Mini Project: Sentiment Analysis and Emotional Comparison of U.S. Presidential Speeches

Name : Archit Dukhande

Course: IST652 - Scripting for Data Analysis

Instructor: Prof. Hernando Hoyos

Date: 04/15/2025

1. The Data and Its Source

For this project, I selected two distinct historical datasets of political speeches sourced from Project Gutenberg, a public domain digital library. The two documents chosen were:

"Speeches of Abraham Lincoln, 1832-1865" – This collection comprises speeches made by President Lincoln during his political career and presidency. It was accessed via: https://www.gutenberg.org/cache/epub/14721/pg14721-images.html

"Speeches of Benjamin Harrison, Twenty-third President of the United States" – This document includes public addresses and communications by President Harrison. It was accessed from: **https://www.gutenberg.org/cache/epub/44682/pg44682-images.html**

Both sources were scraped in HTML format using requests and parsed with BeautifulSoup, allowing for flexible retrieval of paragraph-level speech content.

2. Data Exploration and Cleaning Steps

After fetching the HTML pages for both presidents, I explored the structure to identify how the speech content was organized. I discovered that <p> tags consistently enclosed the main textual content. To ensure only meaningful and full-length sentences were analyzed, I filtered out paragraphs with fewer than 40 characters. This approach effectively removed non-speech elements such as chapter headers, page numbers, and irrelevant lines.

I used get_text(strip=True) from BeautifulSoup to clean and extract text from each paragraph and combined them into single continuous text blocks for each president. These blocks were then passed through Flair's SegtokSentenceSplitter, which is a more advanced sentence tokenizer that accounts for abbreviations, punctuation, and context.

For sentiment analysis, I used Flair's "en-sentiment" model to assess each sentence for polarity (either POSITIVE or NEGATIVE) and intensity (confidence score). These results were stored in a Pandas DataFrame for further analysis and visualization**.**

3. Comparison Questions and Analysis:

These questions were addressed by joining the two processed datasets into a unified format and applying grouped comparisons.

Q1: Which president used more positive vs. negative language in their speeches?

- **Unit of Analysis:** Sentence
- **Method:** Counted sentiment labels grouped by president
- **Finding:** Benjamin Harrison exhibited a significantly greater number of positive sentences compared to Abraham Lincoln, although he also had a higher total of negative sentences due to a larger corpus.
- **Visualization:** A bar chart comparing sentence counts by sentiment and speaker showed Harrison’s language leaning more toward positivity, possibly reflecting formal or ceremonial tones.

Q2: Which president used more emotionally intense language overall?

- **Unit of Analysis:** Sentiment confidence scores (i.e., emotional intensity).
- **Method:** Calculated average sentiment confidence for each president across all sentences.
- **Finding:** Harrison’s speeches displayed a higher average sentiment intensity, particularly for positive sentiments. Lincoln’s scores were more balanced between sentiment types.
- **Visualization:** A bar chart showing average confidence scores emphasized that while both presidents used emotionally engaging

language, Harrison's rhetoric leaned more heavily into positive intensity.

Q3: Is there a difference in emotional intensity between positive and negative sentences?

- **Unit of Analysis:** Sentiment confidence grouped by both president and sentiment
- **Method:** Used a grouped boxplot to visualize the distribution of confidence scores
- **Finding:** Positive sentences generally had higher confidence scores across both presidents. However, Lincoln showed a narrower distribution of intensity, suggesting more emotional balance, while Harrison had a wider range, especially among positive sentiments.
- **Visualization:** The boxplot helped highlight the emotional variation across presidents and sentiment classes, showcasing differences in how sentiment was expressed.

Q4: Word Cloud Analysis of Emotionally Intense Sentences

To supplement the quantitative insights, I also generated word clouds based on the top 100 emotionally intense sentences per president. These visualizations reflect the frequency of emotionally charged words.

- **Lincoln:** His word cloud emphasized terms like “liberty,” “union,” and “people,” aligning with themes of democracy and unity prevalent in Civil War-era speeches.
- **Harrison:** Prominent terms included “government,” “constitution,” and “service,” indicating a focus on civic responsibility and governance.

These word clouds offer a quick visual intuition into each president’s thematic and rhetorical emphasis, making the analysis more interpretable beyond numerical summaries.

4. Description of the Program

This project was built using Python in a VSCode and Jupyter Notebook environment. The program was structured into modular steps:

1. Scraping HTML pages from Project Gutenberg using requests and BeautifulSoup
2. Filtering and cleaning paragraph-level text
3. Sentence segmentation using Flair's context-aware SegtokSentenceSplitter
4. Sentiment analysis with Flair's TextClassifier (model: "en-sentiment")
5. Data transformation with pandas
6. After individual sentiment results were computed for each president, both collections were combined into a single dataset using a structured join. This enabled cross-comparison of sentiment polarity and emotional intensity between Lincoln and Harrison.
7. Data visualization using matplotlib, seaborn, and wordcloud
8. Exporting results to structured .csv and .txt files

Each step includes detailed comments, allowing reproducibility and understanding of the logic used.

5. Output Files

1. **sentiment_analysis_lincoln_harrison.csv**
    - Full sentence-level dataset including original text, sentiment label, confidence score, and president.
2. **sentence_level_sentiments.csv**
    - A simplified version focusing on just the text, sentiment, confidence, and speaker source, ideal for additional summarization or visualization.
3. **lincoln_harrison_raw_text.txt**

- Contains the complete raw scraped speeches of both presidents, separated by labeled sections.

Each file supports a unique purpose: full analysis, visualization-ready summaries, and raw data transparency.

Conclusion:

This project helped me understand how to use text scraping and sentiment analysis tools to explore real-world data. By comparing speeches from two U.S. presidents, I was able to see how emotional tone and sentiment can vary between different leaders. Using Flair made it easy to analyze sentence-level sentiment and confidence. Overall, I learned how to clean, process, and visualize text data effectively, and how to turn raw text into meaningful insights.