A close-up of words

Description automatically generated

Online News Popularity Analysis

10.25.2023

**─**

Sanchit Chaudhary - U48485460

ISM 6136 – Data Mining

Muma College of Business

University of South Florida

**Overview**

**Business Case Analysis: Unlocking the Secrets of Article Popularity**

In the vast world of digital media, understanding what captures everyone's attention on social networks a bit is like solving a captivating puzzle. **Ever wondered** why some articles seem to rule our feeds continuously? Or why a handful of content creators are always in the spotlight, while others struggle to be noticed? What makes some articles stay in the limelight for a long time, while others fade away like shooting stars? And why is it that certain days of the week bring specific topics into the trending spotlight?

The answers to these intriguing questions lie in the **art of predicting article popularity**, particularly the number of times they get shared on social media. Through an in-depth business case analysis, we can unravel the complexities of why some articles become trendy.

Our main goal in analyzing the **"Online News Popularity"** project is to help content creators improve their articles i.e., focus on **content optimization, audience segmentation** etc. for social media. We want to understand what makes articles go viral and use that knowledge to create effective, long-lasting strategies for online success.

# Dataset

This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks (popularity).

* The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them.
* Acquisition date: January 8, 2015.

Attribute Information:

Number of Attributes: 61 (58 predictive attributes, 2 non-predictive, 1 goal field)

## **Details about Data Set:**

Here are details about the column names in the provided dataset:

1. url: URL of the article (non-predictive)
2. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
3. n\_tokens\_title: Number of words in the title
4. n\_tokens\_content: Number of words in the content
5. n\_unique\_tokens: Rate of unique words in the content
6. n\_non\_stop\_words: Rate of non-stop words in the content
7. n\_non\_stop\_unique\_tokens: Rate of unique non-stop words in the content
8. num\_hrefs: Number of links
9. num\_self\_hrefs: Number of links to other articles published by Mashable
10. num\_imgs: Number of images
11. num\_videos: Number of videos
12. average\_token\_length: Average length of the words in the content
13. num\_keywords: Number of keywords in the metadata
14. data\_channel\_is\_lifestyle: Is data channel 'Lifestyle'?
15. data\_channel\_is\_entertainment: Is data channel 'Entertainment'?
16. data\_channel\_is\_bus: Is data channel 'Business'?
17. data\_channel\_is\_socmed: Is data channel 'Social Media'?
18. data\_channel\_is\_tech: Is data channel 'Tech'?
19. data\_channel\_is\_world: Is data channel 'World'?
20. kw\_min\_min: Worst keyword (min. shares)
21. kw\_max\_min: Worst keyword (max. shares)
22. kw\_avg\_min: Worst keyword (avg. shares)
23. kw\_min\_max: Best keyword (min. shares)
24. kw\_max\_max: Best keyword (max. shares)
25. kw\_avg\_max: Best keyword (avg. shares)
26. kw\_min\_avg: Avg. keyword (min. shares)
27. kw\_max\_avg: Avg. keyword (max. shares)
28. kw\_avg\_avg: Avg. keyword (avg. shares)
29. self\_reference\_min\_shares: Min. shares of referenced articles in Mashable
30. self\_reference\_max\_shares: Max. shares of referenced articles in Mashable
31. self\_reference\_avg\_sharess: Avg. shares of referenced articles in Mashable
32. weekday\_is\_monday: Was the article published on a Monday?
33. weekday\_is\_tuesday: Was the article published on a Tuesday?
34. weekday\_is\_wednesday: Was the article published on a Wednesday?
35. weekday\_is\_thursday: Was the article published on a Thursday?
36. weekday\_is\_friday: Was the article published on a Friday?
37. weekday\_is\_saturday: Was the article published on a Saturday?
38. weekday\_is\_sunday: Was the article published on a Sunday?
39. is\_weekend: Was the article published on the weekend?
40. LDA\_00: Closeness to LDA topic 0
41. LDA\_01: Closeness to LDA topic 1
42. LDA\_02: Closeness to LDA topic 2
43. LDA\_03: Closeness to LDA topic 3
44. LDA\_04: Closeness to LDA topic 4
45. global\_subjectivity: Text subjectivity
46. global\_sentiment\_polarity: Text sentiment polarity
47. global\_rate\_positive\_words: Rate of positive words in the content
48. global\_rate\_negative\_words: Rate of negative words in the content
49. rate\_positive\_words: Rate of positive words among non-neutral tokens
50. rate\_negative\_words: Rate of negative words among non-neutral tokens
51. avg\_positive\_polarity: Avg. polarity of positive words
52. min\_positive\_polarity: Min. polarity of positive words
53. max\_positive\_polarity: Max. polarity of positive words
54. avg\_negative\_polarity: Avg. polarity of negative words
55. min\_negative\_polarity: Min. polarity of negative words
56. max\_negative\_polarity: Max. polarity of negative words
57. title\_subjectivity: Title subjectivity
58. title\_sentiment\_polarity: Title polarity
59. abs\_title\_subjectivity: Absolute subjectivity level
60. abs\_title\_sentiment\_polarity: Absolute polarity level
61. shares: Number of shares (target)

# Goals

The primary objective of this business case is to build a predictive model to estimate the number of shares an article published by Mashable will receive on social networks. By creating a robust predictive model, content creators and digital marketing teams can gain valuable insights into the factors that influence article shareability. This project aims to empower content creators to produce more shareable content, improve content marketing strategies, and enhance the overall online presence of Mashable.

**Key Problems and Objectives:**

**1. Predictive Modeling:** Develop a predictive model that can accurately estimate the number of shares an article is likely to receive. This model should be based on historical data and various attributes related to the articles.

**2. Feature Selection:** Identify which attributes (e.g., content length, sentiment, keyword usage, publication day) have the most significant impact on article popularity. This helps in focusing content optimization efforts on the most influential factors.

**3. Algorithm Selection:** Evaluate and compare various machine learning algorithms, such as regression models, decision trees, random forests, or neural networks, to determine which one provides the most accurate predictions for article shareability.

**4. Model Evaluation:** Implement rigorous model evaluation techniques, including metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), or R-squared, to assess the performance and accuracy of the predictive model.

**5. Interpretability:** Ensure that the predictive model is interpretable so that content creators can understand why certain articles are predicted to perform well or poorly in terms of shares.

**6. Content Strategy Recommendations:** Based on the insights from the predictive model, provide actionable recommendations to content creators and editors on how to improve their articles for better shareability. This could involve suggestions for content length, sentiment, keyword optimization, and timing of publication.

**7. Long-term Strategy:** Use the model to inform the long-term content strategy of Mashable. Identify trends and patterns in article popularity over time and adjust content creation efforts accordingly.

In summary, the business case of predicting article popularity aims to leverage data and predictive modeling techniques to enhance content strategy, increase shareability, and achieve better performance for articles published by Mashable on social networks. The insights derived from this predictive model can guide content creators and marketing teams to make informed decisions and optimize their content marketing efforts effectively.

# Methodology:

Python will serve as the primary programming language for constructing solutions, with the utilization of Python libraries like pandas and numpy for data analysis and transformation as needed. Additionally, we will harness open-source libraries from scikit-learn (sklearn) for the implementation of machine learning models, including Logistic Regression, KMeans and others.