Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Can-PolNews: A Multi-Platform Dataset of Political Discourse in Canada

Zeynep Pehlivan, Saewon Park, Alexei Sisulu Abrahams, Mika Jacques Patel Desblancs, Benjamin David Steel, Aengus Bridgman

McGill University

zeynep.pehlivan@mcgill.ca, saewon.park@mcgill.ca, alexei.abrahams@mcgill.ca, mika.desblancs@mcgill.ca, benjamin.steel@mcgill.ca, aengus.bridgman@mcgill.ca

Abstract

For many societies, social media has become the primary venue for encountering and engaging with political discourse. But whereas ideas and conversations span multiple communities and move fluidly between different platforms, publicly available datasets are often limited to a single platform. In this paper, we present a multi-platform social media dataset focused on political discourse in Canada, spanning January 1st, 2023, to January 1st, 2025. Our dataset contains all content posted to social media by Canadian news media (national, provincial, and local) and Canadian politicians (federal and provincial) for 1,852 unique accounts across four major platforms popular among Canadians: Instagram, X/Twitter, TikTok, and YouTube. Politicians are labeled by their political party affiliations and provinces, facilitating comparative analysis of regional political trends and ideological affinities. By covering a two-year time frame, this dataset, containing more than 5 million posts with a normalized schema across four platforms, enables researchers to analyze patterns and trends of digital political engagement, and the interplay of news and political elites on social media in an established democracy.

Datasets — https://figshare.com/s/9dd053304c6fc3ae4879

Introduction

The rise of social media over the past two decades has fundamentally changed how information is disseminated and consumed, establishing them as the primary venue for political discourse in modern democratic societies. These platforms provide a space for politicians, news outlets, civil society organizations, and ordinary citizens, to engage with each other, share ideas, and shape public opinion. While offering unprecedented connectivity and freedom of expression, they are also thought to aggravate political polarization and radicalization. For example, (Benkler, Faris, and Roberts 2018) highlights how social media can amplify existing social divides. Furthermore, during critical events such as elections, social media platforms have been particularly vulnerable to manipulation (Golovchenko et al. 2020; Eady et al. 2023; Moore and Colley 2024). For all of these reasons, political discourse on social media is a priority area of research.

While ideas move fluidly between platforms, each platform presents technical and contextual idiosyncrasies that often end up forcing researchers to specialize, which in turn leads to many single-platform studies and datasets that offer only a fragmented window onto political discourse (Muric, Wu, and Ferrara 2021; Dogdu, Choupani, and Sürücü 2024; Yang et al. 2024; Ai et al. 2024; Sun et al. 2024). When we study only one platform, we fail to see how different platforms interact with each other and how they influence politics as a whole. Additionally, user profiles differ significantly across platforms, with variations in audience demographics, content strategies, and engagement styles. This divergence makes it challenging to develop a cohesive understanding of political discourse in the broader social media ecosystem. It is shown that audiences react uniquely to identical campaign messages, depending on the platform used (Bossetta and Schmøkel 2023). A study by Yarchi, Baden, and Kligler-Vilenchik (2021) discovered distinct patterns of polarization on different social media platforms, underscoring the complexities of online discourse.

These findings underscore the need for comprehensive analysis that considers multiple platforms simultaneously, rather than focusing on individual platforms in isolation. Thus, multi-platform analyses of social media are becoming increasingly popular among researchers who recognize their significant advantages in studying complex phenomena. However, researchers face significant obstacles in accessing comprehensive data from social media platforms. The challenges have been amplified by recent restrictions: X (formerly Twitter) ended free access to its academic API and launched paid tiers, while CrowdTangle, a key tool for accessing data from Meta, platforms like Facebook and Instagram, has been unavailable since August 14, 2024.

One prominent dataset is Aiyappa et al. (2023), which encompasses posts from X/Twitter, Facebook, Instagram, Reddit, and 4chan. This dataset focuses specifically on the 2022 U.S. midterm elections, collecting data over a nearly three-month period and linking posts through a comprehensive keyword list to track 1011 political candidates. Similarly, Lepird et al. (2024) analyzes news-sharing patterns across Facebook, X/Twitter, and Reddit in the context of the 2022 U.S. midterm elections. This dataset explores how different types of news sites are shared across platforms, highlighting the differences in sharing behavior. Pierri, Liu, and Ceri (2023) is another significant resource, focusing on the 2022 Italian general election. It collects millions of posts from

Facebook, Instagram, X/Twitter, TikTok, and YouTube over a four-month period using public APIs and keyword-based searches. This dataset also includes political ads and a list of social media handles associated with political representatives.

All these datasets span a specific period tied to events like elections or are created to address specific topic-centric needs. In contrast, our dataset spans two years, from January 2023 to January 2025, capturing Canadian-related events such as the British Columbia elections, among others, while also offering a broader scope that is not confined to singular events. Additionally, our dataset includes data from news outlet handles, enabling deeper analyses of media influence, political narratives, and regional trends. This extended timeframe and inclusion of diverse sources provide a robust resource for studying the interplay between political discourse, media, and public opinion over a longer horizon.

This paper introduces a novel, multiplatform multilingual social media dataset, Can-PolNews focused on political discourse in Canada. Can-PolNews contributions are listed as follows:

- **Multi-Year Coverage:** Spans two years, from January 1, 2023, to January 1, 2025, enabling longitudinal studies of political discourse and media influence in Canada.
- **Multi-Platform Integration:** Collects data from four major social media platforms, Instagram, X/Twitter, TikTok, and YouTube, capturing diverse engagement patterns and content types.
- **Multi-Lingual Content:** Includes posts in English and French, reflecting Canada's bilingual and multicultural context.
- **Politician-Specific Annotations:** Features detailed annotations of politicians' political affiliations and provinces, facilitating in-depth analyses of regional political trends and party-specific communication strategies.
- **News Outlet Inclusion:** Incorporates content from news outlet handles with their annotations, enabling researchers to study media influence, information dissemination, and the interplay between journalism and political narratives.
- **Cross-Platform Normalization:** Provides normalized data across platforms, harmonizing engagement metrics and content structures to enable seamless comparative analyses.
- **Comprehensive Scale:** Contains over 5 million posts, making it the largest dataset of Canadian political discourse spanning multiple platforms and a two-year period.

These contributions make the dataset a robust resource for studying the interplay between political communication, media influence, and public opinion in Canada, while providing a foundation for cross-platform and longitudinal research.

Methodology

This section outlines the methodology used to construct the dataset, encompassing three key steps: Seed List Creation, Data Collection, and Data Normalization. The process begins with the creation of a comprehensive seed list of politicians and news outlets containing 2028 unique accounts . Next, data is collected from four major platforms: Instagram, X/Twitter, TikTok, and YouTube over a two-year period. Finally, a robust normalization process is applied to harmonize the data across platforms, standardizing engagement metrics, extracting key features such as hashtags and mentions, and ensuring compatibility for cross-platform analysis.

Seed List Creation and Annotations

Our dataset consists of posts collected on X/Twitter, Instagram, TikTok, and YouTube from relevant actors in the Canadian media ecosystem. We refer to these actors as seedlist entities, which we define as a person, group, organization, or media product that is of substantive interest to Canada's media ecosystem. Examples of media products are newspapers and podcasts. To be included in the seedlist, these entities must be alive, in the case of a person, or in operation, for the remaining types at the time of inclusion. In addition, to remain included in the seedlist, the entity must have activity on at least one platform in the last 12 months.

For this dataset project, entities were categorized into two main types: politician and news outlet. Politicians refer to elected and currently serving members from one of the three levels of the Canadian government: federal, provincial, or municipal. Municipal government officials are limited to mayors of cities with at least 100 000 population. News outlets are defined as publishing organizations registered or incorporated in Canada. They must produce original content on a regular basis and should be publicly independent from those it covers. This means that, at least publicly, the news outlet organization should not be promoting certain interest groups, individuals, or causes. The news content that is produced includes print (physical and digital), television, and radio.

We aimed for the identified list of politician entities and news outlet entities to be as exhaustive as possible, meaning that we aimed to collect every single entity that met our definition. For the identification of politicians, the rosters of elected members of the house of commons, each provincial legislative assembly, and each municipal government (of cities above 100 000 population) were used. These rosters came from the official website of each respective federal, provincial, and municipal governments. In addition to the list of politician names, we used the official rosters to also collect the political party and province of each politician. During this process, provincial parties were mapped to their federal party counterparts based on existing public ties to specific federal parties or by their policy similarity to a federal party. Specifically for British Columbia's provincial politicians, we additionally collected the electoral riding for each politician as a part of a study on the 2024 British Columbia provincial election.

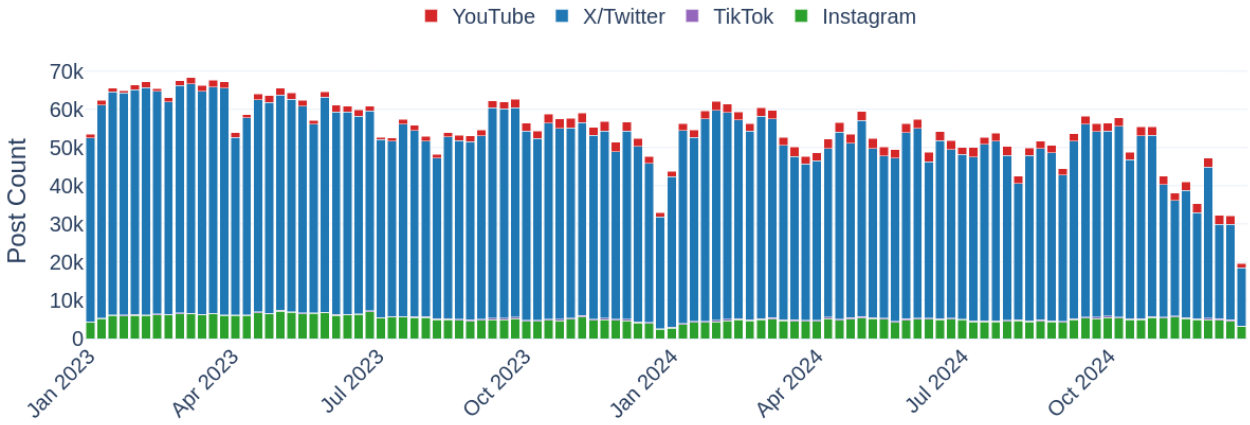

Figure 1: Timeline by platform

Platform	# Politicians (%)	# News Outlets (%)	Total
X/Twitter	992 (92.6%)	721 (92.3%)	1,713
Instagram	930 (86.8%)	423 (54.2%)	1,353
YouTube	329 (30.7%)	204 (26.1%)	533
TikTok	68 (6.3%)	63 (8.1%)	131
Total	2 319	1 411	3 730

Table 1: Distribution of Entities Across Platforms

For the identification of news outlets, we used Media Cloud's[1] "national" and "state-local" datasets of Canadian news outlets. Media Cloud's lists were manually reviewed to confirm that they were still in operation and based in Canada. In addition, we also reviewed each outlet's official website to determine if they were "national", meaning that the outlet covers the majority of Canada's provinces and territories, or "local", meaning that the outlet covers a minority of Canada's provinces and territories. This categorization was recorded as a sub-type of the entity. If a news outlet was labelled as "local", province information was added to the seedlist for that outlet based on the province where its headquarters were located.

Once all the entities were identified and annotated with main type, sub type, provincial party, federal party, electoral riding, and province, social media handles covering as many platforms as available for each entity were recorded. Efforts were made to only include official accounts where news or political content was being posted and to exclude any accounts that only contained personal content. In total, we have 781 unique entities for news outlets and 1071 for politicians. Table 1 shows the number of unique entities by platform and type. The percentages indicate the proportion of politician or news outlet entities that have an account on each platform.

Within the seedlist, each entity is given a unique ID, called "SeedID". This ID remains the same for all post data associated with the same entity, even if the posts are from different social media platforms. This allows us to compare activity and networks between entities across all social media platforms in our dataset. The seed object is defined as follows and appended to each post in the dataset:

```
SeedID: 2486  | 2294
SeedName: CBC News | Justin Trudeau
MainType: news_outlet | politician
SubType: national | member of ...
Province:    | Quebec
NewsOutletCategory: National
ProvincialParty:
FederalParty:    | Liberal
Party:    | Liberal
ElectoralDistrict:
Platform: Tiktok | Instagram
Handle: cbcnews | justinpjtrudeau
```

Figure 2 illustrates the percentage distribution of entities (politicians and news outlets) across four platforms. X/Twitter accounts for the highest share of entities for both politicians and news outlets, reflecting its role as a dominant platform for political and news discourse. Instagram also features a substantial presence, particularly for politicians, likely due to its growing importance in visual storytelling. YouTube and TikTok represent smaller proportions, reflecting their specialized use cases and relatively lower adoption among politicians and news outlets.

Figure 3 illustrates the distribution of entities based on the number of accounts they maintain across platforms. Approximately 44% of news outlets manage only 1 account, implying a high level of dependence on a single platform. Only 4% of news outlets have accounts on all four platforms, and 17% have accounts on three platforms. On the other hand, only 18% of politicians have accounts on a single platform, demonstrating a more diversified presence compared to news outlets. Just over half of the politicians (51%) main-

[1]https://www.mediacloud.org/

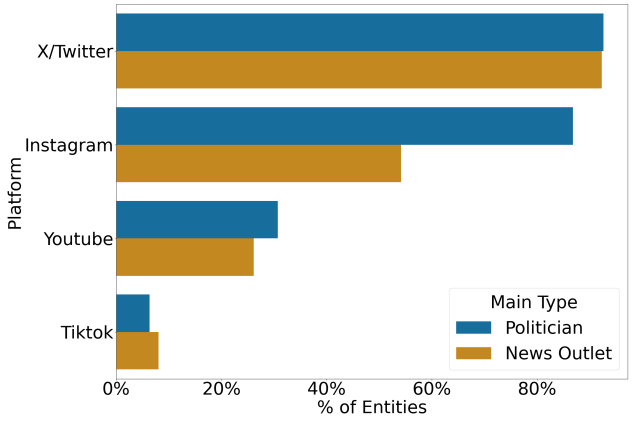

Figure 2: Percent of entities on each platform

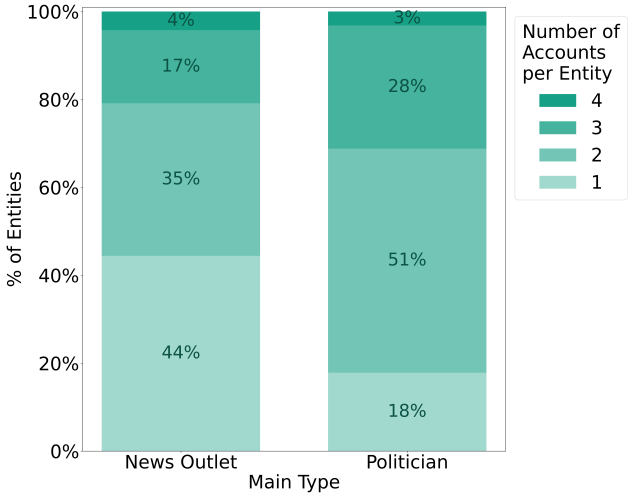

Figure 3: Number of accounts per entity

tain accounts on two platforms, 28% have accounts on three platforms, and 3% have accounts on all four platforms.

Data Collection

To build a comprehensive dataset, we developed platform-specific tools and workflows for collecting data from X/Twitter, TikTok, Instagram, and YouTube. Metadata for each post was preserved in its entirety and stored in JSON format to ensure structural consistency and completeness. While we did not collect media files directly, all available links to media content (e.g., images, videos) were retained in the metadata. This section outlines the specific data collection strategies implemented for each platform:

X/Twitter We developed a web scraping library for X/Twitter that programmatically queries the platform's advanced search functionality to retrieve recent posts from politicians and news outlets. This approach allowed us to collect data on a bi-weekly basis, as well as historical posts from the months prior to an account's inclusion in our seed list, by specifying custom date ranges. Data was collected directly from structured API-like responses rather than rendered web pages, ensuring completeness and consistency of post metadata across the dataset.

TikTok We used a web scraping library for TikTok as described in (Steel, Parker, and Ruths 2024) [2] that allows us to do full reverse-chronological collections for each account. The scraper was run at regular weekly intervals to ensure comprehensive coverage of posts while maintaining the integrity of engagement metrics. While an official TikTok Research API does exist, it does not provide data originating in Canada, and as such, we were not able to use it for this work.[3]

Instagram Until August 14, 2024, Meta provided access to its CrowdTangle platform[4], which we used to collect Instagram data from January 1, 2023, to August 10, 2024. During this period, our collection process relied on a daily scraping schedule. Following the discontinuation of CrowdTangle on August 14, 2024, we transitioned to a custom scraping solution similar to the approach described by (Abrahams 2023), running the scraper on a weekly basis. As a result, Instagram data in our dataset is organized into two separate subsets: *instagram-meta*(CrowdTangle-based) and *instagram*(scraper-based), reflecting the two phases of collection.

YouTube We used the YouTube Data API[5] to collect metadata associated with videos posted by accounts in our seed list. This included post-level details such as titles, descriptions, publication timestamps, and engagement metrics like views and likes. To ensure temporal continuity and capture engagement activity close to the time of posting, we implemented a frequent and regular collection schedule.

For each scraping session, we record the crawl date as: *crawled_date* in our dataset, so that the time at which the metadata was recorded can be known. We store the data in an Elasticsearch [6], following an architecture similar to the one described in Pehlivan, Thièvre, and Drugeon (2021).

Data Normalization

Working with multi-platform social media datasets presents significant challenges due to variations in data schemas. Each platform has its unique data schema, and even the same platform can yield differing schemas when data is collected via different methods (e.g., scraping vs. APIs).

For example, YouTube data includes titles and descriptions, while TikTok includes all caption textual data in a single *desc* field. Instagram data from the scraper primarily relies on the caption field, while CrowdTangle data uses the description field. X/Twitter data maps its textual content to the message field. All these fields are harmonized into a single text_all field, which aggregates content from titles, descriptions, messages, captions, hashtags, and mentions into

[2] https://github.com/networkdynamics/pytok

[3] https://developers.tiktok.com/doc/research-api-codebook/

[4] https://about.fb.com/news/2023/11/new-tools-to-support-independent-research/

[5] https://developers.google.com/youtube/v3

[6] https://www.elastic.co/

a unified textual representation. Additionally, other metadata such as engagement metrics (e.g., likes, shares, views), user identifiers, and platform-specific attributes are standardized to ensure consistency across platforms, facilitating seamless cross-platform analysis. Aligning such disparate structures is critical for uniform analysis. To facilitate meaningful analysis, additional features such as URLs, domains, hashtags, and mentions are also extracted during the normalization process.

By including normalized data with platform-specific data, the dataset becomes a powerful resource for studying political discourse across social media ecosystems. This normalization enables researchers to easily compare engagement metrics, user behaviors, and content trends across platforms, providing deeper insights into the dynamics of multi-platform interactions.


```
date: 2023-08-31T21:40:36Z,
comment_count: 5,
like_count: 16,
view_count: 911
share_count: 2,
seed: {
    ElectoralDistrict: ,
    NewsOutletCategory: ,
    SubType: member of ...,
    Platform: Twitter,
    FederalParty: Liberal,
    SeedName: Justin Trudeau,
    MainType: politician,
    SeedID: 2294,
    Province: Quebec,
    Handle: justintrudeau,
    Party: Liberal,
    ProvincialParty:
},
hashtags: [NationalConvenienceWeek],
user_name: JustinTrudeau,
description: ,
caption: ,
domains: [],
title: ,
message: @ConvenienceCan ...,
platform: twitter,
tags: [reply],
urls: [],
user_id: 14260960,
mentions: [ConvenienceCan,...]
id: 1697363784648568952,
text_all:   NationalConvenienceWeek...
```


Exploratory Data Analysis

The exploratory data analysis provides a comprehensive overview of the dataset's structure and engagement patterns.

Table 2 summarizes the distribution of records across platforms for news outlets and politicians. The data reveals that news outlets dominate the dataset with over 82% of the posts, primarily on X/Twitter. Politicians account for 18% of the posts, with strong representation on both X/Twitter and Instagram. X/Twitter overwhelmingly dominates post volume across both entity types, reflecting its continued centrality in political and news communication in Canada. Therefore, while TikTok's lower volume may be the most visually striking, the dataset more broadly mirrors platform preferences among public figures, with X/Twitter serving as the primary channel for public-facing engagement. Consequently, researchers using this dataset for cross-platform or time-series analysis should consider this representational disparity and apply appropriate weighting or normalization techniques when interpreting results.

Main Type	**Platform**	**# Posts**
News Outlet	X/Twitter	4 258 616
	Instagram	217 116
	Youtube	176 457
	Tiktok	26 640
Total (News Outlet)		4 678 829 (**82.21%**)
Politician	X/Twitter	693 944
	Instagram	299 659
	Youtube	16 515
	Tiktok	2 499
Total (Politician)		1 012 617 (**17.79%**)
Grand Total		**5 691 446**

Table 2: Distribution of Records Across Platforms for News Outlets and Politicians

Additionally, Table 3 showcases the top 5 active handles for both categories, revealing substantial differences in engagement. Notably, "Pierre Poilievre" leads among politicians with over 4232 average likes per post, while "The Score" and "Sportsnet" dominate among news outlets, reflecting their strong appeal in the sports domain. This analysis provides insights into platform-specific usage patterns and engagement dynamics, laying the groundwork for more detailed studies on multi-platform political discourse.

Entity Name	**Post Count**	**Likes(Mean)**
News Outlets		
CTV News	87 263	656.6
The Globe and Mail	84 247	15.1
The Score	72 598	2 093
Journal de Montreal	68 349	7.7
Journal de Quebec	66 484	3.1
Politicians		
Pierre Poilievre	10 996	4 232.5
Dominic Cardy	9 173	10.4
Kevin Vuong	8 842	201.7
Robert Bailey	8 159	0.6
Paul Lane	7 989	6.3

Table 3: Top 5 active handles for News Outlets and Politicians

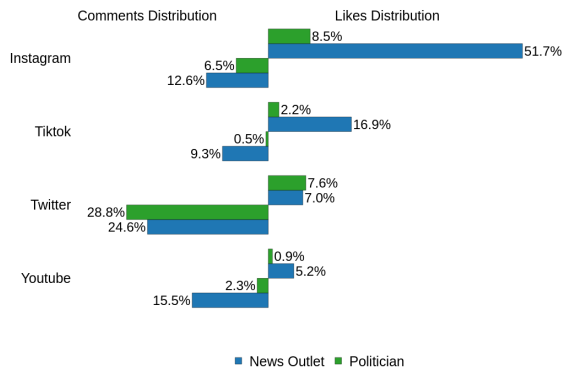

Figure 4: Distribution of Engagement by Platform and Category

A time series chart of posts by platform can be found in Figure 1, representing the data collected over two years. Variations in activity levels may correspond to platform-specific data availability, scraping limitations, or external factors such as significant events or campaigns, as addressed in the section.

Distribution of engagement (like counts and comment counts) is shown in Figure 4, disaggregated by platforms and categories. The right side illustrates the distribution of likes, highlighting Instagram's dominance in likes and the strong presence of news outlets. TikTok also demonstrates a strong presence in likes, reflecting its rapid growth and the viral appeal of its content. The left side depicts the distribution of comments, showcasing X/Twitter as the leading platform, likely due to its real-time, conversational nature, for discussions and politicians receiving substantial engagement compared to news outlets.

Discussion

In this section, we discuss some potential research applications of this dataset, while cautioning the reader regarding several limitations.

Limitations

Despite our best efforts to build a comprehensive dataset, our efforts fell short in a few respects.

Facebook Firstly, and most notably, while Facebook continues to be popular among Canadians, access to Facebook data via Meta's CrowdTangle API was rescinded on August 14, 2024. Whereas for Instagram we were able to compensate for the loss of CrowdTangle by deploying a custom scraper, in the case of Facebook we were not able to meet the technical challenge in time. We may yet release Facebook data for this period once we are able to fill in the gap from August 14, 2024 to January 1, 2025, but for now this remains a limitation of our dataset.

We note that Meta has introduced a new Meta Content Library[7], which may serve as an alternative for researchers seeking access to Facebook and Instagram data. However, approved users are only permitted to analyze the data and use the proposed API within Meta's "clean room" environment with no internet access. This architecture makes it infeasible to integrate Meta data directly into our data pipeline. While researchers can still perform analysis within Meta's environment and export the results, this setup limits the potential for full integration or cross-platform comparisons.

Other platforms We also note that our dataset does not include content from other platforms such as Reddit, Bluesky, Mastodon, nor content from fringe platforms such as Parler or Truthsocial. But survey data collected by the Media Ecosystem Observatory throughout 2023 and 2024 indicates that these platforms account for a vanishing fraction of the Canadian public. To a first approximation, all the major conversations and communities in Canadian political discourse are represented on the major mainstream social media platforms covered in our dataset.

Platform Specific Issues

Each of the four platforms in our dataset presented unique challenges of which researchers should be aware when using the data. Most importantly, the loss of publicly and freely available official APIs for each platform forced us to rely on scraping methods to collect the data. As such, for X/Twitter, Instagram, and TikTok, content was retrieved using scraping methods tailored to each platform's idiosyncrasies, which in turn generated idiosyncratic limitations.

Instagram: Following the discontinuation of the CrowdTangle API on August 14th by Meta which we had been using to collect Instagram posts, we were forced to rely on scraping methods in order to collect Instagram data. Accordingly, the Instagram data in the dataset are from Crowdtangle for the period January 1, 2023 to August 10, 2024, and from an Instagram scraper for the period August 10, 2024 to January 1, 2025. While we are not aware of any discrepancies between these two data sourcing methods, scraping methods generally tend to be less reliable than authorized API access, as they may trigger undetected anti-scraping defense mechanisms on the part of the target web application, leading to data loss. For further information on our Instagram scraper and the issues that can arise with scraping, please see (Abrahams 2023)

TikTok: All data for TikTok was collected using scraping methods; as such, we also note that scraping methods are less reliable than official APIs, and we cannot guarantee the same level of completeness as if we had used an official API.

X/Twitter: Following the discontinuation of Twitter's Academic API access in 2023[8], we initiated alternative collection methods to gather public data from accounts in our seed list. Due to the platform's restrictions and evolving data access policies, we employed programmatic methods that leveraged publicly accessible search functionality. These methods were carefully designed to operate within the limits of available infrastructure while preserving data quality.

[7]https://transparency.meta.com/researchtools/meta-content-library

[8]https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research

We acknowledge that some posts may be missing due to platform-side constraints such as search coverage limits [9] and rate restrictions. However, we believe the dataset still provides a representative sample of posts from the accounts in our seed list. To better assess coverage, we are exploring methods such as comparing our collected data with firehose samples and other benchmark sources to estimate the extent of missing data and validate distributional representativeness. Finally, given that there is no other way to collect this data without significant financial cost, we believe it is of value to the research community.

Since all collection methods were developed independently of one another, at different times, and seed list enrichment efforts were done on an ongoing basis, there are discrepancies between platforms in the dates at which data was collected. We encourage researchers to make use of the *crawled_date* field available for each record when performing their analysis.

Seed list limitations Our seedlist contains a manually annotated list of Canadian politicians and news outlets. Lists of all members of the House of Commons, and of each provincial legislative assembly, are publicly available, and with a mix of manual review and automation, we were able to collect all their official pages on all platforms on which they were active. We did, however, make the decision to not include municipal politicians from towns with less than 100 000 people. There are a total of 5 162 municipalities in Canada and including elected officials for each one of them would have made the seedlist size unmanageable and data collection much harder. By setting a floor on town population size, we limit ourselves to the elected officials with the largest political power.

We developed our seedlist for news outlets based on MediaCloud's Canadian news collections. Accordingly, the completeness of our national and local news outlets depends on the completeness of Media Cloud's dataset, supplemented by our extensive manual review of Canadian news outlets. To our knowledge, our seedlist constitutes the most comprehensive list of Canadian news outlets in the public domain.

Replicability The inherent ephemerality of social media creates challenges for replicating our data collection process. Firstly, users can delete posts, rendering them unavailable for retrieval whether through authorized API access or scraping. Users can also alter the privacy settings on their account, rendering their content invisible to researchers. On the other hand, when content is left up and remains publicly accessible, the engagement counts may change over time as posts accumulate more likes or comments.

Potential Applications

By encompassing a number of important political events in Canada, our dataset contains within it the potential for several event-centric analyses. On the other hand, by comprehensively tracking politicians and news outlets for two years across all major platforms, our dataset offers a baseline against which event-centric analyses can be compared.

Meta news ban In retaliation to Canadian government legislation, since approximately August 9, 2023, Meta has blocked users with Canadian IP addresses from seeing news content on Facebook and Instagram. Given that our dataset begins eight months in advance of this move, and tracks activity for almost a year and a half after the policy's implementation, researchers may find our dataset ideally suited for assessing the impacts of the news ban. For further details, see our preliminary assessment (Parker et al. 2024).

Provincial elections In May 2023 and October 2024, Alberta and British Columbia (respectively) held provincial elections. The Alberta elections saw the left-leaning NDP challenge the conservative UCP in a contest rife with polarization and climate change denial. British Columbia's elections proceeded in the face of fears over election-related disinformation. Our dataset encompasses both of these periods in Canadian politics and researchers interested in investigating them can consult our preliminary findings [10].

Baseline trends During 2024, this dataset underwrote a series of monthly Situation Reports [11] describing basic indicators of the Canadian media ecosystem's health. These situation reports illuminate the power of these data for providing baseline trends against which political incidents can be compared. Researchers and pundits often claim that disinformation or hate speech is on the rise, but our dataset acts as a kind of high-frequency census of the Canadian media ecosystem against which such claims can be judged. Likewise, the period 2023-2024 saw growing concern over the spectre of foreign interference in Canadian politics. With two full years of data at their fingertips, researchers are poised to evaluate whether, indeed, Canadian politicians or news outlets exhibit any signs of adopting the talking points of foreign agendas.

Local versus national news Our dataset covers not only nationally scoped news outlets like the CBC or National Post, but also local news outlets like the Calgary Herald. By capturing all of their content and engagement statistics, our dataset offers researchers a unique opportunity to compare local and national news – how do they differ in terms of content, tone, and (based on engagement statistics) what characteristics of their communication resonate most with their audiences? During a time of widespread hardship in the news media industry, local news is thought to be differentially impacted, but our dataset uniquely offers the kind of comprehensive coverage required to dig into this.

Politicians by region and party A major contribution of our dataset is that politicians are annotated by party affiliation and province, meaning that their choices of topic and engagement behavior can be cross-tabulated. For example,

[9] https://developer.x.com/en/docs/x-api/v1/tweets/search/overview

[10] https://meo.ca/work/cdmrn/the-canadian-information-ecosystem, http://meo.ca/work/british-columbia-election-information-ecosystem-project-baseline-report

[11] https://www.cdmrn.ca/situation-reports

in a preliminary analysis[12], we found that geography appeared to trump ideology, as the social media activity of provincial politicians suggests they engage more with their opponents within the same province than with their fellow party members at the federal level.

Political polarization The fact that politicians are annotated with their party affiliations means that any network mapping of the dataset can reveal whether the social media landscape is fragmented into party-denominated subnetworks. Indeed, our report on the Alberta elections found evidence that NDP and UCP political candidates inhabited distinct subnetworks on Instagram (Bridgman et al. 2024).

Media bias detection Given that politicians can be mapped into separate party-denominated subnetworks, their positionality acts as an implicit ideological embedding of the news media. If, say, the National Post is more densely connected with conservative politicians, while the CBC is situated deep within the Liberal subnetwork, this would indicate that these news outlets are not finding middle ground among all Canadians, but are instead appealing to partisan politics.

Conclusion

In Canada, as elsewhere, social media has become people's primary conduit for political discourse, even as researchers face unprecedented barriers to accessing social media data. In this paper, we have presented a new dataset of social media content and engagement patterns of Canadian politicians and news media, spanning an eventful two-year period and four major social media platforms popular among Canadians. The dataset provides an unprecedented view into Canadian politics at the provincial and national levels, and into Canadian news coverage at both nationally and locally scoped outlets. It is our hope that this multi-platform dataset will facilitate research into Canadian news, politics, and the broader consequences of social media for democracy.

References

Abrahams, A. S. 2023. Social Media Observatory.

Ai, L.; Gupta, S.; Oak, S.; Hui, Z.; Liu, Z.; and Hirschberg, J. 2024. TweetIntent@Crisis: A Dataset Revealing Narratives of Both Sides in the Russia-Ukraine Crisis. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1): 1872–1887.

Aiyappa, R.; DeVerna, M. R.; Pote, M.; Truong, B. T.; Zhao, W.; Axelrod, D.; Pessianzadeh, A.; Kachwala, Z.; Kim, M.; Seckin, O. C.; Kim, M.; Gandhi, S.; Manikonda, A.; Pierri, F.; Menczer, F.; and Yang, K.-C. 2023. A Multi-Platform Collection of Social Media Posts about the 2022 U.S. Midterm Elections. 17: 981–989.

Benkler, Y.; Faris, R.; and Roberts, H. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press. ISBN 978-0-19-092362-4. _eprint: https://academic.oup.com/book/26406/book-pdf/53445042/9780190923648_web.pdf.

Bossetta, M.; and Schmøkel, R. 2023. Cross-Platform Emotions and Audience Engagement in Social Media Political Campaigning: Comparing Candidates' Facebook and Instagram Images in the 2020 US Election. 40(1): 48–68. Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2022.2128949.

Bridgman, A.; Lee-Whiting, B.; Bergeron, T.; Galipeau, T.; Abrahams, A.; Park, S.; Parker, S.; Owen, T.; and Loewen, P. 2024. Indistinct Information Habitats: Information and Attitudes in the 2023 Alberta Election.

Dogdu, E.; Choupani, R.; and Sürücü, S. 2024. Detecting Political Polarization Using Social Media Data. In Han, H.; and Baker, E., eds., *Next Generation Data Science*, 46–59. Cham: Springer Nature Switzerland. ISBN 978-3-031-61816-1.

Eady, G.; Paskhalis, T.; Zilinsky, J.; Bonneau, R.; Nagler, J.; and Tucker, J. A. 2023. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14(1): 62. Publisher: Nature Publishing Group.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-04-27.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Golovchenko, Y.; Buntain, C.; Eady, G.; Brown, M. A.; and Tucker, J. A. 2020. Cross-Platform State Propaganda: Russian Trolls on Twitter and YouTube during the 2016 U.S. Presidential Election. *The International Journal of Press/Politics*, 25(3): 357–389. _eprint: https://doi.org/10.1177/1940161220912682.

Lepird, C. S.; Ng, L. H. X.; Wu, A.; and Carley, K. M. 2024. What News Is Shared Where and How: A Multi-Platform Analysis of News Shared During the 2022 U.S. Midterm Elections. 10(2): 20563051241245950. Publisher: SAGE Publications Ltd.

Moore, M.; and Colley, T. 2024. Two International Propaganda Models: Comparing RT and CGTN's 2020 US Election Coverage. *Journalism Practice*, 18(5): 1306–1328. Publisher: Routledge _eprint: https://doi.org/10.1080/17512786.2022.2086157.

Muric, G.; Wu, Y.; and Ferrara, E. 2021. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. *JMIR Public Health and Surveillance*, 7(11): e30642. Company: JMIR Public Health and Surveillance Distributor: JMIR Public Health and Surveillance Institution: JMIR Public Health and Surveillance Label: JMIR Public Health and Surveillance Publisher: JMIR Publications Inc., Toronto, Canada.

Parker, S.; Park, S.; Pehlivan, Z.; Abrahams, A.; Desblancs, M.; Owen, T.; Phillips, J.; and Bridgman, A. 2024. When journalism is turned off: Preliminary findings on the effects of Meta's news ban in Canada.

[12]https://meo.ca/work/cdmrn/the-canadian-information-ecosystem

Pehlivan, Z.; Thièvre, J.; and Drugeon, T. 2021. Archiving social media: the case of Twitter. In *The past web: Exploring web archives*, 43–56. Springer.

Pierri, F.; Liu, G.; and Ceri, S. 2023. ITA-ELECTION-2022: A Multi-Platform Dataset of Social Media Conversations Around the 2022 Italian General Election. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, 5386–5390. Association for Computing Machinery. ISBN 979-8-4007-0124-5.

Steel, B.; Parker, S.; and Ruths, D. 2024. The Invasion of Ukraine Viewed through Large-Scale Analysis of TikTok.

Sun, Y.; Jia, R.; Razzaq, A.; and Bao, Q. 2024. Social network platforms and climate change in China: Evidence from TikTok. 200: 123197.

Yang, Z.; Imouza, A.; Touzel, M. P.; Amadoro, C.; Desrosiers-Brisebois, G.; Pelrine, K.; Levy, S.; Godbout, J.-F.; and Rabbany, R. 2024. Regional and Temporal Patterns of Partisan Polarization during the COVID-19 Pandemic in the United States and Canada. ArXiv:2407.02807 [cs].

Yarchi, M.; Baden, C.; and Kligler-Vilenchik, N. 2021. Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. 38(1): 98–139. Publisher: Routledge _eprint: https://doi.org/10.1080/10584609.2020.1785067.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes and see the discussion section on the paper
   (e) Did you describe the limitations of your work? Yes, please refer to the limitations sub section on discussion section.
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? No
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, it is included on FigShare description.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Ethical Statement

The proposed dataset adheres to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability. The dataset is hosted on FigShare, where it is assigned a unique identifier to ensure findability and accessibility. The data is accessible in a standardized and well-documented format, enabling interoperability across research tools and platforms. Additionally, we provide detailed metadata and schema descriptions, which ensure reusability and facilitate cross-platform and longitudinal analyses of political discourse.

The dataset was carefully designed with ethical considerations in mind throughout the data collection and processing phases. We only collected publicly available data from X/Twitter, Instagram, TikTok, and YouTube from politicians who are public-facing individuals and news outlets which are public entities. We decided to leave politician and news outlet post data un-anonymized. Furthermore, because they are public individuals or entities, and we only collected publicly available data, we did not deem it necessary to ask for their consent. Finally, in our dataset, the only data not from public-facing individuals or institutions are YouTube comments, and we removed all author-related information from these comments.

In this dataset, we have collected publicly available data from TikTok, Twitter and Instagram using scraping techniques which we understand may not align with the Terms of Service of these platforms. However, we strongly believe that our work, as well as similar work, is important to better understand platforms that are used by billions of users worldwide. Greater platform transparency will lead to higher quality and more trustworthy research, which in turn, will also be an added value for all platforms.

To reduce the potential for misuse of the dataset, we have uploaded it to FigShare under a license that restricts usage to non-commercial, academic, and research purposes.

All of the software tools we've used were used in compliance with their respective licenses.