ABSTRACT

The internet has revolutionized the way news is shared, making it vital for content creators and advertisers to predict the popularity of online articles before they are published. In this study, we employ machine learning techniques to forecast the popularity of news articles using features such as keywords, topics, and publication times. Our dataset, sourced from Kaggle, comprises articles published by Mashable over a two-year period. By analyzing various factors influencing article popularity, we aim to provide insights that can help optimize content creation strategies.

INTRODUCTION

With news spreading quickly online, it's important to know what makes articles popular. This study tries to guess how popular an article will be before it's published using Machine learning methods. We want to find out what things, like keywords and topics, make articles popular. This can help people who make and advertise content make better decisions.

**Advent of the Internet 🡪 Information Boom:**

· The rise of the internet has led to an explosion of information being shared worldwide.

· With easy access to news and articles online, people are constantly consuming and sharing information at a rapid pace.

**Defining Popularity:**

· In the online world, the popularity of content is often measured by metrics like the number of views, shares, likes, and comments.

· The more a piece of content is shared or engaged with, the more popular it is considered to be.

**Predicting Popularity:**

· Using machine learning techniques, it's possible to predict the popularity of online content even before it's published.

· Factors such as the choice of keywords, relevance to current trends, and timing of publication are analyzed to make these predictions.

**Utility – Preemptive Tweaking:**

· Predicting popularity before publication allows content creators to make preemptive tweaks to their content.

· By identifying potential weaknesses or areas for improvement, creators can enhance their content to increase its chances of being popular once it's out in the world**.**

* PRIOR RESEARCH

In our past research, we looked at why some news articles online become really popular and get shared a lot. We wanted to understand what makes people want to share them. Our goal was to give useful tips on how to make content do better on the internet.

* SIGNIFICANCE OF RESEARCH

The significance of this research lies in its potential to enhance our understanding of how online content spreads and becomes popular. By identifying the factors that influence content sharing, we can develop strategies to improve the performance of digital content. This knowledge is valuable for content creators, marketers, and anyone seeking to engage audiences effectively in the digital landscape. Ultimately, it can lead to more impactful communication and better engagement with online audiences.

* RESEARCH QUESTIONS

Question 1: Which types of articles (by category, length, and multimedia usage) consistently receive the most shares, and how can we tailor content production to these findings?

Question 2: What are the key times or days for publishing that correlate with higher shares, and how can we adjust our publication schedule accordingly?

Question 3: What specific recommendations can we make to the marketing team based on engagement?

* **Dataset description and history**

Our dataset comprises articles published by Mashable over a two-year period. It includes various features such as article content, publication dates, and social media shares. The dataset has been curated from Kaggle and is designed for predictive modeling of article popularity.Mashable allows readers to share articles through various social media platforms and other digital channels. Each article typically includes share buttons or links that enable readers to easily share content they find interesting or informative.

* + **Map variables used**

· Url: URL of the article (non-predictive)

· timedelta: Days between the article publication and the dataset acquisition (non-predictive)

· ntokenstitle: Number of words in the title

· ntokenscontent: Number of words in the content

· nuniquetokens: Rate of unique words in the content

· nnonstop\_words: Rate of non-stop words in the content

· nnonstopuniquetokens: Rate of unique non-stop words in the content

· num\_hrefs: Number of links

· numselfhrefs: Number of links to other articles published by Mashable

· num\_imgs: Number of images

· num\_videos: Number of videos

· averagetokenlength: Average length of the words in the content

· numkeywords: Number of keywords in the metadata

· datachannelislifestyle: Is data channel 'Lifestyle'?

· datachannelis\_entertainment: Is data channel 'Entertainment'?

· datachannelis\_bus: Is data channel 'Business'?

· datachannelis\_socmed: Is data channel 'Social Media'?

· datachannelis\_tech: Is data channel 'Tech'?

· datachannelis\_world: Is data channel 'World'?

· kwminmin: Worst keyword (min. shares)

· kwmaxmin: Worst keyword (max. shares)

· kwavgmin: Worst keyword (avg. shares)

· kwminmax: Best keyword (min. shares)

· kwmaxmax: Best keyword (max. shares)

· kwavgmax: Best keyword (avg. shares)

· kwminavg: Avg. keyword (min. shares)

· kwmaxavg: Avg. keyword (max. shares)

· kwavgavg: Avg. keyword (avg. shares)

· selfreferencemin\_shares: Min. shares of referenced articles in Mashable

· selfreferencemax\_shares: Max. shares of referenced articles in Mashable

· selfreferenceavg\_sharess: Avg. shares of referenced articles in Mashable

· weekdayismonday: Was the article published on a Monday?

· weekdayistuesday: Was the article published on a Tuesday?

· weekdayiswednesday: Was the article published on a Wednesday?

· weekdayisthursday: Was the article published on a Thursday?

· weekdayisfriday: Was the article published on a Friday?

· weekdayissaturday: Was the article published on a Saturday?

· weekdayissunday: Was the article published on a Sunday?

· is\_weekend: Was the article published on the weekend?

· LDA\_00: Closeness to LDA topic 0

· LDA\_01: Closeness to LDA topic 1

· LDA\_02: Closeness to LDA topic 2

· LDA\_03: Closeness to LDA topic 3

· LDA\_04: Closeness to LDA topic 4

· global\_subjectivity: Text subjectivity

· globalsentimentpolarity: Text sentiment polarity

· globalratepositive\_words: Rate of positive words in the content

· globalratenegative\_words: Rate of negative words in the content

· ratepositivewords: Rate of positive words among non-neutral tokens

· ratenegativewords: Rate of negative words among non-neutral tokens

· avgpositivepolarity: Avg. polarity of positive words

· minpositivepolarity: Min. polarity of positive words

· maxpositivepolarity: Max. polarity of positive words

· avgnegativepolarity: Avg. polarity of negative words

· minnegativepolarity: Min. polarity of negative words

· maxnegativepolarity: Max. polarity of negative words

· title\_subjectivity: Title subjectivity

· titlesentimentpolarity: Title polarity

· abstitlesubjectivity: Absolute subjectivity level

· abstitlesentiment\_polarity: Absolute polarity level

· shares: Number of shares (target)

*Source:* [*kaggle*](https://www.kaggle.com/datasets/thehapyone/uci-online-news-popularity-data-set)It was crucial to understand the data we had. This dataset was put together to help with a research paper, and you can find more about that paper at the end. After experts pulled out important information from the text, they shared the dataset with the UCI Machine Learning Repository for others to use

* + **Introduce new variables used in our analysis (if we created any)**
  + **DayOfWeek**: We merged the individual columns representing each day of the week into a single column indicating the day on which the article was published. This allowed us to analyze the effect of publication day on article popularity.
  + **article\_channel**: Similarly, we merged the individual columns representing different article channels (e.g., lifestyle, entertainment) into a single column indicating the channel of the article. This helped us analyze the effect of article category on popularity.
  + **target**: We created a new binary variable called "target" to indicate whether an article was popular or not based on the number of shares it received. This variable was used in classification tasks.
  + **target1**: Another binary variable indicating whether an article was popular or unpopular, created for visualization purposes.
  + **HighEngagement**: We binarized the target variable into "HighEngagement" to classify articles as having high or low engagement based on the median number of shares.

METHODOLOGY

* **Data Pre-Processing**

​​We conducted several pre-processing steps on the dataset:

· **Trimming Column Names:** We removed any leading or trailing white spaces from column names to ensure consistency.

· **Handling Missing Values:** We checked for any missing values in the dataset and addressed them appropriately.

· **Column Merging:** We merged weekday columns into a single 'DayOfWeek' column and data channel columns into a single 'article\_channel' column for easier analysis.

· **Column Removal:** We removed non-predictive columns such as 'url' and 'timedelta', as well as columns with zero values in the 'n\_tokens\_content' column.

· **Data Transformation:** We transformed certain categorical variables into factor variables for analysis.

· **Outlier Removal:** We identified and removed outliers using the Interquartile Range (IQR) method to ensure the robustness of our analysis.

· **Correlation Analysis:** We examined correlations between numerical variables and dropped highly correlated columns to avoid multicollinearity issues.

· **Final Check:** Finally, we performed a final check on the dataset's structure and summary statistics to ensure it was ready for exploratory data analysis."

**-Understanding the Data**

In our data understanding phase, we delved into various aspects of the dataset to gain insights:

· **Descriptive Statistics:** We calculated summary statistics such as mean, median, standard deviation, and quartiles for numerical features to understand their central tendency and dispersion.

· **Data Visualization:** Utilizing visual aids like histograms, box plots, and scatter plots, we explored the distributions and relationships among different variables.

· **Feature Importance:** We employed techniques like correlation analysis and feature importance scores to identify key predictors and understand their impact on the target variable.

· **Data Quality Assessment:** We assessed the quality of the data by examining missing values, outliers, and inconsistencies to ensure the reliability of our analysis.

· **Domain Knowledge Integration:** Incorporating domain knowledge, we contextualized the data, identifying patterns and anomalies that might influence our modelling decisions.

· **Data Splits:** We partitioned the dataset into training, validation, and test sets, ensuring that each subset adequately represented the overall data distribution.