Team Details

Team Name	DataWizards		
University	Nanyang Technological University		
Team Members Details	**Full Name**	**Email Address**	**Student ID** **If applicable*
Member 1 (Team Leader)	Zhao Qixian	qzhao010@e.ntu.edu.sg	U2321752L
Member 2	Krystal Pek	kpek001@e.ntu.edu.sg	U2322907F
Member 3	Rishika Mehta	rishika004@e.ntu.edu.sg	U2323133H
Member 4	Mahi Pandey	mahi003@e.ntu.edu.sg	U2321382F

Content Outline

	Page Number(s)
1. Solution Overview	2
2. Solution Features and Implementation Strategy	3
3. Solution Impact	4
4. Solution Architecture	5
5. Appendix	7
6. References	18

OneDrive folder with Pre-Processed Data Files
https://onedrive.live.com/?authkey=%21AEy8W0q0r318oNM&id=CB12D584492A1FAE%21768445&cid=CB12D584492A1FAE

1. Executive Summary

The rapid spread of misinformation has made it difficult to distinguish credible sources from unreliable ones. In response to this growing challenge, our team proposes a news and document similarity analysis tool that leverages Fuzzy Matching and Sentence-BERT models to uncover relationships between journalistic articles and Wikileaks documents. This system provides an interactive platform for users to analyze, compare, and evaluate news excerpts against a database of leaked government documents.

Key Features

- Uses two methodologies (Fuzzy Matching and Sentence-BERT) to compare news articles with Wikileaks documents.
- Extracts and displays key named entities and their relationships.
- Enables users to refine document selection by category.
- Displays top common entities and similarity scores for enhanced interpretability as well as a custom network graph based on the articles being examined
- Includes network graphs and entity-relation heatmaps for deeper contextual exploration.

Our tool can support ISD analysts in identifying misinformation, foreign influence campaigns, and security threats by comparing news reports with leaked government documents. By analyzing named entities, relationships, and similarity scores, ISD can detect:

- Coordinated disinformation efforts that manipulate narratives.
- Leaks of sensitive government information disguised as independent reporting.
- Emerging threats by tracking how certain entities and topics are covered in both mainstream and leaked sources.

The tool's interactive visualizations, including network graphs and entity heatmaps, enable ISD officers to quickly assess connections, uncover anomalies, and enhance intelligence analysis.

This proposal outlines our methodology, including data visualization strategies, solution architecture, implementation roadmap, and expected impact, while also identifying potential areas for improvement to enhance the model's accuracy and scalability.

2. Visualisation Overview

Techniques Used

- **Entity Co-occurrence Heatmaps** – Displays frequently occurring entities and their pairwise relationships to highlight major themes.
- **Network Graphs** – Uses Pyvis to create interactive visualizations that map entity relationships across datasets.
- **Sankey Diagrams** – Shows the flow from subjects to relationships to objects, allowing users to trace connections in extracted entities.

- **Word Clouds** – Highlights the most common terms in news excerpts and Wikileaks documents to reveal dominant topics.
- **Similarity Score Distribution Charts** – Displays the distribution of similarity scores across datasets, offering insights into the strength of document alignment.

Key Insights

- **Heatmaps and Network Graphs** help identify strong entity associations, enabling users to detect key figures, organizations, and trends that frequently appear together.
- **Sankey Diagrams** illustrate directional relationships between subjects and objects, making it easier to understand how different concepts are linked.
- **Word Clouds** provide a high-level overview of frequently discussed topics, facilitating quick thematic analysis.
- **Similarity Score Distribution Charts** highlight the degree of similarity between news articles and Wikileaks documents, indicating where narratives align or diverge.

Outliers and Data Correlations

- **Outliers** – Some news excerpts exhibit extremely high or low similarity scores, suggesting either strong alignment with leaked documents or substantial editorial differences. These outliers can indicate potential misinformation or selective reporting.
- **Correlations** – Frequent entity overlaps between news articles and Wikileaks documents suggest recurring themes, pointing to areas of high public and whistleblower interest.
- **Trends** – By tracking entity occurrences over time, the system reveals evolving narratives and potential shifts in media coverage based on new data leaks.

Conclusions

- **Validated Reports** – News articles with high similarity scores to leaked documents may indicate credible reporting based on prior disclosures.
- **Divergent Narratives** – Significant differences in similarity scores can highlight discrepancies, omissions, or potential biases in news coverage.
- **Emerging Topics** – Frequent entity and topic repetition across multiple news articles and leaks may indicate sustained interest or ongoing investigations.
- **Media Influence** – Certain entities or topics may appear more frequently in mainstream news coverage compared to Wikileaks leaks, hinting at editorial priorities and reporting trends.

3. Solution Features and Implementation Strategy

1. **News Selection and Comparison** – Users can input a news article either by selecting a news link or directly entering a news excerpt to compare against Wikileaks documents.

2. **Dual Similarity Computation Methods** – Two different approaches are used to determine the relationship between texts:
    a. Fuzzy Matching, which measures surface-level similarities based on text overlap.
    b. Sentence-BERT, a deep-learning-based model that analyzes semantic similarities.
3. **Entity and Relationship Mapping** – Named entity recognition is applied to extract key figures, locations, and organizations, and these are further examined in the context of their relationships.
4. **Ranking of Similar Documents** – Based on the similarity scores, the tool provides a ranked list of the most relevant Wikileaks documents related to the selected news article.
5. **Filtering by Category** – Users can refine their searches by selecting document categories to focus their analysis on specific topics.
6. **Rich Visualizations** – The platform integrates various interactive tools, including network graphs, Sankey diagrams, and entity heatmaps, allowing users to explore their findings dynamically.

Technical Implementation

- **Data Preprocessing** (see Appendix, Section 6)
    - The text data undergoes **cleaning, normalization, and entity extraction** to ensure consistent and meaningful comparisons.
    - Similarity scores are computed by **Fuzzy Matching** (for lexical analysis) and **Sentence-BERT** (for deep semantic comparisons).
- **Data Visualization & Frontend Development**
    - The **Streamlit-based UI** enables real-time exploration of similarity scores and entity relationships.
    - Various visual tools like **Seaborn, Matplotlib, Pyvis, and Plotly** are used to create **interactive visualizations** that enhance insight discovery.
- **Optimization & Deployment**
    - **Caching techniques** are used to store processed datasets and optimize the application's responsiveness.
    - The system is deployed as a **web-based interactive tool** ensuring easy access to users.

4. Solution Impact

Key Benefits

- **Enhanced Accuracy** – The combination of heuristic and deep-learning models ensures a nuanced comparison of documents, minimizing false positives and improving detection of nuanced relationships.
- **Actionable Insights** – Users can detect inconsistencies, explore hidden patterns, and validate the authenticity of reports with greater confidence.
- **Scalability** – Designed to handle large datasets efficiently, ensuring the system can scale with growing volumes of news and Wikileaks documents.

- **User-Friendly Interface** – The interactive **Streamlit-based** frontend allows users to filter results, explore entity relationships, and dynamically adjust similarity thresholds.
- **Performance Optimization** – Through caching and backend optimizations, response times remain quick, making real-time analysis practical for end users.

Unique Selling Points

- **Multi-Method Similarity Analysis** – By integrating both **Fuzzy Matching** and **Sentence-BERT**, the system offers a more accurate and comprehensive similarity comparison.
- **Real-Time Visualization** – Users can engage with data through **network graphs, Sankey diagrams, and heatmaps**, enabling dynamic exploration of relationships.
- **Bias Detection and Misinformation Analysis** – By identifying outliers and inconsistent narratives, the system helps highlight possible media biases and misinformation trends.

Key Performance Indicators (KPIs)

- **Accuracy of Similarity Scores** – Measured using **precision-recall metrics** to ensure the relevance of retrieved documents.
- **Processing Speed** – Optimized query response time, ensuring results are generated in real-time without delays.
- **User Engagement** – Analyzed through search frequency, user interaction rates, and time spent exploring visualization tools.
- **Scalability and Performance** – The system's ability to manage increasing dataset sizes without degradation in speed or accuracy.
- **Impact on Decision-Making** – Tracking feedback from journalists and researchers to assess whether the system enhances investigative workflows.

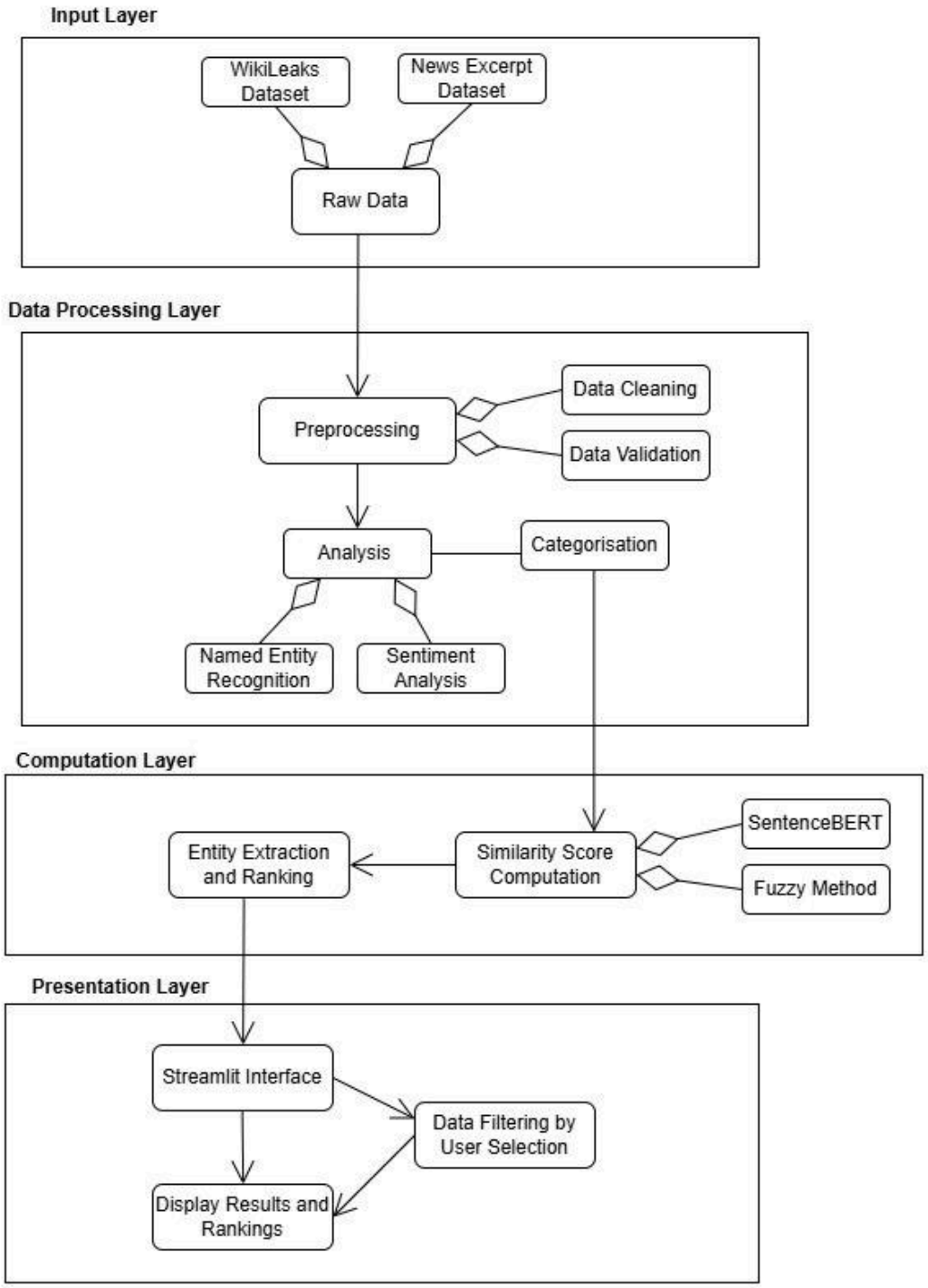
5. Solution Architecture

Architecture Overview

1. **Input Layer**
    - Loads raw data from the **WikiLeaks Dataset** and **News Excerpt Dataset**.
    - Serves as the entry point for document comparison.
2. **Data Processing Layer**
    - Performs **preprocessing**, including **data cleaning** and **data validation** to ensure data integrity.
    - Extracts relevant features through **named entity recognition (NER)** and **sentiment analysis**.
    - Categorizes documents based on extracted information.
3. **Computation Layer**
    - Computes similarity scores using **Fuzzy Matching** and **Sentence-BERT**.
    - Extracts and ranks entity relationships to support deeper analysis.

4. **Presentation Layer**
    - The **Streamlit-based UI** enables user interaction with the data.
    - Allows for **data filtering by user selection** to refine searches.
    - Displays final similarity rankings, entity relationships, and various visualization reports.

Architecture Diagram



6. Appendix

A. Data Screenshots

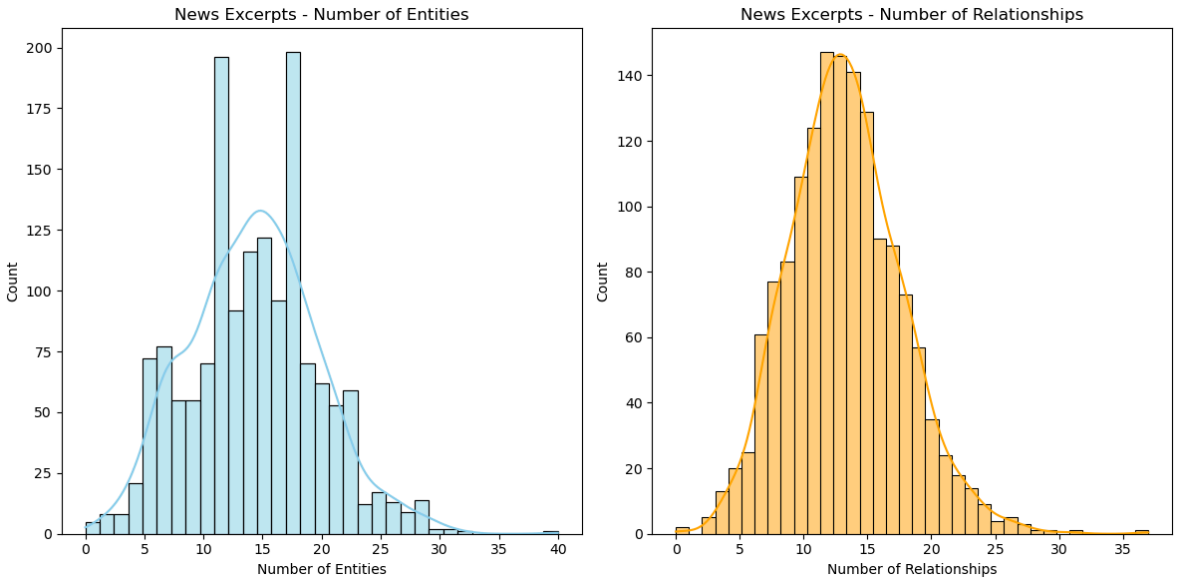


Figure 1.1 Number of Entities & Relationships for News Excerpts

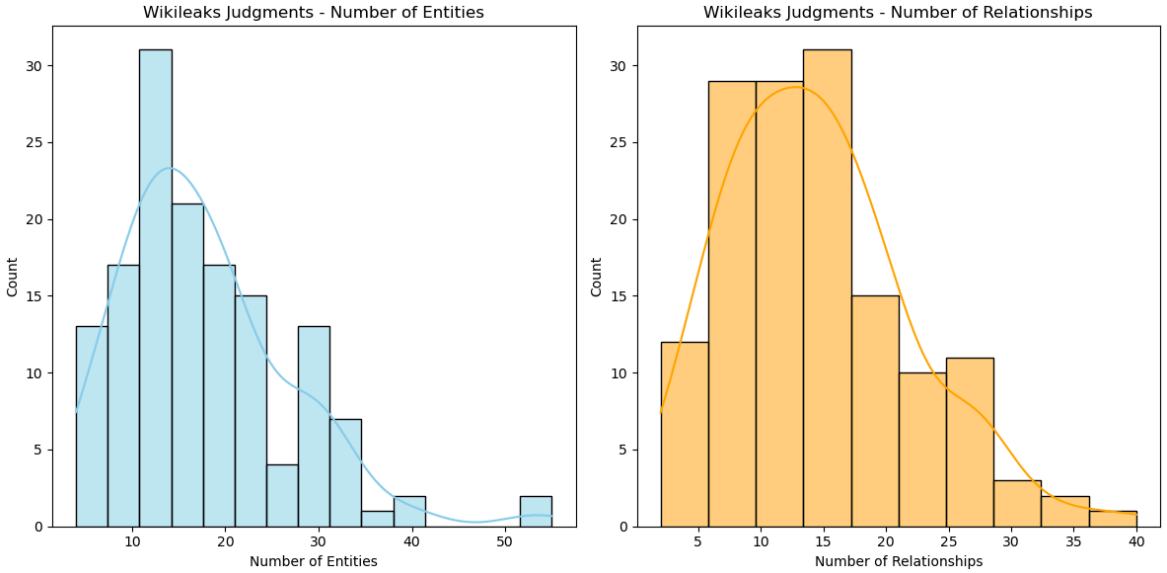


Figure 1.2 Number of Entities & Relationships for Wikileaks

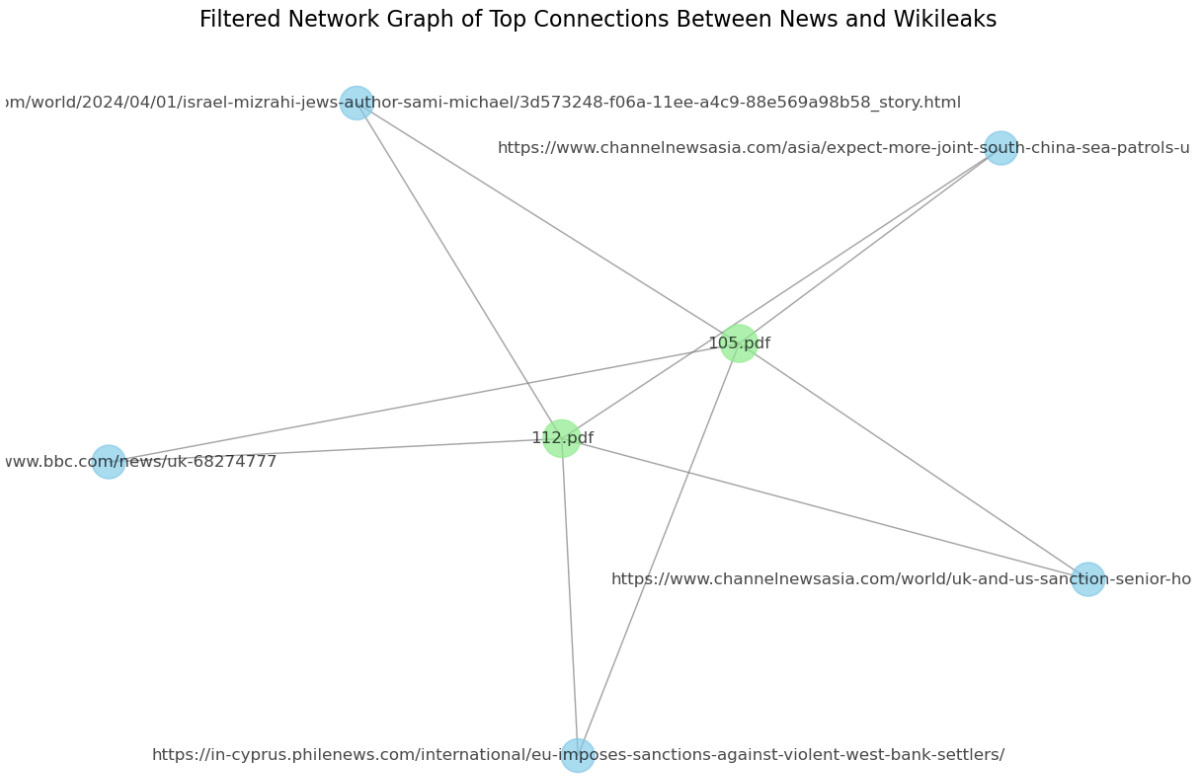


Figure 2 Filtered Network Graph of Top Connections Between News and Wikileaks

Univariate Analysis - News Excerpt

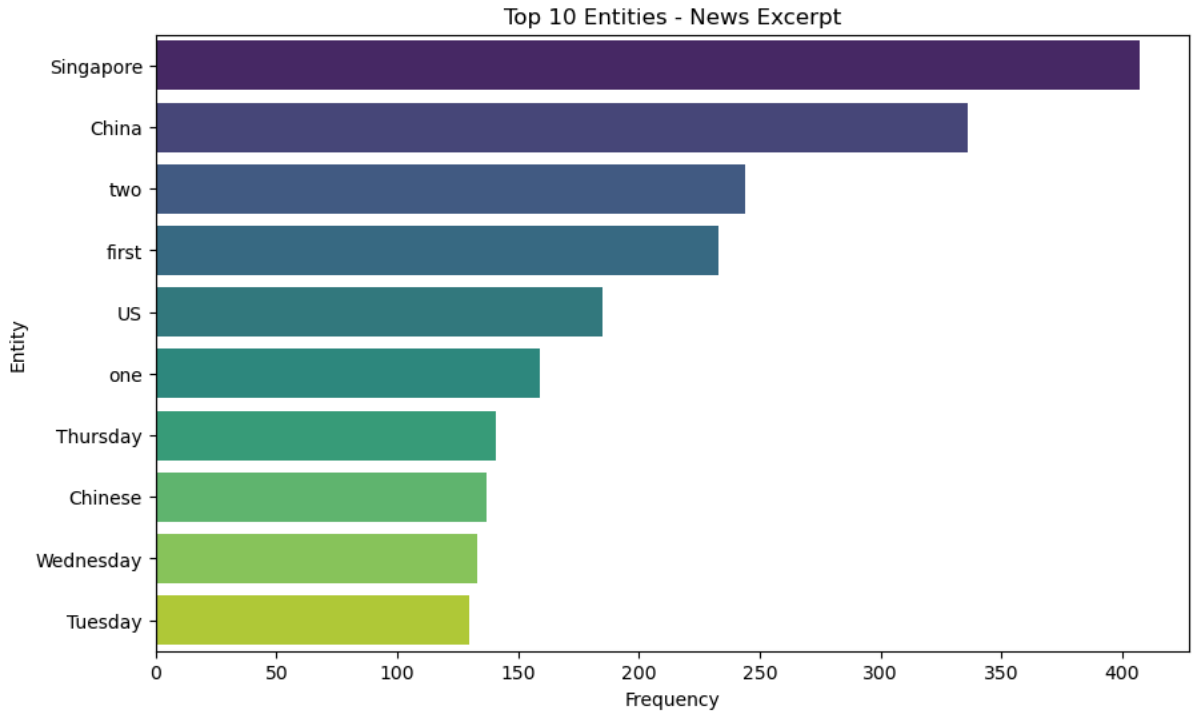


Figure 3.1 Top 10 Entities - News Excerpt

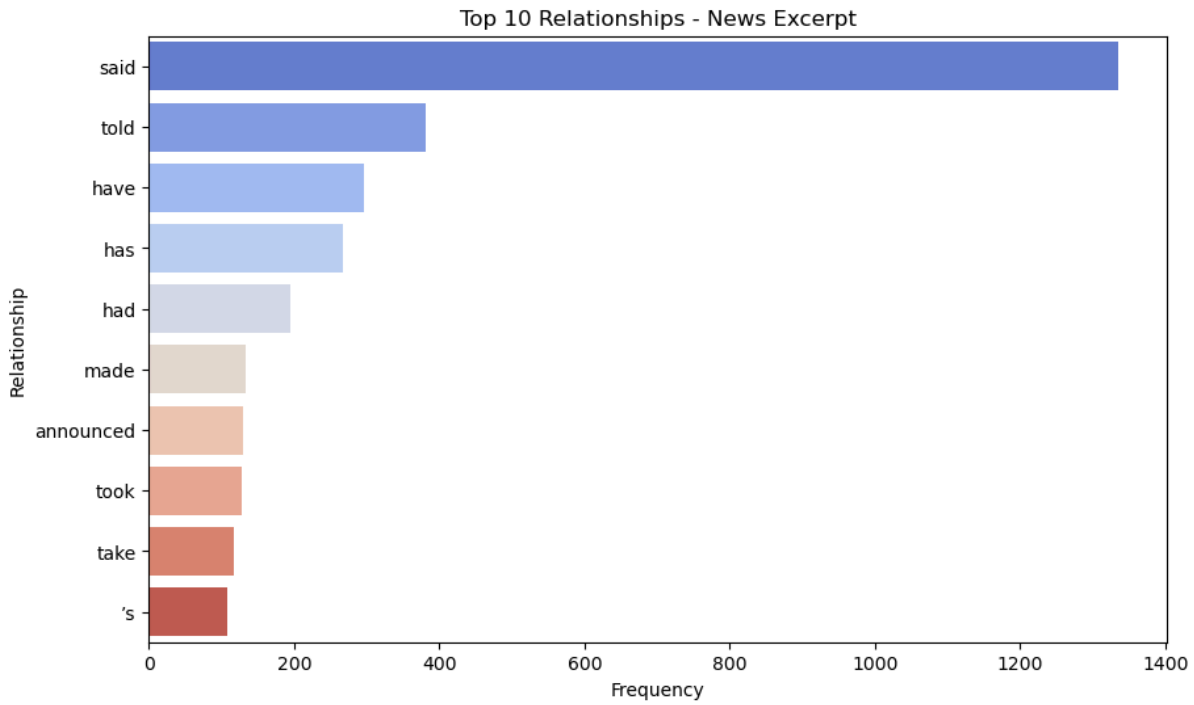


Fig 3.2 Top 10 Relationships - News Excerpt

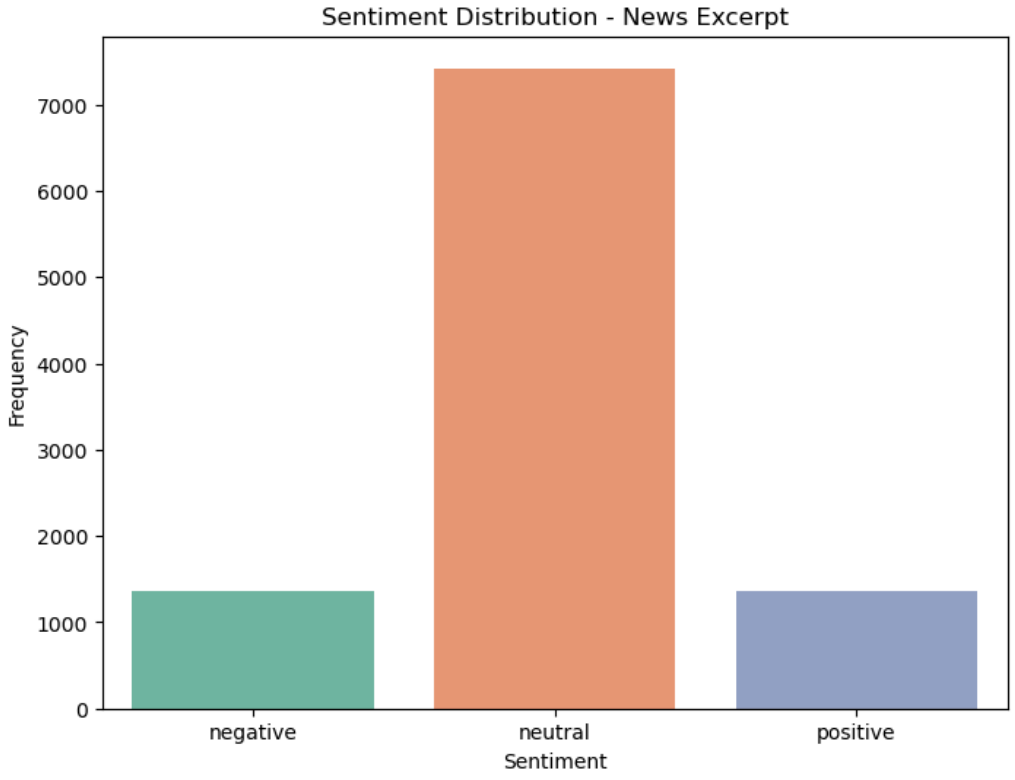


Figure 3.3 Sentiment Distribution - News Excerpt

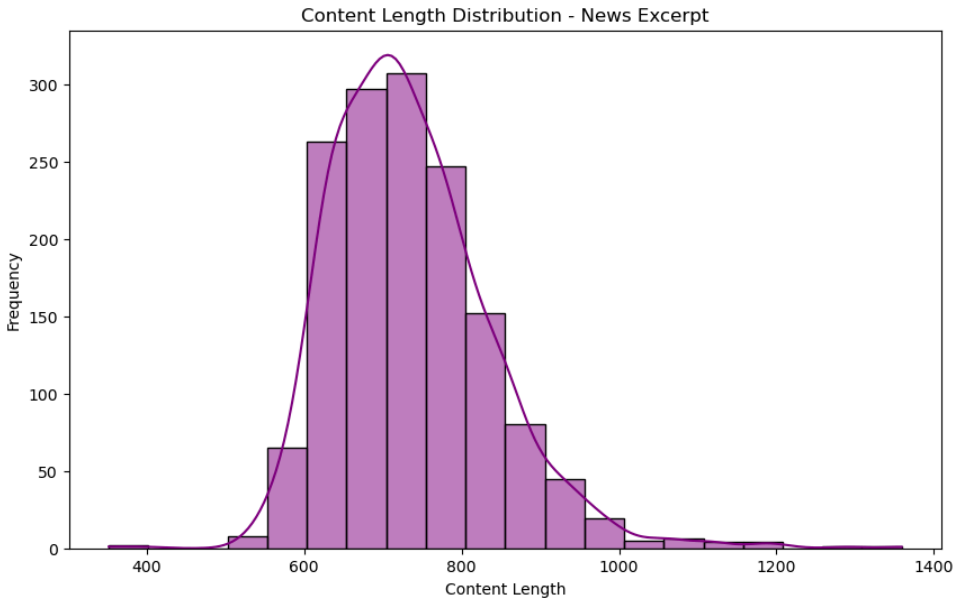


Figure 3.4 Content Length Distribution - News Excerpt

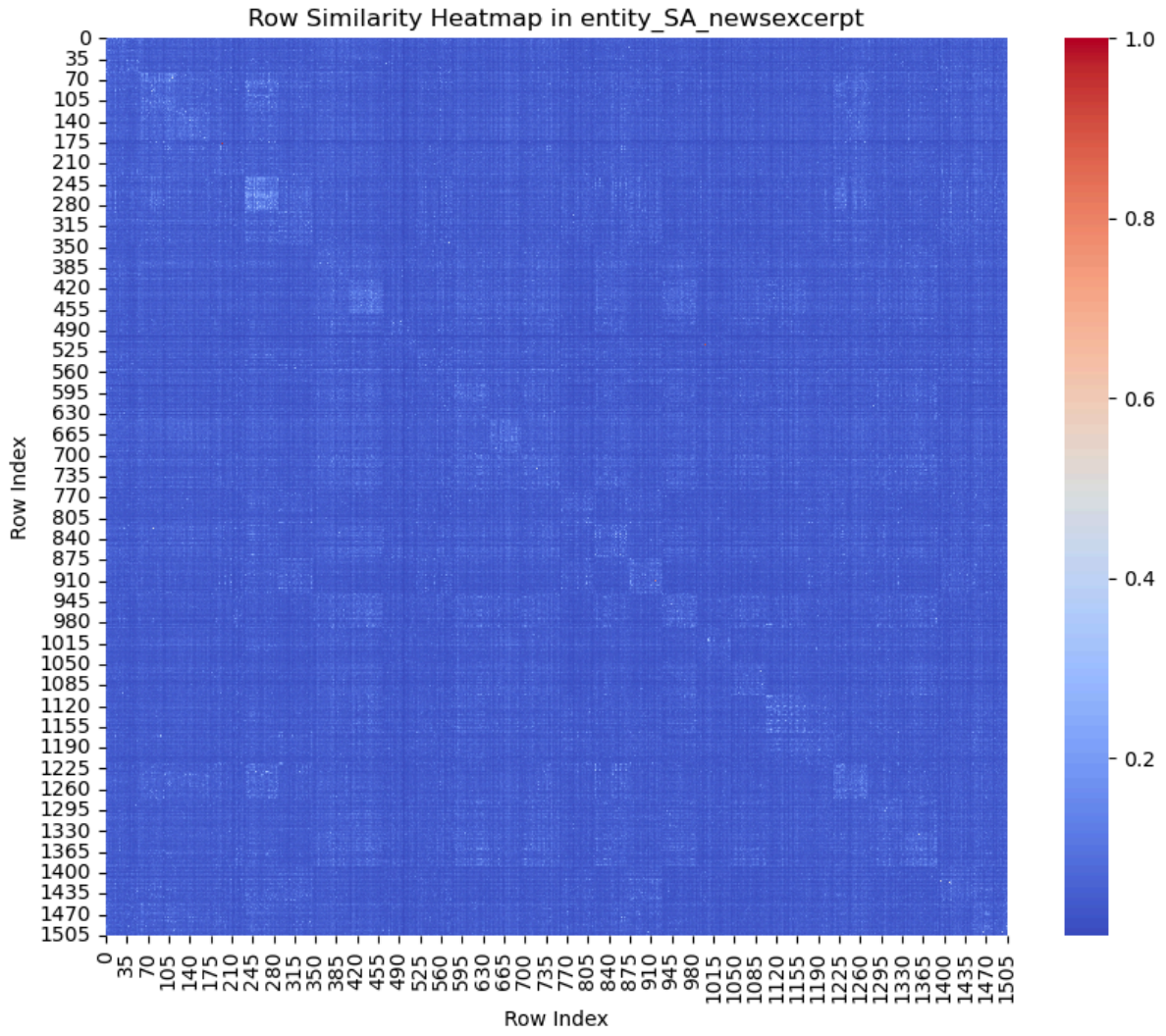


Figure 3.5 Row Similarity Heatmap - News Excerpt

Univariate Analysis - WikiLeaks

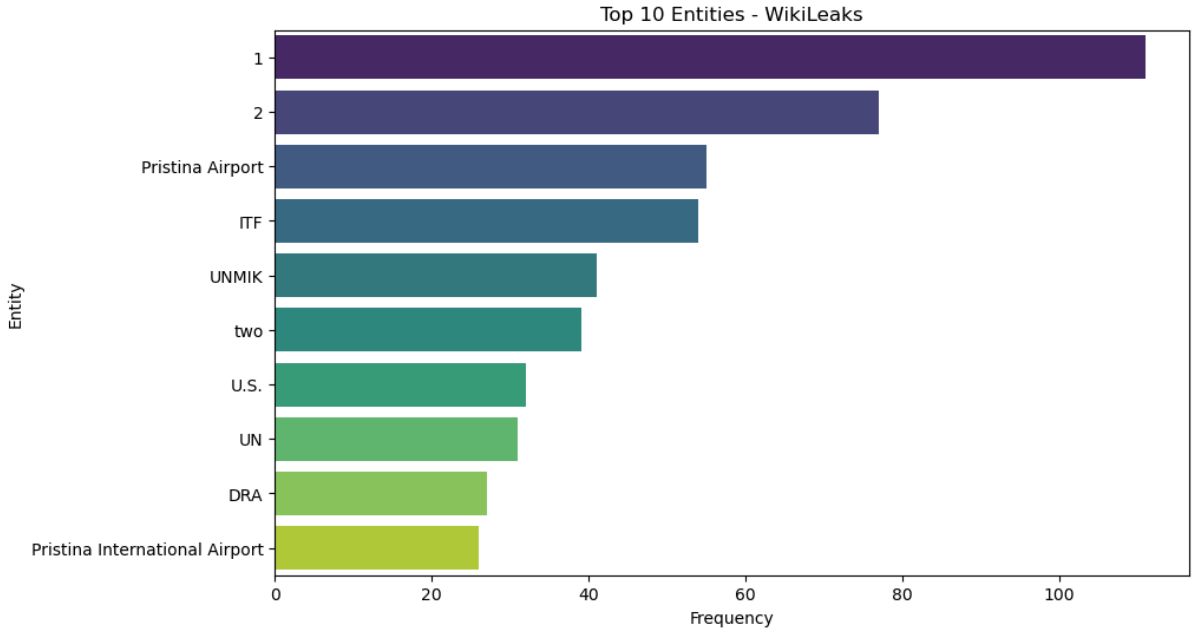


Fig 4.1 Top 10 Entities - WikiLeaks

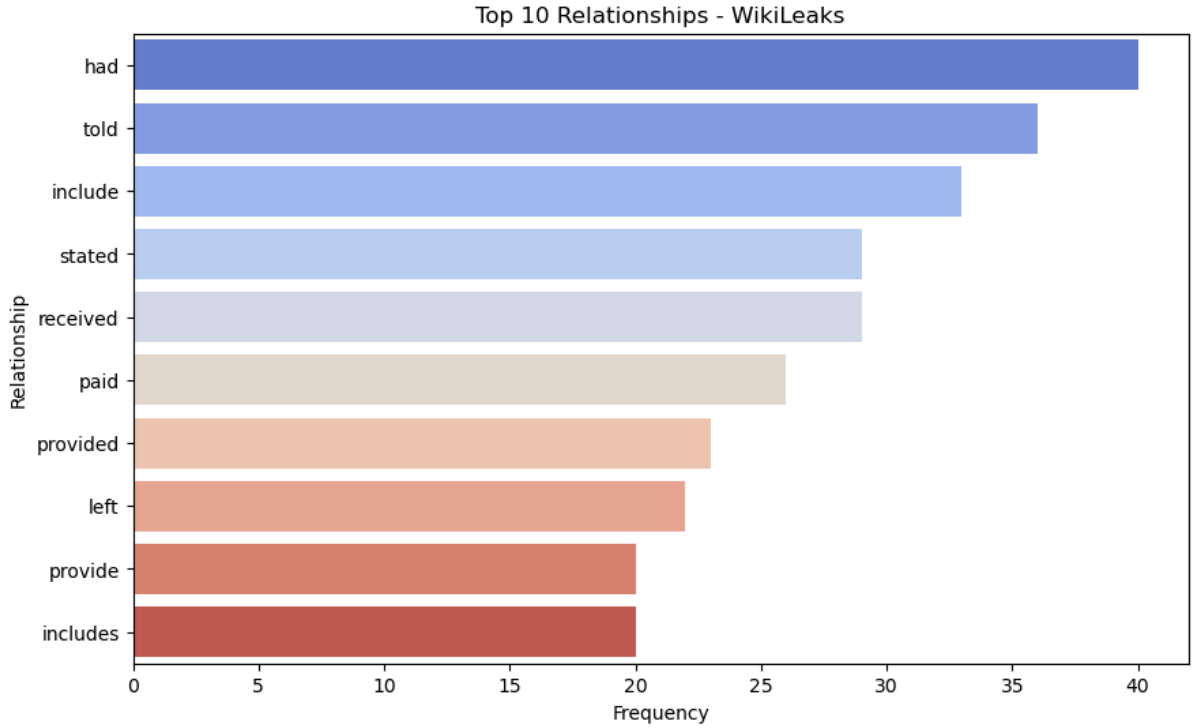


Fig 4.2 Top 10 Relationships - WikiLeaks

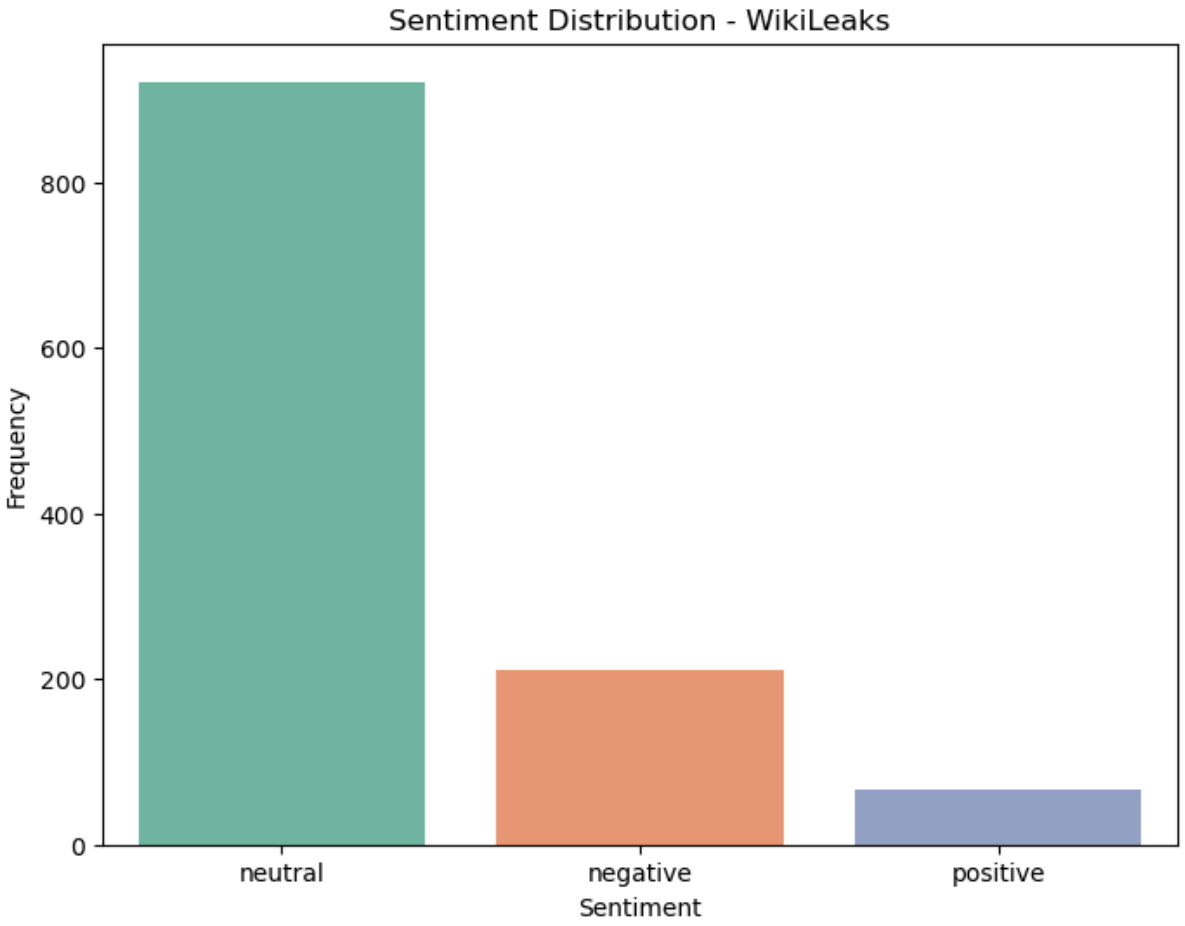


Fig 4.3 Sentiment Distribution - WikiLeaks

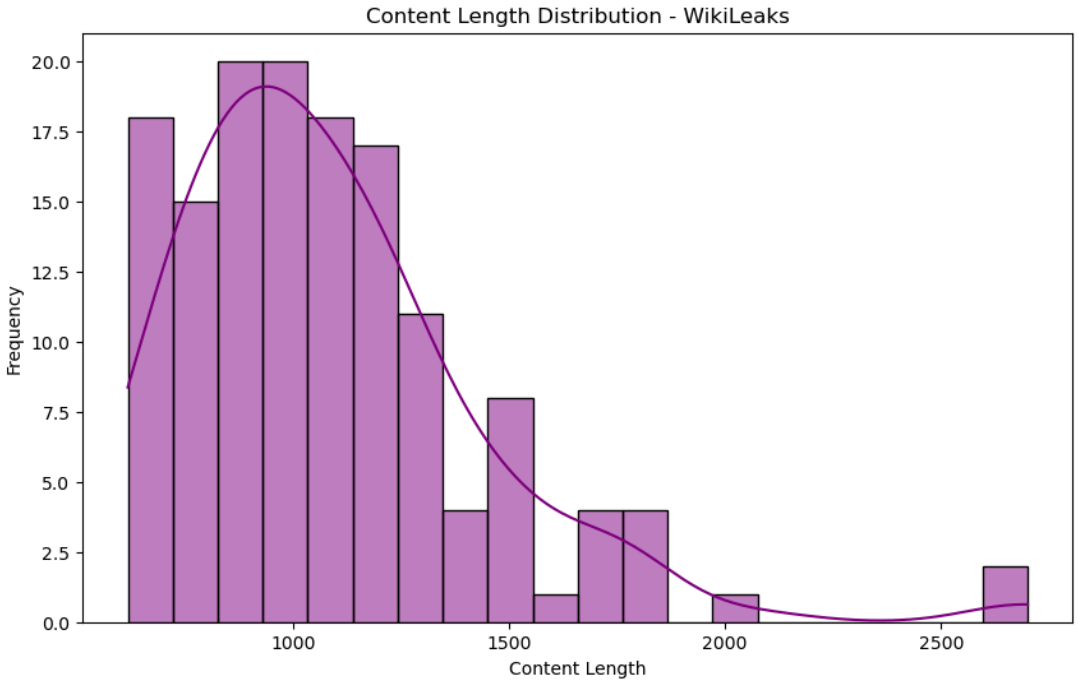


Fig 4.4 Content Length Distribution - WikiLeaks

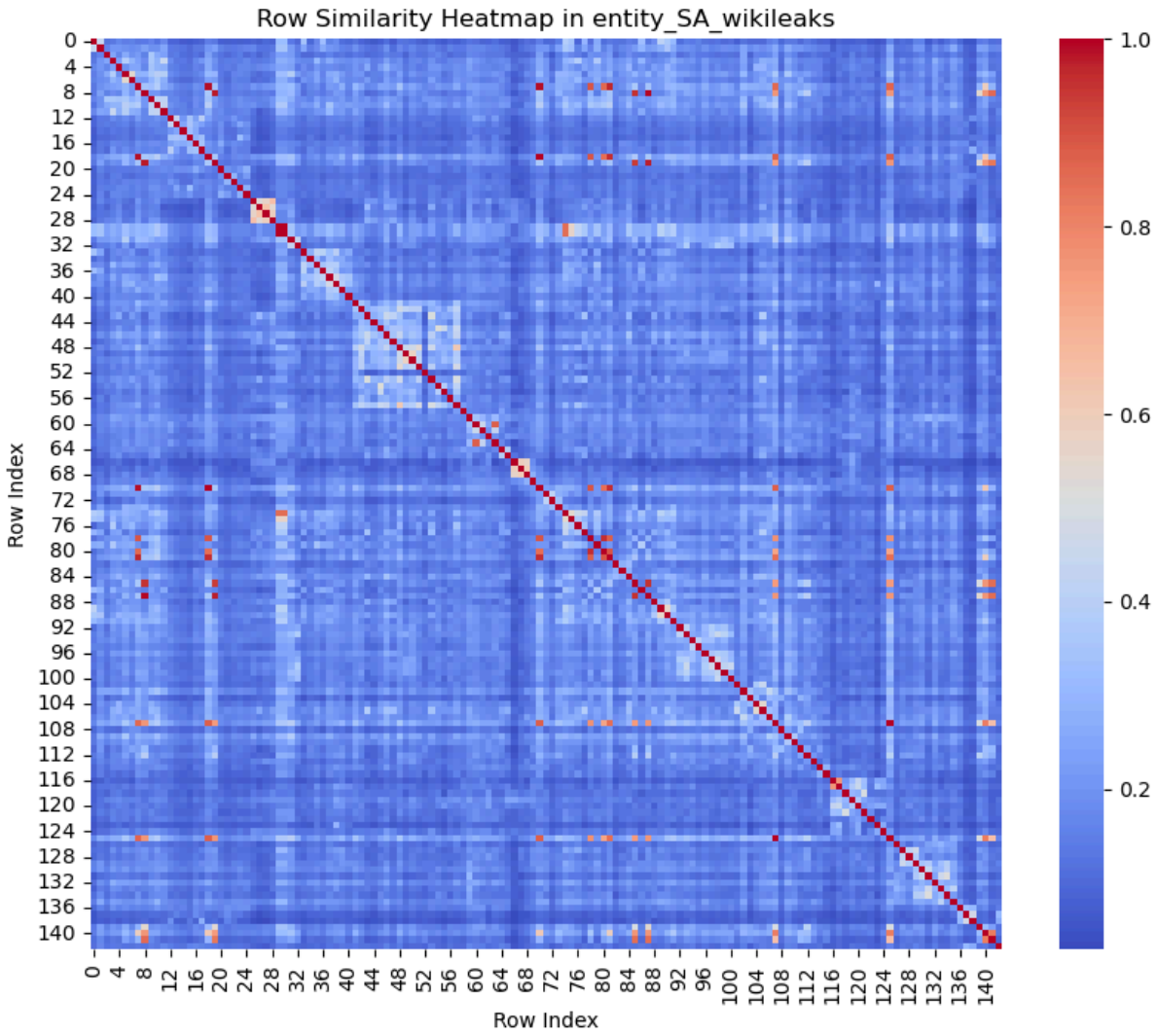


Fig 4.5 Row Similarity Heatmap - WikiLeaks

Visualisations of Similarity Scores using SentenceBERT & Fuzzy Method

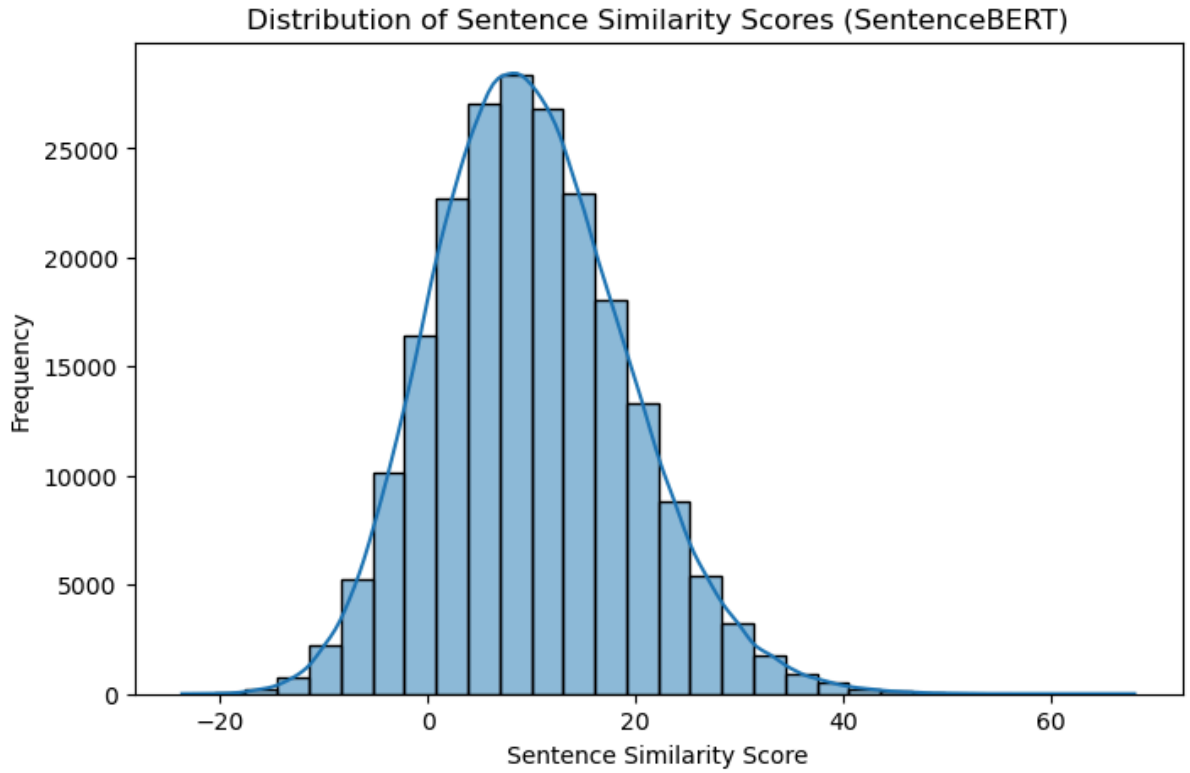


Fig 5.1 Distribution of Sentence Similarity Scores - SentenceBERT

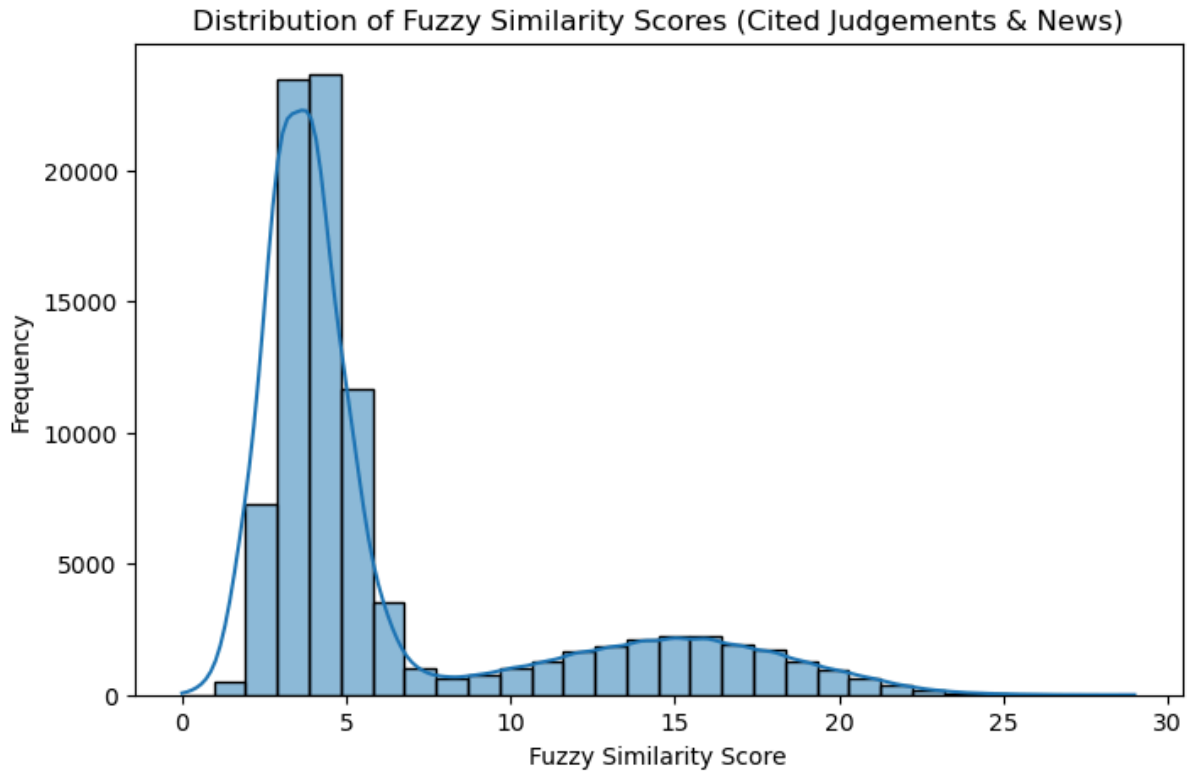


Fig 5.2 Distribution of FuzzySimilarity Scores - Fuzzy Method

Visualisations of Categories found in datasets

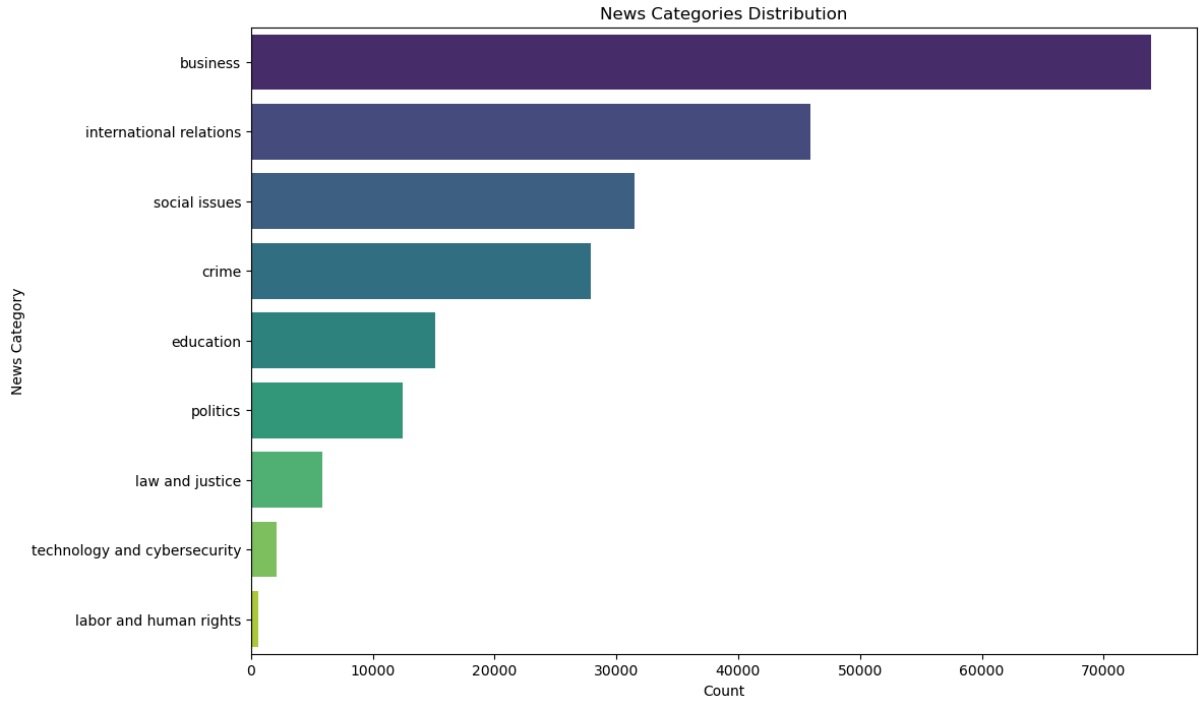


Fig 6.1 News Categories Distribution

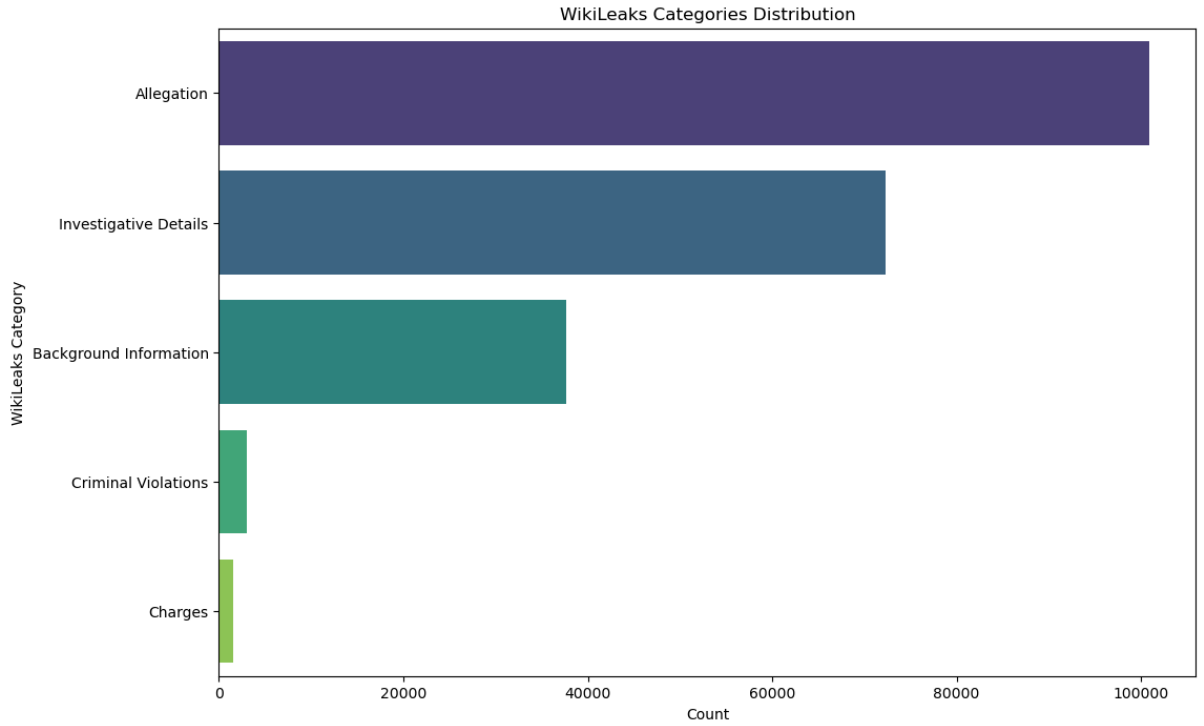


Fig 6.2 Wikileaks Categories Distribution

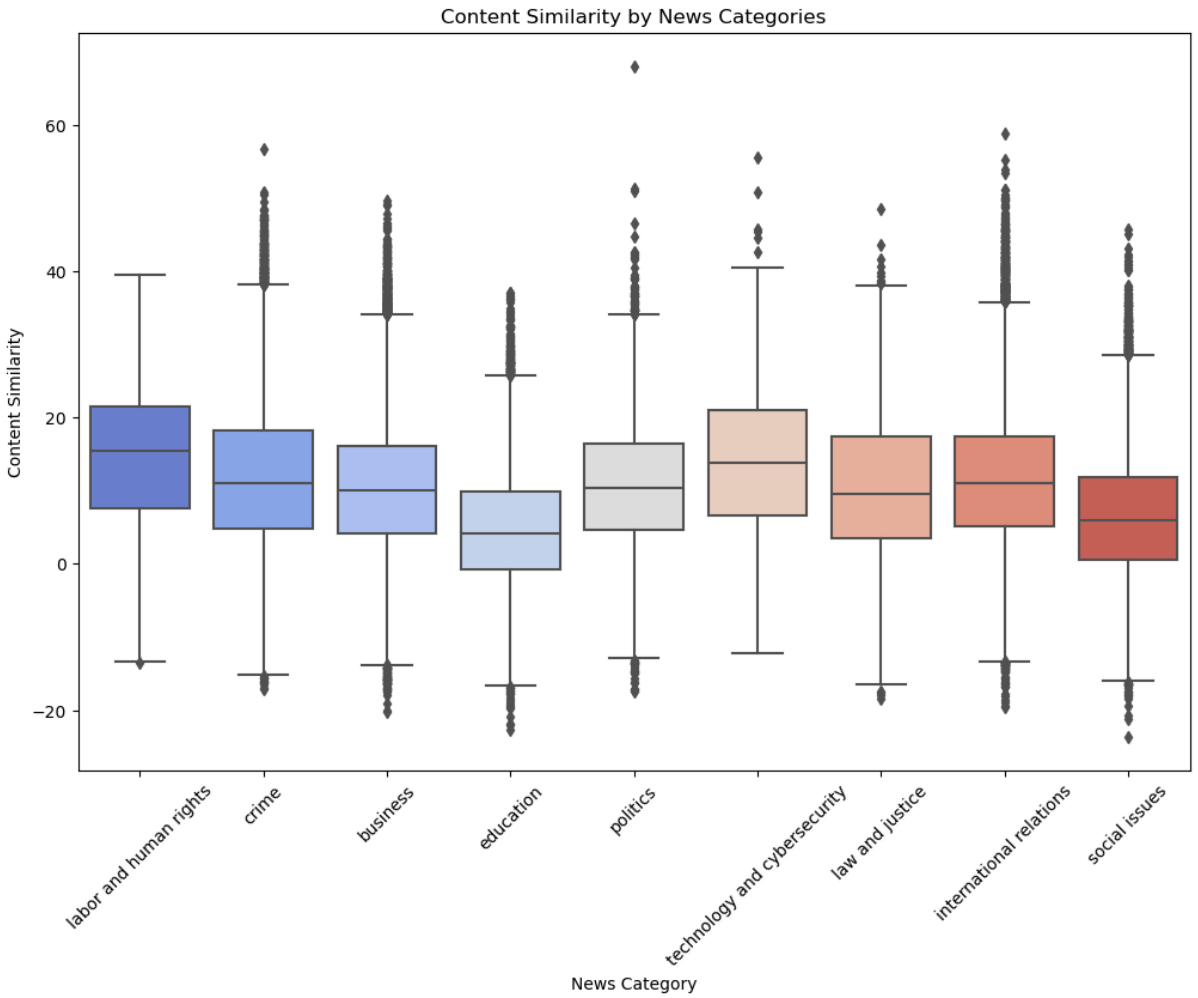


Fig 6.3 Content Similarity by News Categories

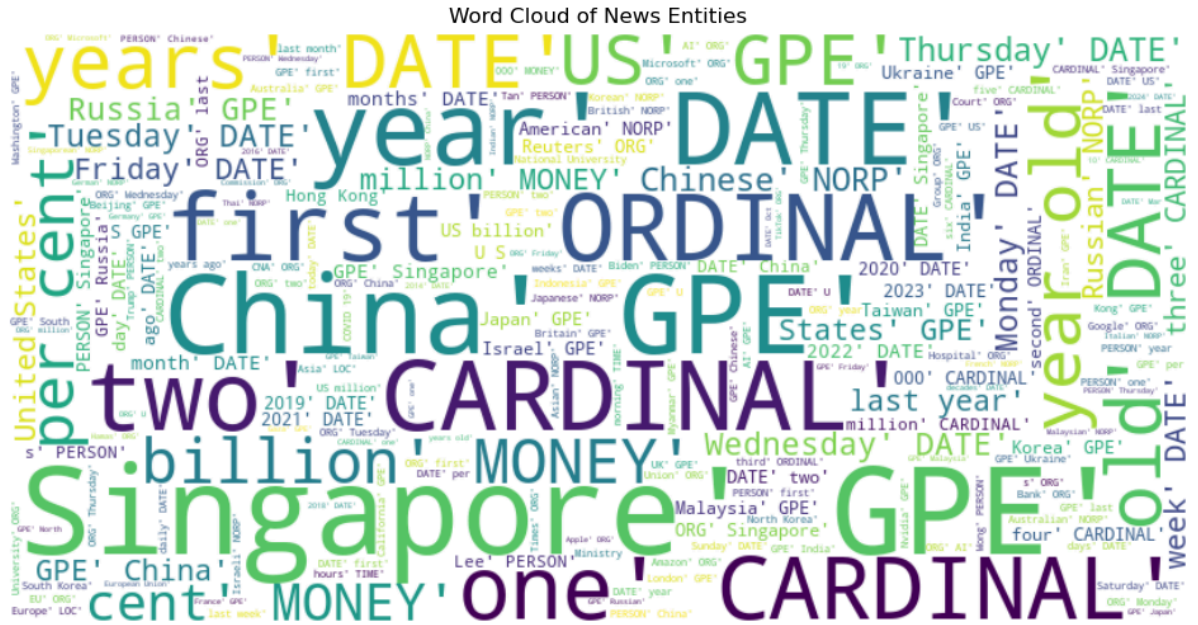


Fig 6.4 Word Cloud of News Entities

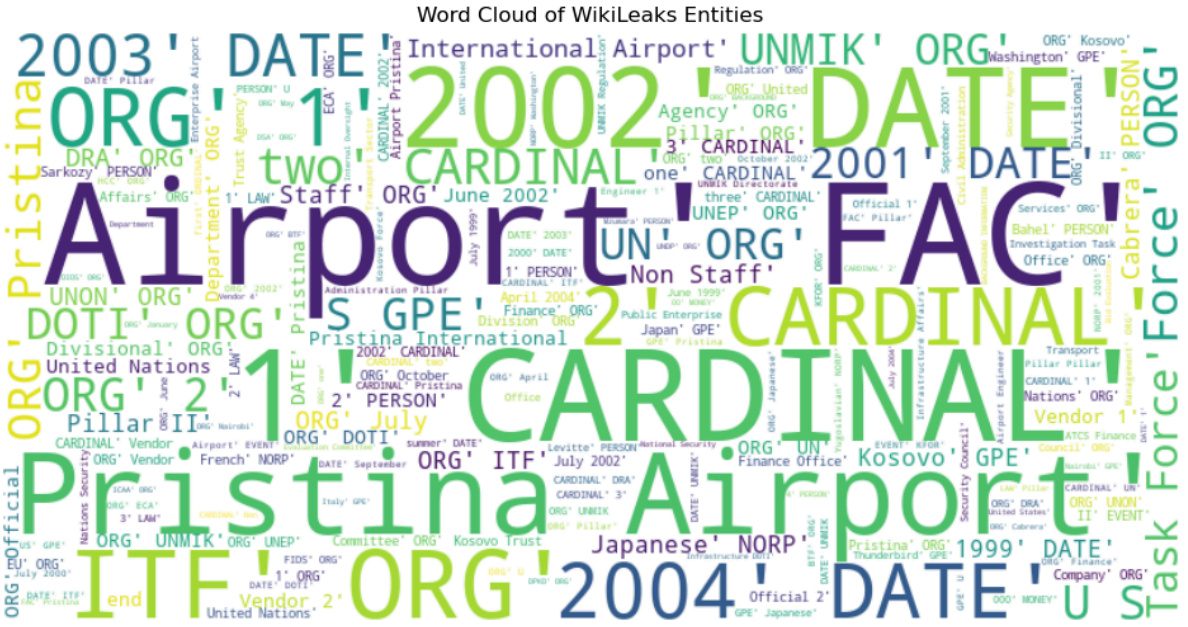


Fig 6.5 Word Cloud of WikiLeaks Entities

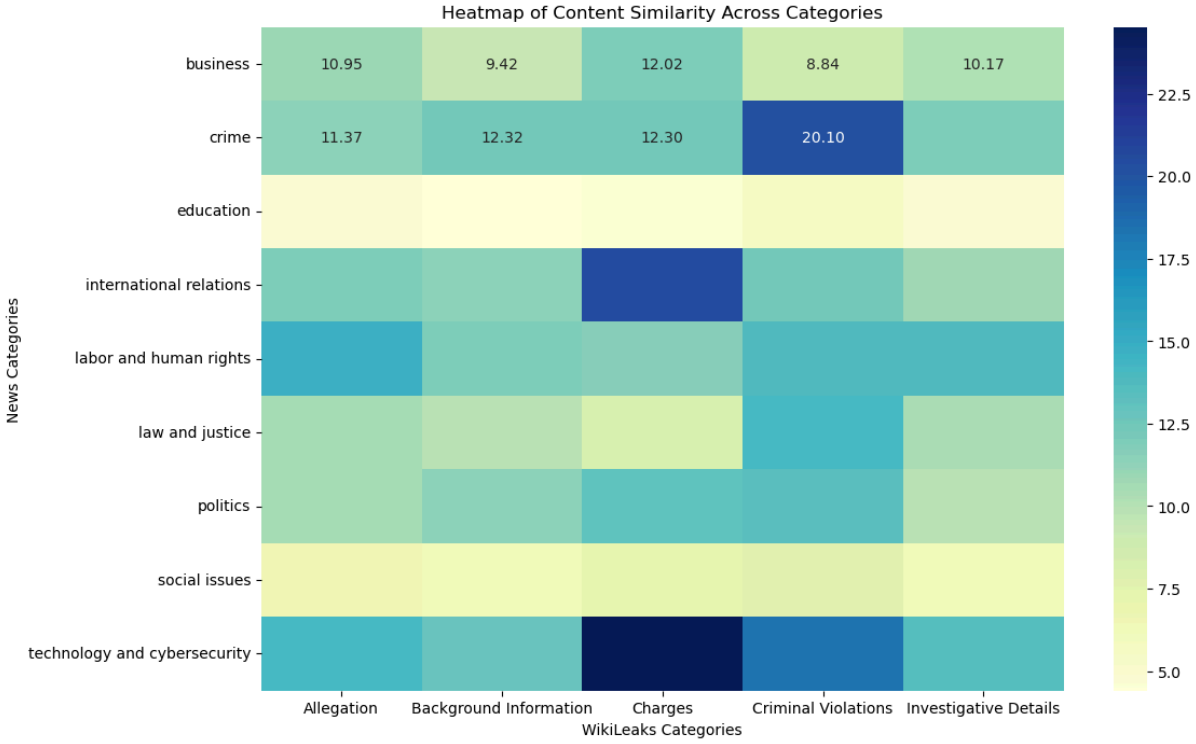


Fig 6.6 Heatmap of Content Similarity Across Categories

Visualisation of timeseries

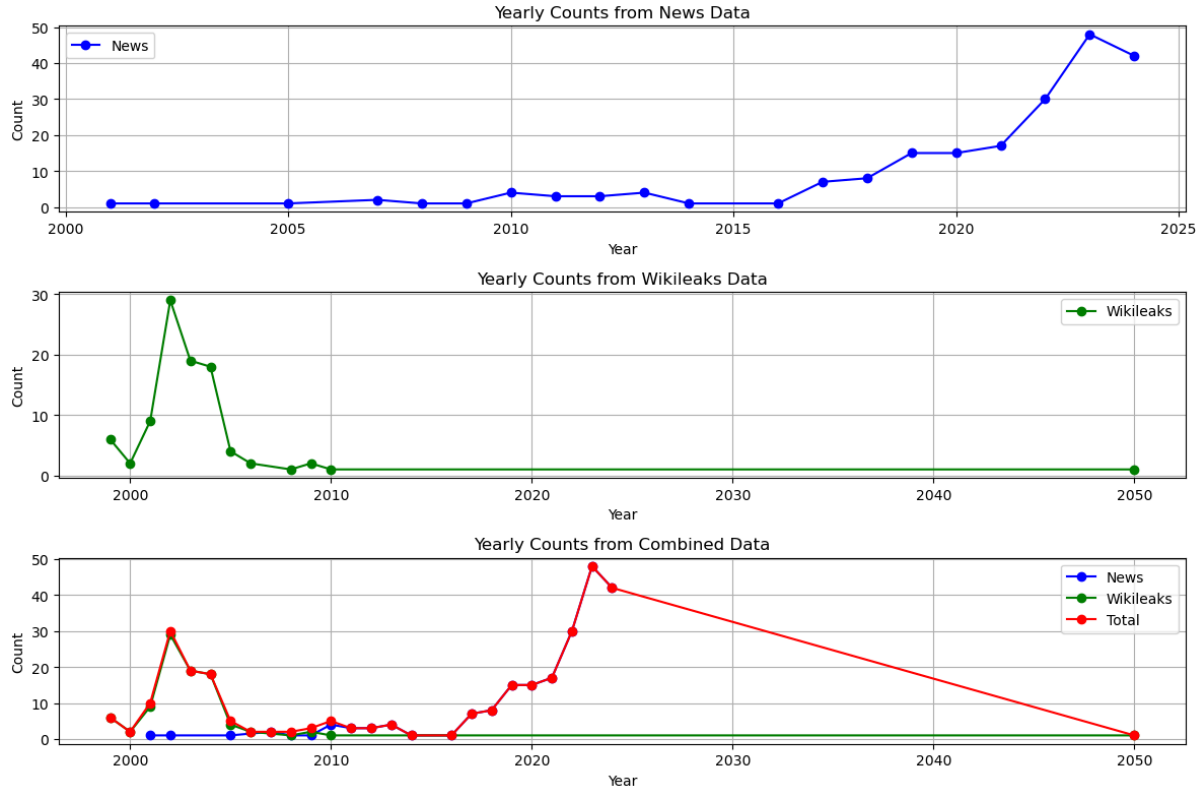


Fig 7 Yearly counts from respective datasets

7 References

fuzzywuzzy. (2020, February 14). PyPI. https://pypi.org/project/fuzzywuzzy/

Reimers, N., & Gurevych, I. (2019, August 27). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv.org. https://arxiv.org/abs/1908.10084

SentenceTransformers Documentation — Sentence Transformers documentation. (n.d.). https://sbert.net/

Streamlit Docs. (2024, September 11). https://docs.streamlit.io/