INTERIM REPORT

Prepared by: Rebka Ayele

Introduction

In this report, I analyze a dataset of news articles with a focus on headline lengths, publisher activity, and publication trends. My goal is to understand the dataset's characteristics and identify any significant patterns.

Data Source

The analysis is based on data loaded from the CSV file named raw_analyst_ratings.csv.

Descriptive Statistics

Headline Lengths

I computed descriptive statistics for the lengths of headlines using the describe() method. The key metrics are as follows:

- **Mean**: The average length of headlines in characters.
- **Median**: The middle value of headline lengths when sorted.
- **Standard Deviation**: A measure of the variation in headline lengths.
- **Minimum**: The length of the shortest headline.
- **Maximum**: The length of the longest headline.

Insights:

- The **mean** and **median** provide insight into the typical length of headlines.
- The **standard deviation** indicates how much headline lengths vary from the mean.
- The **minimum** and **maximum** values highlight the range of lengths observed.

Histogram Visualization

I created a histogram to visualize the distribution of headline lengths. This visualization helps to determine if the lengths are clustered around a central value or if there is a notable skew towards shorter or longer headlines.

Publisher Activity

I analyzed publisher activity by counting the number of articles published by each source using the value_counts() method. The results were sorted in descending order to identify the most active publishers.

Insights:

- This analysis reveals which publishers contribute the most articles, providing a sense of their prominence within the dataset.
- It also helps understand the diversity of sources included in the data.

Daily Publication Frequency

I grouped articles by their publication date and counted the number of articles published each day. This data was visualized using a line chart to identify trends over time.

Insights:

- The analysis shows trends and fluctuations in publication activity.
- It highlights days with unusually high or low numbers of articles, which might correlate with specific events or time periods.
- Further analysis could involve examining these trends in relation to notable events or particular days of the week.

Text Analysis

Sentiment Analysis

I conducted sentiment analysis to assess the overall tone of the headlines, categorizing them into positive, negative, or neutral based on sentiment scores. This involved cleaning the text data and applying sentiment analysis techniques.

Insights:

- Sentiment analysis provides a view of the general tone of news coverage, helping to understand how the sentiment is distributed across headlines.
- Keyword Extraction: I performed keyword extraction to identify commonly occurring terms or entities in the headlines, which can reveal thematic trends. Although this part of the analysis encountered some challenges, I am continuing to refine the results by consulting additional resources and adjusting methodologies.

Conclusion

This interim report summarizes my findings from the analysis of headline lengths, publisher activity, and publication trends. Sentiment analysis has been successfully completed, while keyword extraction is ongoing and being refined. The report provides a foundational understanding of the dataset and sets the stage for further exploration and analysis.