COMP370 Final Project

Daria Vorsina, Karl Wehbe, Monica Li

Group 8

1 Abstract

The 2025 New York City mayoral election was historic, with Zohran Mamdani becoming the city's first Muslim and South Asian mayor-elect. This study analyzes how North American news media covered Mamdani during the election period using a dataset of 810 articles from over 50 U.S. and Canadian outlets. Articles were annotated for sentiment, topic category, and political orientation of the source.

Our findings reveal substantial polarization in coverage: right-leaning outlets produced slightly more articles than left-leaning ones, negative framing appeared nearly twice as often as positive framing, and media attention focused heavily on campaign dynamics and controversies rather than civil policy or social issues. These results highlight broader patterns of media bias and political polarization surrounding progressive candidates.

2 Introduction

The 2025 New York City mayoral election represented a significant moment in American politics, with Zohran Mamdani, a 34-year-old democratic socialist, emerging as the first Muslim and first South Asian mayor-elect in the city's history. His campaign emphasized policies such as rent freezes, police reform, and expanded social services, generating substantial media attention and political debate. This project examines how North American news media covered Mamdani, analyzing patterns in sentiment, topic focus, and framing across outlets with different political orientations.

Our analysis draws on a dataset of 810 articles collected from more than fifty news sources across the United States and Canada. Articles were systematically annotated for sentiment (positive, negative, neutral), topic categories (Campaign, Governance, Controversy, Endorsements, Civil Policy, Social Cause), and source political orientation (left-leaning, center, right-leaning). This dataset enables quantitative analysis of media framing, coverage patterns, and potential bias across the political spectrum.

This report details our methodology for data collection and analysis, examines coverage patterns across political orientations, and discusses the implications for understanding media framing of candidates in contemporary American politics.

3 Data

Articles were collected through multiple news APIs to ensure comprehensive coverage of Zohran Mamdani during and after the 2025 New York City mayoral election period. Data collection utilized four primary sources: Event Registry API, MediaStack, NewsAPI.org, and TheNewsAPI. This multi-platform approach was a key design decision to reduce the unique bias of any single data source, making sure our dataset wasn't distorted by one provider's rules for indexing, regional coverage, or source preferences.

3.1 Data Collection and Design Decisions

To maximize article diversity, we implemented a day-by-day fetching strategy requesting 20-40 articles per day. This strategy was a design decision made to ensure even temporal distribution across the election period, prevent recency bias, and capture coverage evolution from campaign launch through election day. Initial filtering was applied at the API level to ensure relevance and quality, targeting "Zohran Mamdani" with exact phrase matching. Articles were fetched based on relevance scores (based on keyword matching strength) and filtered by source popularity to prioritize coverage from established news outlets over low-traffic or obscure sources.

3.2 Data Processing and Filtering

After collection, API responses were normalized into a standardized schema (title, description, URL, news source). The dataset underwent several cleaning and filtering steps:

- Deduplication: Articles were deduplicated using a two-stage process: normalized URL followed by normalized title (to catch republished content).
- Relevance Filtering: To ensure the core focus of the dataset, articles were subjected to a two-stage relevance check. Initially, we filtered out peripheral mentions by requiring the keywords **"Mamdani"** or **"Zohran"** to appear in the title, description, and content fields. Subsequently, during the manual annotation process, any article deemed irrelevant to the study's scope (i.e., not primarily focused on the candidate or his political activities) was removed by the annotator.

- Schema Normalization: Source name variations (e.g., “CNN International” → “CNN”) were consolidated into a single standardized name to fix inconsistent naming across different APIs and ensure consistency for accurate cross-source comparison.

- Manual annotation was performed on all articles for sentiment and topic categories, with political orientation automatically assigned based on source classification.

This structured dataset enables comprehensive analysis of media framing, sentiment patterns, and coverage differences across the political spectrum during the mayoral campaign and transition period.

3.3 Final Dataset Overview

The final dataset consists of 810 articles from over 50 unique news sources from different political orientations. Key characteristics are outlined below.

News Source Distribution

Table 1 shows the distribution of articles across news sources. The New York Post contributed the most articles, followed by Fox News, Yahoo, CBS News, and The New York Times. The remaining coverage was distributed across 50+ additional sources, with most contributing fewer than 10 articles each.

Table 1: Top 10 News Sources by Article Count

Source	Article Count
New York Post	218
Fox News	167
Yahoo	125
CBS News	59
The New York Times	56

Political Orientation

Figure 1 illustrates the distribution of articles based on the source’s political orientation. Right-leaning sources contributed the most articles (409, 50.5%), followed by left-leaning sources (345, 42.6%), with centrist sources representing a smaller portion (56, 6.9%).

4 Methods

4.1 Question Formulation

Question formulation was in the format of an open floor discussion. The team went through an iterative process: first each teammate individually came up with a few questions individually, then as a group, we discussed question individually generated questions. During our discussion we looked for key terminology, what information must be included in the question, to best suite the type of analysis that we want to conduct.

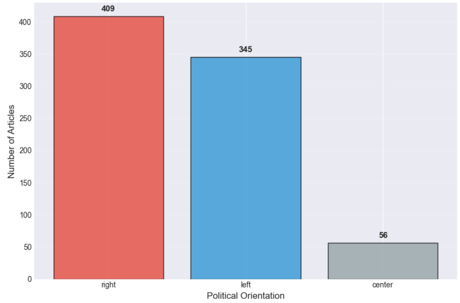

Figure 1: Article Count by Political Orientation

4.2 Annotation

Annotation was performed on all 810 articles in the dataset across three dimensions: topic categories, sentiment, and political orientation of sources.

4.2.1 Typology Development and Optimization

The process began with open coding to develop a comprehensive typology. Each team member performed open coding on approximately 50 articles, developing initial category proposals through an iterative refinement process:

- Initial brainstorming and assignment of provisional category labels (approx. 10 entries).

- Iterative refinement and cross-validation of labels across subsequent entry sets.

The result was that each team member produced their own codebook. The codebooks were analyzed for similarities and differences, during this group discussion we found substantial overlap in our individual interpretations. In a standard annotation setup there should be at least 3 coders, since this is a minimum standard to create consensus. When there is majority coder agreement in it is a good sign of a good typology. In our case, common typology design between teammates was an early sign that we are on a path towards a comprehensive and well defined final typology. Through collective discussion, the team refined definitions, eliminating obscure categories like "Other". We merged overlapping categories based on similarity of definitions or if we found frequent occurrence of edge cases which had enough contextual overlap to bridge two categories. For instance, we originally had a category "Hate Crime", which was for news articles that described cases of crime motivated by hate against a specific group, crimes of "Islamophobia" or "Anti-Semitism" for instance. We then realized that we could dissolve this category between "Controversy" and "Social Cause". If the article texts directly (name) or indirectly (refer to him by his title) featured Zohran Mamdani these would be categorized as "Controversy", otherwise we placed them into the "Social Cause" category. We created more granular distinctions, for example, splitting the initial "Campaign" category into "Campaign" and "Governance", because we realized that a portion of articles were from the time period of Zohran Mamdani's campaign, and others were from after the elections. Our final typology defines six categories: Endorsements, Civil Policy, Controversy, Campaign, Social Cause, and Governance.

4.2.2 Systematic Coding and Dimensions

All 810 articles were manually annotated by the team using the finalized typology and the codebook.

- Topic Categorization: Articles were assigned to a single primary category based on their dominant focus.
- Sentiment Annotation: Each article was classified as Positive (favorable framing), Negative (unfavorable framing), or Neutral (factual reporting without clear bias).
- Political Orientation: Articles were assigned based on source classification: outlets were categorized as left-leaning, center, or right-leaning based on established media bias ratings and editorial positioning.

The application of the typology was optimized using a concise key terms dictionary, which was developed for each category by analyzing occurrences (in the 50 open coded entries) of terminology with strong thematic context. The key terms dictionary, assisted in guiding the annotation process as a supplemental reference to the typology, proving to be an efficient helper tool. This method was used cautiously, keeping in mind the presence of outliers - articles that contained the key term identified for the category, but with more precise analysis ended up being more contextually fit for a different one. This annotation method helped speed up the process and would often be useful in ensuring additional confidence in the annotation.

Similar to the category key terms dictionaries, we used the same approach for sentiment analysis. Generating a dictionary with strong negative or positive connotation.

Manual annotation was not used for political orientation since this is a property of the source rather than individual articles." (e.g., *The New York Post* is consistently right-leaning), not of the specific article's content. This automated, static assignment ensures classification consistency across the entire dataset.

This comprehensive annotation framework enables quantitative analysis of media framing, topic distribution, and sentiment patterns while accounting for source political orientation which is a critical step for understanding how different media outlets covered Zohran Mamdani.

4.3 Topic Characterization Methods

We employed two complementary approaches to characterize topics:

4.3.1 TF-IDF Analysis

Using the full corpus of annotated articles, we computed TF-IDF scores to identify the 10 most representative terms for each category. We used scikit-learn's TfidfVectorizer, configuring it to ignore common English stop words. The vectorizer computes term frequency (TF) as the count of a word within articles in each category and derives inverse document frequency (IDF) from the number of articles in which the word appears across the collection. The vectorizer multiplies these values to produce the TF-IDF score for each word, from which we extracted the top ten per category.

4.3.2 LLM Summarization

Category-specific batching

Before embarking on the LLM categorical summaries, the dataset required some pre-processing due to the practical constraint on token limits for the LLM (Gemini 2.5), and the objective to create better quality prompts to receive better quality outputs.

The practical motivation for the design choice to use sampled data for the LLM was motivated by practical limitations on the volume of text that could be sent via API call to the model used, Gemini 2.5. Each sample file consisted of up to 100 randomly chosen articles per category. This limit was set in consideration of the token limits for the LLM. Difference in sample sizes did not materially influence outcome, since categories were observed and analyzed by the LLM distinctly.

Preprocessing consisted of creating one CSV per category (randomly sampling up to 100 rows per category from the total data set of all 810 collected articles). Aside from respecting token limits, the design choice ensured that the LLM processed coherent thematic groups, which effectively decreased noise in the data; also preventing accidental cross-category contamination, which likely would have happened if the entire dataset was fed at once. Similarly, categories became more distinguishable, contextually independent, resulting in a more accurate summarization by the LLM, which performs best with data that contains less ambiguity and is comprehensively pre-defined. Paired with a prompt that is precise in its instruction, the outcome is reduced contextual drifting, thematic coherence and robust insights.

For each category the script prompted the LLM with a curated message and the category data file separately. As a result, input is wide enough to capture category (each category was a batch) variance but small enough to avoid token limits.

The structure of the prompt included the task statement, specification of tone, highlight of the importance of using the TF-IDF results and data. The prompt also specifies that the LLM be truthful.

- Task: Produce a concise representative contextual summary (about 100-200 words) that explains what the articles in this category are broadly about, the main themes, likely framing and who or what is discussed, and any important context a reader should know.
- "Write in neutral academic tone. Start with a 1-2 sentence summary, then 3-4 short bullet points with notable details."
- "Important: Base your summary only on the examples and the top-10 words above. Do not invent quotes or facts not inferable from the snippets."
- Output format:
  "SUMMARY: <sentence summary>"
  "DETAILS:" "- ...")

Global TF-IDF for LLM Summaries

The design choice to modularize the dataset for the LLM contrasts with the approach of calculating the TF-IDF. Idf values were collected based on all 810 articles rather than category-specific subsets. The design choice of global TF-IDF and categorically-organized data as input to the model, guards against overfitting small categories and anchors the model to empirical evidence from the dataset, reducing hallucinations. Global TF-IDF forces the summary to highlight the true dominant vocabulary journalists used, making the LLM a better tool of interpretation.

Quality of LLM Summaries

The LLM-generated summaries are insightful in bridging the quantitative and qualitative components of this project. The TF-IDF analysis captures accurately the statistics of vocabulary across categories, meanwhile the LLM summaries relies on TF-IDF and textual context to generate descriptive narratives. Summaries are a high-level interpretation of each category's topic coverage. Their true value lies in accessibility, this result can be interpreted by technical and non-technical audiences. In the context of a well formulated prompt, high quality summaries are a valuable tool of communication. The short thematic summary and details rendered by the model closely adhered to the TF-IDF 10 word summaries. The summaries contained recurring actors, discussion topics and references to important events, which would have been difficult and time consuming to formulate manually. The summaries reinforced the validity of the typology by showing how categories manifest in the storytelling of the articles. These design choices:

- Scale manually annotated data into interpretable, category-level datasets
- Produce standardized summaries, allowing for effective cross-category comparison.
- Category-based splitting decreases data noise increasing the accuracy and robustness of summaries
- TF-IDF grounding allows the LLM output tio be realistic and data-driven
- Using textual openings (title + first sentence) reflects real-world media framing practices: these components contain the highest density of framing cues.

5 Results

5.1 Topic Distribution

Table 2 presents the distribution of articles across the six categories defined in the typology, broken down by source political orientation. Overall, coverage was dominated by procedural and competitive topics, with the Campaign (307 articles, 37.9%) and Governance (232 articles, 28.6%) categories accounting for 66.5% of the total dataset. The remaining four categories collectively accounted for 33.5% of coverage. By source, Right-leaning outlets showed the highest volume of coverage for the Campaign (175 articles), Controversy (74 articles), and Civil Policy (27 articles) categories. Conversely, Left-leaning outlets showed the highest volume of coverage for Governance (142 articles).

Table 2: Distribution of Articles by Category and Political Orientation

Category	Total	Percentage	Left	Center	Right
Campaign	307	37.9%	111	21	175
Governance	232	28.6%	142	12	78
Controversy	117	14.4%	35	8	74
Endorsements	59	7.3%	24	6	29
Civil Policy	50	6.2%	19	4	27
Social Cause	45	5.6%	14	5	26

5.2 Sentiment Distribution

Figure 2 presents the overall distribution of sentiment across all 810 articles. The majority of articles were classified as **Neutral** (387, 47.8%). Among the non-neutral coverage, **Negative** sentiment (270, 33.3%) exceeded **Positive** sentiment (153, 18.9%).

Sentiment By Category

Figure 3 displays the distribution of sentiment proportions within each topic category. Controversy stood out with

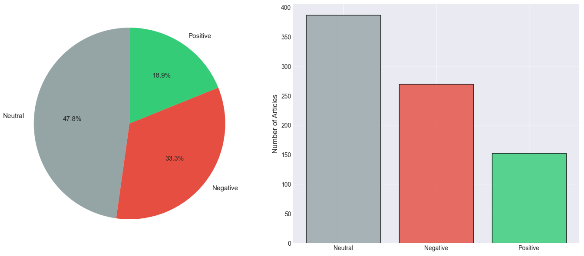

Figure 2: Overall Sentiment Distribution

almost exclusively negative coverage, while Endorsements showed the opposite pattern, dominated by positive sentiment. Civil Policy coverage was polarized between negative and neutral tones, with little positive framing. Campaign and Governance articles were predominantly neutral, though Governance saw a modest positive lean. Social Cause coverage tilted negative but retained a mix of all three sentiment types.

In terms of absolute counts, Controversy contributed the most negative articles followed by Campaign and Governance. Campaign and Governance dominated neutral coverage, while Governance, Endorsements, and Campaign led in positive sentiment.

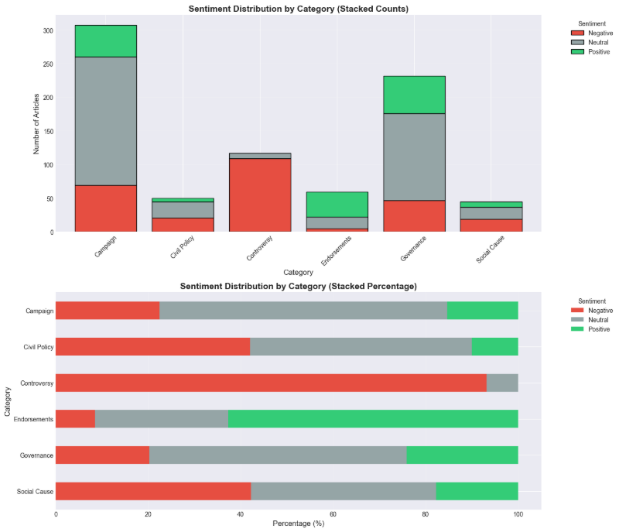

Figure 3: Sentiment Distribution by Category

Sentiment By Political Orientation

Figure 4 displays sentiment distribution separated by source political orientation.

- **Right-leaning sources (n=409):** 49.1% Negative, 41.1% Neutral, 9.8% Positive.
- **Left-leaning sources (n=345):** 13.6% Negative, 58.3% Neutral, 28.1% Positive.
- **Center sources (n=56):** 37.5% Negative, 32.1% Neutral, 30.4% Positive.

Right-leaning sources produced approximately 3.6 times more negative articles than left-leaning sources. Conversely, left-leaning sources produced approximately 2.8 times more positive articles than right-leaning sources.

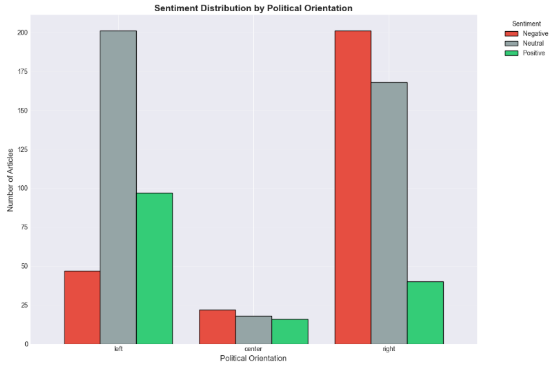

Figure 4: Sentiment by Political Orientation

5.3 Top Terms by TF-IDF

Table 3 lists the top 10 terms with the highest TF-IDF scores per category.

The most frequent and distinguishing terms within Campaign coverage include "Cuomo" and "Andrew" (referencing opponent Andrew Cuomo), alongside "candidate." Controversy coverage features "socialist" and "Trump." Civil Policy includes substantive terms such as "rent," "program," and "plan." Social Cause coverage shows identity-focused terms like "Muslim," "terror," and "Islamophobia." Endorsements includes "endorsement," "hochul," and "jeffries," referencing political supporters. Governance features terms related to administration and transition, such as "elect," "president," and "meeting."

Table 3: Top 10 Terms by TF-IDF Score per Category

	Controversy	Endorsements	Civil Policy	Campaign	Social Cause	Governance
1	mamdani	mamdani	mamdani	mamdani	mamdani	mamdani
2	zohran	new	zohran	zohran	muslim	trump
3	new	york	plan	mayoral	terror	elect
4	york	endorsement	new	new	zohran	mayor
5	city	hochul	free	cuomo	new	new
6	democratic	zohran	program	york	imam	york
7	trump	jeffries	city	city	islamophobia	city
8	mayoral	mayoral	york	andrew	york	zohran
9	mayor	mayor	mayoral	democratic	city	president
10	socialist	endorses	rent	candidate	mayoral	meeting

5.4 LLM Summaries Output

The LLM summaries successfully created descriptive results for all categories, providing a clear overview of the main themes in each. The summaries closely aligned with the TF-IDF key terms. In the **Controversy** category, they emphasized conflicts and the assignment of blame, whereas categories with more factual or process-focused content, such as **Governance** or **Campaign**, produced neutral, policy-focused summaries. Categories with fewer articles, like **Social Cause** and **Civil Policy**, resulted in shorter and less detailed summaries compared to categories with higher coverage. Overall, using category-specific batches together with global TF-IDF proved effective, generating fast, organized, and generalized summaries of media coverage related to Zohran Mamdani's campaign.

6 Discussion

Our analysis of the dataset reveals distinct patterns in how North American media framed Zohran Mamdani. By synthesizing the topic distribution, sentiment analysis, and lexical features (TF-IDF), we can draw several conclusions about the nature of the coverage.

6.1 Contrasting Editorial Strategies: Left vs. Right

The different strategies used by opposing media outlets show how political orientation shapes both the amount and tone of coverage. Across the full dataset, negative sentiment appeared far more frequently than positive sentiment, and this imbalance was driven primarily by right-leaning outlets which fits with the expectation that media tend to be more critical of candidates who challenge their ideological stance. Since these outlets accounted for over half of all articles, their consistently negative tone had a strong impact on the overall distribution of sentiment in the corpus. This concentration of negative coverage reflects how partisan media tend to respond to ideologically distant candidates by emphasizing criticism, controversy, and skepticism rather than policy substance.

Right-leaning articles relied heavily on frames designed to portray Mamdani as ideologically extreme, with controversy-oriented topics and criticism rather than policy, and when policy is mentioned it's typically framed negatively. In contrast, left-leaning outlets contributed fewer negative articles and instead maintained a more neutral and favorable treatment , softening the overall sentiment pattern but not enough to offset the volume produced by the right.

- **Right-Leaning Strategy (Negative Amplification):** Right-leaning outlets produced the largest share of articles (50.5% of the dataset) and emphasized negative sentiment (49.1% negative overall). Their coverage centered on controversy and critique, reflected in the substantial negative sentiment in the **Controversy** and **Civil Policy** categories.

- **Left-Leaning Strategy (Neutralization and Validation):** Left-leaning outlets prioritized neutrality (58.3% neutral overall) while using positive sentiment in more targeted ways. Positive framing appeared mainly through establishment validation, particularly in **Endorsements**, **Governance** and **Campaign**. This reflects an editorial approach that normalizes the candidate through factual reporting, reserving overly positive sentiment for specific instances.

Together, these patterns reveal a key difference in how outlets covered Mamdani: right-leaning sources emphasized ideological opposition and controversy, while left-leaning sources focused on neutral, fact-based reporting. These contrasting approaches reflect broader tendencies in partisan media, though both involve editorial choices that shape how candidates appear to their audiences.

6.2 Dominance of "Horse-Race" Narratives

The data suggests a strong media preference for the competitive aspects of politics over substantive policy analysis. Media coverage overwhelmingly focused on polling results, campaign strategies, and political conflicts rather than Mamdani's proposed policies. As shown in Table 2, the **Campaign** and **Governance** categories combined to form 66.5% of the total dataset, while **Civil Policy** accounted for only 6.2%. This imbalance indicates that media attention was primarily fixed on the mechanics of winning and the transition of power rather than the candidate's governing platform. This interpretation is supported by the TF-IDF results (Table 3), where the Campaign category is defined by terms like "Cuomo," "Andrew," that reference Andrew Cuomo, Mamdani's primary opponent and former Governor of New York, emphasizing the horse-race narrative.

6.3 "Spectacle" Framing

The sentiment data reveals that coverage of Mamdani was largely negative particularly regarding controversy. The **Controversy** category was the third most frequent topic (14.4%) and was almost entirely negative (92.3%) which is expected given the nature of the category. The specific vocabulary associated with this category in the TF-IDF analysis "socialist" indicate that media outlets used provocative labels to boost audience interest. These controversies were not presented as thoughtful debates but rather as sensational spectacles designed to generate attention.

Right-leaning sources were responsible for the majority of Controversy articles (74 vs. 35 from left-leaning sources), establishing a pattern that would manifest across multiple coverage dimensions.

6.4 Critical Framing of Policy and Identity

A key finding is the prevalence of negative sentiment even within topics dedicated to discussing substantive policy and actionable plans. The **Civil Policy** category, despite focusing on "rent," "programs," and "plans," exhibited 48.0% negative sentiment compared to only 8.0% positive. This implies that Mamdani's progressive policy proposals were more frequently met with skepticism or critique than validation in the press. Similarly, the **Social Cause** category contained high-frequency terms such as "terror," "Islamophobia," and "Muslim," paired with around 44% negative sentiment. Media often portrayed Mamdani's identity in a negative light, emphasizing social issues through a lens of threat rather than recognizing positive cultural contributions. This framing shifted the narrative toward a defensive stance, overshadowing constructive aspects of his identity and the broader community.

6.5 Balanced and Positive Reporting: Notable Exceptions

While most data supports trends of negativity and strategic framing, notable exceptions add complexity to our findings. The **Endorsements** category stands out as the only one dominated by Positive sentiment (over 80% positive). This suggests that positive coverage occurs predominantly when established political figures validate one another, rather than through independent journalistic assessment.

The **Governance** category displays a more balanced sentiment distribution between Positive and Negative, with considerable Neutral coverage. When discussions shift to routine political operations and after the election is over, sensationalism diminishes, allowing for more objective reporting approaches.

6.6 LLM Summaries Interpretation

The LLM summaries gave important insights that support the statistical and TF-IDF results.

The differences in summaries show that media framing is influenced not just by how often positive or negative sentiment appears but also by how stories are structured. For example, the **Controversy** summaries' focus on conflict and blame matches the high negative sentiment in that category (92.3% Negative), supporting the idea of "spectacle framing". On the other hand, **Governance** and **Campaign** summaries were neutral and procedural, reflecting the media's focus on horse-race narratives over policy details.

The shorter, less organized summaries for **Social Cause** and **Civil Policy** show a **structural imbalance in coverage**, meaning that media attention tends to favor conflict-driven or election-focused stories, leaving less space for detailed discussion of policies or social issues.

6.7 Implications and Conclusions

This study shows that Zohran Mamdani is covered within a highly divided media environment. Coverage volume, topic selection, and sentiment varied noticeably by political orientation. Right-leaning outlets produced the most coverage overall and framed Mamdani more negatively, particularly in stories emphasizing controversy and campaign conflict. Left-leaning outlets contributed fewer articles but focused more on campaign and governance, offering comparatively positive or neutral framing. Across all outlets, there was relatively little detailed coverage of policy or social issues, as most attention stayed on campaign events and politically charged topics.

Rather than offering in-depth analyses of major social and political issues, media coverage leaned heavily on strategic updates and scandal-driven narratives, creating an environment where audience engagement is driven more by controversy than informed policy discussion.

These patterns observed reflect broader trends in political journalism, where candidate identity, conflict-oriented storytelling, and ideological alignment shape how narratives are constructed and circulated. Understanding these dynamics highlight the importance of examining media environments when studying how progressive candidates are portrayed and how the public talks about them.

The heavy imbalance between campaign-focused coverage and real policy discussion raises questions about how well the media helps people make informed decisions. When coverage prioritizes electoral horse-race narratives and sensational controversies over actual policies, it becomes unclear whether voters make decisions based on limited policy information, other sources, or the strategic stories the media emphasizes.

Mamdani's success despite heavily negative coverage is notable. It may indicate the limited influence of oppositional media or voters growing ability to critically evaluate biased coverage. This analysis is limited to traditional news media and does not capture the increasingly important role of social media platforms and direct candidate-to-voter communication. Future research should examine how progressive candidates use alternative platforms to bypass traditional media gatekeepers and counterbalance negative framing.

7 Group Member Contributions

Daria conducted open coding and annotations, produced the LLM summary analysis, and authored the Methods (typology/open coding/annotation/LLM summaries), Results (LLM summaries) and part of Discussion sections (LLM summaries).

Karl focused on data collection, annotations, visualization development and wrote the Data section and part of the Method, Result and Discussion sections.

Monica performed open coding and annotations, implemented the TF-IDF analysis and distribution visualizations, and wrote the Method and Results sections regarding TF-IDF, Discussion and Group Contributions.