Data Visualization Project

Introduction

By identifying the factors that contribute to an article's virality, content creators can optimize their articles to increase the likelihood of them being shared and viewed by a larger audience. This can involve tweaking headlines, article length, multimedia elements, or writing style to better engage readers.

Viral articles typically result in a significant increase in website traffic, which can lead to higher ad revenue and increased opportunities for monetization. Understanding the factors that drive virality can provide a competitive edge in the fast-paced world of digital media. Content creators who consistently produce viral content will likely stand out from the competition and attract more readers and followers.

Project Description

Using the Online News Popularity dataset, we hope to investigate the factors that have a major impact on an article's virality. This dataset provides an extensive assessment of numerous parameters relevant to Mashable articles published over a two-year period, with the goal of forecasting the number of social network shares (popularity). The original dataset had 61 columns and 39,644 rows, with 60 float data types, one text column, and one integer column serving as the target variable. Although the dataset was created for regression problems, we will change our modeling strategy by changing it to a classification challenge. We will divide the number of shares in the target column into five percentile bins to enable for more precise analysis and interpretation. Among other things, we will seek to:

1. Investigate the impact of article length on social media shares; analyzing the relationship between the number of words in an article and its propensity to be shared across various platforms.
2. Assess the correlation between title length and article popularity: evaluating the effectiveness of concise versus extended titles in driving article virality.

3. Evaluate the influence of stopwords on article shareability; examining the role of common stopwords and determining their effect on the overall shareability of an article.
4. Compare the effectiveness of external versus internal links in driving article virality; analyzing the impact of incorporating external resources and references in comparison to internal links on reader engagement and social media sharing.
5. Examine the role of multimedia elements in enhancing article appeal; investigating the influence of images, videos, and other multimedia components on their potential to boost article popularity and shareability.
6. Analyze the optimal word length for maximizing reader engagement; identifying the most effective word length range to maintain reader interest and increase the likelihood of an article being shared.
7. Identify the most popular content genres for attracting social media shares; examining various content categories to determine which genres resonate the most with audiences and have the highest potential for virality.
8. Explore the relationship between article subjectivity and its propensity to go viral; investigating the impact of an article's tone, sentiment, and emotional appeal on its potential for virality.
9. Assess the impact of polarizing language on article shareability; examining the use of provocative or controversial language and its effect on driving social media shares.
10. Evaluate the effects of employing high-performing, low-performing, and average-performing keywords on article popularity; investigating the strategic use of keywords to optimize shareability.
11. Identify the most influential attributes that impact an article's virality; deduce from the data the critical factors that drive an article's popularity and optimize its shareability.

Members

Name	Tool
Ghiță, Alexandru Ștefan	ggplot
Tomeh, Lara	Tableau

Wali, Ahmad Mustapha	Plotly

Appendix

Feature Information

1. url: The article's web address.
2. timedelta: Days between the article publication and dataset acquisition.
3. n_tokens_title: Word count in the title.
4. n_tokens_content: Word count in the content.
5. n_unique_tokens: Ratio of unique words in the content.
6. n_non_stop_words: Ratio of non-stop words in the content.
7. n_non_stop_unique_tokens: Ratio of unique non-stop words in the content.
8. num_hrefs: Number of links in the article.
9. num_self_hrefs: Number of links to other Mashable articles.
10. num_imgs: Number of images in the article.
11. num_videos: Number of videos in the article.
12. average_token_length: Average word length in the content.
13. num_keywords: Number of keywords in the metadata.

14-18. data_channel_*: Binary flags for data channel categories (e.g., Lifestyle, Entertainment, Business, etc.).*
*19-27. kw_*: Metrics related to the worst, best, and average keywords (based on shares).
28-30. self_reference_*shares: Min, max, and average shares of referenced Mashable articles.*
31-37. weekday_is: Binary flags for days of the week the article was published.

38. is_weekend: Binary flag for weekend publication.

39-43. LDA_*: Closeness to LDA topics 0 to 4, based on Latent Dirichlet Allocation, a topic modeling technique.
44. global_subjectivity: Text subjectivity, which measures the degree of personal opinion, emotion, or judgment in the content.

45. global_sentiment_polarity: Text sentiment polarity, which measures the overall positivity or negativity of the content.
46. global_rate_positive_words: Ratio of positive words in the content.
47. global_rate_negative_words: Ratio of negative words in the content.
48. rate_positive_words: Ratio of positive words among non-neutral tokens.
49. rate_negative_words: Ratio of negative words among non-neutral tokens.
50. avg_positive_polarity: Average polarity of positive words.
51. min_positive_polarity: Minimum polarity of positive words.
52. max_positive_polarity: Maximum polarity of positive words.
53. avg_negative_polarity: Average polarity of negative words.
54. min_negative_polarity: Minimum polarity of negative words.
55. max_negative_polarity: Maximum polarity of negative words.
56. title_subjectivity: Subjectivity of the article's title.
57. title_sentiment_polarity: Sentiment polarity of the article's title.
58. abs_title_subjectivity: Absolute subjectivity level of the title.
59. abs_title_sentiment_polarity: Absolute polarity level of the title.
60. shares: Number of shares of the article on social media (target variable).