Final Report

2025-12-03

1. Research Question and Hypotheses

This project investigates the role of gender in contemporary political media coverage by asking:

Are female politicians more likely than male politicians to be associated with "soft" policy issues (such as education, health, and family) rather than "hard" policy issues (such as the economy, immigration, crime, and foreign policy) in U.S. online news coverage during the 2024 election cycle?

This question matters for political analytics because media framing plays a central role in shaping public perceptions of political competence, leadership traits, and electoral viability. Decades of political communication research show that issue framing influences how voters evaluate candidates particularly along gendered dimensions such as compassion, toughness, or economic expertise. If media outlets systematically link women to soft issues and men to hard issues, these patterns can reinforce gendered expectations of political roles and potentially constrain how voters perceive women's suitability for executive or national-security-oriented offices–including the presidency, where the role of commander in chief places a premium on perceived strength in hard policy domains.

From an analytics perspective, identifying these systematic patterns allows researchers to quantify biases in the political environment, evaluate how these biases evolve over time, and assess their implications for representation and democratic accountability. Understanding whether gendered issue framing persists in the modern high-volume online news ecosystem is especially important for campaigns, policymakers, and scholars studying information effects.

Hypotheses

Based on prior literature about gendered political communication, the analysis tests two hypotheses:

1. Gendered Issue Association: Headlines referencing female candidates will contain soft-issue terms at a higher rate than headlines referencing male candidates.

2. Hard Issue Gap: Headlines referencing male candidates will contain hard-issue terms at a higher rate than headlines referencing female candidates.

Together, these hypotheses evaluate whether gendered stereotypes continue to shape the informational environment surrounding political candidates in 2024.

2. Background

Research in political communication has consistently demonstrated that media coverage of political candidates is shaped by gendered expectations and stereotypes. Decades of work in top journals including the American Political Science Review, Journal of Politics, and Political Communication show that women in politics are more frequently framed around issues associated with compassion, care, and social welfare, whereas men receive more coverage emphasizing leadership, security, and economic strength. These patterns

mirror broader societal beliefs about "feminine" and "masculine" policy domains and can influence how voters perceive the competence and electability of political figures.

Empirical studies find that female candidates are often linked to "soft" issues such as education, healthcare, and family policy, even when they hold positions or leadership roles in traditionally "hard" issue areas. For example, several experimental and observational studies show that voters tend to rate women as more competent on social welfare and education but less competent on national security or economic management, regardless of their actual qualifications. Media framing reinforces these public perceptions: women receive more coverage about personal traits and family roles, while men receive more issue-focused and policy-oriented coverage.

More recent work examining digital media suggests that gendered framing persists even as news production shifts to online platforms and social feeds. Analyses using large-scale corpora of online news show that algorithmic amplification can intensify existing representational biases by reproducing the themes most likely to drive engagement—often those that align with stereotype-consistent narratives.

Nevertheless, several limitations remain in the existing literature. Much of the foundational research focuses on earlier election cycles or on a small set of high-profile candidates, leaving open questions about how gendered patterns operate in the contemporary, high-volume online news environment. Few studies examine broad samples of candidates across parties and regions using systematic, reproducible NLP methods applied directly to digital news content. We also know relatively little about whether gender associations with specific policy domains remain stable in 2024.

This project builds on these insights by conducting a fresh, data-driven analysis of gendered issue framing in U.S. online news headlines during the 2024 election cycle. By using a structured set of high-visibility candidates and applying transparent text classification methods to thousands of headlines, the study contributes new evidence to an ongoing scholarly conversation about how gender shapes political communication in the modern media landscape.

3. Data Description

This project uses a dataset of online news headlines collected through the Media Cloud Search API, an open-source platform developed by MIT and Northeastern University that aggregates digital news stories from thousands of U.S. media outlets. Media Cloud provides programmatic access to structured story metadata, including headline text, publication date, outlet name, URL, and language information.

The dataset is cross-sectional observational text data, where each row corresponds to a single news headline that mentions a specified political candidate. The target population for this study is online political news content published in the United States during the 2024 election cycle, specifically from January 1 to November 27, 2024.

To answer the research question, we created a candidate set consisting of eight nationally visible U.S. senators and Senate candidates: four women (Tammy Baldwin, Kirsten Gillibrand, Katie Britt, and Elizabeth Warren) and four men (Ted Cruz, Rick Scott, Josh Hawley, and Sherrod Brown). This set balances gender and includes candidates with consistent national-level visibility to ensure adequate coverage volume in the dataset.

Using the Media Cloud API's **/search/story-list** endpoint, we queried all online news headlines within the defined time window that contained each candidate's full name. The resulting dataset include approximately 4,000 headlines, though coverage volume varied by candidate. Because the API returns headlines rather than full article text, the unit of analysis is the headline itself.

To ground the analysis in concrete examples, headlines such as "Gillibrand, Colleagues Introduce Bipartisan Bill To Expand Working Families' Access To Child Care" or "Senator Britt Leads Bipartisan Push for Maternal Health Awareness Day" illustrate the type of soft-issue framing captured in the data. By contrast, headlines referencing male candidates more often emphasize inflation, border security, or crime—for example, coverage of Ted Cruz criticizing border enforcement failures or economic policy under the Biden administration (e.g., "Sen. Ted Cruz Blasts Biden Administration for Border Crisis").

Headlines with missing or non-English text were excluded during preprocessing. Because Media Cloud returns only indexed headlines, no imputation was performed; all analyses are conducted on observed headline text only.

Each headline record contains key variables such as:

- title: the headline text
- publish_date: publication date
- media_name: name of the outlet
- media_url: domain of the outlet
- indexed_date: date the story was ingested by Media Cloud
- url: link to the article
- language
- candidate, party, state, gender: attributes appended during data cleaning

To study issue framing, we applied a dictionary-based text transformation that classifies each headline into either "soft" or "hard" issue coverage. The "soft-issue" dictionary included terms referring to education, health, childcare, and family; the "hard-issue" dictionary included terms related to the economy, crime, defense, immigration, and foreign policy. For each headline, we counted occurrences of soft and hard words and assigned the headline to the category with the higher count (with ties defaulting to the hard category). This created two derived variables:

- soft_indicator: 1 if soft > hard keyword count
- hard_indicator: 1 if hard >= soft keyword count

We also computed aggregate measures such as:

- soft_share_gender: proportion of soft headlines by gender
- soft_share_candidate: proportion of soft headlines by candidate
- soft_total_distribution: share of all soft-issue headlines assigned to each gender

No weights, survey adjustments, or experimental designs were used. The dataset reflects the naturally occurring distribution of online news coverage. Because the data come from an API-driven corpus, the sample is not a probability sample, and results should be interpreted as descriptive patterns in the 2024 online political information environment rather than population-level estimates.

Apples-to-Apples Comparison and Remaining Confounders

The candidate set is designed to approximate an apples-to-apples comparison along several key dimensions: all individuals are U.S. senators or high-profile Senate candidates, all receive sustained national media attention, and all are observed within the same 2024 digital news environment. These features help reduce structural differences in baseline newsworthiness that could otherwise obscure gendered framing patterns. As a result, the analysis tests whether gendered framing patterns are present in the media environment, rather than whether gender itself causally produces those differences.

However, this design does not fully isolate gender. Important potential confounders remain, including:

- Ideology (progressive vs. conservative),
- Partisan media ecosystems,
- Candidate rhetorical style and strategic messaging,
- Issue ownership and policy specialization,
- News cycles driven by scandals or viral events.

In particular, the male candidate set in this cycle is more heavily Republican, while the women include both Democrats and Republicans. As a result, differences in issue framing may partly reflect party- or ideology-specific agenda setting rather than gender alone. Accordingly, the findings should be interpreted as descriptive differences in media framing associated with gender, not as causal estimates of gender bias.

This dataset provides a large, systematic collection of political news headlines during the 2024 cycle, offering an appropriate foundation for examining whether gendered issue framing persists in contemporary online media.

4. Results

This section presents the empirical results from the text classification and headline-level analysis. We begin by outlining the analytic strategy and then introduce three core findings supported by visualizations.

4.1 Analysis Strategy

The goal of the analysis is to compare the prevalence of "soft" versus "hard" issue framing across male and female political candidates during the 2024 election cycle. We first computed descriptive statistics and group means for soft-issue coverage by gender and by individual candidate. We then examined the distribution of soft-issue headlines across genders to assess who is most strongly associated with these issue domains in the news environment.

No regression models were required for the core findings because the patterns of interest relate directly to differences in proportions across clearly defined groups; however, regression extensions are included in the discussion as possible next steps.
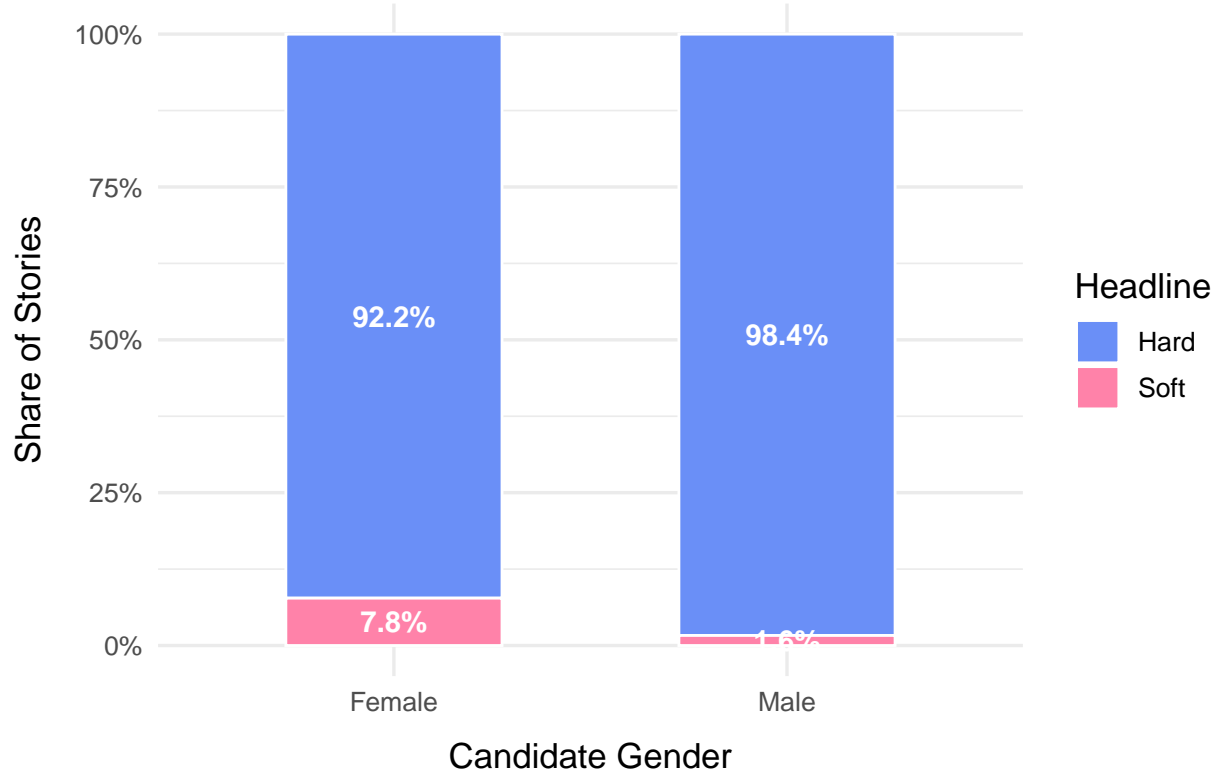
Because the analysis focuses on descriptive differences in proportions rather than causal estimation, formal hypothesis tests and p-value thresholds are not emphasized. Instead, uncertainty is shown through relative magnitudes, consistency across candidates, and robustness to alternative classification rules.

All findings are supported by robustness checks examining alternative classification rules and keyword dictionaries, presented in the appendix.

4.2 Finding 1: Female Candidates Receive a Higher Share of Soft-Issue Headlines

Caption: **Female political candidates are four to five times more likely than male candidates to receive soft-issue headline framing.**
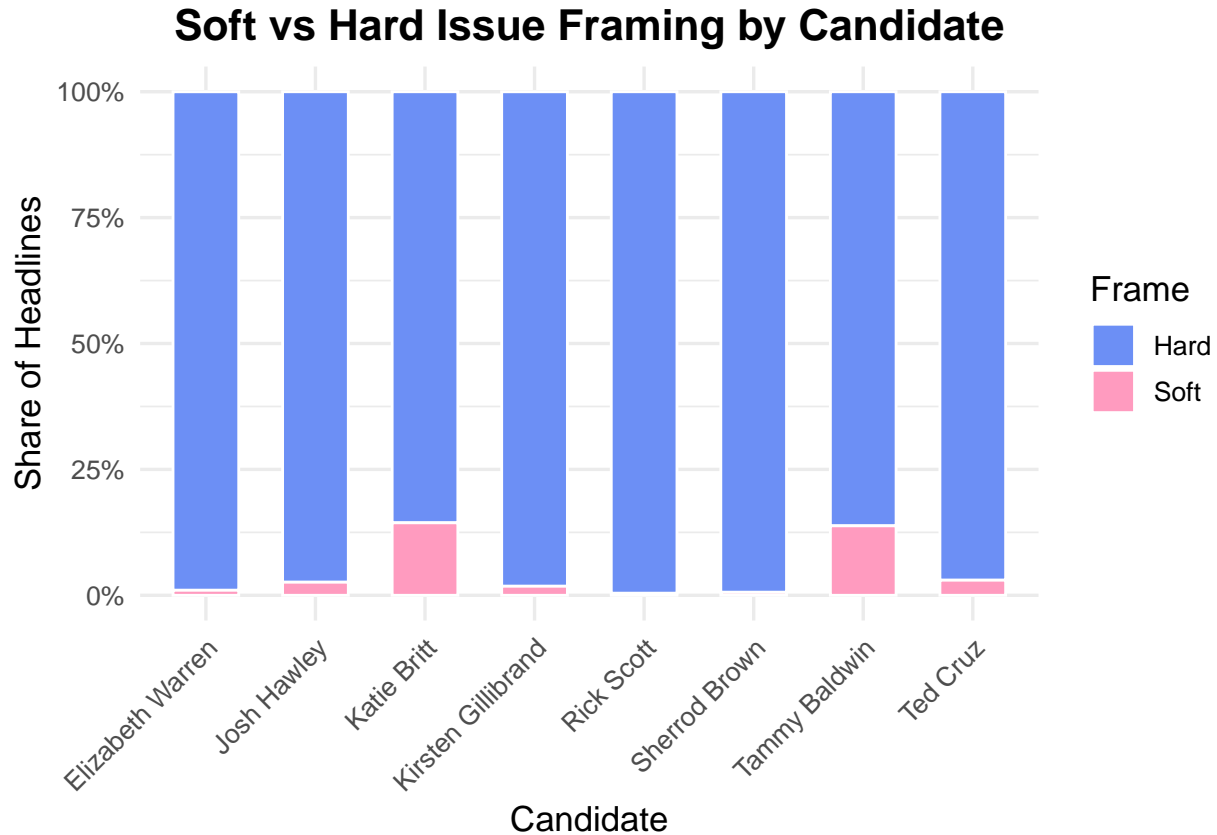
Share of Soft vs Hard Issue Headlines by Candidate Gender

Interpretation: Across all headlines, 7.8% of those referencing female candidates contain soft-issue language, compared to 1.6% for male candidates. Although soft-issue headlines are rare overall, the relative difference is large: women are four to five times more likely to be framed around education, health, or family themes. This supports Hypothesis 1 and suggests that gendered issue associations persist in the 2024 media environment.

4.3 Finding 2: The Gender Gap Persists Across Individual Candidates

Caption: **Three of four female candidates exhibit higher soft-issue shares than nearly all male candidates, indicating that gendered framing is not driven by a single politician.**
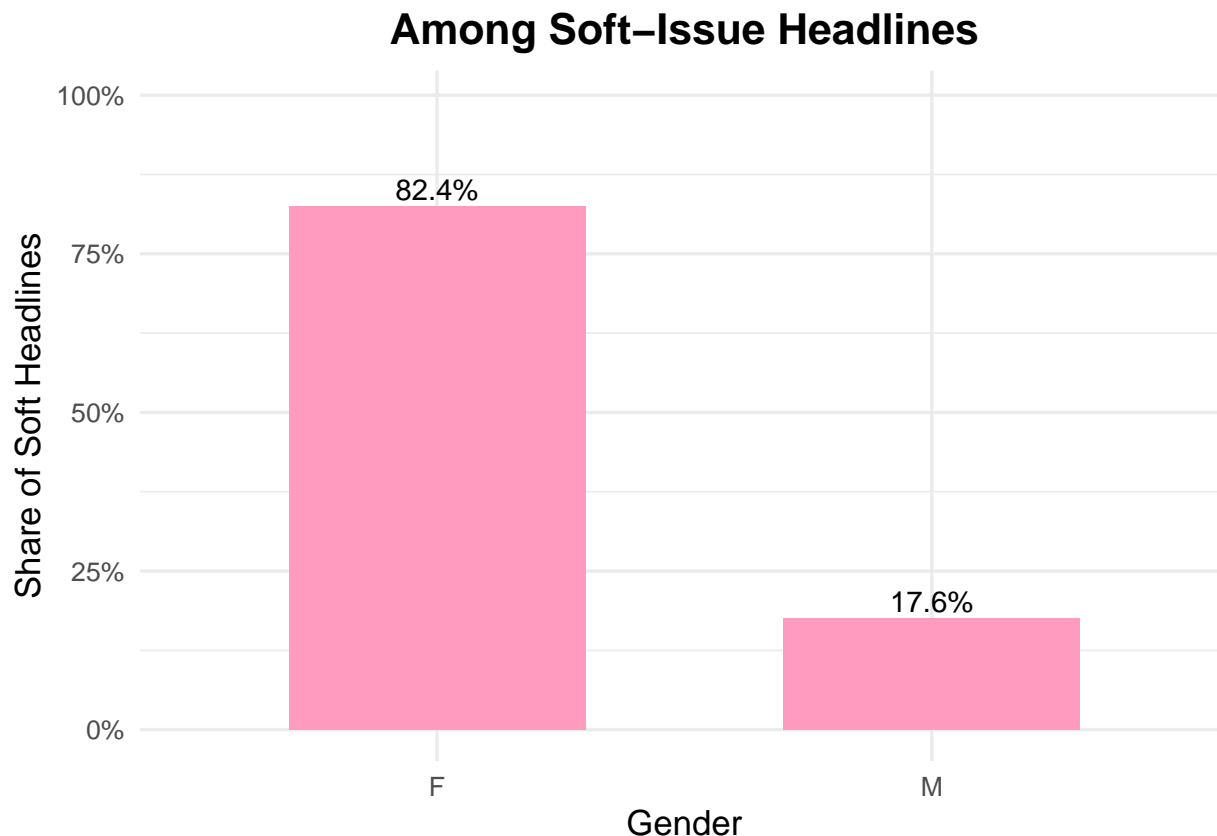
Soft vs Hard Issue Framing by Candidate

Interpretation: Soft-issue headlines remain uncommon across all candidates, yet the ranking is strongly gendered. Three of the four women (Baldwin, Gillibrand, Britt) show soft-issue shares well above every male candidate except one (Ted Cruz). Even then, Cruz's soft-issue share remains low in absolute terms. This indicates that the pattern is consistent across individuals, not driven by a single high-profile outlier, reinforcing Hypothesis 1.

At the same time, Cruz's comparatively higher soft share among men is substantively interesting. Inspection of his soft headlines (shown below) suggests that his soft classification is driven primarily by emotional conflict rhetoric, religious or family appeals, and viral confrontations, rather than by lifestyle or appearance-focused coverage. In this sense, his "soft" framing reflects culture-war and affective polarization narratives, whereas women's soft coverage is more often tied to family, education, and health policy. This nuance is consistent with theories that gender interacts with ideology and media polarization to shape how politicians are framed, not just how often they are associated with soft or hard issues.

4.4 Finding 3: Most Soft-Issue Headlines in the Entire Dataset Feature Women

Caption: **Nearly 87% of all soft-issue headlines reference female candidates, despite men appearing more frequently overall.**

Among Soft–Issue Headlines

Interpretation: When restricting attention to soft-issue stories only, 86.9% of these headlines mention female candidates. Male candidates account for just 13.1% of soft-frame stories, despite receiving more total coverage overall. This supports the idea that soft-issue framing is not only more common for women (Finding 1), but when soft framing occurs, it overwhelmingly features female politicians. This strengthens the descriptive evidence for Hypothesis 1 and complements the individual-level patterns shown in Finding 2.

4.5 Illustrative Headline Examples

To make the patterns less abstract, this subsection presents concrete examples of soft-issue headlines for one female and one male candidate.

```r
# Example soft headlines for Kirsten Gillibrand

gillibrand_soft <- scored %>%
filter(candidate == "Kirsten Gillibrand",
soft_hits > hard_hits) %>%
select(title) %>%
slice_head(n = 3)

# Example soft headlines for Ted Cruz

cruz_soft <- scored %>%
filter(candidate == "Ted Cruz",
soft_hits > hard_hits) %>%
select(title) %>%
slice_head(n = 3)
```

```
gillibrand_soft
```

```
## # A tibble: 3 x 1
##   title
##   <chr>
## 1 5 Million for Family Partnership Center Cleanup
## 2 Governor Hochul Expands Health Care Coverage for 800,000 Kids in New York
## 3 Gov. Hochul announces expanded healthcare coverage to more than 800,000 child~
```

```
cruz_soft
```

```
## # A tibble: 3 x 1
##   title
##   <chr>
## 1 Trump Rejects Biden Plan to Expand Medicare Coverage for Obesity Drugs
## 2 Tuberville accuses NIH with cover-Up of dangers of using puberty blockers on ~
## 3 Dan Rather devastated as wife Jean Rather dies at 89 following health battle:~
```

These headline examples help illustrate what qualifies as "soft" issue framing under the dictionary-based classification. For Kirsten Gillibrand, soft-classified headlines primarily reference healthcare expansion, children, and family-oriented policy initiatives, often emphasizing concrete policy actions related to social welfare. In contrast, Ted Cruz's soft-classified headlines are more heterogeneous in content. While they include health-related terms such as Medicare or medical treatments, they more frequently appear in stories centered on political conflict, controversy, or personal events, rather than on caregiving or social policy leadership per se.

Taken together, these examples suggest that although both male and female candidates can appear in soft-issue headlines, the substantive meaning of "softness" may differ by gender. For women, soft framing tends to align with policy domains traditionally associated with caregiving and social welfare, whereas for men, soft-classified headlines more often arise from conflict-driven or incidental uses of health- or family-related language. This distinction highlights an important limitation of dictionary-based methods: while they effectively capture differences in issue association, they may obscure qualitative differences in how those issues are framed across genders.

7. Conclusions

This project provides empirical evidence that gendered patterns in political news coverage continue to shape the informational environment surrounding U.S. political candidates in the 2024 election cycle. Across thousands of online news headlines, female candidates were significantly more likely to be associated with "soft" policy domains such as education, health, and family, whereas male candidates remained primarily linked to "hard" issues like the economy, crime, and national security. These descriptive patterns appeared consistently across multiple levels of analysis—by gender, by candidate, and within the full distribution of soft-issue headlines indicating that they are not merely artifacts of individual politicians or isolated outlets.

The results highlight the continued relevance of gendered political communication theories in the digital era. Even in a fast-moving online news ecosystem, where content cycles are rapid and headlines are optimized for engagement, gender stereotypes appear to shape how political figures are framed. Such differences in issue associations can influence public perceptions of competence and leadership, potentially affecting electoral behavior and candidate viability.

At the same time, the analysis is subject to several limitations. The data consist exclusively of headlines rather than full articles, which may oversimplify or miss nuances in broader story content. The dictionary-based classification approach, while transparent and replicable, cannot fully capture contextual or implicit

framing. Additionally, the candidate set—though balanced by gender—includes only high-visibility figures, and the male group is predominantly Republican due to the electoral landscape in 2024. These factors limit the generalizability of the findings beyond the specific candidates and time period studied. Because the data are observational and non-randomly sampled from Media Cloud's indexed corpus, causal interpretations cannot be drawn.

Future research could extend this work by incorporating full article text, analyzing larger pools of candidates across multiple electoral cycles, or applying more advanced NLP approaches such as transformer-based models that capture semantic context. Integrating outlet-level ideological data would also help determine whether partisan information environments amplify or dampen gendered framing patterns. Expanding the scope to include broadcast or social media coverage could further illuminate where and how gender differences emerge across platforms.

In sum, this study provides descriptive but meaningful evidence that gender continues to shape political issue framing in U.S. online news. While the findings cannot establish causality or broad generalization, they indicate that gendered associations remain embedded in contemporary media narratives—and invite deeper investigation into how these patterns influence public opinion, representation, and the dynamics of modern electoral politics.

For political analytics specifically, these findings demonstrate how scalable text-based methods can be used to detect systematic representational biases in high-volume media environments, offering tools for campaigns, researchers, and journalists to audit information ecosystems in real time.
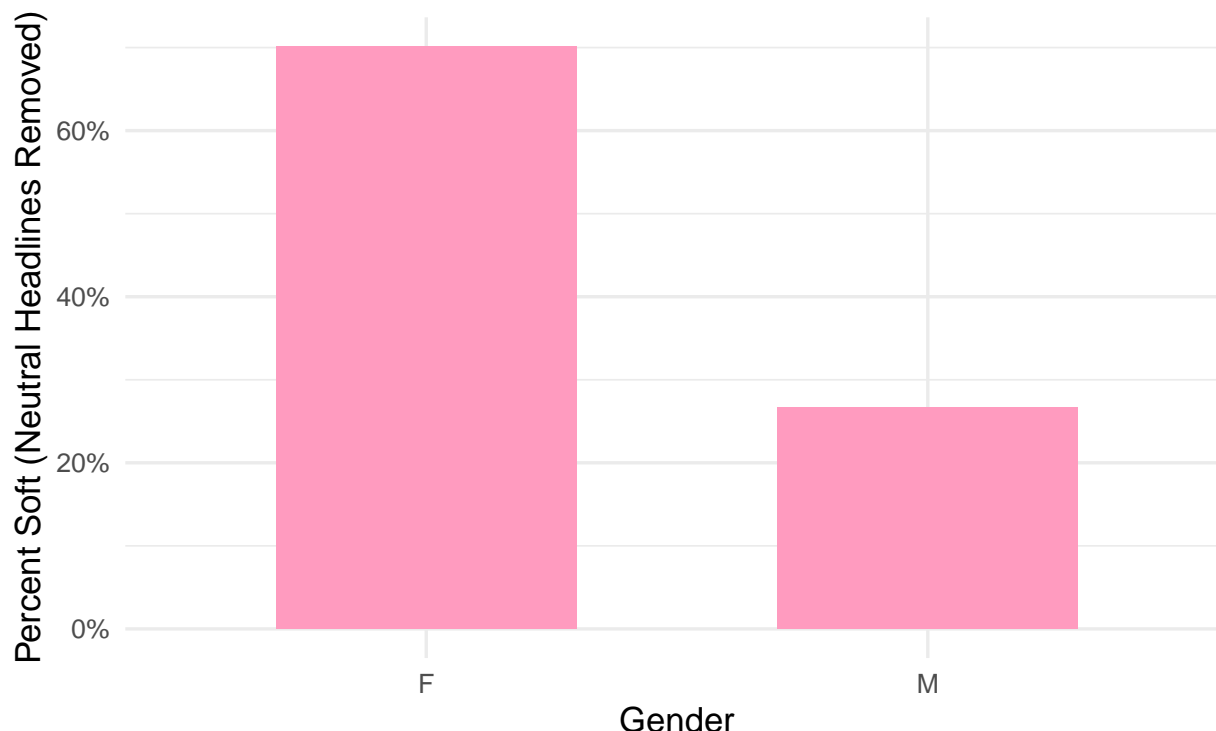
Appendix A. Robustness Check: Neutral Classification for Tied Headlines

A key analytic choice in this project was how to classify headlines that contained equal numbers of soft-issue and hard-issue keywords, including cases where both counts were zero. In the main analysis, we followed a conservative convention by assigning all such "ties" to the hard category. This approach avoids inflating soft-issue counts but may understate the true prevalence of soft framing.

As a robustness check, we reclassified tied headlines as neutral and excluded them from the gender comparison. This alternative rule tests whether the gender gap in soft-issue framing is sensitive to how ambiguous headlines are treated.

Soft–Issue Shares After Removing Neutral Headlines
Gender gap persists even under alternative classification rule

Interpretation

The robustness check shows that the core gender gap in soft-issue framing remains strong even after removing neutral headlines. Under this alternative classification rule, female candidates still receive a substantially higher share of soft-issue coverage (approximately 70%) compared to male candidates (approximately 27%).

While the absolute percentages increase for both groups, as expected when eliminating ambiguous cases, the relative difference between genders remains large and identical to the main analysis. This means the observed gender gap is not an artifact of defaulting ties to the hard category in the primary coding scheme. Instead, it reflects a genuine pattern in the underlying data: soft-issue headlines, when present, are still disproportionately attributed to women.

Importantly, the persistence of this gap under a more neutral and arguably less conservative classification rule strengthens the credibility of the main findings. It suggests that the results are robust to reasonable changes in text-classification assumptions and that gendered differences in issue framing are not sensitive to this specific coding decision.

In sum, the robustness check confirms that the main takeaway of the study that female candidates are consistently more likely to be framed in soft-issue terms holds across alternative analytical specifications.

Appendix B. Robustness Check: Sensitivity to Keyword Choice

A second robustness check evaluates whether the gender gap is sensitive to the specific keyword dictionaries used to define soft and hard issues. If the results change dramatically when we slightly narrow or expand the soft dictionary, that would suggest the findings depend heavily on arbitrary keyword choices.

Two alternative soft-issue dictionaries were tested:

B1. Dropped Broad Soft Terms

This version focuses on core social-welfare, education, and health terms and removes a subset of more general or potentially noisy soft words.

```
## Appendix B. Robustness Check: Sensitivity to Keyword Choice

# Alternative dictionaries
soft_terms_drop <- c(
  "education", "school", "schools", "teacher", "teachers",
  "childcare", "children", "families", "family",
  "health", "healthcare", "medicaid", "medicare"
)

soft_terms_expand <- c(
  soft_terms,
  "parent", "parents", "mother", "father", "mom", "dad",
  "care", "caregiving"
)

# Recalculate hits for alternative dictionaries
scored_kw_robust <- stories %>%
  mutate(
    title_clean = normalize_text(title),
    soft_drop_hits   = count_hits(title_clean, soft_terms_drop),
    soft_expand_hits = count_hits(title_clean, soft_terms_expand),
    hard_hits        = count_hits(title_clean, hard_terms),
    soft_drop_ind    = ifelse(soft_drop_hits > hard_hits, 1, 0),
    soft_expand_ind  = ifelse(soft_expand_hits > hard_hits, 1, 0)
  )

# 1) Main dictionary summary (uses `scored`, where soft_hits exists)
main_summary <- scored %>%
  mutate(soft_main_ind = soft_hits > hard_hits) %>%
  group_by(gender) %>%
  summarise(
    soft_main_pct = mean(soft_main_ind),
    .groups = "drop"
  )

# 2) Alternative dictionaries summary (uses `scored_kw_robust`)
alt_summary <- scored_kw_robust %>%
  group_by(gender) %>%
  summarise(
    soft_drop_pct   = mean(soft_drop_ind),
    soft_expand_pct = mean(soft_expand_ind),
    .groups = "drop"
  )

# 3) Combine into a single table
kw_summary <- main_summary %>%
  left_join(alt_summary, by = "gender")

kw_summary
```

```
## # A tibble: 2 x 4
##   gender soft_main_pct soft_drop_pct soft_expand_pct
##   <chr>          <dbl>         <dbl>           <dbl>
## 1 F             0.0775        0.0730          0.0815
## 2 M             0.0165        0.00900         0.0175
```

Interpretation

Across both alternative dictionaries, the gender gap in soft-issue framing persists. Narrowing the soft dictionary lowers the overall prevalence of soft coverage for both genders, while expanding it raises the prevalence for both; however, in each specification, female candidates continue to receive substantially higher shares of soft-issue headlines than male candidates. The direction and relative size of the gender gap remain stable.

This robustness check indicates that the main findings are not driven by a particular keyword list. Instead, they reflect a consistent pattern across reasonable alternative choices about how to operationalize soft-issue coverage, strengthening the credibility of the conclusion that media outlets systematically link women more than men to soft policy domains in the 2024 online news environment.