**Introduction**

In the world of news and media, certain words often come to define the stories we read. Whether it's a name, a place, or a recurring theme, these words can give us clues about the subject of an article before we even dive into the details. For instance, when you think of articles about political figures like Donald Trump, Kamala Harris, Tim Walz, or J.D. Vance, certain words or phrases are likely to pop up more frequently in association with each of them. By examining which words appear most often in news articles, we can start to see patterns that might tell us who or what the article is really about.

But it’s not just about identifying the topic. The words used in popular news sources can also reveal potential biases or a kind of groupthink about a specific person. Certain descriptors or phrases might be used repeatedly across different articles, shaping our perception of these figures in subtle ways. By analyzing word frequency, we can gain insights into not only the subject of the articles but also how the media portrays these individuals, possibly reflecting underlying biases or collective opinions that influence public perception. This approach provides a window into understanding both the content and the context of how news is delivered.

**Analysis**

**Data Preparation**

This section focuses on the process of preparing the data for analysis. The steps include loading and cleaning the data, labeling it appropriately, and visualizing it to gain initial insights. Proper data preparation is crucial to ensure that the subsequent analysis is accurate and meaningful.

1. **Overview of the Data:** The dataset used in this analysis consists of news articles retrieved from the NewsAPI. These articles cover topics related to specific public figures, including Donald Trump, Kamala Harris, Tim Walz, and J.D. Vance. The data includes various attributes such as the publication date, source, title, description, and content of each article.An example of the raw data in JSON format can be found below:

**A screenshot of a computer

Description automatically generated**

1. **Data Loading and Cleaning**: The first step in the process is loading the data from various sources and cleaning it to remove any inconsistencies or irrelevant information. The following tasks were performed:
   1. **Loading Data**: Articles are retrieved using the NewsAPI for each public figure using their last names as a search criteria and limiting results to the top 100 most popular articles since 6 August 2024.
   2. **Cleaning Data**: The data is cleaned by:
      * Removing unnecessary characters, punctuation, and whitespace.
      * Normalizing the text by converting it to lowercase.
      * Removing stopwords and words directly related to the figures’ names (i.e. their first and last names) to prevent skewing the analysis.
      * Dropping irrelevant columns including author, source ID, url, and urlToImage
      * Reformatting publishedAt to just a date
      * Renaming ‘description’ to ‘headline’ and ‘publishedAt’ to ‘Date’

Data Before Cleaning:

A screenshot of a computer

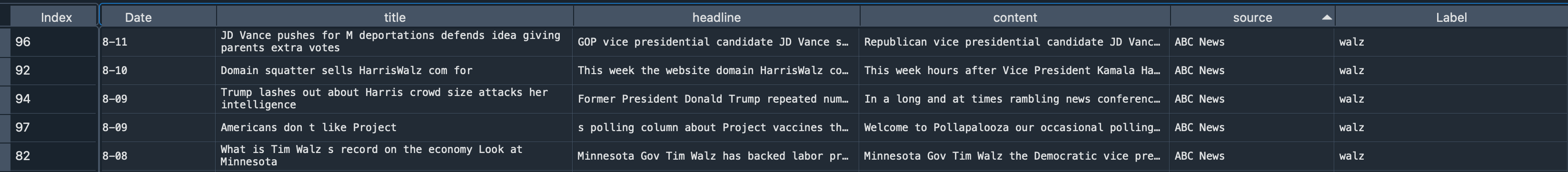
Description automatically generated

Data After Cleaning:

A screenshot of a computer screen

Description automatically generated

* 1. **Labeling**: After cleaning, each article is labeled according to the public figure it primarily discusses. This labeling allows for the categorization of the data, making it possible to identify patterns specific to each figure.
     + **Labeling Process**: A label is assigned to each article based on the query used to retrieve it. For example, articles fetched with the query "Trump" are labeled as "trump."



1. **Data Visualization**: To better understand the distribution of words associated with each public figure, word clouds are created. Word clouds are visual representations where the size of each word indicates its frequency in the dataset. This visualization helps to quickly identify the most prominent words associated with each label.

|  |  |
| --- | --- |
| A close up of words  Description automatically generated | A close up of words  Description automatically generated |
| A close-up of words  Description automatically generated | A close up of words  Description automatically generated |

* 1. J.D. Vance:
     + Prominent Words: "president," "presidential," "campaign," "vice," "republican"
     + Analysis: The word cloud for Vance is heavily centered around terms related to a presidential campaign, indicating that much of the media coverage focuses on his involvement in the presidential race as a vice-presidential candidate. Words like "republican" also suggest the political context, reflecting his alignment with the Republican Party.
  2. Tim Walz:
     + Prominent Words: "minnesota," "president," "running," "governor," "vice"
     + Analysis: The word cloud for Walz is clearly dominated by references to his role as the governor of Minnesota. The prominence of words like "minnesota," "governor," and "presidential" suggests that the coverage often revolves around his political actions and ambitions within his state and on a national level. The repeated mention of "vice" indicates that he is frequently discussed in the context of a vice-presidential role.
  3. Donald Trump:
     + Prominent Words: "president," "assassination," "attempt," "biden," "crooks," "silicon," "musk"
     + Analysis**:** The Trump word cloud highlights associations with dramatic events and figures, such as "assassination" and "attempt," which suggests a focus on significant, possibly controversial, incidents in the news. The frequent appearance of "biden" indicates ongoing coverage of the political rivalry between Trump and Biden. The inclusion of "musk" and "silicon" hints at discussions involving influential tech figures or industry, possibly reflecting Trump's impact on or opinions about tech leaders and companies.
  4. Kamala Harris:
     + Prominent Words: "president," "assassination," "crooks," "musk," "biden," "social"
     + Analysis: Harris’s word cloud also features dramatic and charged terms such as "assassination," which might reflect media focus on high-stakes events. The recurring names "musk" and "biden" suggest that her media coverage is often linked with other prominent figures, possibly indicating collaborations, political alliances, and comparisons. The word "social" may point to discussions around social issues or her engagement on social media platforms.

**Model/Methods**

The models and techniques applied include K-means clustering, Cosine Similarity and Hierarchical Clustering, Decision Trees, and Multinomial Naive Bayes (MNB). Each method provides unique insights into the structure and classification of the articles, helping to differentiate content related to Donald Trump, Kamala Harris, Tim Walz, and J.D. Vance.

1. **K-Means Clustering:** K-means clustering is an unsupervised learning algorithm that partitions data into a set number of clusters based on feature similarity. This model was applied to the cleaned and tokenized data to explore natural groupings of articles without any prior labels.
   1. **Process:**
      1. The algorithm was run with different values for the number of clusters (k) to identify the optimal number of clusters that best group the articles.
      2. The **elbow method** was used to determine the optimal k by plotting the sum of squared distances (inertia) for different values of k and identifying the point where the rate of decrease sharply slows down, indicating the "elbow" point.
   2. **Elbow Plot**: The inertia sharply decreases from k=1 to k=3, indicating that the data points are becoming more tightly grouped as the number of clusters increases. After k=4, the rate of decrease in inertia slows down, which suggests that adding more clusters beyond this point provides diminishing returns.

Therefore, k=4 might be considered the optimal number of clusters, as it balances the trade-off between having a low inertia and avoiding overfitting with too many clusters.

**A graph with a line

Description automatically generated**

* 1. **Findings:** When initially clustering the articles into 4 clusters using the K-means algorithm, the results did not show a clear distinction between the different subjects of the articles. The stacked bar plot reveals that articles related to Kamala Harris, Donald Trump, JD Vance, and Tim Walz are intermixed within each of the clusters.

A graph with different colored bars

Description automatically generated

Cluster 2, in particular, is densely populated with articles from all four subjects, indicating that the K-means algorithm struggled to separate these topics effectively. The remaining clusters (0, 1, and 3) also show a mixture of articles from different labels, further highlighting the difficulty in distinguishing the articles based solely on their content using this method.

This suggests that there is significant overlap in the vocabulary used across articles about these different figures, which the K-means algorithm could not differentiate. The assumption of spherical clusters inherent to K-means may not be well-suited to the structure of this text data, leading to overlapping clusters that do not align well with the intended subject distinctions.

Given these results, it may be worthwhile to explore alternative clustering techniques, such as hierarchical clustering using cosine similarity, which could potentially provide a better separation of the topics based on their content. Alternatively, refining the feature selection process could also help to improve the clustering outcomes.

1. **Hierarchical Clustering with Cosine Similarity:** Hierarchical clustering offers an alternative approach to grouping data, one that does not require a predefined number of clusters. Unlike K-means, which assumes spherical clusters and requires the number of clusters as an input, hierarchical clustering builds a hierarchy of clusters that can be interpreted at various levels of granularity. This is particularly useful when the natural structure of the data is unknown or when exploring different levels of similarity is of interest.
   1. **Process:** For this analysis, cosine similarity was used to measure the distance between articles. Cosine similarity is particularly well-suited for text data because it focuses on the orientation (or angle) between vectors rather than their magnitude. This makes it more robust in handling variations in document length and word frequency, which are common in textual datasets.
      1. **Cosine Similarity and Ward's Method**

The distance matrix, computed as 1 −cosine similarity, provides a measure of dissimilarity between pairs of articles. The smaller the cosine distance, the more similar the articles are in terms of the words they contain. This distance matrix was then used as input to perform hierarchical clustering using Ward's method.

Ward's method is an agglomerative clustering technique, meaning it starts with each article as its own cluster and then iteratively merges the closest clusters. The goal is to minimize the variance within each cluster. Ward's method tends to create clusters of relatively equal size, which is ideal for exploratory data analysis.

* 1. **Results and Analysis**

The dendrogram generated from hierarchical clustering provides a visual representation of the clusters at different levels of similarity. By cutting the tree at four clusters, the articles are grouped into four distinct clusters.

A drawing of a city

Description automatically generated

The stacked bar plot visualizes the distribution of articles in each cluster by label (Kamala, Trump, Vance, Walz). Interestingly, Cluster 1 contains the majority of the articles across all labels, indicating that this cluster groups together articles that share common themes or vocabulary, regardless of the specific person they discuss. This could be indicative of general political topics or words frequently used across multiple figures.

Clusters 2, 3, and 4 are smaller and more specific, but they still show a mix of articles about different people. This suggests that, while hierarchical clustering provides a more nuanced grouping than K-means, there is still significant overlap in how articles about different figures are described in popular media.

**A graph with different colored squares

Description automatically generated**

This overlap might be due to the nature of political reporting, where similar themes, topics, and language are used across different figures, reflecting broader narratives or biases within the media. Alternatively, it may indicate that the words used in these articles are not sufficiently distinctive to separate the articles into cleanly distinct groups, which could be an artifact of the content or the feature selection process.

Overall, hierarchical clustering with cosine similarity and Ward's method provides a flexible and insightful approach to exploring the structure of text data. However, as with K-means, the results highlight the challenges of clustering articles based on word frequency alone, particularly when dealing with overlapping topics or figures that are frequently discussed together in the media.

1. **Multinomial Naive Bayes:** The Multinomial Naive Bayes (MNB) model is a popular choice for text classification tasks, particularly when the features are word counts or frequencies. This probabilistic classifier assumes that the features are independent of each other, and it uses Bayes' theorem to predict the probability that a given set of features belongs to a particular class. Despite its simplicity, MNB is often effective for text data, especially when the features are word counts, as it can handle the high-dimensional nature of textual data well.
   1. **Model Training and Prediction**

In this analysis, the dataset was first split into a training set and a test set, with 70% of the data used for training the model and the remaining 30% for testing. After splitting, the labels were separated from the data, leaving only the feature matrix (word counts) to be used for training.

The MNB model was instantiated and fitted to the training data. The trained model was then used to predict the labels for the articles in the test set, and the model's confidence in each prediction was calculated and displayed as prediction probabilities.

* 1. **Confusion Matrix and Analysis**

The performance of the MNB model was evaluated using a confusion matrix, which shows the number of correct and incorrect predictions for each class. The rows represent the actual labels, while the columns represent the predicted labels.

**A blue squares with white text

Description automatically generated**

The confusion matrix, shown above, reveals the following insights:

* **Vance**: The model correctly predicted 11 out of 29 articles labeled as "Vance." However, it incorrectly classified 14 of these articles as "Trump," indicating some overlap or similarity in the word usage between articles about Vance and Trump.
* **Trump**: Only 5 out of 32 articles labeled as "Trump" were correctly classified, with the majority (22) being incorrectly predicted as "Vance." This further suggests that articles about Trump and Vance share common vocabulary or themes, which the model struggled to distinguish.
* **Walz**: For articles about Walz, the model correctly classified 10 out of 29. The misclassifications were spread across the other labels, indicating that Walz-related articles have some distinguishing features but still share similarities with the others.
* **Kamala**: The model performed best on articles about Kamala, correctly classifying 18 out of 30. This suggests that articles about Kamala have more distinct vocabulary or themes compared to the other classes, making them easier for the model to identify.
  1. **Prediction Probabilities**

The prediction probabilities provide additional insight into the model's confidence in its predictions. For example, the model was very confident (with probabilities over 80%) when predicting certain articles as related to Kamala, but it was less confident when distinguishing between Vance and Trump, often assigning close probabilities to both classes.

A screen shot of a computer screen

Description automatically generated

* 1. **Summary:** While the model was able to accurately classify articles about Kamala with relatively high confidence, it struggled to distinguish between articles about Vance and Trump. This overlap in word usage likely reflects broader narratives or common themes in the media coverage of these figures. The confusion matrix highlights areas where the model could be improved, such as by refining the feature selection process or incorporating additional context into the analysis.

1. **Decision Tree:** This model is a simple and easily interpretable method used for classification tasks. The model uses the frequency of specific words in the headlines to make its classifications.

### **Decision Tree Visualization**

### The Decision Tree diagram generated from this model showcases how the tree splits the data based on different features (words). Each node in the tree represents a decision rule based on the presence or absence of a particular word, leading to a classification at the leaf nodes. The visualization is color-coded to show the different classifications made by the tree at various stages. Given the high-dimensionality of the dataset, this tree is not easily interpreted visually.

A line of colorful squares

Description automatically generated

### **Confusion Matrix**

### The matrix highlights some of the challenges the model faced, particularly in differentiating between articles about Trump and Vance, where the model often misclassified the articles.

A blue squares with white text

Description automatically generated

* **Correct Predictions:** The model performed relatively well in predicting articles about Kamala, with 13 correct predictions. This suggests that the words associated with Kamala are distinct enough to allow the model to differentiate these articles from those about other figures.
* **Misclassifications:** A significant number of misclassifications occurred, particularly with articles about Trump. For example, many articles about Trump were incorrectly classified as being about Vance. This indicates an overlap in the language used in articles about these two figures, making it difficult for the model to draw clear distinctions. Similarly, articles about Walz were often misclassified as being about Trump or Kamala, suggesting that the language used in these articles may not be unique enough to ensure accurate classification.
* **Overall Accuracy:** The confusion matrix shows that while the model has some ability to classify articles correctly, there are still substantial challenges, particularly in differentiating between figures who may be discussed in similar contexts or with similar terminology.

### **Feature Importance**

### **Top Features:** Words like "president," "minnesota," "running," and "presidential" were identified as the most important features. This suggests that articles frequently mentioning these terms were more likely to be classified into specific categories. For instance, the term "minnesota" likely contributed heavily to identifying articles about Walz, while "president" and "presidential" were influential across multiple categories, reflecting the high frequency and significance of these terms in political discourse.

### **Influence of Common Words:** Some words, despite being common across multiple categories, were still among the top features. This highlights the model's reliance on certain key terms that are prevalent in the headlines. However, this may also explain some of the misclassifications seen in the confusion matrix.

A screen shot of a computer

Description automatically generated

* 1. **Summary:** The Decision Tree model's performance suggests that while certain key terms can be strong indicators of an article's subject, the overlap in language across political figures poses a challenge for accurate classification. The model struggles with articles that discuss similar topics or use similar language for different figures, leading to misclassifications.

**Conclusions**

The analysis of news articles about Donald Trump, Kamala Harris, Tim Walz, and J.D. Vance revealed several key findings that shed light on how these figures are portrayed in the media:

1. **Distinct Themes for Each Figure**:
   * **Kamala Harris** was often associated with words like "president," "assassination," and "crooks." This suggests that her coverage frequently touches on high-stakes political events and controversies. The word "social" also indicates her engagement with social issues or media.
   * **Donald Trump** shared some common terms with Harris, like "president" and "assassination," but also had unique associations with figures like "Musk" and topics like "silicon." This points to Trump's ongoing involvement in significant technological and business-related discussions.
   * **Tim Walz** was closely linked to words like "Minnesota," "governor," and "running," highlighting his role in state politics and his involvement in broader political aspirations. The emphasis on his state-specific leadership contrasts with the more national focus seen in coverage of Harris and Trump.
   * **J.D. Vance** was heavily associated with terms like "presidential," "campaign," and "republican," indicating that his media portrayal centers on his political ambitions, particularly within the Republican Party.
2. **Overlapping Narratives**:
   * Despite these distinctions, there was a significant overlap in the language used across articles about these figures. For instance, words like "president" and "vice" appeared frequently in articles about multiple individuals, reflecting common themes in political coverage.
   * The clustering analysis showed that articles about these figures often shared enough similarities to be grouped together, suggesting that the media might be employing similar narratives or framing across different individuals.
3. **Media Bias and Groupthink**:
   * The analysis also hints at potential biases or a collective viewpoint within the media. For example, the consistent association of Trump with terms like "crooks" and "assassination" could reflect a particular framing that might influence public perception. Similarly, the portrayal of Kamala Harris alongside terms like "Musk" and "social" suggests a focus on her connections with influential figures and social issues, which could shape how she is perceived in the public eye.
4. **Challenges in Differentiation**:
   * The difficulties encountered in cleanly separating articles by subject matter using clustering techniques underline the challenge of distinguishing between figures who are frequently discussed together or who share common themes in the media. This points to the complexity of media coverage, where the lines between different public figures are often blurred by shared narratives and vocabulary.

These findings underscore the importance of critically analyzing media content, as the words and themes chosen by journalists play a crucial role in shaping public opinion and potentially perpetuating biases. Understanding the patterns in language use can help readers navigate the complex landscape of news coverage with a more informed perspective.