Moderation and Misinformation on Social Media

Riya Aggarwal
University of California, Santa Cruz

Sallar Farokhi
University of California, Santa Cruz

Joey Perrello
University of California, Santa Cruz

1 Abstract

In the more recent annals of political history, social media platforms have often been found to be in the epicenter of interaction, from libelous scandals and general elections, to global-scale conflicts [2]. This, in turn, has inadvertently placed a tremendous amount of authority in the hands of those who maintain it. This has become a recent issue, with Elon Musk's numerous involvements with President Donald J. Trump, the 2024 presidential election, and other country's political recourse enacted on the newly coined X. With the consolidation of platform after platform into a select few, and visible proof that a shareholder of Twitter can directly benefit from the outcome of the election, the sudden possibility of political and media tampering from a third-party has gone through the roof. While we understand that certain organizations and political institutions have been known to inflict censorship on citizens, based on how they speak on political issues current events, this trend has been sparsely studied for larger, more relevant mediums.

In this paper, we investigate three different social media applications: Twitter, Reddit, and Threads, and the mechanisms through which these platforms may or may not be influencing political discourse, such as moderation practices, fact-checking systems, and engagement algorithms. Over the course of 20 days, we accrue tens of thousands of posts from these websites, and perform semantic analysis on data presented when searching specific topics and keywords, all of which were related to the 2024 presidential election. As a result, we have managed to draw a parallel between online political discussion and the platform on which it occurs.

2 Introduction

Social media has become one of the biggest methods of communication and public discussion across the world, with titans such as Instagram and Twitter dominating over the majority of US citizens. Message boards, hashtags, chat rooms; these mediums have permeated into the fabric of society, and as a result, they have a direct impact on the way hundreds of millions of users engage in discussion, especially when it comes to politics.

The United States in particular has made it a point to encourage the usage of social media in order to reach the masses. With major news providers, political figures, and organizations reliant on the mechanisms of social media to transmit their outreach, one must begin to see the responsibility that these platforms are entrusted with, and more so, the amount of power these platform's developers have. While practically every platform would hope to claim that they keep their censorship and authoritative reach to a minimum, this is a fairly worthless promise in this day and age. For concerned citizens and researchers, the problem at hand is figuring out how the political biases of these corporations influences the propagation of information on their platforms, or if they even do. If they were to enact such a policy, a censorship model would have to be considered where company-approved content is promoted over other content, while still maintaining user freedoms. This could be given away through how moderation is conducted, the frequency of fact-checking software and its usage, and engagement metrics on specific posts. Unlike the more traditional national censorship architecture, as seen with The Great Firewall and Russia, censorship in the United States takes on a greater subtlety, which leaves us with a lot of assumptions to make.

In this study, we constructed our test parameters around the political figures and individual accounts that were relevant at the time. In light of the recent presidential election, the elevated message rates serve as a perfect opportunity to quickly gather a lot of data within a small scope of topics. From this data, we performed an analysis of the popularity metrics of these posts, depending on message sentiment and content. From there, were were able to make connections between the political skew of the message and its average popularity and message sentiment. If our proposed experiment can shed some light on whether or not there is politically-charged censorship happening on these platforms, then it becomes entirely possible for them in the future to receive scrutiny for their undisclosed methods, and be held accountable for their wrongdoings. Moreover, a clear elaboration on the methods by which these platforms censor content can serve as an example for future companies to regulate their content more transparently.

3 Related Work

There is a lot of research on how misinformation is spread online [4], yet there seems to be minimal research on how these corporations themselves try to push their own agenda. The academic war on censorship has led a majority of its research towards censorship detection, but this cannot help us detect content promotion. Work in the past [3] has attempted to gauge the flow of popularity on popular message boards, such as Twitter, through the use of natural language processing. These studies were more focused on general characteristics of the tweet and its likelihood that it would be retweeted [1], which is foundation but ultimately non-sequitur to the focus of this paper, which hopes to determine a linearity between the political content and its popularity.

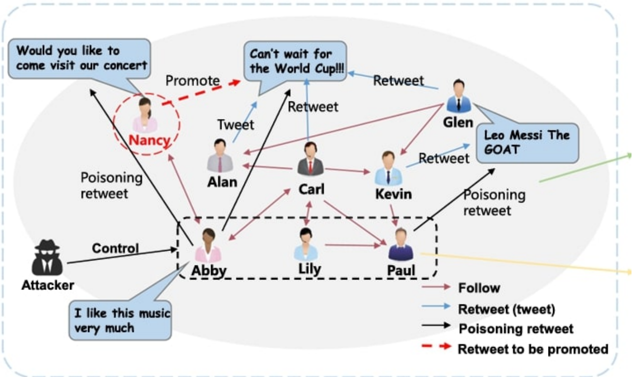

Figure 1: A toy example of attacking social media via behavior poisoning, where the attacker desires to influence Nancy to spread sports-related content

In addition, there have been attempts to analyze and characterize the method of behavior poisoning [5] that occurs on social media, through actions such as retweeting and profile modifying. Figure 1 above shows the process in which an attacker attempts to influence Nancy by promoting and retweeting other topics with a series of accounts that the attacker controls. This method of influencing platforms and users is much more accessible to an attacker, as account creation isn't a limitation for most platforms. While this study does not attempt to detect behavior poisoning, fake accounts and artificial promotion are two of the major issues that plague popular websites, and the data we collected shows signs of influence from these accounts.

In both of these related studies, the main focus was on the lower level of conceptual operations, which was the tweet and its relation to its popularity, or how it could be used to affect popularity. This is more of a bottom-up approach to the issue we presented, meaning that a researcher would attempt to use the sentiment and characteristics of tweets in order to make observations about regulation policies on Twitter, partisan or not. Our contrasting study takes a top-down approach to the problem: we attempt to determine the sentiment and popularity that follows a tweet given the political genre presented. As a result, we can forgo the less relevant details about the tweets and gather a broad perspective on the "state" of the social media platform, mainly in comparison with other platforms. We also hope that work like this can inspire further, more fine-grained analysis of social media, especially considering the recently introduced paradigm of "fake news" [4].

4 Methodology

In order to perform an analysis of social media, Our work faced many obstacles from the beginning, even with the fortuitous timing of the election. Observing data from November 2nd to the 21st, we missed a lot of key moments in the electoral race, such as every presidential debate, the primary election, and other key events in the news. In light of these setbacks, we managed to collect data from each of our test platforms on an hourly interval, collecting data from preset search options and message boards. The following sections provide individualized details for each platform.

4.1 Twitter

With the analysis of Twitter, one of the most important things to keep in mind was the context of the observation period. Within the weeks leading up to the final vote, Elon Musk was holding million dollar giveaways, funded by his right-wing super-PAC, exclusively within swing states. Following President Trump's victory, he is now poised to assume a role within the United States Government. It's no secret that Musk wanted Trump to win, but the real question is how do we prove it?

Political Topic	Keywords
News Providers	@MSNBC, @FoxNews, @ABC, @america ...
Project 2025	Trump, involve, ban, MAGA, abortion ...
Voting Demographics	population, white, black, latino, election map, survey, poll ...
International Affairs	migrant, immigration, policy, foreign, illegal ...

Figure 2: Search parameters for Twitter

As previously mentioned, we used a Selenium bot that would log tweets from the feed created by specific search parameters. These parameters were tailored towards certain political topics of discussion that were popular at the time, such as the conspiracy surrounding immigrants in Springfield. Figure 2 above shows the four most relevant test parameters, and the keywords that correspond with them. Note that these were fruitful keywords on Twitter specifically, and may yield varying results on other social media platforms.

After the observation period ended, we were quickly granted access to a wealth of tweets and their corresponding metadata. In addition to the challenge of tying the sentiment of the tweet to its popularity, we first needed to figure out the political view itself. One of the details of Twitter that worked in our favor was the strongly polarizing nature of politics within user tweets, allowing for as little middle ground as possible. This allowed us to, given a certain political topic of discussion, determine the sentiment of the tweet and draw one of two conclusions, Democratic or Republican. The results will be discussed in 5.

4.2 Threads

A custom scraper was developed to collect data from Threads, focusing on posts related to Kamala Harris and Donald Trump. The scraper was designed to extract and process data efficiently while minimizing detection by the platform. Below is a breakdown of the scraper's key components:

- *Human Behavior Simulation:* The scraper utilized the Playwright module to mimic user actions such as logging in, navigating through the feed, and scrolling. These behaviors ensured the scraping activity appeared natural, reducing the chances of triggering anti-bot mechanisms.
- *Random Sleep Intervals:* To emulate human browsing patterns, the scraper introduced randomized delays between interactions. This approach helped avoid detection by creating variability in request timings.
- *Post Extraction and Metadata Collection:* Posts were extracted from targeted URLs, along with their embeded JSON data. The scraper retrieved essential metadata, including timestamps, URLs, likes, comments, reposts, and textual content. The collected data was stored in CSV format for analysis.
- *Keyword Filtering:* Posts were filtered using a predefined set of keywords, such as "Trump," "Harris," "vote,""swing states," "win," "cheat," "recount," "election," etc. This filtering ensured the dataset remained relevant to the research objectives and focused on political discourse.

The collected posts were analyzed using the Google Cloud Natural Language API. Sentiment scores were assigned to each post, ranging from -1.0 to 1.0, and categorized into three groups: *Positive:* Score > 0.2 *Neutral:* $-0.2 \leq$ Score ≤ 0.2 *Negative:* Score < -0.2 .

The sentiment analysis allowed for the classification of posts and provided insights into the emotional tone of the discussions. The results were aggregated to identify overall trends in public sentiment. A comparative analysis between Kamala Harris and Donald Trump highlighted differences in sentiment polarization and engagement metrics. The correlation between sentiment scores and user interactions, such as likes, comments, and reposts, was examined to understand how public sentiment influenced engagement with posts.

While the scraper incorporated advanced techniques, it encountered several limitations:

- **Detection Issues:** Despite implementing random sleep intervals and behavior simulation, automation was occasionally detected, leading to interruptions in data collection.
- **Platform Limitations:** Threads' relatively low user activity and lack of a public API restricted the volume of data that could be extracted.
- **Candidate Presence:** Kamala Harris had a verified account on Threads, whereas Donald Trump did not, leading to disparities in data availability and engagement metrics.

4.3 Reddit

To understand potential bias within Reddit's discussions, a comprehensive data scraping and analysis methodology was implemented. Using Reddit's API tools, a bot was created with the aim of maintaining a low profile and performing passive data collection. This bot scraped data across numerous political subreddits at hourly intervals, with specific keywords designed to pull relevant political information. The primary focus was to gather posts that could indicate trends in content moderation, user interaction, and the general direction of political discourse. These collected posts, along with their engagement metrics, were stored in CSV files for subsequent analysis.

An important aspect of this methodology involved establishing parameters to ensure the bot avoided undue detection and maintained ethical data collection. Strategies such as simulating human behavior through randomized sleep intervals were used to make the bot appear more like a natural user. The scraped data encompassed post details like upvotes, comments, and any fact-checking metadata, which served as key inputs for subsequent analysis.

Once data collection was complete, the focus shifted to understanding the correlation between content popularity, moderator activity, and fact-checking, with particular attention given to posts that could be perceived as politically biased. Natural language processing (NLP) techniques were applied to infer ideological leanings from post content, which allowed for the categorization of posts into Republican-leaning, Democrat-leaning, or unbiased. A semi-supervised learning approach was adopted to iteratively refine the accuracy of the model, using high-confidence predictions to train and improve the model's performance.

The analysis also looked at the effectiveness of fact-checking in general versus ideological subreddits. Subreddits dedicated to ideological perspectives tended to exhibit echo chamber behaviors, where fact-checking was inconsistent and discussions often leaned heavily toward one political perspective. This observation hinted at the limitations of relying solely on community moderators for content oversight, suggesting the need for a third-party fact-checking system that could provide a more unbiased review of popular posts.

5 Results

5.1 Twitter

As mentioned earlier in 4.1, we collected the data based on genre and applied a sentiment analysis to the collected tweets. To process the tweets, we utilized the Google Natural Language API to determine the sentiment, magnitude, toxicity, and insult levels of each tweet, giving us another metric to judge the test parameters. From this, we found a noticeable increase between toxicity and negative sentiment, and on topics of discussion which President Trump has directly struck out against. Figure 3 shows that the majority of tweets, regardless of view count, exhibited high levels of toxicity, based on the Jigsaw API.

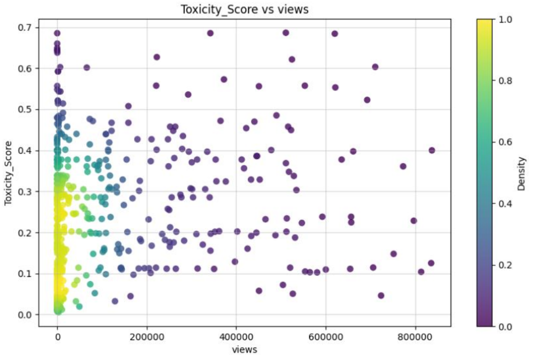

Figure 3: Immigration Tweets

We also notice that there is a considerably smaller trace of activity following topics of discussion that the Democratic party popularized, such as Project 2025. While it could be presumed that Twitter became a decidedly right-leaning platform after Elon's acquisition, it is very unlikely that the sudden surge of popularity surrounding the GOP's points of criticism would mirror the also sudden decline in Democratic discourse *without a change in Twitter's popularity algorithm*. This supports our hypothesis that Twitter was politically altered.

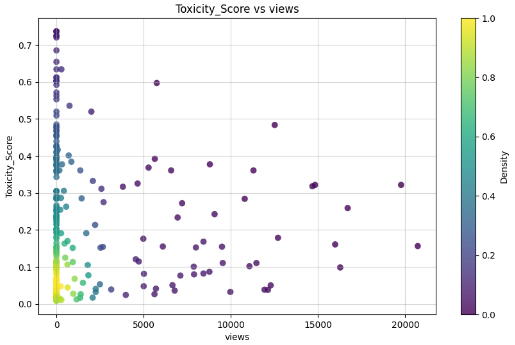

Figure 4: Project 2025 Tweets

5.2 Reddit

On the side of Reddit, the scraping experiment provided insights into the complex dynamics of content promotion, community moderation, and potential biases within the platform.

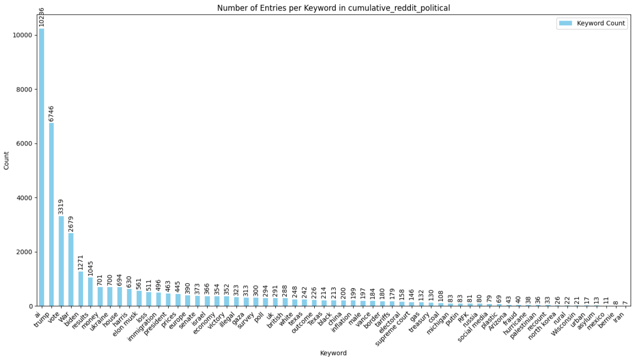

Figure 5: Cumulative Keyword Searches across all three weeks

Over the span of three weeks, we can see the most discussed keywords. Surprisingly, the most popular keyword was completely unrelated to politics: ai. This shocking result reminded us that Reddit is not a website primarily focused on politics. People who visit Reddit tend to be more interested in their own sub-communities; politics is not the focus of Reddit. Still with this in mind there are other important observations to note about this graph. Primarily, we see right-wing candidate Donald Trump taking the spot as the second most searched keyword by nearly double that of the next most popular keyword, "vote." Lastly, the democratic nominee for the 2024 election has not even half the amount of posts as former president Joe Biden, showing the general disinterest and lack of discussion related to left-wing politicians.

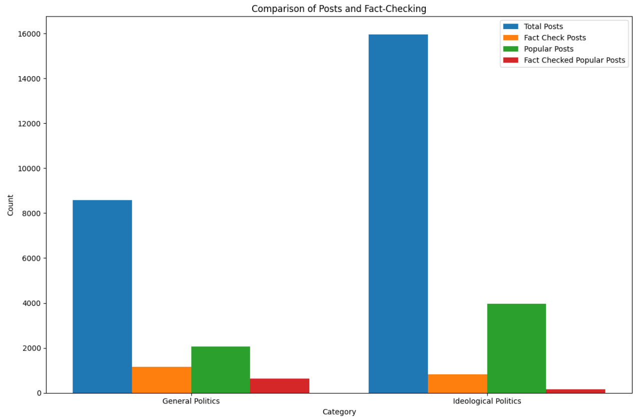

Figure 6: Fact Checked Popular posts across General and Ideological Politics

With the above figure, we can observe the absolute dominance of ideological subreddits over general. Additionally, we can see the pitiful amount of fact checking; not only with the total posts, but popular posts as well. The presence of fact checking is not close to half the amount of total or popular posts. Most graphs related to our test cases followed this trend as well, with a few graphs breaking this trend.

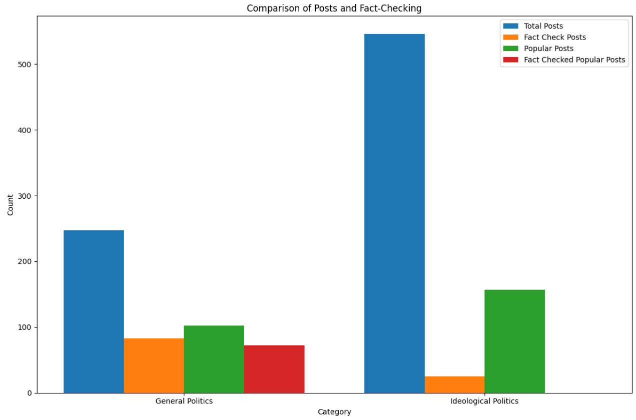

Figure 7: Immigration Week Three Analysis

Here are the same fact checking scripts as before ran on our tests related to immigration in week three. This is one of the few examples where we saw our fact checking script pull different trends than what we saw previously. Immigration is a subject that is prone to a lot of bigotry and misinformation. This is why we were pleased to see such a high increase of fact checking in the general politics subreddit section. The importance of this cannot be understated as these subreddits are meant to be non biased, so it is important that fact checking is prevalent related to controversial topics. On the other side, the Ideological subreddits had virtually no fact checking when it came to this subject, but double the amount of total posts. This shows the sheer amount of posts that are being posted with no supervision, flooding the already biased subreddits with potentially wrong or hateful content.

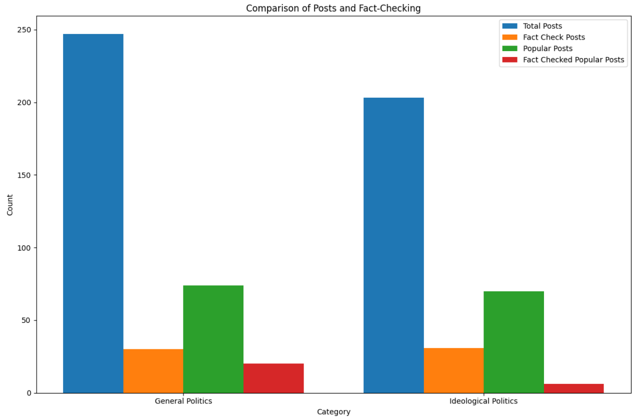

Figure 8: Week One: International Affairs Test

Here is the last test parameter that displayed a different trend in data than the trend we saw in the cumulative file. This test was related to international affairs, so any American policy that is not domestic, or any issues related to countries outside of the US. Here, we see that general politics received far more posts than ideological politics. The fact checking and popularity trends however did not change. What this tells us is that ideological subreddits tend to be related towards US politics, and they generally show less interest than that of general politics. People all around the world may visit these general subreddits for news, but avoid ideological ones due to their lack of interest in international affairs.

These results suggest that Reddit, like many social media platforms, faces challenges in maintaining balanced discourse, especially within politically sensitive communities. The findings highlight the importance of improved fact-checking mechanisms and the need for more rigorous oversight to mitigate the risks of misinformation and echo chamber effects.

5.3 Threads

There were three main takeaways from the observations made for Threads:

Popularity Comparison:

A marked disparity in popularity between the candidates was observed during the pre- and post-election periods. Before the election, Donald Trump accounted for 71% of the discourse, while Kamala Harris garnered 29%. This gap widened significantly after the election, with Trump dominating 95% of the discussion and Harris's share shrinking to just 5%.

Engagement Metrics:

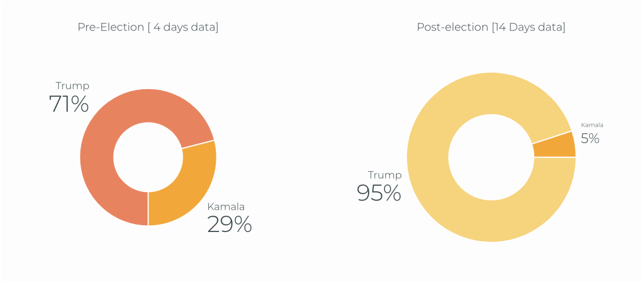

Figure 9: Comparison of discussion volume between Donald Trump and Kamala Harris on Threads - Pre-election and Post-election.

Posts related to Kamala Harris often achieved high engagement, attributed to her verified presence on Threads. Her posts received more likes than comments and reposts, with positive sentiment posts (average sentiment score ∼0.5) achieving the highest peaks in likes. In contrast, Donald Trump's posts, often polarizing in nature, attracted significant engagement across all sentiment categories. Some negative sentiment posts about Trump have amassed more than 10,000 likes, reflecting his capacity to provoke strong reactions and maintain audience attention.

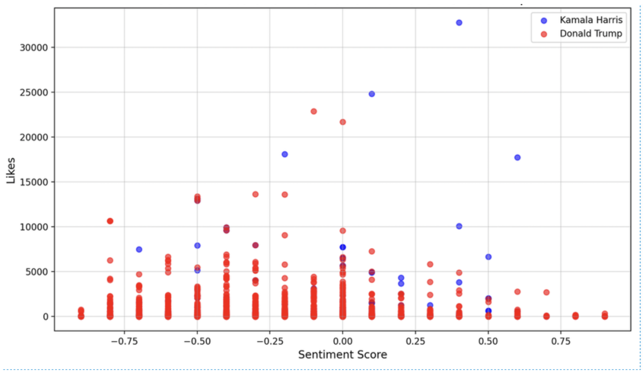

Figure 10: Sentiment Score vs. Likes: Kamala Harris vs. Donald Trump

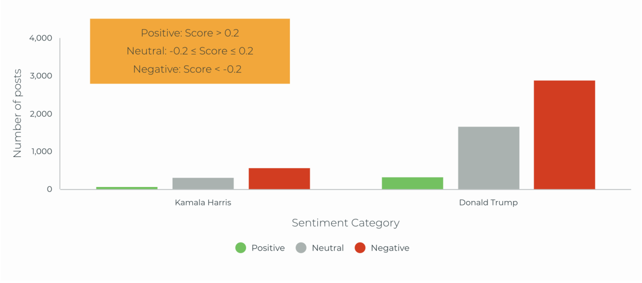

Figure 11: Comparative Sentiment Trends Across Candidates

Sentiment Distribution:

The sentiment analysis revealed stronger polarization in Trump-related posts, which resulted in consistently high engagement across a wide range of sentiments as shown in figure 11. Posts about Kamala Harris performed best when associated with positive sentiment but were less frequent overall. This trend indicates that while Harris achieved notable engagement in positive contexts, Trump dominated the conversation volume and sentiment diversity on Threads. Consequently, the platform did not show an apparent bias toward Harris, as Trump maintained a significant presence in both discussion volume and engagement metrics.

6 Discussion

The impact of this work is significant as it provides a deeper understanding of how online discussions are shaped by both platform dynamics and user behavior. By identifying biases within community moderation and content promotion, this research highlights critical areas where improvements could be made to foster a more balanced and informed public discourse. The applicability of this work serves as a reference for other social media platforms facing similar challenges with content moderation and bias. In practice, the implementation of third-party fact-checking and enhanced moderation strategies could help mitigate misinformation, reduce polarization, and ultimately create a healthier online environment for discussions on politically sensitive topics.

7 Future Work

This research highlights several areas where further investigation and methodological refinement could improve our understanding and the robustness of our conclusions. One significant limitation of the current study lies in its temporal scope. The data collection period was restricted to three weeks, which limits the ability to capture broader trends and variability over time. Expanding the time frame for data collection to include real-time updates over a longer duration would provide a more comprehensive dataset and allow for the detection of temporal patterns that may not have been observable within the short time frame of this study.

In addition, a valuable direction for future work would involve incorporating historical data into the analysis. This approach would enable researchers to observe long-term trends and shifts across platforms, offering insights into how specific events or changes in platform governance influence user behavior and content dynamics. For instance, examining Twitter data spanning before and after Elon Musk's acquisition could reveal critical shifts in trends such as content toxicity, fact-checking practices, or user engagement. This type of longitudinal analysis would also be particularly insightful when paired with an investigation of election-related trends, providing a clearer picture of how sociopolitical events influence and are influenced by social media discourse.

Another avenue for future exploration involves interdisciplinary collaboration, particularly with researchers in political science. Political science experts could offer valuable insights into current political trends and ideological underpinnings that may not be fully captured through computational analysis alone. Such collaborations could improve the development of machine learning models by integrating contextual knowledge. For example, political science students or scholars may be better equipped to discern the ideological nuances in posts, leading to more accurate classification and analysis. This interdisciplinary approach would help bridge gaps in understanding and improve the applicability of findings in various academic and practical domains.

References

[1] Swati Aggarwal et al. “Media bias detection and bias short term impact assessment”. In: *Array* 6 (2020), p. 100025. ISSN: 2590-0056. DOI: https://doi.org/10.1016/j.array.2020.100025. URL: https://www.sciencedirect.com/science/article/pii/S2590005620300102.

[2] Hans W. A. Hanley, Deepak Kumar, and Zakir Durumeric. *"A Special Operation": A Quantitative Approach to Dissecting and Comparing Different Media Ecosystems’ Coverage of the Russo-Ukrainian War*. 2023. arXiv: 2210.03016 [cs.CY]. URL: https://arxiv.org/abs/2210.03016.

[3] Nasir Naveed et al. “Bad news travel fast: a content-based analysis of interestingness on Twitter”. In: *Proceedings of the 3rd International Web Science Conference*. WebSci ’11. Koblenz, Germany: Association for Computing Machinery, 2011. ISBN: 9781450308557. DOI: 10.1145/2527031.2527052. URL: https://doi.org/10.1145/2527031.2527052.

[4] Kai Shu et al. “Fake News Detection on Social Media: A Data Mining Perspective”. In: *SIGKDD Explor. Newsl.* 19.1 (Sept. 2017), pp. 22–36. ISSN: 1931-0145. DOI: 10.1145/3137597.3137600. URL: https://doi.org/10.1145/3137597.3137600.

[5] Chenwang Wu et al. “Attacking Social Media via Behavior Poisoning”. In: *ACM Trans. Knowl. Discov. Data* 18.7 (June 2024). ISSN: 1556-4681. DOI: 10.1145/3654673. URL: https://doi.org/10.1145/3654673.