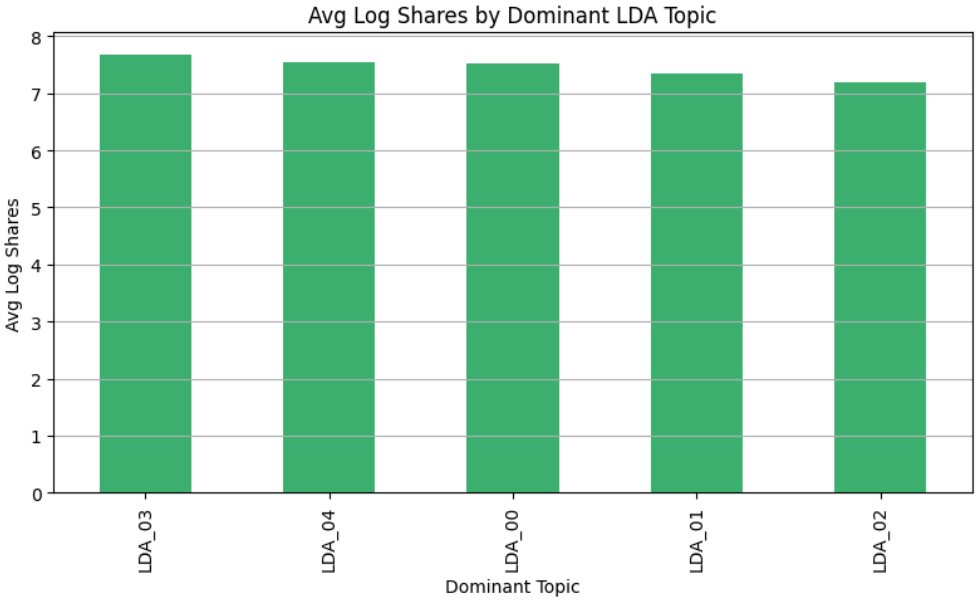

Figure 2.C.23 Dominant Topic

III. Insights Deep-Dive

A. *Significant Features*

Based on the Spearman Correlation with *log_shares*, the most significant features are: *shares, log_shares, kw_avg_avg, kw_avg_avg, self_reference_avg_shares, self_reference_min_shares, self_reference_max_shares, is weekend, data_channel_is_socmed, and global_subjectivity.*

On the other hand, based on a random forest test, the most significant features are: *kw_avg_avg, self_reference_avg_shares, kw_max_avg, self_reference_min_shares, LDA_01, num_hrefs, n_non_stop_unique_tokens, kw_avg_max, and global_subjectivity.*

Lastly, on an XGBoost test, the most significant features are: *data_channel_is_tech, kw_avg_avg, data_channel_is_entertainment, data_channel_is_socmed, self_reference_avg_shares, kw_max_max, is_weekend, self_reference_min_shares, weekday_is_saturday, and kw_max_avg.*

The different approaches yield different results due to their different algorithm for getting the correlation. However, there are three common features among the approaches that indicate that these features play a significant role in the *log_shares*, namely *kw_avg_avg, self_reference_avg,_shares,* and *self_reference_min_shares*. Four additional features are present in 2 out of 3 methods: *kw_max_avg, data_channel_is_socmed, global_subjectivity, and is_weekend.*

B. *Channel Popularity*

The significance of the feature *data_channel_is_socmed* is better highlighted when it's compared to other content channels. The order of the channels based on average log shares is s*ocmed, lifestyle, tech, bus, entertainment, and world.* Socmed has 7.78 shares, followed by 7.61, 7.58, 7.41, 7.31, and 7.20, respectively.

C. *Content Structure and Shares*

The article’s title length shows a dip in its shares when the title length is around 10 - 15 words. On the contrary, it rises when the word ranges from 0 - 5 and 15 - 20. However, it should be taken into account that the dip is 7.46 and the peak is 7.54.

On the other hand, the shares show an upward trend when the article has more content length. This indicates that lengthier articles perform better compared to shorter articles. Articles with no images and rich image (6 images) presence show hiring sharing, while an article with only 1 image shows a significant drop in the rating. Additionally, articles with no videos show lower performance. Lastly, articles with 0 links show significance to the shares, but this drops when there is only 1 link. This gradually goes up when more links are added and peaks at 21 links. This shows how self-reference plays a significant role in shareability.

D. Day of Week

The day of week trend analysis shows that shareability gradually increases when it's near the weekend, specifically Saturday. This shows how releasing an article on the weekend plays a significant role in its shareability.

E. Keyword Significance

The keyword significance analysis shows different results among the keyword features. For the maximum number of shares worst-performing keyword *(kw_min_max)* shows that it peaks at a certain range. This means that even the worst-performing keyword can significantly contribute to the article's shareability. The same can be said of the *kw_avg_avg* and *kw_max_avg*.

On the other hand, self-reference shows high significance to the average log shares but drops when the number of references is too high.

F. Positive and Negative Word Usage

The positive and negative word usage analysis shows that articles are normally distributed; however, it leans slightly towards the negative word rate as seen in *Figure 2.C.18.*

G. Sentiment and Subjectivity

The sentiment and subjectivity analysis shows the performance of an article when it's more emotionally engaging and generic. A title with more sentiment tends to do better, as shown in *Figure 2.C.19* and *Figure 2.C.20.* Additionally, a more negative sentiment is engaging, as shown in *Figure 2.C.21.*

Next, a title that tends to be more subjective performs better with its shareability. This yields the same result when compared to the content of the article; however, it performs well when it's also not subjective.

H. Topic Modeling LDA Impact

For the LDA impact, the results don't differentiate a lot from each other. However, LDA_03 performs the best with 7.68 shareability, followed by

LDA_04, LDA_00, LDA_01, and LDA_02 with 7.55, 7.52, 7.35, and 7.20, respectively.

IV. Conclusion

Among the 61 attributes/columns, there are only a few significant features that affect the shareability of Mashable Inc.'s articles. These features are, namely, *kw_avg_avg, self_reference_avg,_shares, self_reference_min_shares*, *kw_max_avg, data_channel_is_socmed, global_subjectivity, and is_weekend.*

To further understand these features, 7 analyses were conducted: *Channel Popularity Analysis, Content Structure and Shares Analysis, Day of Week Trends Analysis, Keyword Significance Analysis, Positive and Negative Word Usage Analysis, Sentiment and Subjectivity Analysis,* and *Topic Modeling LDA Impact Analysis.*

The *Keyword Significance Analysis* showed the importance of using proper keywords in the article. It showed that the use of proper keywords yields greater article shareability. Additionally, it showed that self-reference within the article significantly affects shareability. This can also be seen in the *Content Structure and Shares Analysis.* The *Channel Popularity Analysis* showed that the channel performs better when it is about social media. On the other hand, the *Day of Week Trends Analysis* showed that people share content more during the weekends, especially on Saturdays. Lastly, an article that is globally subjective performs better, which is shown during *Sentiment and Subjectivity Analysis.*

These different analyses showed a better visualization and reasoning on the few significant features identified during the *EDA Analysis.* From this, we can draw recommendations for Mashable Inc. for better article performance.

V. Recommendation

A. Optimize Keyword Strategy

- Focus on high-performing keywords based on *kw_avg_avg* and *kw_max_avg*. These have consistently shown a high correlation with shareability.
- Avoid keyword overuse. Moderate levels of self-reference (not too low, not too high) lead to higher shares.
- Use keyword analytics to guide writing, ensuring keywords align with previously successful terms and maintain semantic relevance.

B. Include Internal Article References

- Articles with a higher *self_reference_avg_shares* and *self_reference_min_shares* perform better.
- Link internally to other Mashable articles with high share counts to improve content relevance and boost visibility.

C. Target the Right Channel

- *data_channel_is_socmed* had the highest average log shares, proving that social-media-focused content resonates best.
- Allocate more editorial resources toward Social Media and Lifestyle categories.

D. *Publish Toward the Weekend*
- Shares increase on Saturdays and generally trend upward toward weekends.
- Schedule high-potential articles to go live between Thursday and Saturday for maximum reach.

E. *Enhance Visual Engagement*
- Articles with 0 or 6 images perform better than those with 1–5.
- Videos also improve shareability when used in moderate amounts.
- Aim for either highly visual content (5+ images) or text-heavy content with zero images. Avoid minimal imagery.

F. *Use Emotion and Subjectivity*
- Articles with moderate to high subjectivity and emotional tone (especially negative or controversial ones) perform better.
- Use subjective titles and content that trigger emotional reactions (e.g., curiosity, surprise, concern).

G. *Leverage Longer Articles*
- Articles with greater content length tend to have higher share counts.
- Produce in-depth, well-structured articles of above-average word count, avoiding overly short posts.

H. *Apply Topics Modeling Insights*
- While topic effects were subtle, LDA_03 and LDA_04 were linked to better performance.
- Use LDA to pre-label future content and prioritize writing topics that align with LDA_03 themes.