COMP 370 Final Project: Politician in the Media

Tarek Gohar and **Timofey Galyanov** and **Jikael Gagnon**
McGill School of Computer Science
3480 Rue University, Montreal, QC H3A 2A7
tarek.gohar@mail.mcgill.ca
timofey.galyanov@mail.mcgill.ca
jikael.gagnon@mail.mcgill.ca

Introduction

Background and Motivation JD Vance's rise to stardom happened during the 2016 election cycle, during which he released his book *Hillbilly Elegy*. The book details Vance's upbringing, which provided a direct account of the lives of working class white Americans - a critical subject in the 2016 election. At this time, however, Vance was a harsh critic of Trump's policies, going as far as to write that he's "back and forth between thinking Trump is a cynical a–hole like Nixon who wouldn't be that bad (and might even prove useful) or that he's America's Hitler"[4]. Vance would later apologize for these comments in 2021; in the same month, he would announce his Senate campaign and would win a seat by early 2022.

In early 2023, Vance endorsed Donald Trump in the 2024 primaries, and was chosen as Trump's running mate in July of 2024. From the moment of his nomination, Vance's run as VP candidate was immediately shrouded in controversy, due in part to his prior criticism of Trump. Since then, Vance's run as VP nominee has been a jumble of ups and downs. In this light, we have been hired by a media company to understand how Vance is being covered in the media leading up to the 2024 United States presidential election, with a particular focus on the sentiment of the coverage and the topics they focus on.

Key Findings After analyzing over 500 news articles surrounding JD Vance, our key findings were:

- The Vice President-elect entered the final month of the election with an overall negative sentiment, as nearly 70% of the non-neutral coverage portrayed him unfavorably.
- JD Vance was found to be frequently associated with polarizing terms often appearing in media coverage related to his communications and public rhetoric.
- Media tends to cover highly controversial topics like scandals and political attacks.

Data

Data Source

Data for this project was collected via *NewsAPI's* `/v2/everything` endpoint; this API limits results to articles from the past 30 days, and only returns a maximum of 100 articles per request. For each article, the API provides multiple features, including the title, description, and content.

Filtering and Bias Mitigation

Initial Filtering To prioritize the most impactful and widely-read articles, the requests were sorted based on the popularity of the sources and publishers. This approach ensures that the dataset focuses on articles from the most influential and highly-viewed outlets, maximizing the relevance and significance of the collected information. The data was then filtered to only include english articles, to ensure consistency, and removed those that did not specifically contain the keyword "JD Vance" in the description. In the case that a particularly noteworthy event occurred, requests for articles around that event might return a set of articles all focused on that event, since articles are sorted by popularity. To mitigate bias towards any particular event and broaden the scope of our findings, requests were made in non-overlapping three day intervals across the 21 days leading up to the election (October 10th and November 4th 2024, both inclusive).

Post-Collection Filtering Results returned by the API were erroneous in three primary ways:

1. Articles are marked as `[removed]`
2. The description does not contain the required keyword ("JD Vance")
3. The article is completely unrelated (eg. one article warned of an incoming sci-fi apocalypse in 2030)

Filtering of cases 1 and 2 were automated, while case 3 was filtered manually; this phase left a total of 506 articles.

Features

The dataset was annotated both manually and automatically, leading to the following features:

1. `source`: The news source. Eg. "CNN"
2. `author`: The author of the article.
3. `title`: The headline or title of the article.
4. `description`: A description or snippet from the article.

5. `url`: The direct URL to the article.
6. `publishedAt`: The date and time the article was published, in UTC
7. `content`: The unformatted content of the article, where available.
8. `topic`: The topic the coverage focuses on.
9. `negative/neutral/positive`: Whether the topics were positive, negative, or neutral
10. `align`: The political alignment of the news source.

Methods

Manual Annotation

The manual annotation phase had two objectives: first, creating a typology to classify the topics discussed in the articles; second, identifying whether articles spoke of Vance in a positive, negative, or neutral tone. For the typology, a first pass of reading the articles was made over the entire dataset (specifically the titles and descriptions, and sometimes the actual articles themselves as well). In this pass, articles were labeled with one or more primary topics. Before starting the second pass, all primary topics were collected in a set. The topics were partitioned into eight categories to build the typology. Then, a second pass was done, where the types were annotated in a new column by mapping the topics to the types in the typology.

Initially, the first pass was intended to be done only on the first 200 articles. However, after finishing the first 200, it was noticed that overtime, as events during the timeline occurred, the topics in the articles changed drastically. Therefore, any typology created based on the first 200 articles might not be relevant to the rest of the data set. So, the first pass was continued until the end of the document.

Additionally, during the first pass, the second objective was completed by indicating in a separate column the sentiment of the article towards JD Vance.

Automated Annotation

To identify possible sources of bias in the sentiments of each article, we added an additional `align` feature that describes the political alignment of the source of each article.

MediaBias/Fact Check (MBFC) is a database of over 8800 media sources that are scored on a scale from 0 to 10 based on their political bias; scoring is defined based on various qualitative and quantitative metrics including: wording, factual sourcing, story choices, and political affiliation. Based on these scores, news sources are labeled with a bias and ratings for their factuality and credibility. To access this database, we used a third-party API that returns the database as a JSON file [1]. The following is the data returned for CNN:

```
"name": "CNN - Bias and Credibility",
"url": "www.cnn.com/",
"bias": "left",
"factual": "mostly",
"credibility": "high credibility"
```

For each article returned from the data collection phase, the bias was determined by extracting the domain name from the URL and cross-referencing with the database; alignment for articles for which database entries did not exist were marked as "N/A".

TF-IDF Scores

To compute TF-IDF scores, the content of each article was converted to lowercase and then tokenized using `TweetTokenizer` from Python's Natural Language Toolkit (NLTK). This tokenizer was selected over the default tokenizer because it preserves contractions as a single token, rather than splitting it into different tokens:

```
# default tokenizer
word_tokenize("don't cry!")
# output: ['do', "n't", 'cry', '!']

# TweetTokenizer
TweetTokenizer().tokenize("don't cry!")
# output: ["don't", 'cry', '!']
```

More specifically, this allows for distinction between words like "do" and "don't", which can provide context about the sentiment in the article. For example, articles might frequently mention that Vance's economic policies *don't* make sense; this distinction is lost in the case of the default tokenizer. Further, common stop words were filtered out using NLTK's english stop word list. Next, tokens for articles in the same category were concatenated together and counted using Sci-Kit Learn's `CountVectorizer` to compute tf-idf scores.

Results

The analysis identified several important findings about media coverage of the Vice President-elect's sentiment leading into the United States elections and resulted in an eight-topic typology.

Typology Definitions

The following is a list of definitions for the eight topics:

Domestic Policy Defined as covering discussions about internal government policies, social services, and legislative measures affecting the country. Subcategories such as immigration, healthcare, and constitutional rights suggest topics related to governance, social justice, and public services. Commonly seen terms include *projects*, *illegal*, *immigration*, *agenda*, *deport*, and *immigrants*. The topic is seen roughly 10% of the time.

Economic & Financial Issues Defined as encompassing topics related to the economy, market trends, and financial practices. Subcategories such as inflation, personal wealth, and manufacturing policy indicate a focus on economic health, labor issues, and fiscal responsibility. Commonly seen terms include *strike*, *gazette*, *crossing*, *economy*, and *picket*. The topic is seen roughly 5% of the time.

Elections & Political Strategies Defined as involving the analysis, updates, or narratives surrounding elections, campaigns, and political alignments. Subcategories such as campaign rallies, voter trends, and election fraud allegations reflect themes of political engagement, strategic movements, and public opinion. Commonly seen terms include *lost*, *polls*, *senate*, and *McConnell*. An edge case for this type includes "2020 election fraud speculations". This is included under this category as it pertains to the 2020 presidential elections despite it being also related to the type of Scandals & Controversies. Another edge case for this type is "campaign rally", since it is related to Media & Public Engagement. However, it is more accurately related to the elections. The topic is seen most often at 21% of the time.

Foreign Policy & National Security Defined as pertaining to international relations, defense strategies, and global security. Subcategories such as cyber attacks, US-China relations, US-Russia relations and peace efforts underlining a focus on diplomacy, threats, and cooperative international efforts. Commonly seen terms include *chinese*, *hackers*, *targeted*, *phones*, and *cellphones*. The topic is seen roughly 11% of the time.

Media & Public Engagement Defined as relating to depiction of public figures or topics in the media. Subcategories such as news appearances, podcasts, and personal anecdotes indicate themes of communication, narrative shaping, and public relations. Commonly seen terms include *McCain*, *Raddatz*, *Packers*, and *bar*. The topic is seen roughly 13% of the time.

Political Attacks & Accusations Defined as including any content focused on criticisms, accusations, or negative portrayals of individuals or groups in the political sphere. Commonly associated with negative sentiment. Subcategories such as inter-party tensions, intra-party tensions, misinformation, and attacks on political figures highlight a broader pattern of conflict, defamation, or defense in political narratives. Commonly seen terms include *Khosla*, *tapper*, *Wilson*, *garbage*, and *CNN*. The topic is seen roughly 16% of the time.

Scandals & Controversies Defined as covering discussions about scandals, allegations and controversies that have sparked public debate or reactions. It includes offensive remarks, perceived misconduct, and contentious claims related to individuals, groups, or events. Commonly seen terms include *Aaron*, *Kofsky*, *adviser*, *heard*, *Haitian*, and *drugs*. The topic is seen roughly 16% of the time.

Social Issues Defined as capturing societal trends, cultural movements, and demographic concerns. Subcategories such as parenthood, transgender issues, and climate change encompass broader societal challenges, cultural debates, and public health concerns. Commonly seen terms include *normal*, *gay*, *childless*, *women*, *ladies*, and *trans*. The topic is seen roughly 9% of the time.

Typology Results

The typology displayed an even distribution, averaging 63.75 articles per category. *Elections & Political Strategies* emerged as the leading category, accounting for more than 20% of the articles, while *Economic & Financial Issues* ranked the lowest, representing only 4% of the total (see Table 1).

Topic	Proportion
Elections and Political Strategies	0.213
Scandals and Controversies	0.156
Political Attack and Accusations	0.154
Media and Public Engagement	0.130
Foreign Policy and National Security	0.113
Domestic Policy	0.101
Social Issues	0.091
Economic and Financial Issues	0.042

Table 1: Distribution of Topics Across Political Categories

Applying TF-IDF analysis to each topic produced Figure 1, which highlights the ten most significant words for each category. The top scoring words in each category were `chinese`, `Aaron`, `project`, `lost`, `McCain`, `normal`, `Khosia`, and `strike`.

Sentiment Results

After manually annotating the articles, the overall sentiment was determined based on their titles and descriptions. Out of the total articles, 278 were categorized as neutral, 153 as negative, and 75 as positive. Excluding the neutral articles, the analysis revealed that negative articles outnumbered positive ones by a ratio of 2:1, indicating that our collected articles expressed predominantly negative sentiment toward JD Vance leading up to the election. Notably, the high frequency of neutral articles are a result of many short articles that simply state a fact. A representative example is an article titled "Vance tells Rogan he initially thought Trump had been killed in July assassination attempt: 'I was so pissed'", which is simply a series of direct quotes from Vance's interview with no commentary from the author [2].

Certain topics were particularly associated with negative sentiment. As shown in Figure 2, categories such as *Scandals & Controversies*, *Domestic Policy*, and *Social Issues* exhibited strikingly high proportions of negative sentiment, with dominance rates of 84.9%, 70%, and 83.3%, respectively. In contrast, *Media & Public Engagement* and *Foreign Policy & National Security* were the only topics where positive sentiment prevailed, accounting for 86.4% and 89.8%, respectively.

Political Bias Results

Roughly 20% of all articles were marked as N/A, meaning their political biases were not listed in the MBFC database; since this database mainly includes popular news sources, these articles likely came from sources outside of the mainstream media. Excluding these articles, right/right

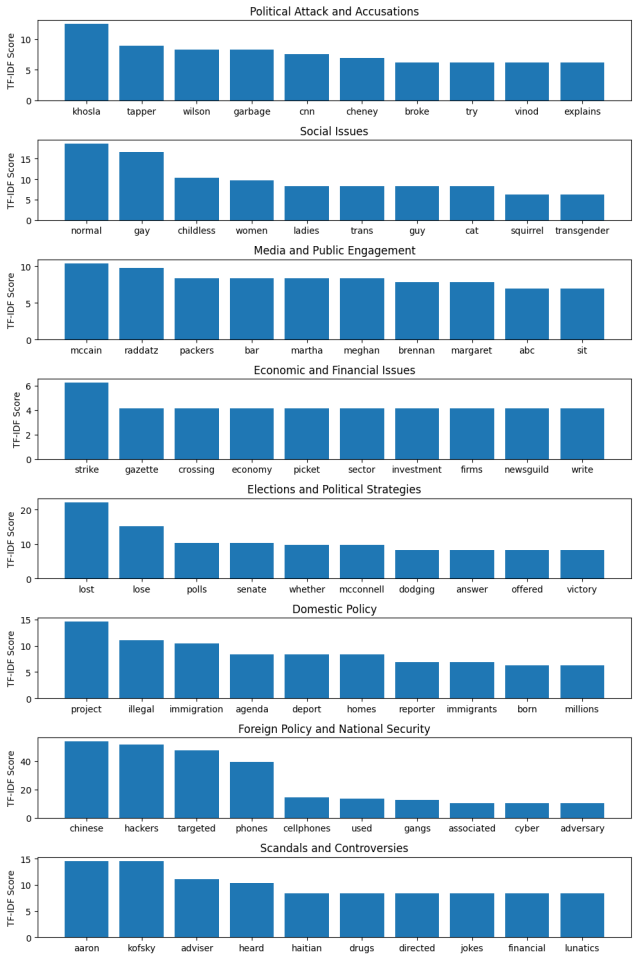

Figure 1: Top 10 TF-IDF Scored Words Across Key Political Categories

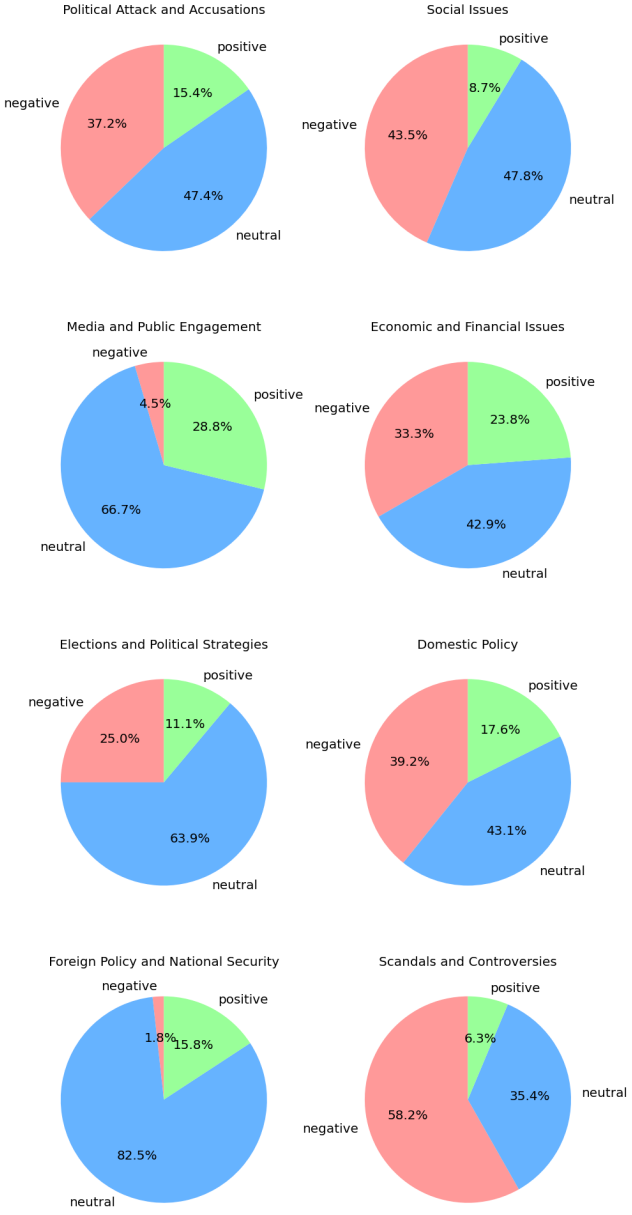

Figure 2: Composition of sentiments for each topic

of center/extreme right articles accounted for roughly 43%, left/left of center articles accounted for 40%, center for 13%, and satire, conspiracy, and pro-science accounting for the remaining 4%. As a sanity check for the sentiment analysis, Figure 3 displays the number of articles for each sentiment organized by political alignment; as one would expect, left leaning sources tended to be more negative than positive, and right leaning sources tended to be more positive than negative.

Discussion

Takeaways

Media coverage of JD Vance has been primarily negative, with language across various topics shaping a public perception of him as deeply flawed. This negativity underscores a consistent trend in how he is portrayed, suggesting a significant impact on his public image. Upon verifying the political biases of the media sources, a clear divide emerges. Left-aligned outlets overwhelmingly emphasize negative aspects of JD Vance, reinforcing a critical narrative, while right-aligned sources focus on portraying him in a more favorable light. This polarization in media coverage highlights the role of political bias in shaping public perceptions.

The language used to describe the Ohioan politician plays a pivotal role in shaping his public persona. This trend is evident in the most frequently used words across various topics. For instance, in discussions surrounding social issues, terms such as “normal,” “gay,” and “guy” appeared often, usually referencing his appearance on *The Joe Rogan Experience*, a podcast known for featuring friends and public figures in discussions about contemporary issues. During the podcast, Rogan remarked that “[JD Vance] and Trump won just the normal gay guy vote”[3]. While the comment was likely intended as harmless, the ambiguous phrasing of “normal gay guy” led to negative reactions from the media, further tarnishing his public image.

Other frequently used terms include “childless,” “cat,” and “lady,” referencing his controversial remarks about women without children, whom he labeled as “childless cat ladies” [5]. This choice of generalizing and disparaging language once again contributed to a damaged reputation, reinforcing the predominantly negative narrative in our results.

An analysis of the topics covered by the media regarding JD Vance in Table 1 reveals that *Scandals & Controversies* and *Political Attacks & Accusations* are among the most frequently discussed. This indicates a media focus on polarizing and negative stories, often emphasizing the more sensational aspects of JD Vance’s public and political life. Furthermore, including *Elections & Political Strategies*, they account for over half of the total coverage. Given that these topics predominantly feature negative sentiment, it suggests that the majority of the public is exposed to unfavorable narratives about JD Vance.

As seen in Table 2, the data reveals a near balance in the proportion of left-wing and right-wing articles, with approximately 23.5% aligning with the political views of JD Vance and the GOP, and around 28.4% favoring the Democratic Party. This suggests an equitable distribution of sources.

Alignment	Proportion
left	0.235
right	0.19
left-center	0.17
right-center	0.152
center	0.132
extreme-right	0.094
satire	0.02
conspiracy	0.005
pro-science	0.003

Table 2: Proportions of Media Biases

However, significant differences emerge in sentiment analysis.

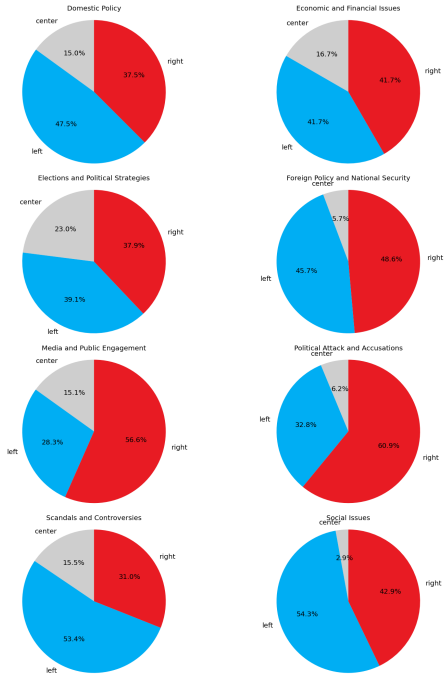

Figure 3: Alignment Proportion by Topic

As illustrated in Tables 3 and 4, left-aligned media sources exhibited a predominantly negative tone, with nearly 46% of their articles classified as negative and only 3.75% as positive. In contrast, right-aligned media sources presented a more favorable perspective, with 34.9% of their articles categorized as positive and just 9.89% as negative. This disparity underscores how political bias influences the portrayal of JD Vance.

In analyzing comparing figure 2 and figure 3, it is evident that the top three topics discussed in the media, which have the highest proportion of negative sentiment, also show the highest proportion of left-wing coverage. This suggests a potential correlation between the prevalence of negative sentiment in media coverage and the alignment of the reporting sources.

Sentiment	Proportion
neutral	0.506250
negative	0.456250
positive	0.037500

Table 3: Left-Winged Article Sentiment Proportions

Sentiment	Proportion
neutral	0.552326
positive	0.348837
negative	0.098837

Table 4: Right-Winged Article Sentiment Proportions

An analysis of media sources with minimal political bias and centrist views reveals that sentiment towards JD Vance remains predominantly negative as seen in Figure 3. This indicates that even when examined through an unbiased lens, the overall coverage and sentiment regarding JD Vance lean negative.

To conclude, media coverage of JD Vance is mostly negative, with a focus on controversial topics and critical language. Left-leaning outlets are represent the politician almost entirely negatively, while right-leaning sources are more supportive. Centrist outlets lean negative, pointing to a consistent trend in how he is portrayed. This widespread negativity plays a big role in shaping how the public views him, directly influencing their vote during the 2024 United States presidential election.

Limitations Notably, the conclusions drawn from this report should be regarded cautiously, since the number of articles (500) is too small to draw conclusions with any degree of certain. Further, as is especially the case when dealing with political data, our annotations and analyses are both susceptible to underlying biases. Other sources of bias include the algorithm used by NewsAPI to retrieve the articles, and the political biases of annotators contributing to the MBFC database.

Group Member Contributions

All members of the team contributed equally and fairly. All members of the team discussed and collected data to be used in the project. All design choices were agreed upon by the team. Timofey handled the manual annotation of the dataset for both topics covered and sentiment analysis. Jikael handled the data analysis portion of the project, both exploring and visualizing the data to be included in the report. Finally, Tarek wrote the report.

References

[1] Alberto Escobar. *Political Bias Database*. 2024. URL: https://rapidapi.com/albertoescobar/api/political-bias-database.

[2] Alexander Hall. *Vance tells Rogan he initially thought Trump had been killed in July assassination attempt: "I was so pissed"*. 2024. URL: https://www.foxnews.com/media/vance-tells-rogan-he-initially-thought-trump-had-been-killed-july-assassination-attempt-i-so-pissed.

[3] Joe Rogan. URL: https://youtu.be/fRyyTAs1XY8?si=KTDBoXd5v5TH9mD_.

[4] Gram Slattery. URL: https://www.reuters.com/world/us/jd-vance-once-compared-trump-hitler-now-they-are-running-mates-2024-07-15/.

[5] Savannah Walsh. *JD Vance's "Childless cat ladies" remark was a reliable punchline at emmys 2024*. 2024. URL: https://www.vanityfair.com/hollywood/story/jd-vances-childless-cat-ladies-remark-was-a-reliable-punchline-at-emmys-2024.