Mid-term Project Report

By Uday, Surya, Dhrithi, and Akhil

1. Data Collection and Preprocessing

In this project, we have used web scraping to collect data on companies of interest from the Hindenburg Research website. The data was collected using Python, and the BeautifulSoup library was employed for parsing the HTML content. Once the data was collected, the text content was preprocessed by converting it to lowercase, removing special characters and punctuation, and filtering out stopwords using the NLTK library.

2. Exploratory Data Analysis

The Exploratory Data Analysis (EDA) focused on the following aspects:

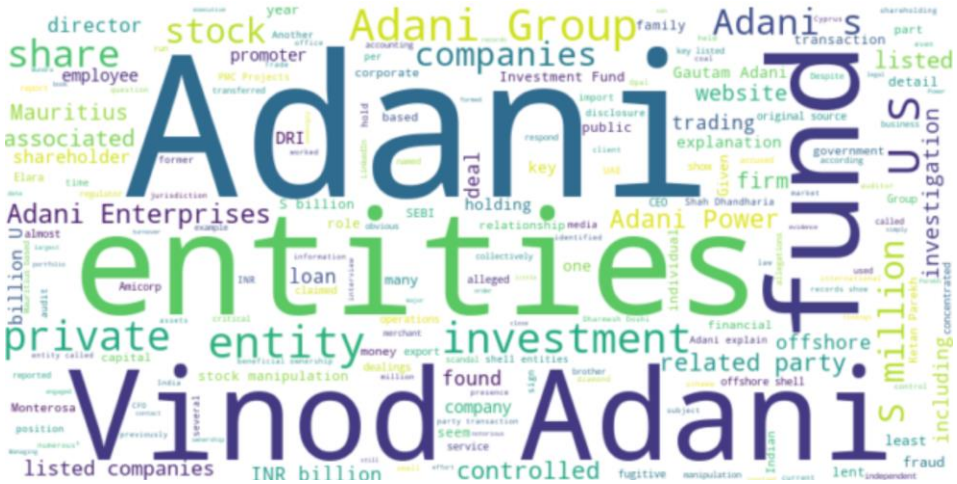
- Identifying the most common words within the text.
- Visualizing the distribution of words using word clouds.
- Performing sentiment analysis on the text.
- Most Common Words

From the output of the EDA, we have identified the following insights:

- The word "Adani" appears to be the most frequent across all categories, followed by "group", "entities", and "companies". This suggests that the Adani Group and its associated companies are a significant focus of the articles.
- Other frequently mentioned words include "investment", "power", and "us", indicating that the articles might also discuss investments in power projects, and their impact on or relation to the United States.
- The presence of words like "Vinod" suggests that specific individuals related to the Adani Group, such as Vinod Adani, may also be discussed in these articles.

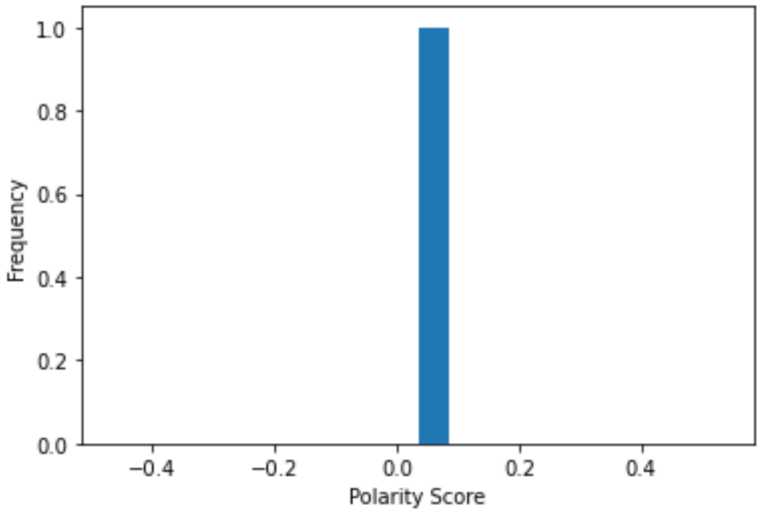
2-1. Word Clouds

The word clouds generated for different categories provide a visual representation of the distribution of words. The larger the word appears in the word cloud, the more frequently it occurs in the text. This reinforces the insights derived from the most common words analysis, with "Adani", "group", "entities", and "companies" appearing prominently.



3. Sentiment Analysis

Sentiment analysis was performed using the TextBlob library, and the resulting polarity scores were plotted in a histogram. Since the text has only one sentence with a very large number of words, the histogram does not provide any meaningful insights. To improve this, preprocessing should be adjusted to split the text into meaningful sentences before performing sentiment analysis.



4. Updated Problem Statement and Methodology

Based on the initial analysis, the problem statement can be refined to investigate the nature of investments in power projects by the Adani Group and its associated companies and the impact of these investments on the United States. The methodology should be updated to include a more detailed analysis of the most frequently mentioned companies and individuals, as well as sentiment analysis on properly preprocessed sentences.

5. Project Plan

Completed Steps:

- Data collection using web scraping - Completed by Uday
- Preprocessing and EDA - Completed by Surya

Next Steps:

- Refine preprocessing to split text into meaningful sentences - Assigned to Dhrithi
- Perform detailed analysis on the most frequently mentioned companies and individuals - Assigned to Akhil
- Perform sentiment analysis on properly preprocessed sentences - Assigned to Surya
- Draft the final report based on the refined analysis - Assigned to Uday
- Review and finalize the report - Assigned to the entire team