First Visualization & Analysis

2025-11-22

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

Media framing influences perceptions of candidate competence, strength, and electability. If women are systematically linked to soft issues while men dominate hard-issue coverage, media narratives may reinforce gendered expectations of political leadership. Understanding these patterns is important for analyzing bias in political communication and voter information environments.

Hypotheses

- **Gendered Issue Association**: Female politicians will appear more often in articles containing soft-issue terms.
- **Hard Issue Gap**: Male politicians will appear more often in articles containing hard-issue terms.
- **Outlet Moderation**: Right-leaning news outlets will show larger gender gaps in issue associations than centrist or left-leaning outlets.

Data Description

The data comes from Media Cloud, an open-source, nonprofit media research platform developed by the Massachusetts Institute of Technology (MIT) and Northeastern University. Media Cloud maintains a large searchable archive of online news stories published by mainstream, local, and digital-native outlets across the United States.

The dataset used in this project is a cross-sectional, observational collection of online news story metadata, obtained through the Media Cloud Search API. The data were collected programmatically by querying the search/story-list endpoint for stories mentioning each of eight prominent U.S. political candidates during the 2024 election cycle.

The target population for this dataset is the universe of U.S. online news stories that mention any of the eight selected political candidates. Because Media Cloud aggregates content from thousands of online news sources, the dataset reflects broad coverage patterns across the U.S. media ecosystem rather than any specific subgroup of voters or individuals.

The time span of the data collection covers January 1, 2024 through November 27, 2024, corresponding to the bulk of the 2024 election year.

The unit of analysis is the individual news story, as indexed in Media Cloud's database. Each row represents one story mentioning one of the target candidates.

The sample size consists of all stories returned by the Media Cloud API for the eight analyzed candidates. Because the platform enforces rate limits and returns only metadata rather than full text, the total number of stories varies by candidate, but the combined dataset contains N = 5000 stories.

The dataset contains the following relevant variables:

- id — unique Media Cloud story identifier
- indexed_date — date when the story was indexed by Media Cloud
- language — story language (e.g., "en" for English)
- media_name — news outlet name (e.g., "The New York Times")
- media_url — base URL of the outlet
- publish_date — publication date of the story
- title — story headline
- url — direct URL to the story
- candidate — which candidate's query returned the story
- party — candidate's political party (Democrat/Republican)
- state — candidate's state
- gender — candidate gender

These variables provide enough information to examine coverage patterns, including which candidates are discussed most frequently, which outlets cover each candidate, and the temporal distribution of political news coverage during the 2024 election cycle.

No survey weights are included, and the dataset does not represent a sample of individuals or voters. Instead, it represents an observational census of online news coverage for a specific set of queries (candidate names in quotes) within a defined time window.

Because Media Cloud provides metadata only (not full article text), the dataset is used primarily to measure coverage volume, outlet characteristics, and story-level sampling for subsequent issue-framing analysis.

Preliminary Visualization & Analysis

To construct a preliminary measure of issue framing, We apply a hand-coded dictionary of "soft" (education, health, family, welfare) and "hard" (economy, defense, crime, foreign policy) terms to the headline text of each article. For each story, I count the number of soft-related and hard-related keywords that appear in the title, and classify the story as "soft-framed" when soft terms are more frequent than hard terms, and "hard-framed" otherwise. This approach is conservative and headline-based only; it does not use full article text.

```
soft_terms <- c(
  "education", "school", "schools", "teacher", "teachers",
  "childcare", "children", "families", "family",
  "health", "healthcare", "medicaid", "medicare",
  "abortion", "reproductive", "maternity",
  "welfare", "food assistance", "snap"
)

hard_terms <- c(
  "economy", "economic", "inflation", "jobs", "tax", "taxes",
  "budget", "deficit",
  "defense", "military", "border", "immigration",
  "crime", "policing", "police",
  "terrorism", "foreign policy", "ukraine", "israel", "china"
)
```

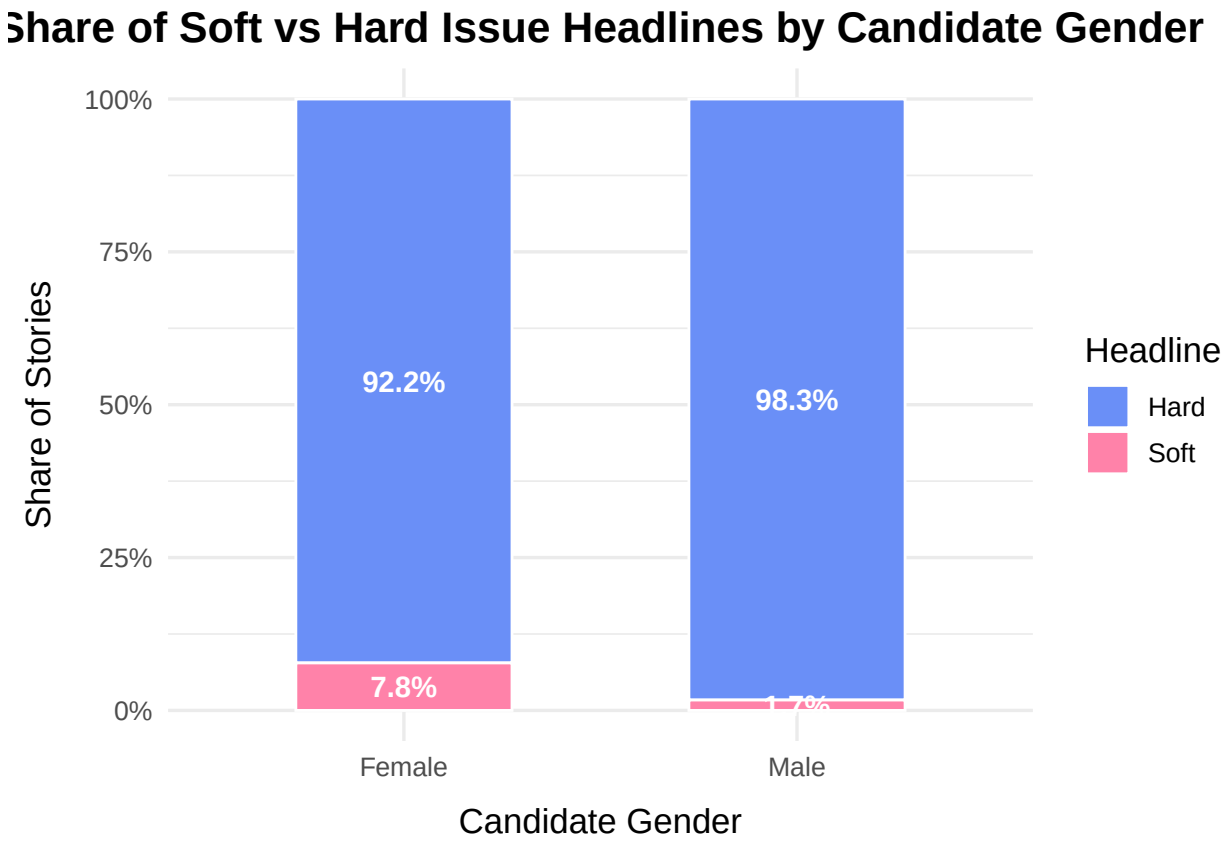
We'll treat title as the "text" field and count how many times words from each dictionary appear.

```r
normalize_text <- function(x) {
  x %>%
    str_replace_all("\\s+", " ") %>%
    str_squish() %>%
    str_to_lower()
}

count_hits <- function(text, terms) {
  pattern <- str_c("\\b(", str_c(terms, collapse = "|"), ")\\b")
  str_count(text, regex(pattern, ignore_case = TRUE))
}

scored <- stories %>%
  mutate(
    title_clean = normalize_text(title),
    soft_hits   = count_hits(title_clean, soft_terms),
    hard_hits   = count_hits(title_clean, hard_terms)
  )
```

Plot

```r
viz_data <- scored %>%
  mutate(frame_simple = ifelse(soft_hits > hard_hits, "Soft", "Hard")) %>%
  group_by(gender, frame_simple) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(gender) %>%
  mutate(
    pct = n / sum(n),
    pct_label = percent(pct, accuracy = 0.1),
    gender_label = ifelse(gender == "F", "Female", "Male")
  )

ggplot(viz_data, aes(x = gender_label, y = pct, fill = frame_simple)) +

  geom_col(width = 0.55, color = "white") +
  geom_text(
    aes(label = pct_label),
    position = position_stack(vjust = 0.5),
    color = "white",
    size = 4,
    fontface = "bold"
  ) +
  scale_fill_manual(
    values = c(
      "Soft" = "#FF82A9",   # pink
      "Hard" = "#6A8FF7"    # blue
    )
  ) +

  scale_y_continuous(labels = percent_format()) +
```

```r
labs(
  title = "Share of Soft vs Hard Issue Headlines by Candidate Gender",
  x = "Candidate Gender",
  y = "Share of Stories",
  fill = "Headline"
) +

theme_minimal(base_size = 13) +
theme(
  plot.title = element_text(face = "bold", hjust = 0.5),
  legend.position = "right",
  axis.title.x = element_text(margin = margin(t = 10)),
  axis.title.y = element_text(margin = margin(r = 10))
)
```



Interpretation:

Next Steps

Models

- Estimate a logistic regression predicting whether a headline is soft-framed (1 = soft, 0 = hard) from candidate gender, controlling for party and state.
- Test an interaction between candidate gender × party to assess whether gendered issue framing varies by partisan context.

- Explore subgroup effects by examining models separately for female-only and male-only candidates to check consistency.

Variables

- Create final binary outcome variable: soft_frame = 1 if soft_hits > hard_hits else 0.
- Add a party indicator (Democrat vs Republican) as a control.
- Add state fixed effects if needed to account for regional variation in coverage.

Visualizations

- **Coefficient plot** for logistic regression results showing the effect of candidate gender on probability of soft framing.
- **Predicted probabilities plot**, comparing:
  - Male vs Female candidates
  - Democratic vs Republican candidates
- **Subgroup bar chart** showing soft framing shares by candidate

Robustness Checks

- Repeat analysis using a stricter dictionary, removing ambiguous words to test stability of soft/hard classification.
- Check results using headline-only vs. title+URL text (e.g., including URL slug keywords).
- Re-run models excluding outlier candidates (those with unusually high or low headline volumes).