Data Appendix

Group 4: Dylan Beyerlein, Kelly Kohout, Logan Lee, Shreeja Tangutur (leader)
DS 4002: Project 1

Section 1: News Articles (Uncleaned Dataset)

1.1 Unit of Observation

Each row represents one News article published by *The Cavalier Daily* after 2024-2025.

1.2 Dataset Description

This dataset contains the original News article metadata and titles. It includes 735 rows and 5 columns: doc_id, authors, published_date, section, and title.

1.3 Data Variable Explanations

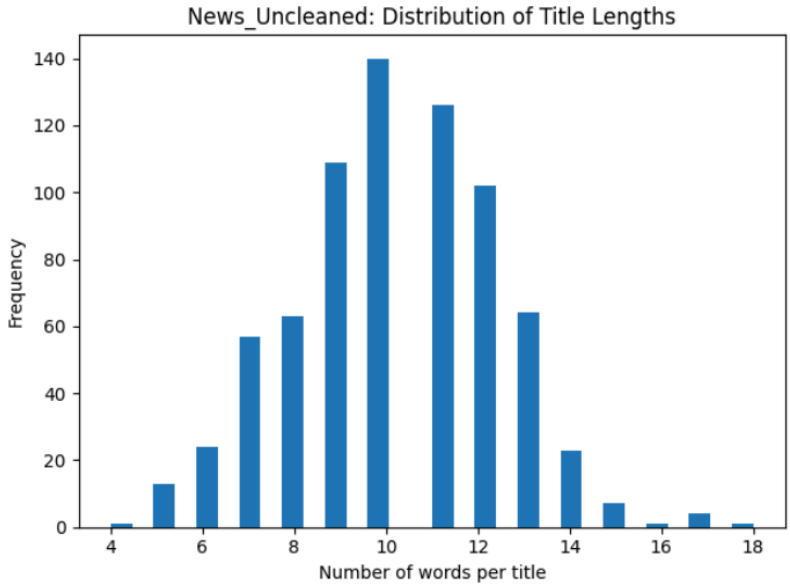
Column Name	Description	Data Type	Example Value
doc_id	Unique identifier for each article	String	“0037577d-a27b-4bd4-aeb2-eea7203ee82a”
authors	Author(s) of the article	String	“Cecilia Mould”
published_date	Standardized publication date (Month DD, Year)	Date/String	“7-May-24”
section	Newspaper section label	String	“News”
title	News article title	String	“Board of Visitors Finance Committee hears presentation on budget process”

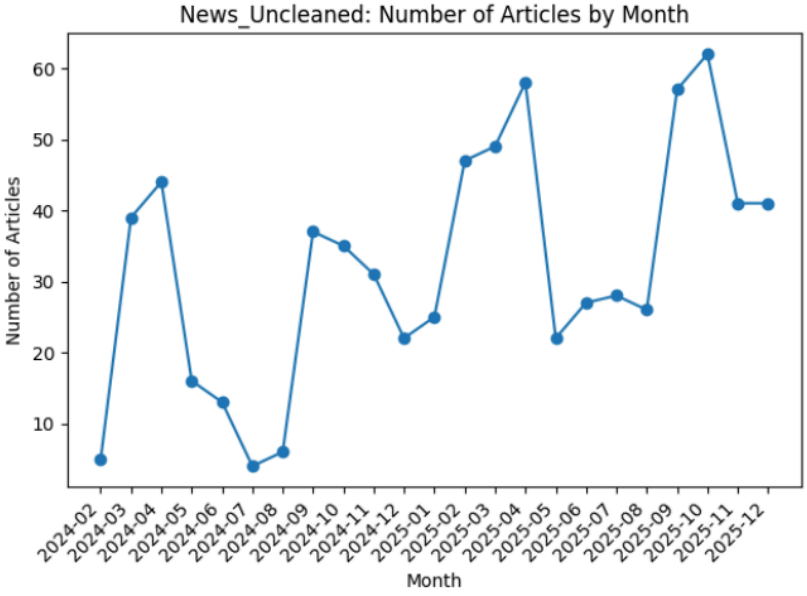
1.4 Statistics

- Total number of articles: 735
- Total number of unique authors: 71
- Average title length (words): 10.16
- Median title length (words): 10.0

- Shortest title length (words): 4
- Longest title length (words): 18
- Data range covered: February 2024 - December 2025

1.5 Figures

- Distribution of Title Lengths
  - 

- Number of News Articles by Month
  - 

Section 2: News Articles (Cleaned Dataset)

2.1 Unit of Observation

Each row represents one News article published by *The Cavalier Daily* from 2024-2025.

2.2 Dataset Description

This dataset contains the News articles post cleaning. It includes 735 rows and 3 columns: title, section, and clean_title.

2.3 Data Variable Explanations

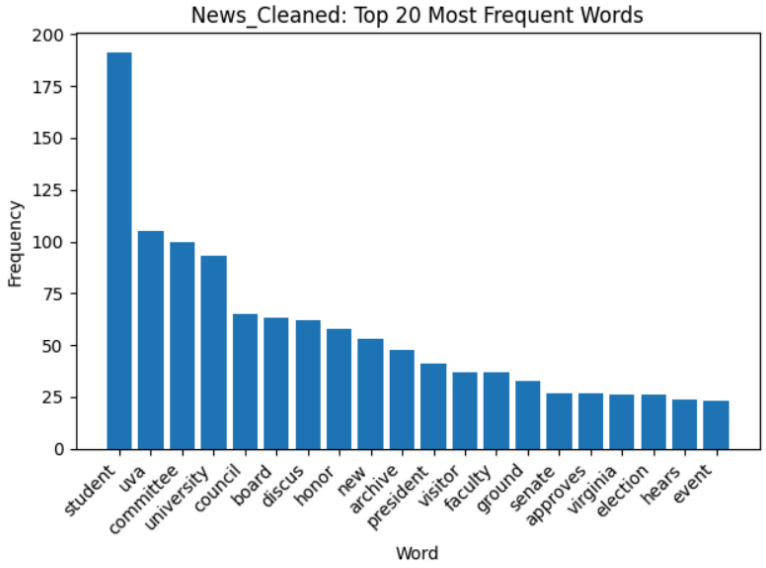
Column Name	Description	Data Type	Example Value
title	The original headline of the article.	String	“Board of Visitors Finance Committee hears presentation on budget process”
section	The newspaper genre label.	String	“News”
clean_title	The title of the article after lowercasing and removing punctuation.	String	“board of visitors finance committee hears presentation on budget process”
tokens	List of tokens derived from the cleaned title.	List[String]	[‘board’, ‘visitor’, ‘finance’, ‘committee’, ‘hears’, ‘presentation’, ‘budget’, ‘process’]

2.4 Statistics

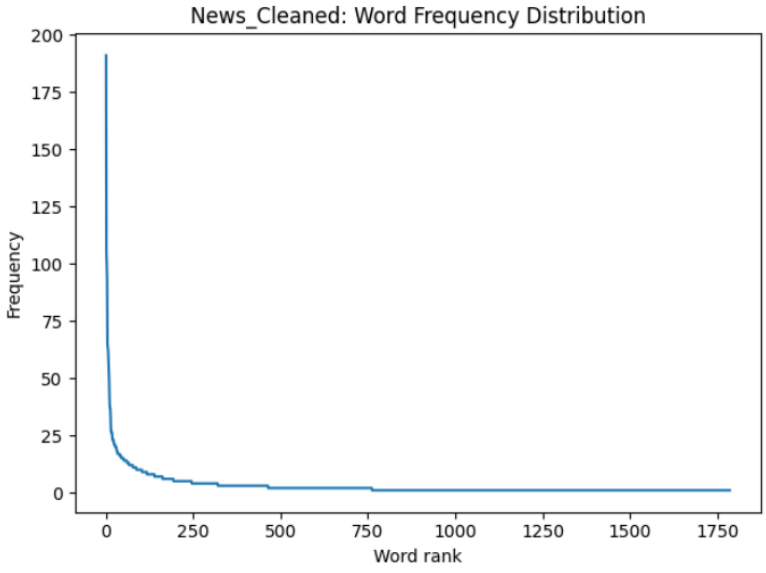
- Total number of News titles: 735
- Total number of tokens: 5557
- Total number of unique words: 1785
- Average tokens per title: 7.56
- Median tokens per title: 8.0
- Minimum tokens per title: 2
- Maximum tokens per title: 14
- Percentage of words appearing only once: 57.37

2.5 Figures

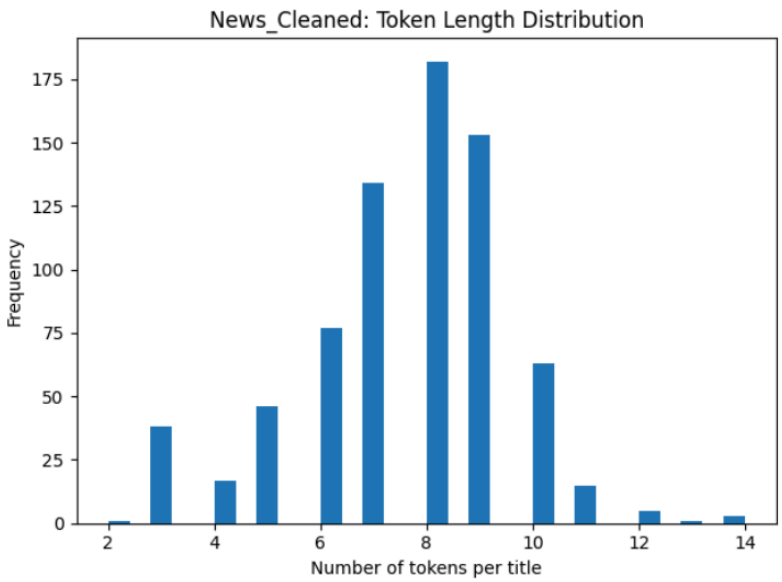
- Top 20 Most Frequent Content Words



- Word Frequency Distribution



- Token length distribution


News_Cleaned: Token Length Distribution
Frequency
Number of tokens per title
0
25
50
75
100
125
150
175
2
4
6
8
10
12
14

Section 3: Opinion Articles (Uncleaned Dataset)

3.1 Unit of Observation

Each row represents one Opinion article published by *The Cavalier Daily* after 2024-2025.

3.2 Dataset Description

This dataset contains cleaned Opinion article metadata and titles. It includes 268 rows and 5 columns: doc_id, authors, published_date, section, and title.

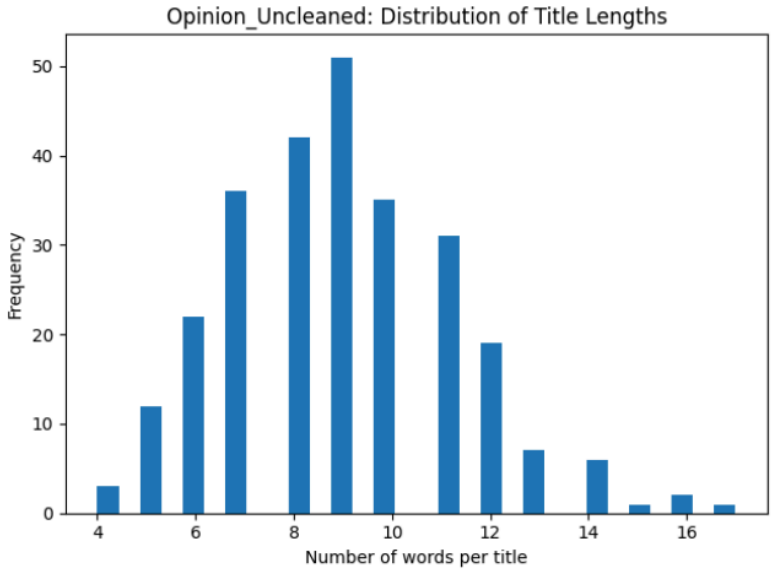
3.3 Data Variable Explanations

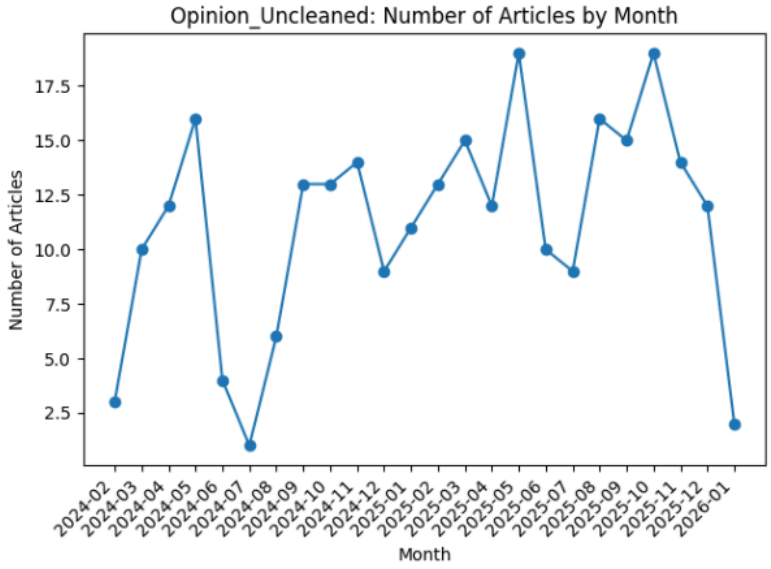
Column Name	Description	Data Type	Example Value
doc_id	Unique identifier for each article	String	“00337a3a-5eed-4ff4-9a8e-663c59de852d”
authors	Author(s) of the article	String	“Nathan Onibudo”
published_date	Standardized publication date (Month DD, Year)	Date/String	“17-Sep-24”
section	Newspaper section label	String	“Opinion”
title	News article title	String	“ONIBUDO: Give us back our Hill”

3.4 Statistics

- Total number of articles: 268
- Total number of unique authors: 119
- Average title length (words): 8.99
- Median title length (words): 9.0
- Shortest title length (words): 4
- Longest title length (words): 17
- Data ranged covered: February 2024 - January 2026

3.5 Figures

- Distribution of Title Lengths
  - 

- Number of Opinion Articles by Month
  - 

Section 4: Opinion Articles (Cleaned Dataset)

4.1 Unit of Observation

Each row represents one Opinion article published by *The Cavalier Daily* from 2024-2025.

4.2 Dataset Description

This dataset contains the Opinion articles post cleaning. It includes 268 rows and 4 columns: title, section, clean_title, and tokens.

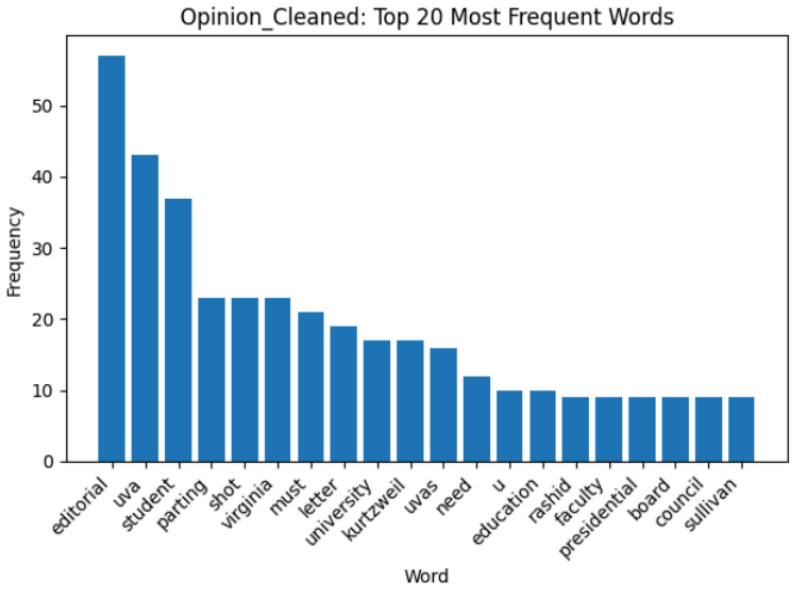
4.3 Data Variable Explanations

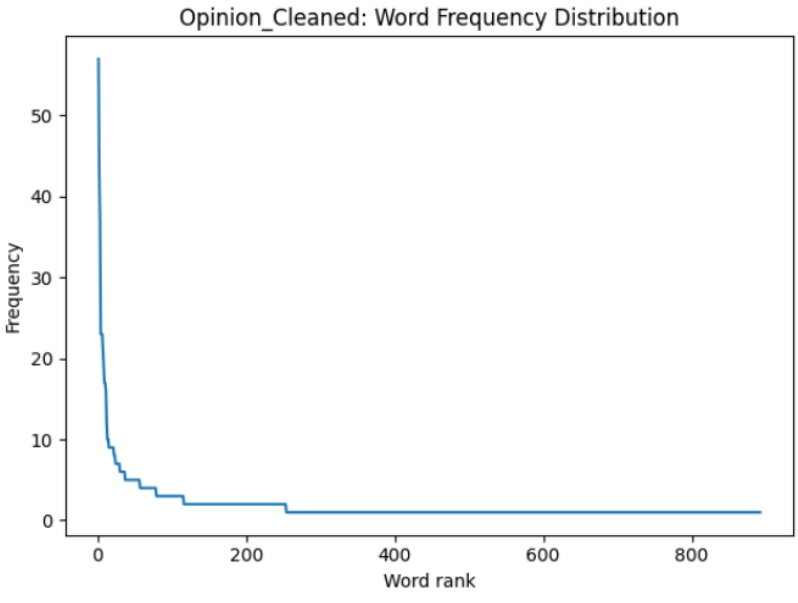
Column Name	Description	Data Type	Example Value
title	The original headline of the article.	String	“ONIBUDO: Give us back our Hill”
section	The newspaper genre label.	String	“Opinion”
clean_title	The title of the article after lowercasing and removing punctuation.	String	“onibudo give us back our hill”
tokens	List of tokens derived from the cleaned title.	List[String]	[‘onibudo’, ‘give’, ‘back’, ‘hill’]

4.4 Statistics

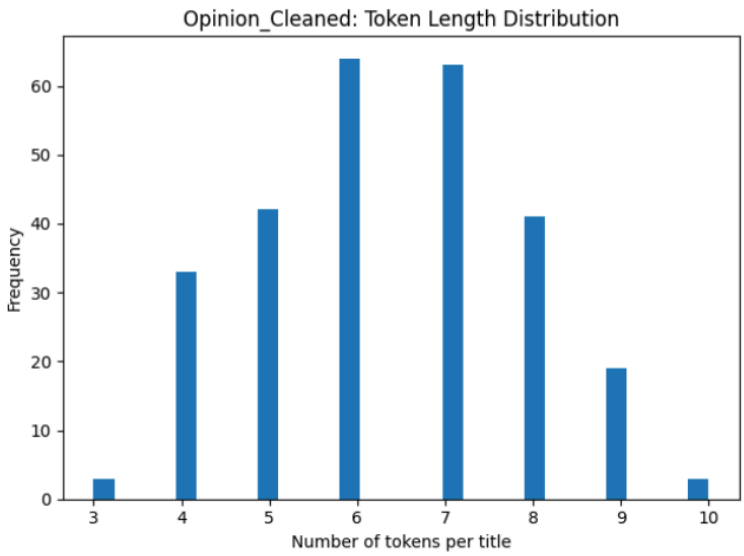
- Total number of Opinion titles: 268
- Total number of tokens: 1705
- Total number of unique words: 892
- Average tokens per title: 6.36
- Median tokens per title: 6.0
- Minimum tokens per title: 3
- Maximum tokens per title: 10
- Percentage of words appearing only once: 71.64

4.5 Figures

- Top 20 Most Frequent Content Words
    - 

- Word Frequency Distribution
    - 

- Token length distribution



- Distribution of Title Length by Section