**Comprehensive Analysis of** “Mashable” News Articles Dataset



INSTITUTE OF BUISNESS ADMINISTRATION

KARACHI

**PROJECT REPORT**

(APPLICATION DEVELOPMENT)

**GROUP MEMBERS**

Musab Bin Jamil & Hussain Diwan

Abstract

This study systematically analyzes the Mashable news articles dataset, focusing on news genres, sentiment analysis, and sharing patterns. Employing robust methodologies, the research categorizes genres, evaluates sentiment, and scrutinizes user engagement. Utilizing statistical techniques, the study aims to provide nuanced insights into news dissemination and audience reception to improve the shareability of future posts by Mashable.

**Table of Contents**

1. Introduction
2. About the Dataset
3. Our Work
4. Results and Discussion
5. Conclusion
6. References
7. Appendices:

Introduction

The dataset under scrutiny originates from Mashable, a prominent online news source, and is created by a team of dedicated content creators. Analyzing news article data, especially from Mashable, is pertinent due to its significant influence on digital media consumption patterns. Understanding the inherent structures of news genres and sentiments within this dataset provides valuable insights into content preferences and audience engagement. This report aims to elucidate these aspects through a comprehensive analysis, focusing on categorizing news genres and conducting sentiment analysis. By delving into the scope of this dataset, we endeavor to contribute to a nuanced comprehension of the dynamics shaping digital news consumption.

Background

Mashable, a reputable digital media outlet, stands as a comprehensive news source known for its diverse coverage across entertainment, business, technology, lifestyle, world affairs, and social media. The dataset encompasses these distinct news genres, each representing a unique facet of contemporary information. The data collection process employs the Random Forest classifier, a robust algorithm facilitating accurate categorization. Additionally, this study employs Natural Language Processing (NLP) techniques to delve into sentiment analysis, aiming to discern the emotional tone within the articles. This background sets the stage for a thorough exploration of Mashable's varied content and the analytical tools applied in elucidating patterns and sentiments within the dataset.

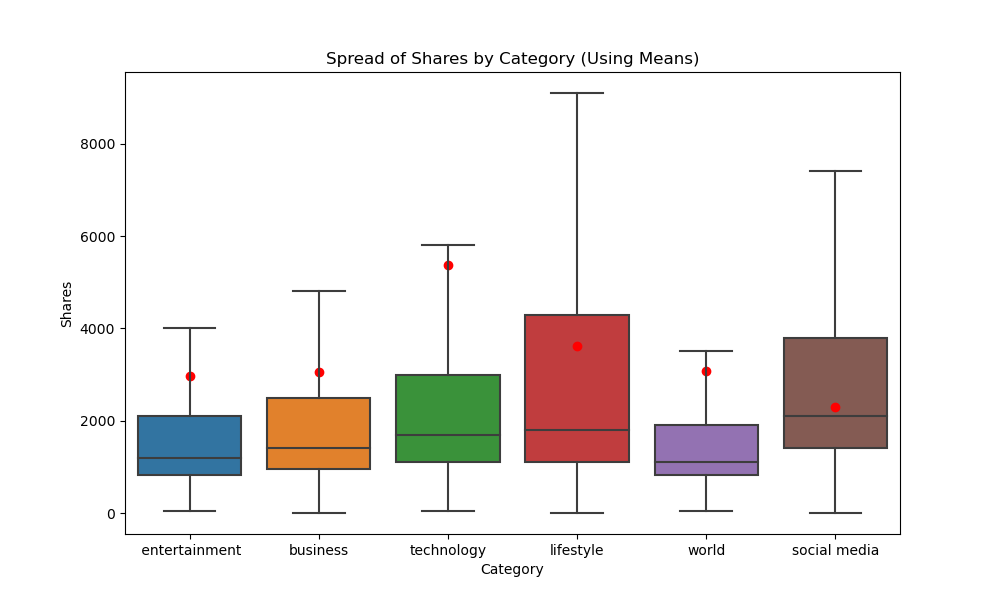
Data Handling and Preprocessing

Data handling and preprocessing involved a meticulous approach for refining the initial dataset structure, comprising 60 columns. The preprocessing steps aimed at enhancing analysis efficiency included the implementation of Python code snippets. The initial structure was streamlined through reverse encoding, and dropping a few columns which reduced the dataset to 12 columns. This reduction was strategically undertaken to focus on pertinent variables, optimizing computational resources and improving the interpretability of subsequent analyses. Cleaning and normalization procedures were integral, addressing potential inconsistencies and outliers in the data. The Python code snippets employed for these procedures can be found in the appendix.

Data Analysis

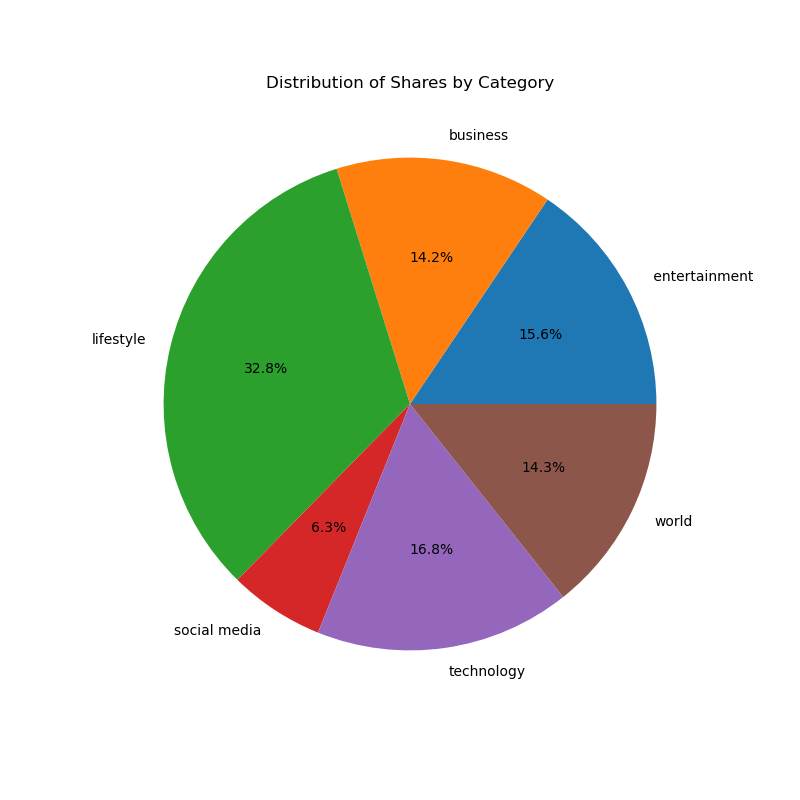
Boxplot (Shares Distribution by Genre)

The boxplot visually represents the distribution of shares across different news genres. It reveals variations in audience engagement, highlighting potential variations in the popularity of articles within each genre. From this plot are initially mislead to think that lifestyle articles are more common but upon analysis of averages we can see (through the red dots) that technology news is shared more often than others.



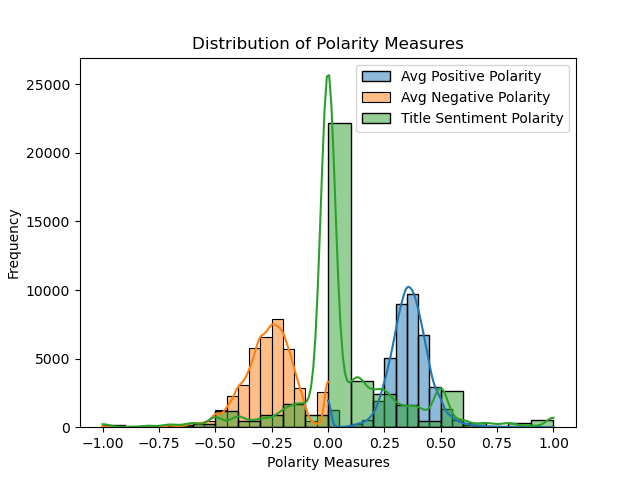
Pie Chart (Genre Article Distribution):

The pie chart illustrates distribution of articles across genres. This visualization provides insights into the generation patterns of article publication. As we can see the number of posts for lifestyle articles out weigh every other category by at least twice the article count. Further on we will discuss how this is should be handled.



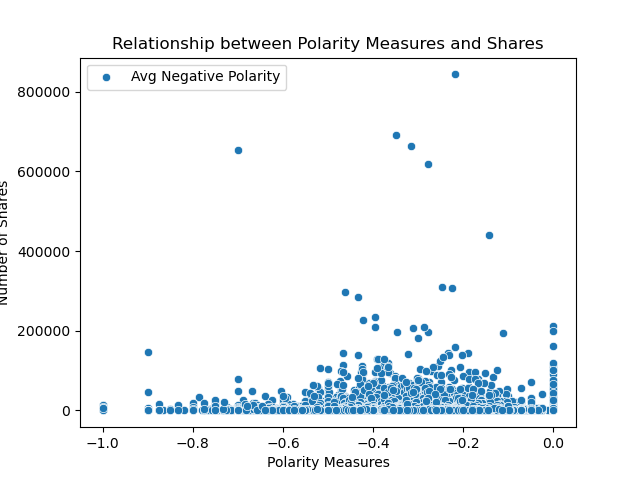
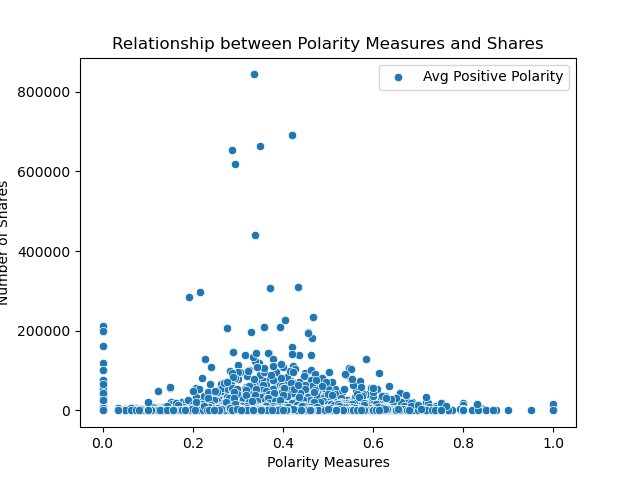
Sentiment Analysis

The sentiment analysis evaluates the positive and negative polarity of the articles, shedding light on the overall tone of the content. Findings indicate the prevailing sentiments within the dataset, influencing reader perception. By analyzing viral posts, we were able to identify that the majority comprises of posts with neutral polarity.



A graph of a relationship between polarity measures and shares

Description automatically generated



Trend Analysis:

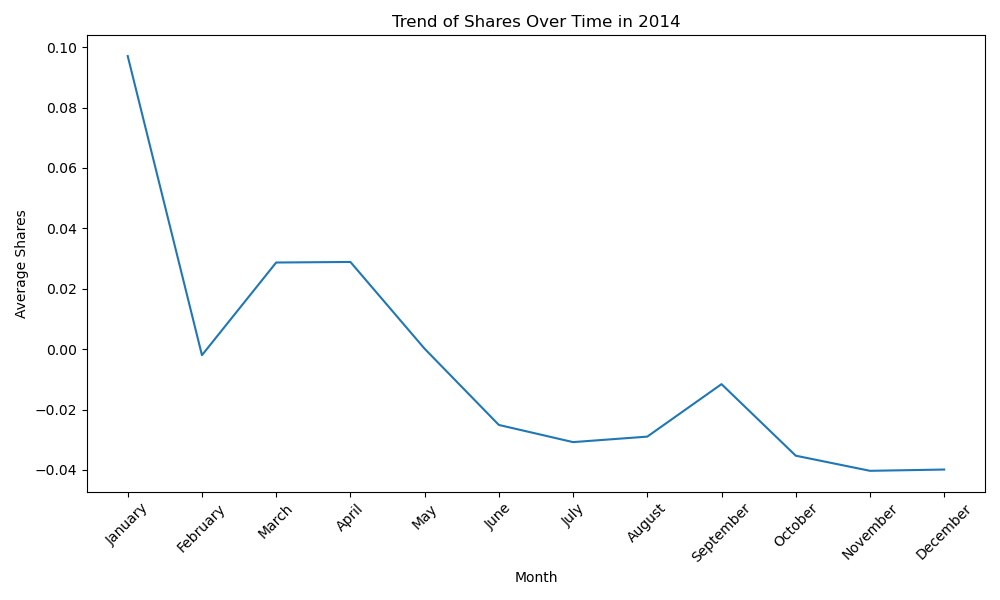
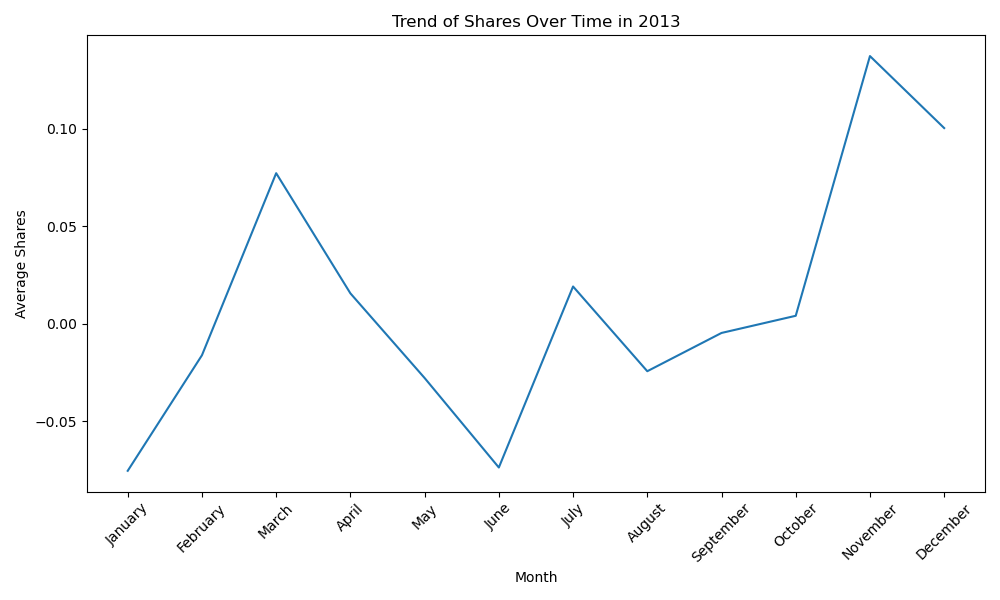
Exploring article share patterns over two years unveils trends in audience engagement. As we previously identified that technology posts have more average shares, Mashable should focus on producing more tech related articles compared to their current focus on lifestyle.

A bar graph with different colored bars

Description automatically generated

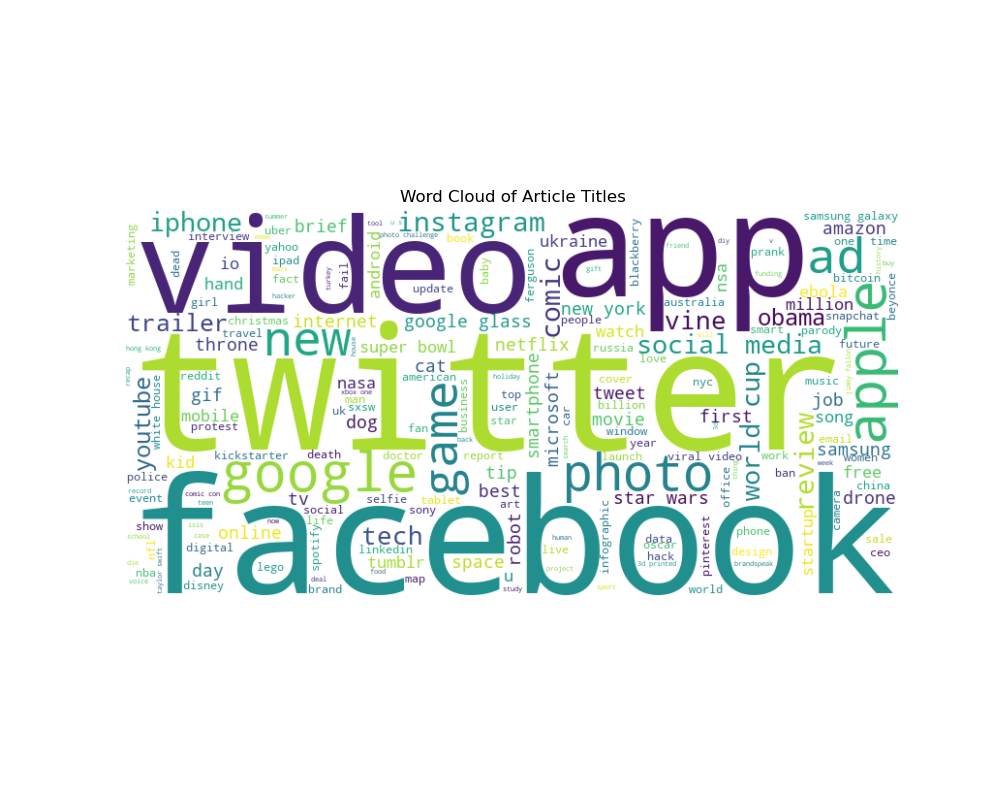
Share pattern over yearly span

The x-axis representing individual months and the y-axis denoting the corresponding average share values. The graph provides a visual narrative of how sharing patterns fluctuate throughout the year, offering insights into potential seasonal variations or noteworthy events influencing audience engagement. The rising or falling trajectory of the line signifies the general trend in average shares, allowing for a quick and intuitive understanding of the evolving dynamics in content dissemination over the specified timeframe.



Word Cloud:

A word cloud, as a future addition, can visually represent the most common words in the articles. This adds to the overall appeal of the website whilst showing there usage of common and uncommon words and therefore providing insights into vocabulary improvements for the future.



Discussion:

The boxplot analysis reveals distinct patterns in shares distribution across news genres. Notably, genres such as technology and world news appear to garner higher shares on average, suggesting a heightened audience interest in these topics. Conversely, lifestyle and social media genres exhibit comparatively lower average shares, indicative of potential variations in reader engagement.

The pie chart depicting article distribution by genres provides insights into content publishing trends. A consistent growth is observed for lifestyle articles, which is not the way to go. We suggest an adaptive content release strategy by Mashable, ensuring a continuous presence for readers.

Sentiment analysis offers nuanced insights into the emotional tone of articles. Positive sentiment, for instance, might correlate with higher shares, indicating a potential preference for optimistic content. Conversely, negative sentiment could influence sharing patterns differently. Correlational analysis, however results in no such relation between posts. In observation neutral posts are more commonly shared. Further investigation on sentiment polarity and article shares would shed light on the interplay between emotional tone and audience engagement.

The trend analysis over two years unravels evolving patterns in article sharing. Identifying peaks and troughs in share counts may correlate with major events, revealing the impact of external factors on content virality. Unexpected insights may emerge, prompting further exploration into specific periods that deviate from expected trends.

In summary, the combination of these analyses provides a comprehensive understanding of Mashable's news articles dataset, offering valuable insights into genre preferences, temporal dynamics, emotional resonances, and the evolving nature of article sharing patterns. Further investigation into specific correlations and unexpected findings could yield deeper insights into audience behavior and content virality.

Conclusion

The analysis unveils technology and world news as consistently popular, while lifestyle and social media genres attract more modest engagement. Despite minor correlation between sentiment and shares, the conclusion implies that sentiment polarity may not wield a significant impact on overall share counts. For content creators and marketers, tailoring content to popular genres and maintaining a positive tone emerges as pivotal for enhancing virality. Strategic planning, particularly around temporal publishing trends, remains crucial for optimizing audience reach. The call for further research into sentiment's intersection with specific content types within genres suggests avenues for deeper exploration. These insights extend beyond the study, offering valuable guidance for refining strategies across diverse content platforms and marketing campaigns. Additionally, the referenced statement "We show that a negative news post receives a higher number of comments and shares compared to a positive news post, which gets a higher number of likes" adds a nuanced layer, highlighting the complex dynamics of user engagement based on sentiment in news posts.

References:

Cite the dataset creators, any methodologies, and additional research sources.

https://www.sciencedirect.com/science/article/abs/pii/S2468696418300260

Appendices:

Data Tables:

Table 1: Original Data (60 Columns)

Table 2: Final Project Data (12 Columns)

Code Snippets:

Code 1: Data Cleaning

Code 2: Discretization, Normalization and Train-Test split

Code 3: Data Cleaning

Dash App:

Code 1: Dash App

Code 2: home

Code 3: about

Code 4: share

Code 5: sentiment

Code 6: trend