Parliamentary Submissions Analysis

Manh Tuan Nguyen

Applied Natural Language Processing - Spring 2025

Assignment 1

UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Table of contents

Executive Summary

This report details the application of **Natural Language Processing (NLP)** techniques to analyze public submissions regarding the **Workplace Gender Equality Amendment (Setting Gender Equality Targets) Bill 2024**. The transition from voluntary reporting to mandatory targets for **Designated Relevant Employers (DREs)** has elicited diverse responses from the Australian public and professional sectors. By analyzing a corpus of 31 written submissions, this project utilized **Latent Dirichlet Allocation (LDA)** and **K-Means Clustering** to deconstruct the discourse. The analysis revealed a structural divide in the debate: **Unions** primarily frame gender equality through the lens of **workplace safety and rights**, while **Industry Bodies** focus heavily on **legislative compliance and administrative burden**. The findings provide policymakers with data-driven evidence that the successful implementation of the Bill requires bridging the gap between the "technical compliance" concerns of employers and the "lived experience" priorities of the workforce.

1. Business Understanding

a. Business Use Cases

In the context of public policy, the volume of textual data generated during parliamentary inquiries can be overwhelming. This project applies NLP to the **legislative scrutiny process**. The specific use case is **Automated Stakeholder Analysis**: converting unstructured PDF submissions into structured insights to assist parliamentary committees. The primary challenge motivating this project is the **"Sensemaking Gap."** While submissions are publicly available, synthesizing the conflicting arguments of 31 distinct stakeholders—ranging from individual citizens to multinational corporations—requires significant manual effort. **Machine Learning (ML)** algorithms are relevant here as they can objectively identify latent patterns and thematic clusters that human readers might miss due to cognitive bias or volume fatigue.

b. Key Objectives

The primary objective of this project is to determine how different stakeholders frame the shift from voluntary reporting to mandatory targets.

- **Stakeholders:** The Senate Finance and Public Administration Legislation Committee (primary consumer), the Workplace Gender Equality Agency (WGEA), and submitting entities (Unions, Industry Bodies).
- **Requirements:** To identify dominant themes, assess the polarity of views (supportive vs. cautious), and highlight specific implementation concerns.
- **Methodology:** The project addresses these requirements by using **Topic Modeling** to uncover hidden themes and **Clustering** to group stakeholders based on semantic similarity rather than just their self-identified categories.

2. Data Understanding

The dataset consists of **31 public submissions** provided by the Parliament of Australia. These documents were originally in PDF format, representing a mix of digital-native text and scanned images.

Data Source Characteristics

- **Volume:** The corpus contains approximately 38,000 processed tokens.
- **Distribution:** As visualized in the initial inspection, the data is **strongly right-skewed**. Submission lengths vary from brief cover letters (~220 words) to comprehensive policy reports (~14,000 words).
- **Stakeholder Balance:** The dataset is polarized, with **Union/Worker Representatives** (9 submissions) and **Industry/Employer Bodies** (9 submissions) comprising nearly 60% of the corpus.

Exploratory Data Analysis (EDA)

Initial analysis (See *Figure 4.1* in analysis output) identified that the most frequent substantive terms were **"women"**, **"employers"**, and **"data"**. This indicates that the discourse is anchored in the *mechanics* of the legislation rather than abstract ideology. Furthermore, the **Bigram Network Graph** (*Figure 4.6*) revealed distinct semantic clusters: a "Safety Cluster" linking *sexual harassment* and *violence*, and a "Compliance Cluster" linking *reporting requirements* and *relevant employers*.

3. Data Preparation

Data preparation was critical to ensuring the validity of the NLP models, particularly given the noise inherent in PDF extraction.

Techniques Applied

1. **Hybrid Extraction:** `pdfplumber` was used for digital PDFs, while a fallback to **OCR (Optical Character Recognition)** via `pytesseract` was implemented for image-based submissions (e.g., from the Australian Retailers Association).
2. **Artifact Removal:** A custom cleaning function was developed to resolve specific character-encoding artifacts found in bolded headers (e.g., correcting *"WWoorrkkppllaaccee"* to *"Workplace"*).
3. **Stopword Filtering:** A domain-specific stopword list was compiled to remove procedural terms (e.g., "submission", "inquiry") and high-frequency terms from the Bill's title ("gender", "equality", "setting") to force the model to find underlying themes rather than restating the topic.
4. **Tokenization:** Text was normalized to lowercase and tokenized. Tokens shorter than 3 characters and numeric digits were removed to reduce noise.

Data Splitting Strategy

As this is an unsupervised learning task (Clustering/Topic Modeling) rather than a supervised classification task, a train-test split was not required. The entire corpus was utilized to maximize the density of the term-document matrix.

4. Modeling

Two primary unsupervised machine learning algorithms were selected for this analysis:

1. **Latent Dirichlet Allocation (LDA):**
    - **Rationale:** LDA is the industry standard for **Topic Modeling**. It is a probabilistic model that assumes each document is a mixture of topics. This is ideal for parliamentary submissions, which often touch upon multiple aspects of a Bill (e.g., a single submission might discuss both *pay equity* and *compliance costs*).
    - **Implementation:** The model was trained using `Gensim` with `NUM_TOPICS=5`. The hyperparameters `alpha` and `eta` were set to 'auto' to allow the model to learn the optimal document-topic and word-topic distributions.
2. **K-Means Clustering with TF-IDF:**
    - **Rationale:** To group documents based on overall content similarity. **TF-IDF (Term Frequency-Inverse Document Frequency)** was used for vectorization to downweight common words and highlight unique, identifying terms.
    - **Implementation:** `Scikit-learn` was used to project the 31 documents into vector space and cluster them into `N_CLUSTERS=4`. Principal Component Analysis (PCA) was then used to reduce dimensionality for visualization.

5. Evaluation

a. Results and Analysis

The LDA model achieved a Coherence Score of **0.4150**, converging on five distinct and interpretable topics:

- **Topic 0 (Business Diversity):** Focused on *sector, procurement, diversity*.
- **Topic 1 (Pay Gap):** Focused on *gap, workers, recommendation*.
- **Topic 2 (Legislative Scheme):** Focused on *review, law, scheme, proposed*.
- **Topic 3 (Data & Leadership):** Focused on *data, leadership, equity*.
- **Topic 4 (Worker Reality):** Focused on *health, leave, violence, harassment*.

Key Insight: The **Stance Heatmap** (*Figure 5.2*) revealed a stark thematic divergence. Stakeholders classified as **"Cautious/Conditional"** (mostly Industry Bodies) were **67% focused on Topic 2 (Legislative Scheme)**, indicating their primary concern is the mechanism of implementation. Conversely, **"Supportive"** stakeholders were spread across Topic 1 (Gaps) and Topic 3 (Data), while Unions heavily dominated Topic 4 (Safety).

b. Business Impact and Benefits

The final model provides high value to policy analysts by quantifying the "gap" between stakeholders.

- **Quantifiable Insight:** The analysis proves that resistance to the Bill is not ideological but administrative. The clustering shows that Industry bodies are not linguistically similar to the "Critical" individual submission; they are a distinct cluster focused on *feasibility*.
- **Efficiency:** The model automatically categorized 31 dense legal documents into clear thematic buckets, saving hours of manual review time.

c. Data Privacy and Ethical Concerns

- **Privacy:** The source data is public parliamentary record. However, one submission was titled "Name Withheld." The analysis respected this by not attempting to deanonymize the author, treating the text purely as a "Critical/Opposed" data point.
- **Indigenous Considerations:** While the dataset did not contain specific submissions from Indigenous peak bodies, there is a risk in NLP that minority dialects or specific cultural concerns (e.g., intersectional data on First Nations women) are "washed out" by stopword removal or high-frequency filtering. This limitation must be acknowledged in policy recommendations.

6. Conclusion

This project successfully demonstrated that **Natural Language Processing** can transform raw parliamentary data into actionable policy intelligence. The analysis confirms a **"Two Worlds" narrative** regarding the *Workplace Gender Equality Amendment Bill 2024*.

1. **The "Compliance" World:** Inhabited by Industry Bodies, defined by vocabulary regarding *schemes*, *reviews*, and *legislation*.
2. **The "Rights" World:** Inhabited by Unions and Advocacy groups, defined by vocabulary regarding *safety*, *harassment*, *data*, and *leadership*.

Recommendations

To ensure the Bill's success, the government must address the specific anxieties of the "Compliance" cluster—likely through clear, simplified reporting guidelines—without compromising the structural data requirements demanded by the "Rights" cluster. Future work should involve a longitudinal analysis, comparing these submissions to those from the 2012 Act to track how the discourse on gender equality has evolved over the decade.

7. References

- **Bird, S., Klein, E., & Loper, E.** (2009). *Natural Language Processing with Python*. O'Reilly Media.
- **Blei, D. M., Ng, A. Y., & Jordan, M. I.** (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.
- **Parliament of Australia.** (2024). *Workplace Gender Equality Amendment (Setting Gender Equality Targets) Bill 2024*. Retrieved from aph.gov.au.
- **Pedregosa, F., et al.** (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- **Workplace Gender Equality Agency (WGEA).** (2024). *Gender Equality Indicators*. Retrieved from wgea.gov.au.

8. Appendices

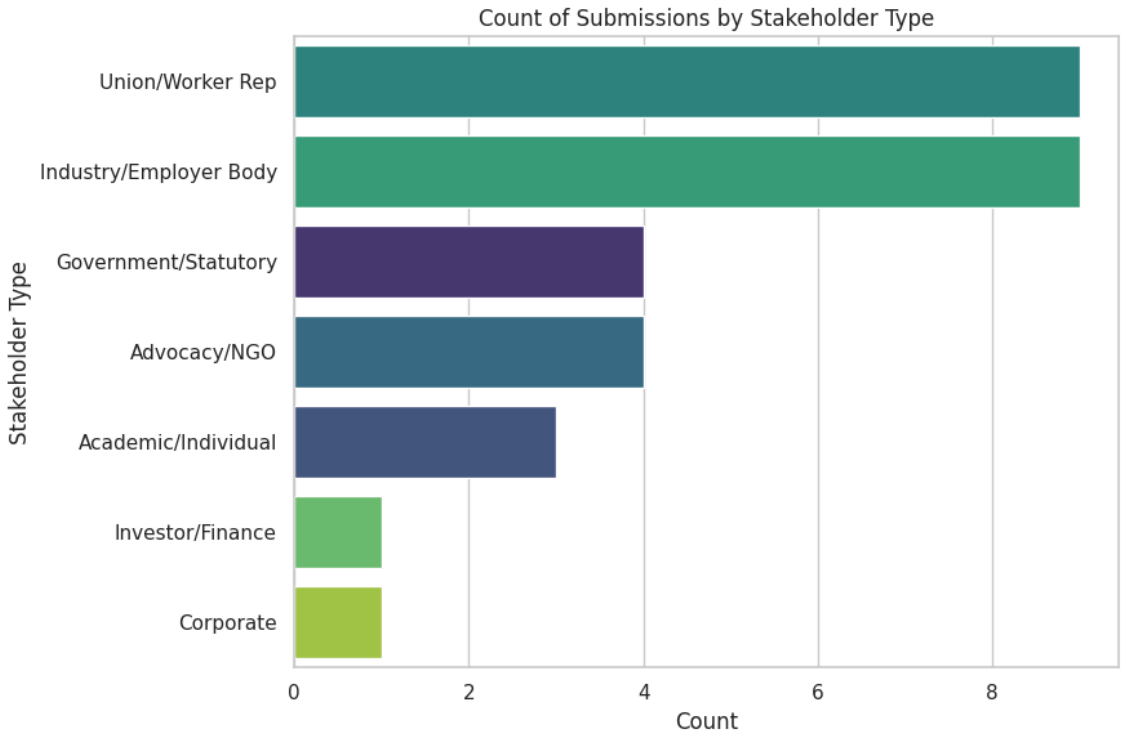

Figure 2.1: Stakeholder Distribution Analysis – Bar chart showing the count of submissions categorized by stakeholder type (e.g., Union, Industry Body, Government).

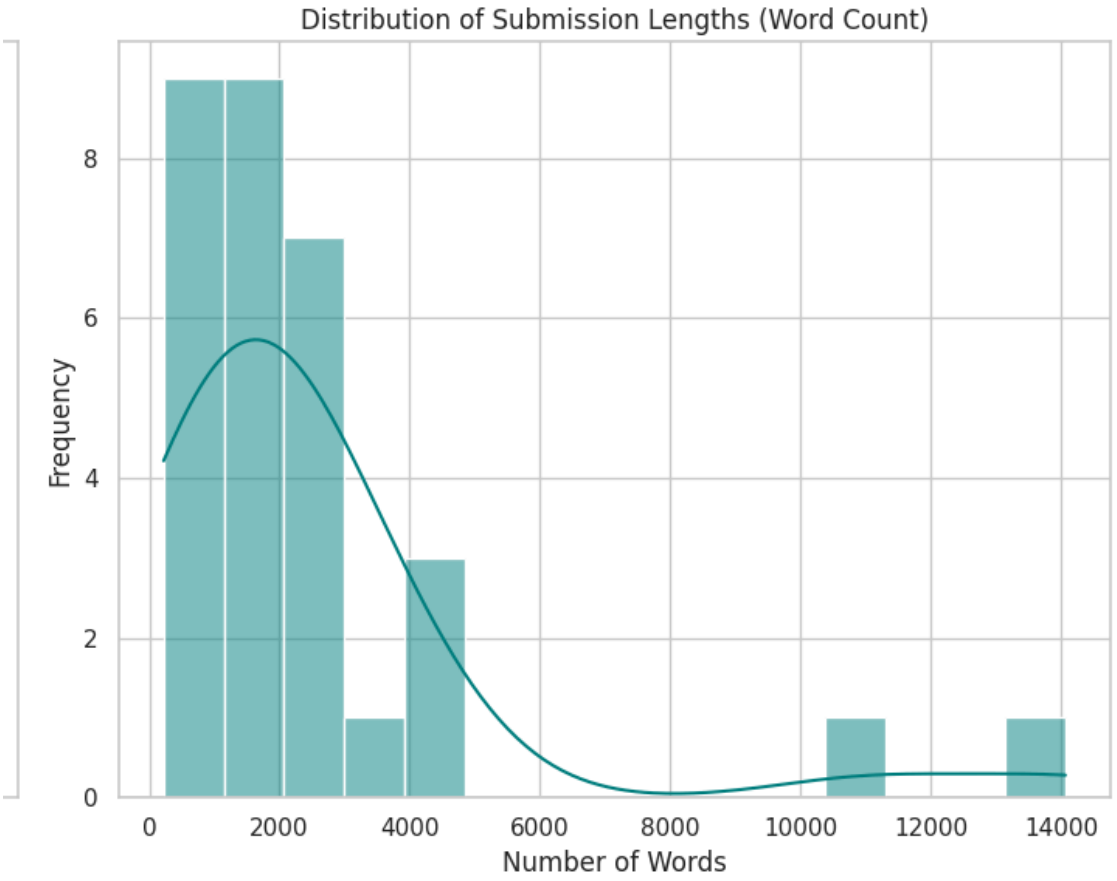

Figure 2.2: Distribution of Raw Submission Lengths – Histogram and KDE plot displaying the word count distribution of the raw PDF texts.

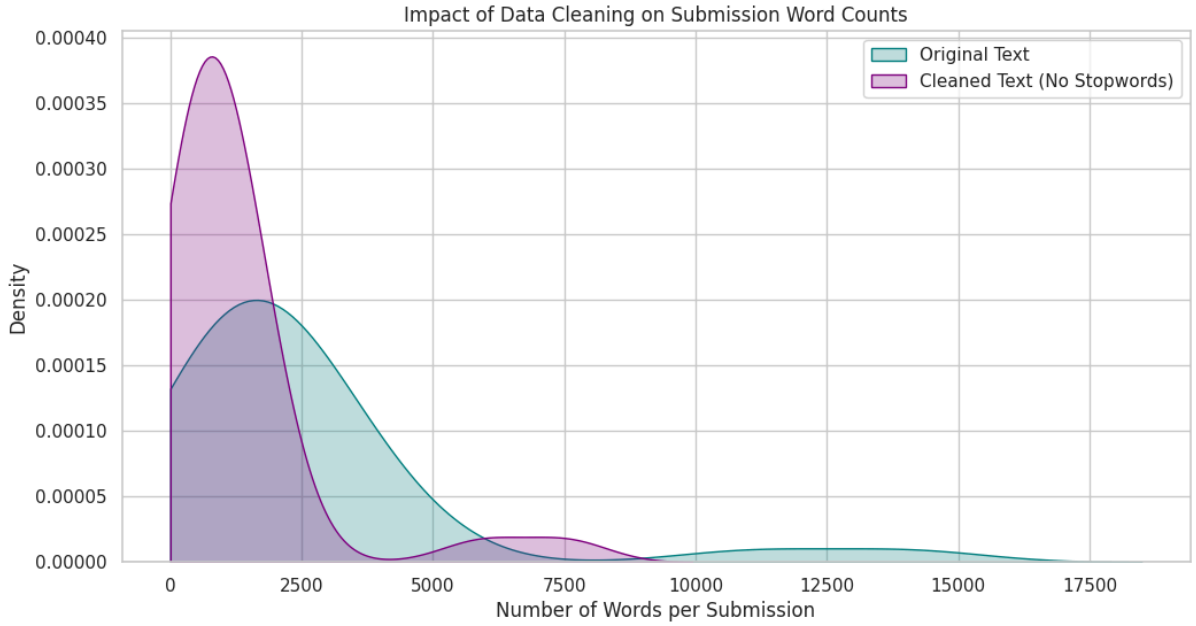

Figure 3.1: Impact of Data Cleaning on Text Volume – Comparative density plot showing the reduction in word count after removing stopwords and artifacts.

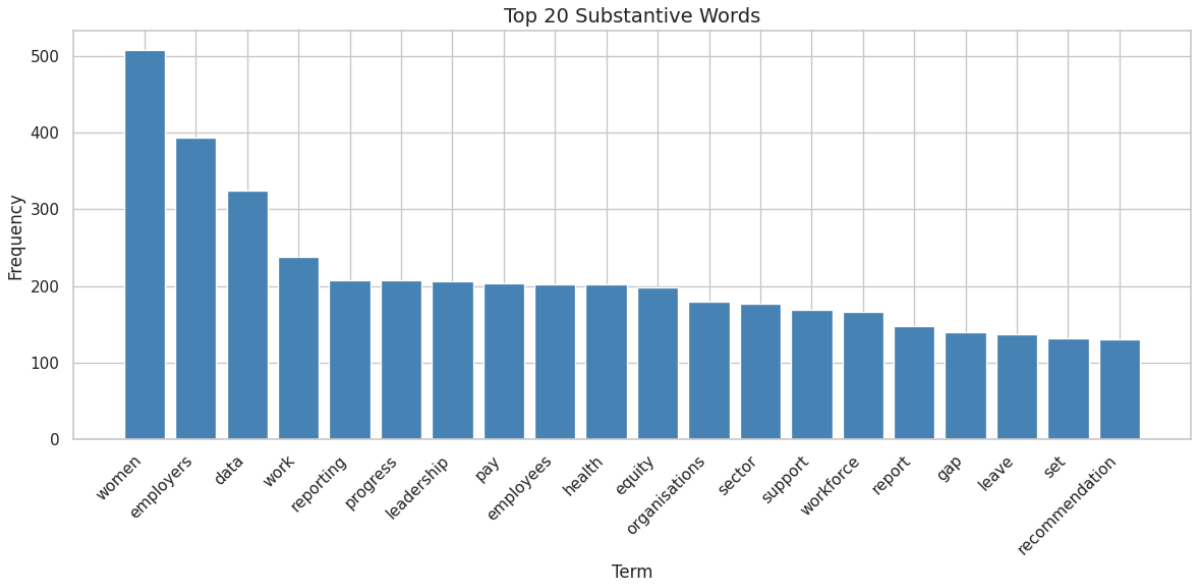

Figure 4.1: Top 20 Substantive Words – Bar chart ranking the most frequent meaningful terms in the cleaned corpus.

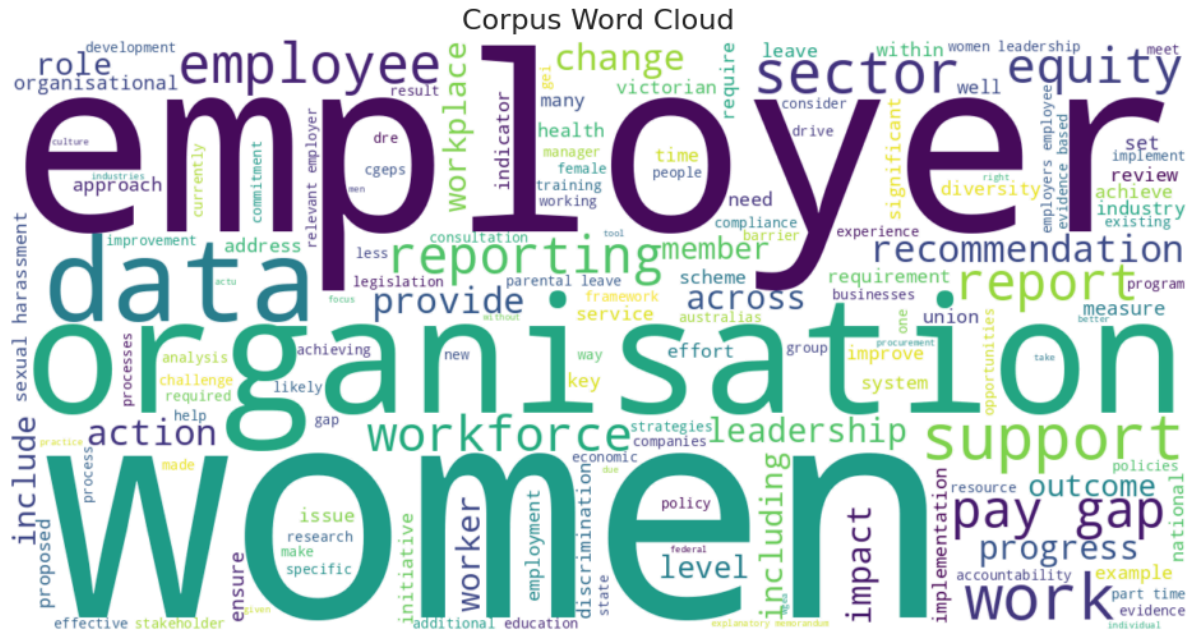

Figure 4.2: Semantic Word Cloud – Visual representation of word frequency, highlighting dominant terms like "Employer," "Women," and "Reporting."

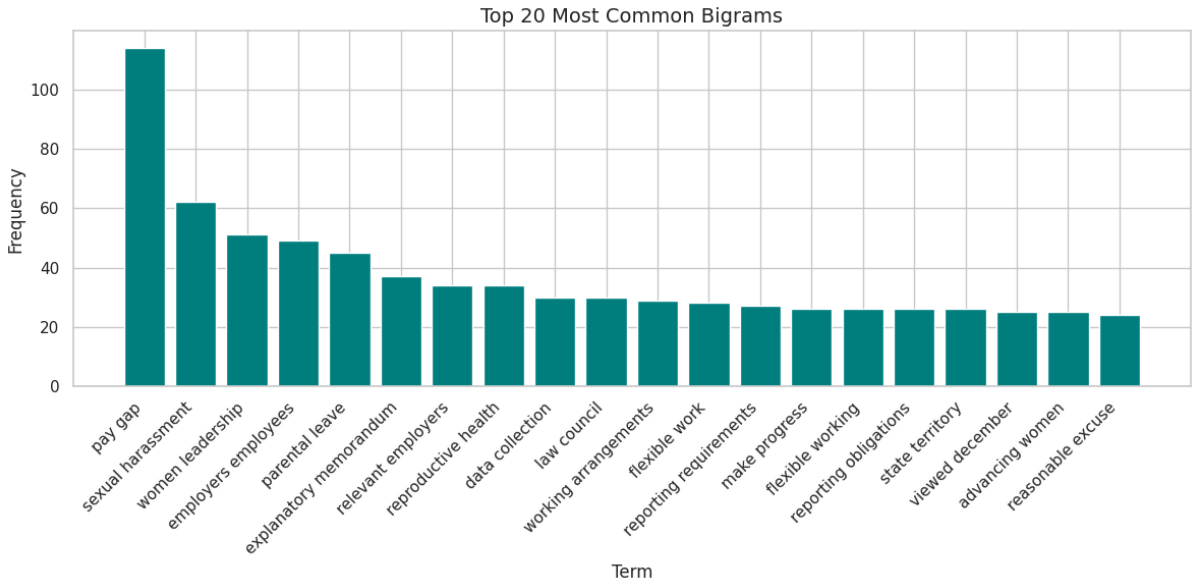

Figure 4.3: Top 20 Frequent Bigrams – Bar chart showing the most common two-word phrases (e.g., "Pay Gap," "Sexual Harassment").

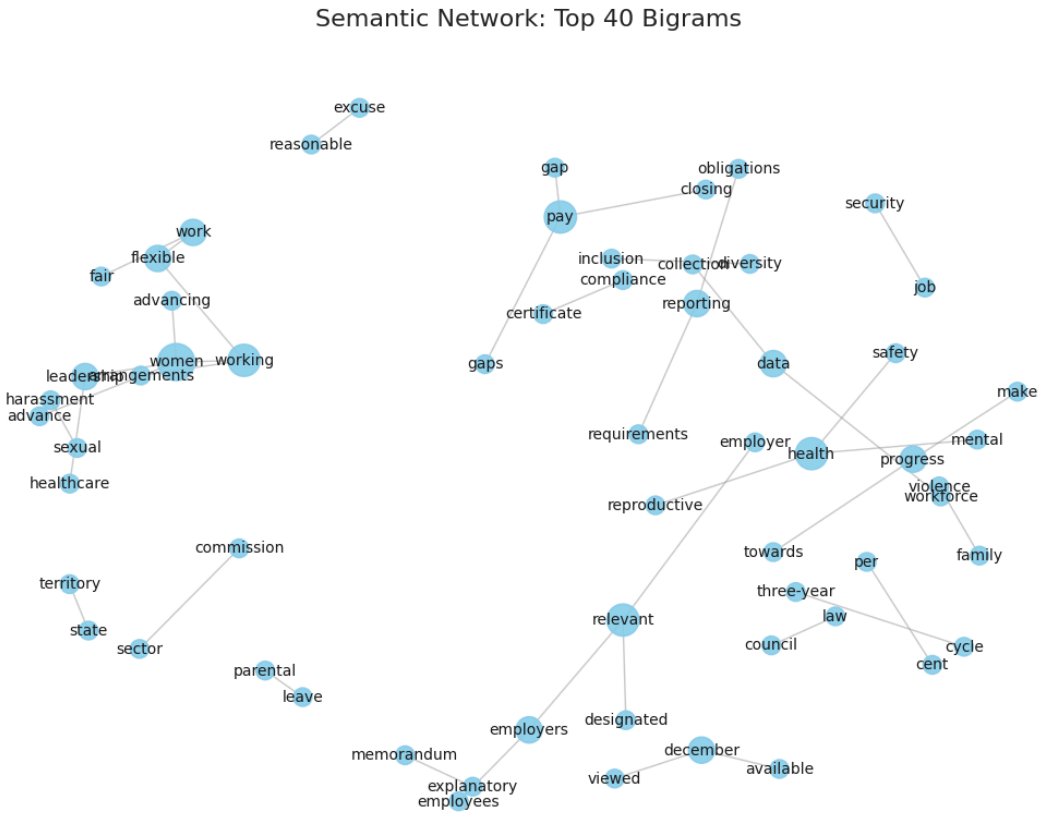

Figure 4.4: Semantic Network Graph – Network diagram visualizing the co-occurrence relationships between the top 40 bigrams.

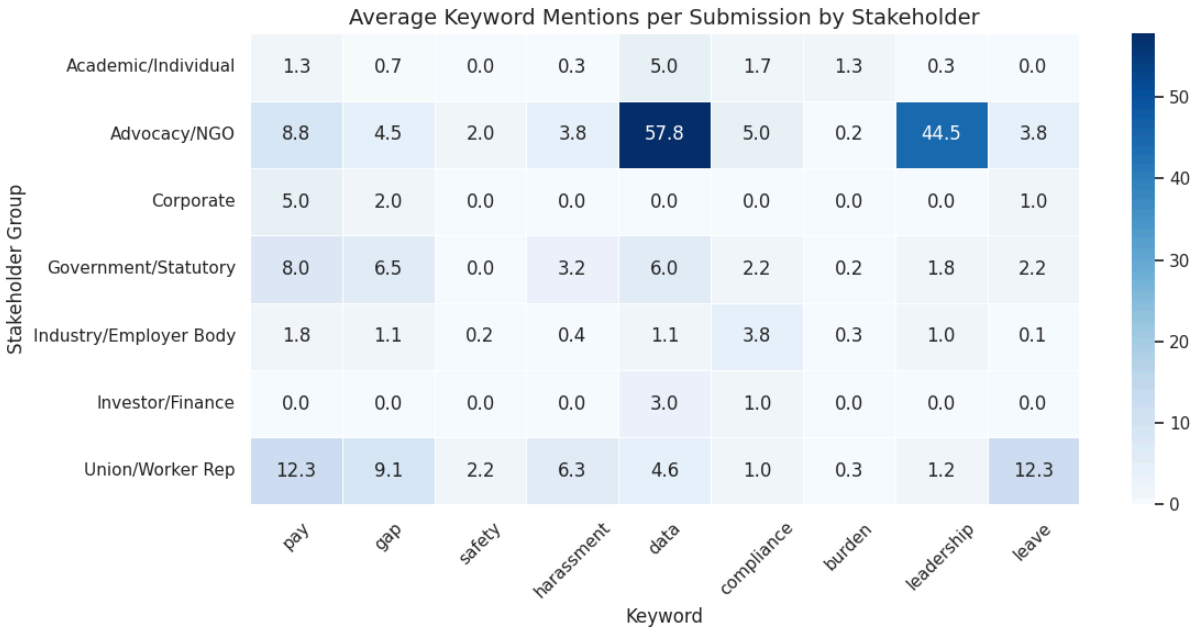

Figure 4.5: Stakeholder Keyword Heatmap – Matrix showing the average frequency of specific policy terms (e.g., "Pay," "Data") across different stakeholder groups.

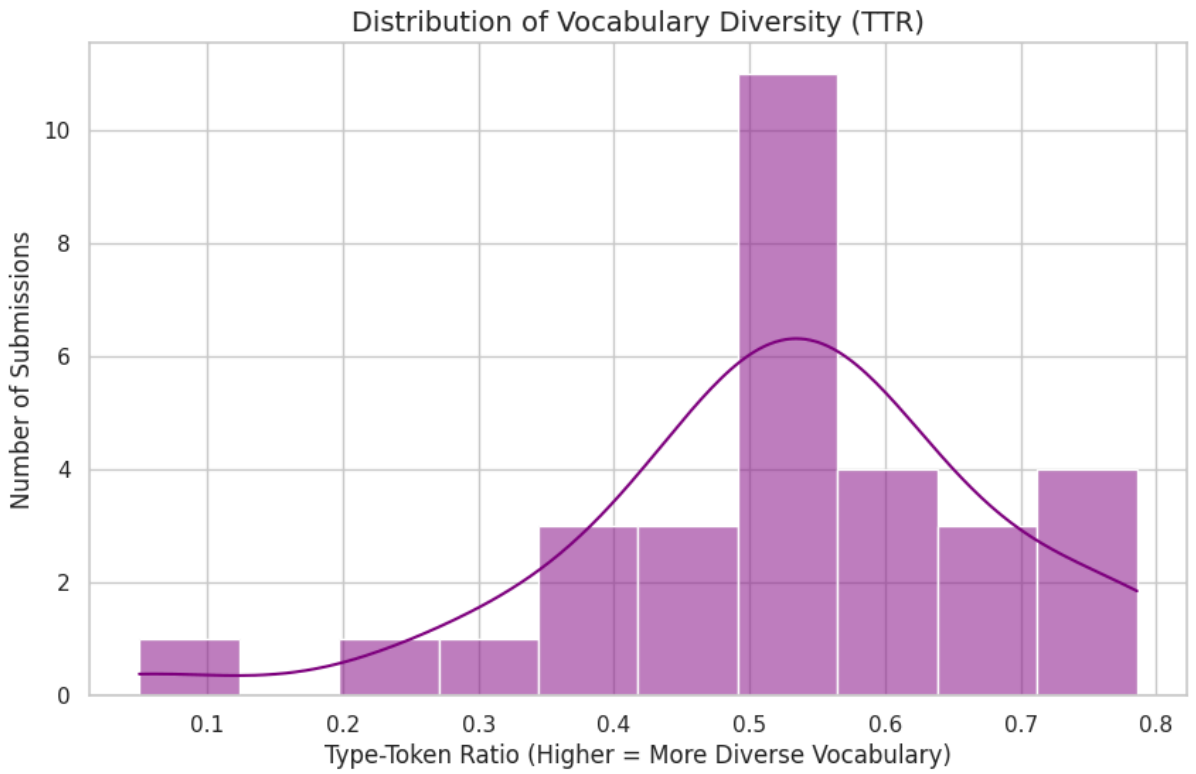

Figure 4.6: Vocabulary Diversity Distribution – Histogram of Type-Token Ratios (TTR) indicating the lexical richness of submissions.

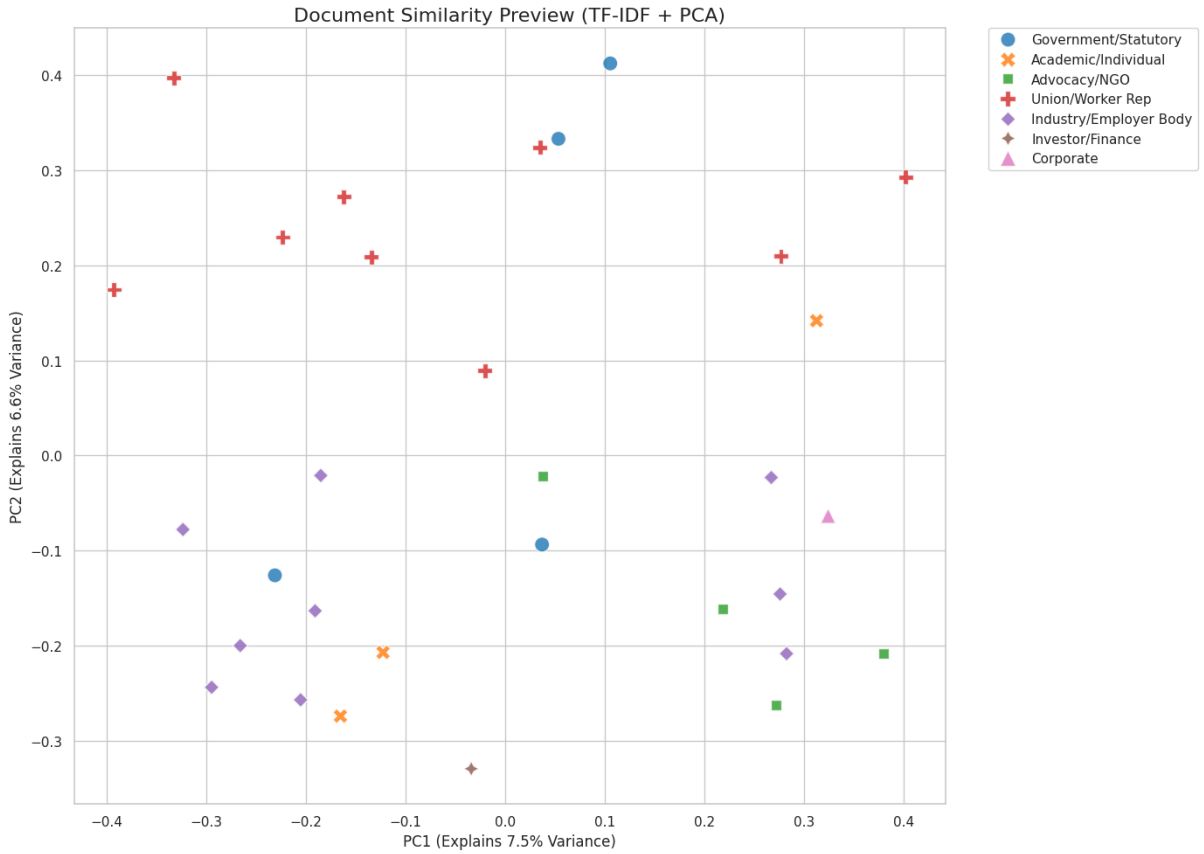

Figure 5.1: Document Clustering (PCA Projection) – Scatter plot visualizing K-Means clusters in 2D space using Principal Component Analysis.

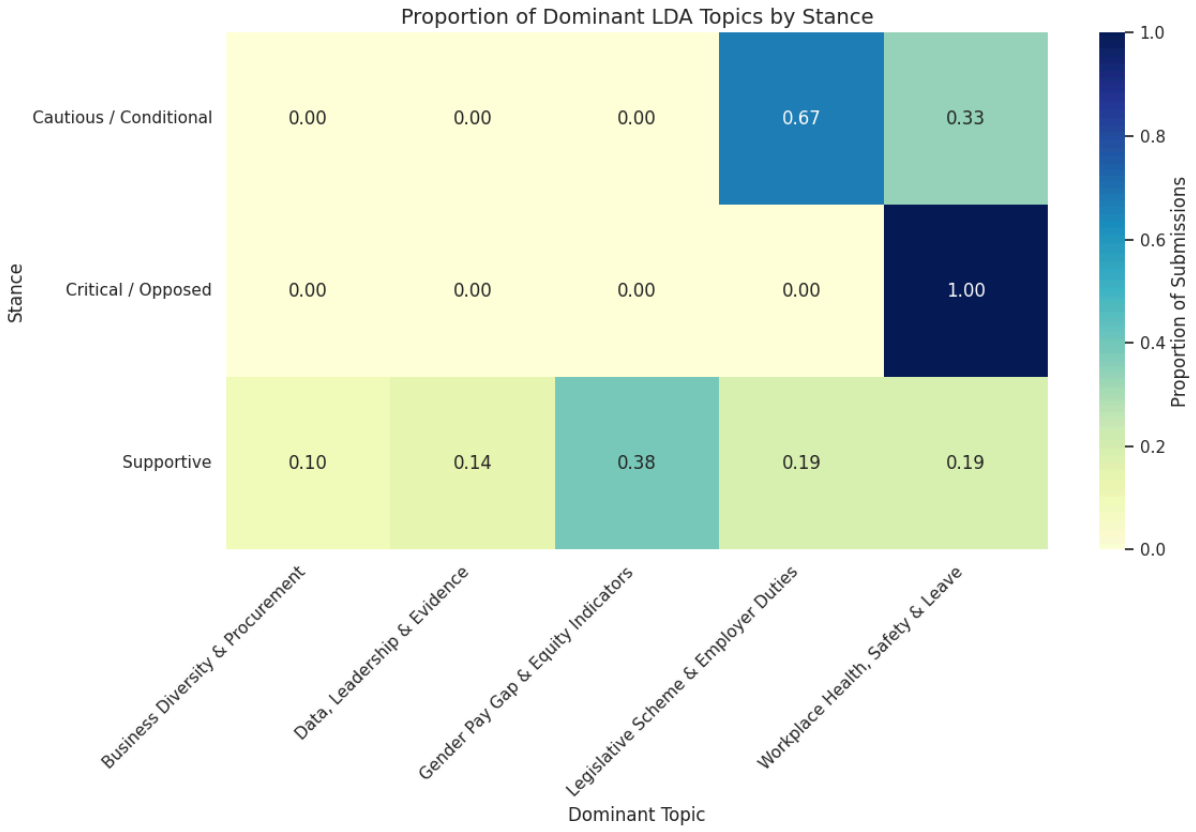

Figure 5.2: Stance Analysis Heatmap – Correlation matrix showing the relationship between stakeholder stance (Supportive vs. Cautious) and dominant latent topics.