

Podcasts as Data: Mapping the Canadian Podcast Ecosystem and its Economic and Housing Debates

SCOTT ROBERT PATTERSON¹
Mila – Quebec AI Institute, Canada

COLIN CHIA
University of Victoria, Canada

AENGUS BRIDGMAN
McGill University, Canada

Podcasts are central to media ecosystems, yet no research program engages them systematically as digital trace data. We address this deficiency by demonstrating how podcasts' RSS-based distribution infrastructure enables podcast data collection and analysis at scale. With a dataset of 6,055 unique Canadian podcasts comprising 655,123 episodes, we provide an initial description of the Canadian podcast ecosystem in terms of its production patterns and political content. Our descriptive analysis unfolds in three parts. First, at the podcast level, we find that the Canadian ecosystem is substantively diverse, primarily urban, and biased towards English. Second, at the episode level, we document exponential growth through the 2010s, a pandemic-era surge in new podcast launches, and that the top 10% of podcasts produce nearly half of all episodes. Third, in terms of content, we find that online-first podcasts exhibit greater variation in sentiment than broadcast formats, that discussions of economic topics during 2023–2024 were more sentimentally negative when focused on inflation, and show how word embedding methods can be used to map how contested issues like housing are discussed across the ecosystem. In addition to our findings and data, we include open tools that others can use to expand this research agenda. Our approach demonstrates that podcasts can and should be treated as data.

Keywords: podcasts, media ecosystem, housing

¹ Scott Robert Patterson: scott.robert.patterson@umontreal.ca

Date Submitted: 2025-12-05

Introduction

Podcasts occupy a central place in political media. They reach large audiences, carry news and commentary, and give politicians, journalists and influencers space for extended conversation. Campaigns schedule podcast appearances alongside television and radio hits, and many listeners treat podcasts as their main source of political information. While scholars have acknowledged the importance of podcasts as a social and political phenomenon, little attention has gone towards cultivating podcasts as a source of digital trace data.

This paper makes two contributions. First, we illustrate how podcasts can be assembled into textual corpora with rich metadata for social scientific inquiry. Second, we provide an initial description of the Canadian podcast ecosystem through analysis of both metadata and content. Our metadata analysis reveals a heterogeneous, urban-concentrated ecosystem characterized by varied publication schedules and significant output concentration among frequent producers. The content analysis suggests that podcast ecosystems are more stylistically and substantively diverse than traditional audio media. Additionally, we demonstrate how podcast data can be used to map debates across a variety of policy issues. While our substantive results are limited to the Canadian context, our work makes it clear that podcasts are worthy of more systematic scholarly attention. To facilitate the development of this agenda, we include a series of open tools that can be used to build podcast datasets.

We proceed in five parts. First, we review existing literature on podcasts as a media format. Second, we discuss how the most common podcast distribution protocol - the RSS feed - is amenable to data collection because it is standardized, serialized, and rich with metadata. Third, we describe our dataset of Canadian podcasts. Fourth, we analyze the Canadian podcast ecosystem at three levels: podcast-level and episode-level metadata, as well as transcribed content. We conclude by discussing limitations and future research directions.

Background

Podcasts have become an important component of contemporary media systems, including for politics. Emerging in the early 2000s, the term “podcast” combines “iPod” and “broadcast,” reflecting the format’s origins in Apple’s distribution infrastructure (Hammersley 2004). Since then, podcast consumption has grown by fits and starts. Commentators often point to the “*Serial* effect” to describe the medium’s shift from niche to mainstream, a reference to the 2014 true-crime series that introduced millions of listeners to the format (Sherrill 2022). That growth has only accelerated in the 2020s. In the United States, nearly half of adults reported listening to a podcast in the past month in 2022, with 41 per cent tuning in to shows devoted to politics and government (Shearer et al. 2023). In Canada, 38 per cent of the population listened to at least one podcast per month in 2023, and the average listener devoted nearly six hours per week to the medium (Triton Digital 2024). In China, podcast usage has also expanded rapidly, reaching an

estimated 86 million users by 2021 and placing the country second only to the United States in audience size (Zhu and Yang 2025).

Politicians have adjusted accordingly. Appearances on popular podcasts now sit alongside television hits and radio interviews as standard items in campaign media strategies. In Canada, several candidates during the 2025 federal election campaign, including Mark Carney and Pierre Poilievre, used podcast interviews to reach specific demographic segments and voters less likely to follow traditional news. During the 2024 U.S. presidential campaign, Donald Trump and Kamala Harris likewise targeted key audiences through high-profile podcast appearances. In a similar vein, Tan (2025) shows that political leaders in Singapore use podcasts to soften their technocratic or elitist images, taking advantage of the informal and conversational tone that characterizes much of the medium.

Beyond public officials, podcast hosts and influencers have become political actors in their own right (Pehlivan et al. 2025). Their programs often blend news, commentary, and personal narrative, building parasocial relationships that can shape political attitudes and participation. The career of Dan Bongino—who hosted a popular political podcast prior to his appointment as deputy director of the FBI—is one prominent illustration of how podcast celebrity can intersect with formal political institutions.

Recognizing this growing importance, scholars have begun to examine podcasting as a social and political phenomenon. On the audience side, Chan-Olmsted and Wang (2022) document how listeners choose podcasts for the mix of informative and entertaining content. Johnson and McCall (2025) illustrate how podcast listeners exhibit high levels of trust with the hosts of their favorite shows. Another factor is that podcasts are hands-free and can be consumed on-demand, making it easy to listen while commuting, cleaning, or exercising (Groot Kormelink et al. 2025; Beckstead and Llinares 2025).

On the production side, scholars highlight podcasts' capacity to create a sense of intimacy between hosts and listeners. Perks and Turner (2019) argue that podcasting affords a more direct and personal relationship than many earlier broadcast formats. García de Torres et al. (2025) similarly note that podcast journalists not only convey information but also share emotions and experiences, echoing McLuhan's (1964:137) claim that auditory media are especially emotive and intimate. Podcasts typically differentiate themselves from broadcast radio through an informal, vernacular, and conversational tone, which in turn heightens the feeling of closeness between hosts and their audiences (Florini 2015; Swiatek 2018).

An emerging consensus suggests that podcasts should be treated as a distinct media format rather than a mere extension of radio (Chan-Olmsted and Wang 2022). Thematically, podcasts tend to serve niche interests, engage deeply with niche topics, and build communities around shared identities. At the same time, they also interact with legacy media, in that podcasters may surface

issues or scandals that traditional news organizations then investigate and amplify (Jiang and Shen 2023). This dual role as an alternative space for political talk and a feeder into mainstream news positions podcasts as a locus of contemporary political communication.

The same features that make podcasts attractive to listeners and producers can also foster fragmentation and polarization. As with other forms of online media, podcasts can reinforce the biases and narratives of the audiences that they cater to. Building on trends already visible in partisan broadcast outlets (Prior 2013), podcasts and other digital media may encourage selective exposure to ideologically congenial content, contributing to echo chambers and more polarized information environments (Steppat, Castro Herrero, and Esser 2022; Asker and Dinas 2019). The intimate, personalized character of podcast listening may also encourage individualized and affectively charged modes of political engagement (Bratcher 2022; Rae 2023).

Despite this growing body of work, the research program that treats podcasts systematically as a form of digital trace data is limited. This contrasts sharply with other digital media, such as microblogging platforms (e.g., X/Twitter or TikTok) or video platforms like YouTube, which have become standard sources for computational social science (Muenger et al. 2025). One of the main reasons is technical, in that the audio distribution of podcasts introduces friction at every stage of analysis. Collecting, storing, and processing large volumes of audio is resource-intensive; extracting content for analysis requires an additional transcription step; and methodological work in political science and communication has largely focused on text, not sound (Bridgman et al. 2025; Grimmer, Roberts, and Stewart 2022).

Notable exceptions exist. For example, Morris et al. (2019) initiated a project on podcast preservation, Lotfian and Busso (2019) used podcasts to build an emotional speech corpus, DeMets and Spiro (2025) conducted an analysis of podcast guest networks, and Brandt et al. (2023) used podcasts to study Russian narratives about their invasion of Ukraine. Closer to our work, Litterer et al. (2025) created a large corpus of podcast transcripts that span a two month period. Additionally, Vilceanu (2025) analyzes podcast reviews, and Dietrich et al. (2019) show that tonal shifts in oral argument are predictive of legal outcomes. Still, as a form of digital trace data, podcasts are underexplored.

Our approach for analyzing podcasts uses podcast metadata and transcriptions of audio files. In principle, audio-specific methods could also be useful. In practice, however, transcription remains the most straightforward way to integrate podcasts into existing text-as-data workflows. Until recently, large-scale transcription was prohibitively expensive for most research projects. This constraint has eased with the emergence of open-weight automatic speech recognition (ASR) models that can be run locally or in the cloud at relatively low cost. OpenAI's Whisper family of models is a prominent example ("OpenAI 2023). Released under a permissive Apache 2.0 license, Whisper models can be downloaded, deployed on standard consumer hardware, and scaled up on accelerator hardware when available. While compute and storage still carry costs,

open ASR models have driven transcription prices down to a level compatible with large-scale academic research.

What remains is the question of how to collect audio and metadata in a systematic, transparent, and reproducible way. Here, podcasts have an underappreciated advantage over many other media formats: they rely on a standardized distribution mechanism based on RSS feeds. RSS feeds expose episode-level information and direct links to audio files in a machine-readable form. Combined with modern ASR, this infrastructure makes podcasts amenable to the construction of large-scale text corpora for social scientific analysis.

The Podcast Distribution Model

In this section, we detail how technical features of the podcast distribution system can be manipulated to facilitate systematic data collection. By mimicking the functions of a podcast streaming application, researchers can collect and transcribe troves of freely-available podcasts for use in research. Moreover, with speech-to-text language models, researchers can build corpora from transcribed episodes.

Podcasts are distributed through a highly standardized protocol known as “Really Simple Syndication”, or RSS. An RSS feed is a text file written in XML that follows a consistent structure across platforms and hosting services. A podcast’s RSS file contains structured metadata about the podcast as a whole and about each published episode. When using a podcast streaming application, users search for content that is contained in the RSS files. Additionally, most podcast applications offer offline listening because RSS feeds contain download links to audio files. Podcasts can be collected for research by mimicking the functions of a podcast streaming application.

RSS feeds encode information about podcasts at two levels. The first level contains information about the podcast as a whole, such as a description of its primary theme, a link to the show’s website, the language used by producers, and other structured variables. We refer to this as the **podcast-level**. The second level contains information about individual episodes. **Episode-level** features include an episode title, a publication date, a description or summary of the episode’s contents, and a download link. Many RSS feeds include additional data, such as location, whether an episode has explicit language, and more. However, these can be less consistent than more common features. At both levels, content is stored in a predictable and machine-readable manner, which enables the same RSS feed to be recognized across different streaming applications.

The standardization of RSS feeds makes podcasts unusually amenable for data collection. Since the structure of the feed does not depend on website layout or proprietary platform features, all of the podcast’s freely-available digital traces can be collected with a simple parsing script. To support future work in this area, we developed a set of tools that allow researchers to move from

RSS-feed discovery to transcription without building a full pipeline from scratch. These tools are intended to make the collection process accessible and to lower the technical overhead required to work with podcast data.²

Metadata stored at the podcast-level and episode-level contain many useful features for analysis. However, to analyze the contents of podcasts, we recommend transcribing the audio files. We ran transcription by loading the Whisper model onto a GPU cluster that was accessible via the Digital Research Alliance of Canada. Using an open weight model on a public GPU cluster allowed us to run transcription at no direct cost. Transcription is possible on off-the-shelf hardware, including a CPU on a standard computer. However, transcribing a 40-minute podcast on a standard CPU can take more than 10 minutes, while the same can be done on a powerful GPU in less than 30 seconds.³ Our experience highlights the role of computing resources in building podcast corpora. The methods we present can be adapted to many contexts, but large-scale work depends on access to hardware that can handle significant volumes of audio.

Data and Methods

We expedited our data collection process by purchasing a seed list of Canadian podcasts from Feedspot (feedspot.com), a commercial media aggregator that compiles podcast directories from public RSS indexes and platform listings, on February 8, 2024. The list cost approximately USD \$200 and was delivered as a spreadsheet within 48 hours of the request. Researchers seeking to replicate this approach can contact Feedspot directly through their website, though comparable lists could also be assembled from Apple Podcasts charts, Spotify's public catalog API, or open directories such as Podcast Index (podcastindex.org). The purchased dataset contained RSS feed links for 6,055 podcasts with a Canadian location tag in the podcast-level metadata. From this list, we were able to extract 655,123 podcast episodes. Then, from within this group, we transcribed a subset of 15,546.

Table 1. Number of podcasts, episodes, and transcribed episodes.

Total Podcasts	Total Episodes	Transcribed Episodes
6,055 ⁴	655,123	15,546

² These tools are detailed in appendix 3.

³ Transcription can also be done with API-based speech-to-text models, though the user will incur costs. The approximate current costs for Whisper transcription via the OpenAI API are US\$ 0.006/minute, or US\$ 0.24 for a 40-minute podcast.

⁴ The number of active podcasts may change over time. These results reflected the number of active podcasts in February 2024.

Our description of the Canadian podcast ecosystem proceeds reflects these levels of analysis. First, we describe the ecosystem using variables that are nested at the **podcast** level of analysis. Second, we do the same using variables at the **episode** level of analysis. Third, we conduct a quantitative content analysis on the text of episode **transcripts**. This nested approach enriches our discourse analysis by shedding light on structural features, such as geography, language, and asymmetries in the presentation of policy themes.

Results

Our description of the Canadian podcast ecosystem proceeds according to our three levels of analysis: podcast, episode, and transcript.

Podcast-level Features of the Canadian Podcast Ecosystem

The podcast-level metadata reveal several structural features of the Canadian ecosystem. Categories are diverse (322 unique labels), though the top 10 account for 46% of podcasts; “Canada Podcasts” alone covers 30%, functioning more as a geographic marker than a topical label. The ecosystem is geographically concentrated in urban centres — Toronto, Vancouver, and Ottawa together account for approximately 19% of located podcasts — though inconsistent granularity in the Location variable (335 unique labels) limits deeper attribution. English dominates at 88%, while French podcasts are underrepresented at 11% (whereas francophones are 22% of Canada’s population). Finally, two-thirds of podcasters maintain at least one social media presence, with the majority active across multiple platforms. We provide further detail on each of these variables in Appendix A.

Episode-level Features of the Canadian Podcast Ecosystem

This section focuses on a description of episode-level podcast variables.⁵ How many episodes do podcasters tend to produce? Our findings suggest a mean of 120.3 episodes per podcast and a median of 51 episodes per podcast. The episode production concentration curve below illustrates that episodes are asymmetrically distributed across podcasters.

⁵ Additional descriptions of episode-level variables are presented in Appendix B.

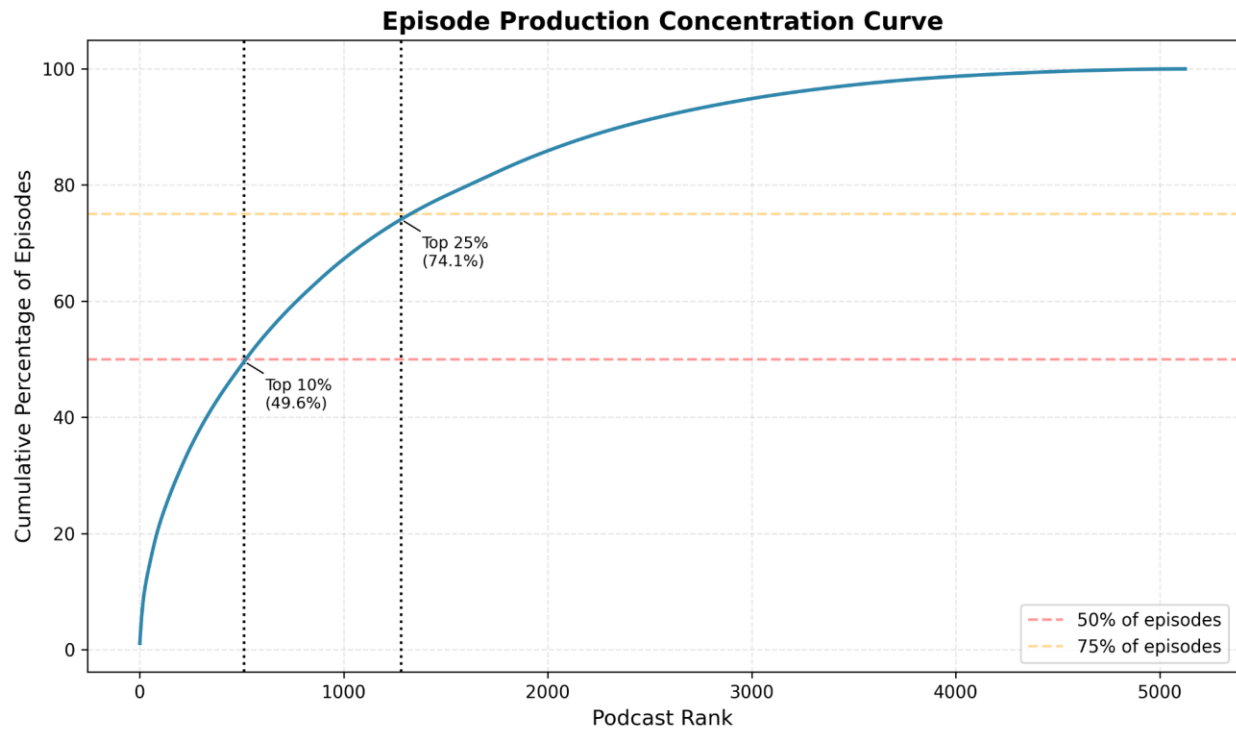


Figure 1. Episode production concentration curve.

The curve shows the cumulative percentage of episodes as a function of podcast rank (ordered from most to least prolific). The steep initial rise indicates substantial concentration: the top 10% of podcasts produce 49.4% of all episodes, while the top 25% produce 74.0% of episodes. This indicates that a small group of prolific podcasters produce the bulk of the ecosystem's content. It is possible that non-prolific podcasters release content on a less frequent basis or that they give up on podcasting after a short time.

How frequently do podcasters release podcasts? Our results illustrate the following distribution patterns.

Table 2. Episode release patterns.

Frequency	Percentage
Daily	12.4%
Several per week	4.6%
Weekly	46.1%
Bi-weekly	19.9%
Monthly	14.0%
Irregular	3.0%

The results indicate that the average podcast is released weekly, followed by bi-weekly, monthly, and daily schedules.

Thematic Patterns Among Top Producers

What kind of content do prolific podcasters produce? We address this by examining thematic patterns among top producers. The results are depicted below.

Table 3. Top themes for prolific podcasters.

Theme	Name	Number of Episodes
News & Commentary	Sophie Durocher	6,733
	Antoine Robitaille	3,446
	Rebel News	3,281
	Toronto Today with Greg Brady	2,802
	Franchement dit	2,400
	CBC Front Burner	1,926
Sports	Inside Sports with Reid Wilkins	5,000

	Oilers	3,000
	Edmonton Elks	2,763
	The Rod Pedersen Show	2,512
Religious	Back to the Bible Canada	2,194
	Jewish Faith & Jewish Facts	2,479
	indoubt Podcast	2,194
	Daily Torah Schmooze	2,139
Niche professional	Real Estate Espresso	2,853
	The Lead Pedal Podcast for Truck Drivers	2,363
	TSX Quarterly	1,946
Self-help	Optimal Health Daily	3,358
	Self Discovery Media Community	2,540

The most prevalent theme among top producers is *News & Commentary*, which includes many daily news podcasts. This is followed by sports, religious, and niche professional subjects. The diversity of content among top producers, spanning news, sports, religious programming, and niche professional audiences, demonstrates that sustained podcast production is not limited to a single institutional model or content strategy. Three of the five most prolific podcasts in the news and commentary category (*Sophie Durocher*, *Antoine Robitaille*, and *Franchement dit*) are podcast versions of QUB Radio programming, a network owned by the private broadcasting and telecoms conglomerate Quebecor. The podcast with the third-most episodes in this category is produced by Rebel Media, a right-wing extremist outlet funded through a combination of embedded advertising on platforms like YouTube that host their content, audience donations, and subscriptions (Elmer and Burton 2022). *Inside Sports with Reid Wilkins*, the podcast with the second-most episodes overall, is produced by CHED, an Edmonton radio station owned by Corus Entertainment. In the niche professional category, *Real Estate Espresso* is a short daily podcast by a real estate investor, while *The Lead Pedal Podcast* is part of an independent media organization which lists several sponsors within the trucking industry on its website. A fuller

analysis of patterns in ownership, funding, and business models in the wider podcast ecosystem is a promising topic for future research.

How has the Canadian podcast ecosystem developed over time? To investigate this, we examine the date for the first episode of each podcast in our dataset. The figure below illustrates the temporal distribution of podcast launches.

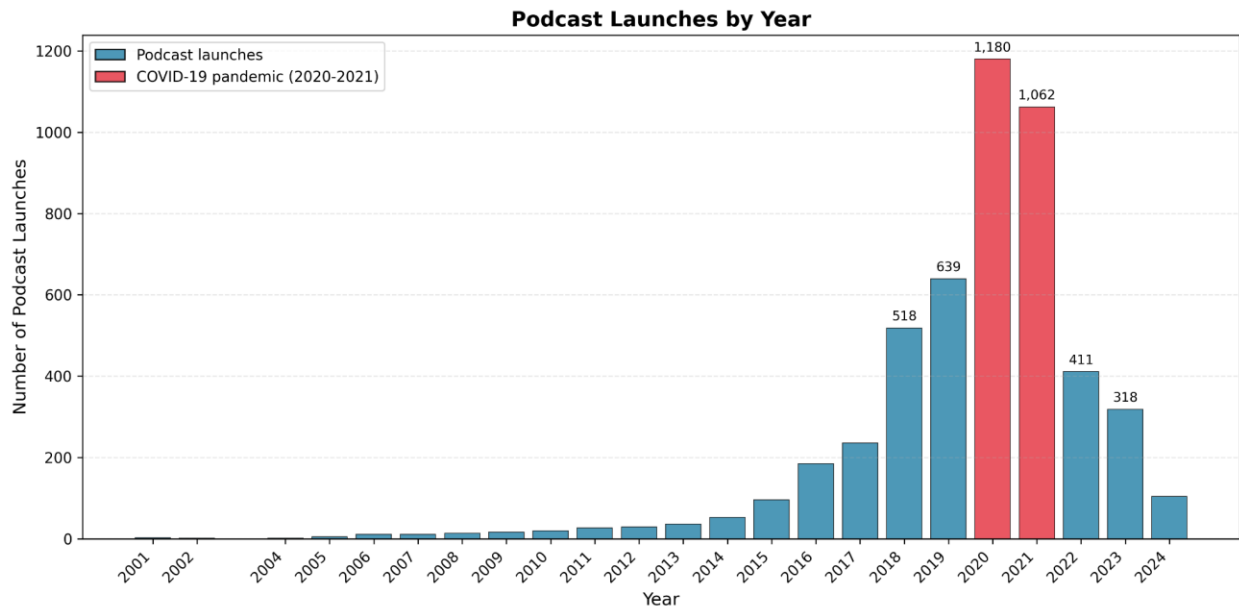


Figure 2. Podcast launches by year.

This figure reveals several distinct phases in the development of the ecosystem. Growth proceeded gradually through the 2000s and became exponential in the 2010s. Then, new podcasts spiked during the COVID-19 pandemic. The post-pandemic period saw a rapid decline in new podcasts, approximating levels of the mid-2010s. This suggests that the ecosystem may have become saturated during the pandemic era.

Episode-level metadata can be used to illustrate production patterns in the podcast ecosystem. Our findings illustrate that the Canadian ecosystem is populated with several prolific podcasts and many with more modest production schedules. Popular podcasts cover a range of themes, though prolific podcasters tend to cluster around just a few.

The variables discussed thus far are useful for understanding the structure of the Canadian podcast ecosystem. However, these measures are distant from the actual content of the episodes. In the next section, we illustrate how podcast and episode level metadata can be used together to identify and map the structure of substantive themes in the ecosystem.

Tracking substantive themes in the Canadian podcast ecosystem

This section shifts from the overall metadata to examining transcribed content from a subset of politics-focused podcasts. The analysis focuses on three questions about how political issues are discussed in podcast media: whether podcasts exhibit stronger sentiment than traditional broadcast formats; whether sentiment varies with economic policy discussions during a period of economic uncertainty; and how housing issues are discussed substantively.

The content analysis draws on transcripts from **6,213 episodes** from 53 podcasts selected from the larger dataset (full list in appendix 2). Two researchers collaboratively reviewed podcast names and descriptions, reaching consensus on those likely to discuss broad political or current affairs topics.⁶ These podcasts span both online-first productions (2,727 podcasts), audio recordings of traditional radio and television broadcast programs (2,645 broadcasts), and podcasts produced by major media outlets (841 “media podcasts”) enabling comparison between formats. This subset represents the portion of the ecosystem most directly engaged with political communication and policy debate. The date range covered in this analysis is January 2023 to July 2024.

Do podcasts have stronger sentiment than broadcasts?

Podcasts are varied in genre. The format most often associated with the medium’s current popularity is one where each episode is a long and wide-ranging interview with a single interviewee, or a conversation between like-minded hosts (Adler Berg 2023, Florini 2015). Other podcasts follow the traditional mass-media format of short segments that cover a range of topics within one episode. While the former are often conversational and informal in tone, the latter can be hard to distinguish from typical radio or television broadcasts. The distinction between these archetypes permits a basis for comparison.

The notion that podcasts cater to niches and social enclaves implies they are likely to express stronger sentiments than mass-media broadcasts targeted at broader audiences, which are likely to be more neutral and equivocal in sentiment (Bratcher 2022; Rae 2023). Media fragmentation is considered to be associated with political polarization, and a particular concern is how this contributes to affective polarization where individuals hold hostile feelings towards other members of society on the basis of political affiliation (Goel and Merkley 2024). In parallel, rising populism emphasizes antagonistic anti-elite discourse (Mudde 2014), and is associated with stoking anger and negative sentiment through segmented digital and social media (Yilmaz

⁶ This selection approach may under-represent political content from less recognizable or non-English-language podcasts. Researchers drew on familiarity with the Canadian media landscape; the resulting list (appendix 2) illustrates what was considered political or current affairs content for this purpose. The subset still enables meaningful comparison across formats and provides a well-defined basis for the content analysis questions under study.

et al. 2026), including against traditional broadcast media itself (Wasibord 2018). To the extent that some podcasts are part of this phenomenon of polarization and populism, we should find that podcast content is on average more strongly emotive – whether positively or negatively – than media aimed at broader audiences.

We explore these intuitions through a comparative sentiment analysis of podcasts, seeking to examine variations in the strength of positive and negative sentiment using natural language processing tools. We probe two widely-held notions about media. Firstly, we examine the notion that the content of media is intrinsically shaped by their formats and production. Secondly, we examine more closely the idea that podcasts cater to specific niches, and thus that some podcasts may be part of broader trends of emotionally-charged political polarization and populism. Among podcasts produced by major commercial and public media organizations, we also distinguished between those produced as online-first content (841 “media podcasts”), and those which are recordings of radio or TV programmes (2,645 “broadcasts”). This allows a comparison between the traditional broadcast and podcast formats, as well as between podcasts produced by major media companies and by independent outlets. For instance, *The House* is a traditional-format broadcast radio program focused on Canadian parliamentary politics, interviewing multiple politicians and pundits on a range of topics for each week’s episode, aired on CBC Radio at a regular time slot. In contrast, *Canadaland* is an online-first podcast that positions itself as non-mainstream and critical, including of the established news media. Its podcast description states, “We hold the powerful to account, and we scrutinize institutions and individuals that others won’t”. Illustrating the blurred line between these formats is *Front Burner*, a CBC production that adopts a podcast format of in-depth conversation between the host and a guest on a specific topic, and QUB Radio, an outlet owned by major private media company Quebecor which adopts a talk-radio format but is primarily available as an online live stream. These shows take the name of their regular host, such as *Antoine Robitaille*.

The sentiment analysis is carried out using the Vader module, distributed as part of Natural Language Toolkit (NLTK) for Python (Hutto and Gilbert 2014). Vader is a pre-trained model optimized to detect sentiment in social media posts, and produces three separate scores for negative, neutral, and positive sentiments expressed in a chunk of text. It is commonly used by businesses to monitor public opinion about their brand, classify product reviews and customer feedback, and by investment analysts to predict financial market movements based on news and corporate financial reports.

The analysis is conducted on transcripts of podcast episodes, with the score for each podcast being the average of all of its episodes in the sample. A comparison of podcasts based on whether they are produced by major media outlets or independently (such as by podcast-focused businesses like Canadaland or Rebel Media) also suggests traditional media tends to be more consistently measured and equivocal in sentiment, regardless of whether their content is online-first or primarily intended for radio or television broadcast. Conversely, independent podcasters

are sometimes more emotive. Table 4 reports that in two of the three sentiment models used, independent podcasters show a greater standard deviation in their average sentiment scores than podcasts produced by major broadcasters. The BERT-based Tabularis model shows that traditional media has slightly greater variation, although this result is driven by an outlier (see figure A1 in appendix). While thus not definitive, this overall analysis emphasizes the variation within rather than between categories, and cautions against attributing political polarization or populism to large and diverse categories of media formats or producers as some of the literature suggests.

Table 4. Standard deviations of podcast average sentiment scores, by model. “Major broadcasters” includes the broadcast and media podcast categories.

	Flair	DistilBERT (Tabularis)	Vader
Independent podcasters	0.520	0.550	0.0642
Major broadcasters	0.397	0.580	0.0518

Figure 4 shows the distribution of positive and negative sentiment scores by podcast, taken as the average of all their episodes in the sample, using the Vader model which outputs both positive and negative scores. This is presented as a kernel density estimation, with the contours representing a topographical map of how the respective types of podcasts cluster. Podcasts towards the top right of the figure show more of both positive and negative sentiment, while a podcast closer to the bottom left has on average lower scores for both positive and negative sentiment, implying a more formal and objective tone. This illustrates how podcasts created by independent, online-first producers may express stronger sentiment, with wider dispersion than media-produced podcasts.

Overlap between the peaks of each type suggests that most podcast content is similar in sentiment across independent producers or major media outlets. However, as indicated by the wider stretch towards the top and right of the diagram, some independent, online-first podcast media does tend to express stronger sentiment in both positive and negative directions. In the appendix we report similar results from repeating this analysis using different sentiment analysis models, namely the Tabularis multilingual sentiment analysis model, based on DistilBERT and designed to aid analysis of social media and customer feedback, and the Flair NLP library. These were chosen based on their ability to analyze texts that, like our podcast transcripts, concern general subjects and are in both English and French. Furthermore, these represent three major distinct NLP paradigms in current use. Visual comparisons at podcast-level using the Tabularis and Flair models are reported in the appendix.

The average scores for podcasts produced by major media organizations are more clustered than podcasts. This implies they take a more consistently equivocal and neutral tone than podcasts. This could be a function of having a more neutral and measured tone overall, or a journalistic practice of presenting multiple views on topics, which may result in positive and negative views cancelling each other out. One outlier is *All in a Day*, CBC Ottawa's afternoon drive radio broadcast, which scored as having high positive and very low negative sentiment. Independently produced podcasts show a greater range in their tone, with many being similar on sentiment measures to major media outlets but also dispersing towards the top and right of Figure 3, implying they tend to express stronger sentiment overall. An illustrative outlier here is *Uncolonized*, which had both highly positive and negative sentiment scores, and describes itself as "a funny and blunt podcast about race, politics, and culture".

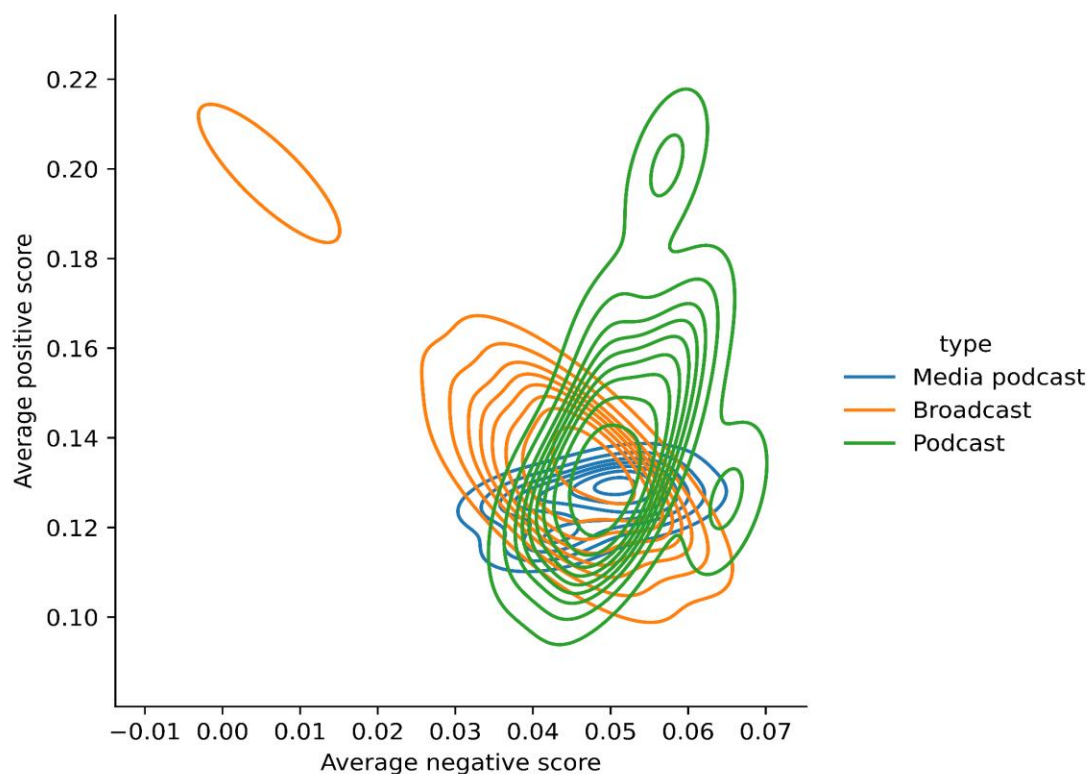


Figure 3. Podcast average Vader positive and negative scores (kernel density estimation).

An episode-level analysis, visualized in figure 4, gives greater weight to podcasts with more episodes, and is largely consistent with the findings based on podcast averages. Individual episodes of independent podcasts have a small skew towards being both more sentimentally positive and negative compared to content produced by major media organizations. These suggest that who produces the content appears to be a more important factor in its sentiment valence than whether the format is traditional radio or podcasting. The more informal, intimate, and personalistic characteristics that are typically associated with podcasts and podcast

journalism (Dowling 2024) thus can, but do not necessarily create a major difference from radio and TV broadcasts in their strength or direction of sentiments expressed.

Overall, these findings suggest that while the bulk of independent podcasts are similar in sentimental tone to those from major media outlets, they are more variable. This indicates the podcast format facilitates producing content with stronger overall sentiment, which some but not all independent producers take advantage of. Conversely, major media organizations, preferring to work within journalistic norms and seeking mass audiences, retain similar sentiment profiles between their podcasts and traditional broadcast productions.

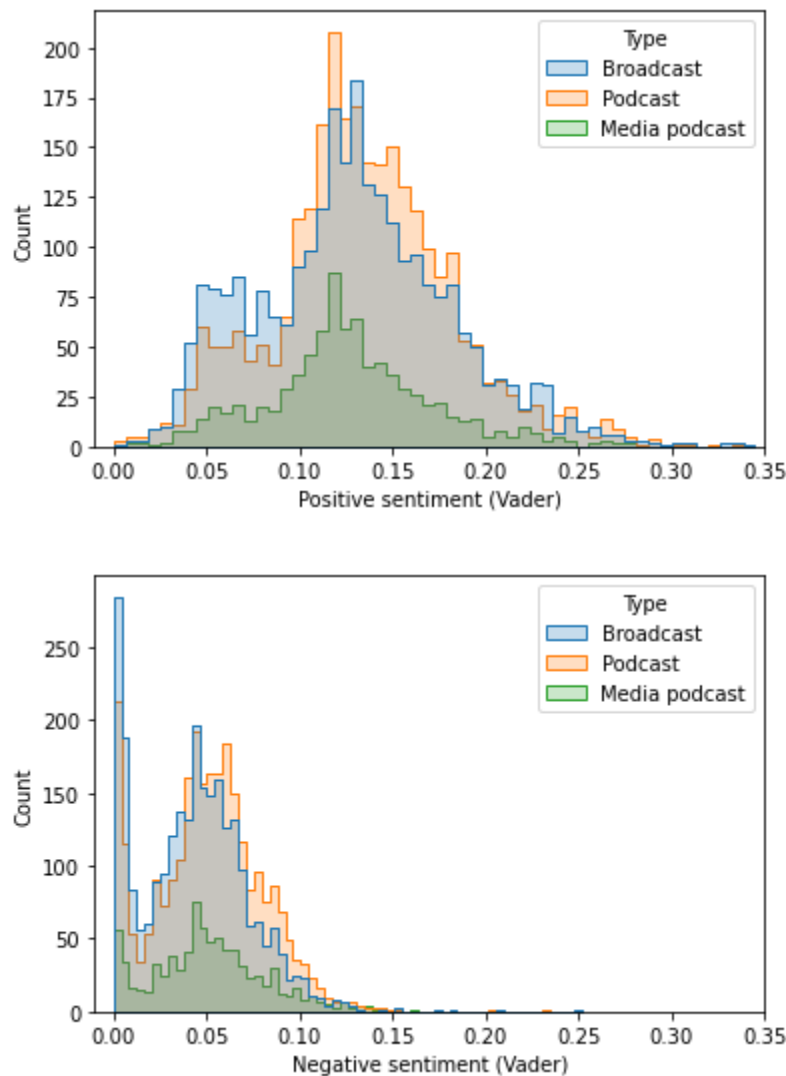


Figure 4. Histograms of positive and negative sentiment of politics podcasts, by episode.

Economic sentiment and policy topics

This section examines whether podcast episodes with more discussion of economic topics also exhibit more negative sentiment, and whether this is associated with the ideological stance of the episode overall. Popular perception about economic performance is known to be influential in public approval of incumbent governments and voting behaviour. This in turn is likely to be influenced by the media, particularly as individuals cannot directly observe macroeconomic performance at national scale. Research has examined how media reporting influences public perceptions of the economy, and how these may depart from the picture given by statistical indicators like inflation rates or gross domestic product (Kayser and Peress 2026). Perceptions about economic conditions are influenced by political partisanship in the U.S., with individuals identifying with the incumbent president's party tending to have more positive economic perceptions, and vice versa (Anson 2016; Prior, Sood, and Khanna 2015). Furthermore, economic expectations and sentiment among the general public are widely tracked through surveys of business, investor, and consumer sentiment. These factor into central bank decisions on monetary policy, as well as business decisions in the private sector. The significance of sentiment in economic and political dynamics is thus well-established in theory and practice.

The goal of this analysis is to descriptively examine whether discussions of economic topics are associated with negative (or positive) sentiment, if there is a difference in sentiment when discussions focus more on a particular topic (in this case, inflation), and whether this is associated with particular stances on economic ideology, as well as type of media producer. For instance, if sentiment is more negative when the topic of inflation is discussed relative to economic topics more broadly, this may indicate it is a particular focus of political discontent. Discussion may vary in positive or negative sentiment when it expresses different economic ideology or focuses on specific topics, particularly when economic policies or situations vary in their impact on different segments of the population, or are ambiguous, such as when strong GDP and wage growth occur at the same time as elevated inflation. The intention here is not to classify podcasts in terms of economic ideology or policies which they support, but to examine overall patterns in whether they are associated with identifiable skews in sentiment. The analysis also does not distinguish between speakers, but rather classifies podcast episodes overall, as it is reasonable to expect that listeners are exposed to entire episodes and therefore to any disagreement between speakers (which would tend to give an episode a lower score on both left and right ideology).

We utilized the Political DEBATE base model (Burnham et al. 2026), a natural language inference (NLI) classifier to run zero-shot classification of the same subset of 6,213 episodes of politics-focused podcasts to detect topics and ideology. To classify the podcast episodes by ideology, we specified the hypothesis “This text expresses {} economic ideology” with the possible classifications of left, neutral, or right. The results were then passed to a dichotomous variable as left or right based on which received the higher score. The topic classification used the hypothesis “This text is about {}” where the possible labels were economy, inflation, or other. The model classified 86 and 1264 episodes as having inflation and the economy as their topics, respectively. The ideology classification coded 4,330 episodes as expressing right-wing

and 1,885 left-wing stances. For sentiment scores we utilize FinBERT, a domain-specific model widely used to study financial and economic communications (Araci 2019), such as central bank speeches and corporate earnings calls.

The analysis suggests that episodes where discussions where inflation was the more prominent topic were slightly more negative in sentiment than discussions about the economy more broadly. While such wider discussions were more neutral, sentiment was negative if inflation was the main topic and in episodes expressing left-wing economic ideology. During the period covered by these podcast episodes, opinion polling and provincial election results in Canada demonstrated negative perceptions of the economic situation and incumbent governments of both left and right (Angus Reid 2024). Commentary in the wider global context postulated a wider international “vibecession” during the 2023-24 time period studied wherein negative economic sentiment reflected a disconnect in the measured and perceived state of advanced economies. This was often attributed to the early 2020s inflation spike (Burn-Murdoch 2023), although inflation in Canada had fallen to the conventional 2 per cent target by 2024 (OECD 2025). While the overall average sentiment was close to neutral for all episodes expressing more right-wing ideology, as well as for general economic topics on both left and right, expressions of left-wing ideology were more sentimentally negative when inflation was the main topic.

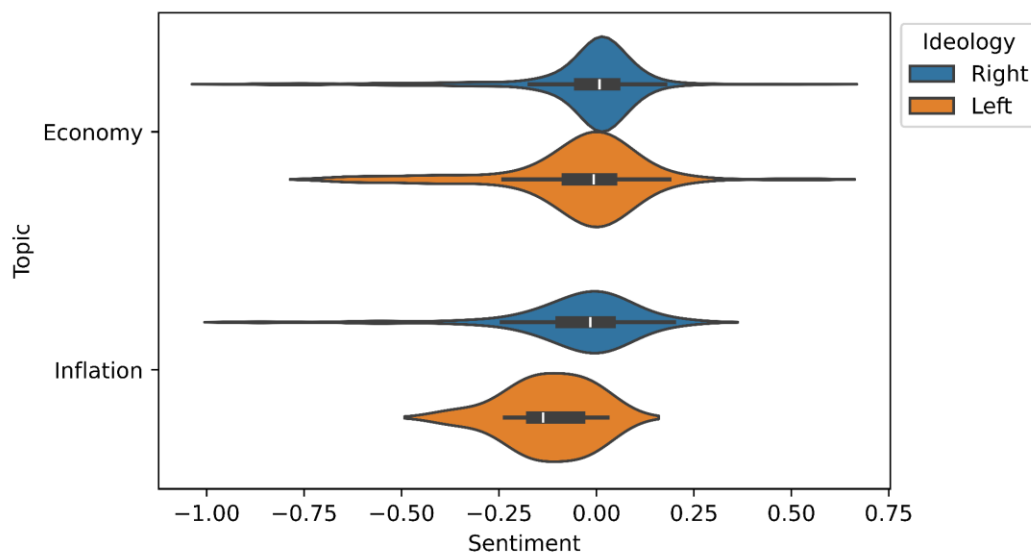


Figure 5. Violin plot showing distribution of podcast episodes by detected topic, economic ideology, and sentiment. Boxes and whiskers indicate quartiles.

An important caveat is that political discourse may be shifting in ways that are hard to capture on a left-right axis. While “left” ideological stances have traditionally been critical of economic elites and big business, populism has more recently seen right-wing politics adopt strident anti-elite rhetoric attacking institutions and corporations as well. In Canada, the Conservative Party leader vocally criticized the central bank for not being austere enough by facilitating increased fiscal spending and failing to reduce inflation, and in May 2022 promised to fire the Bank of

Canada governor. This analysis points to the significance of cost-of-living issues in political discussion, another aspect to which we now turn.

Probing a key issue with word embeddings

Word embeddings capture the proximity between words in a corpus, by representing them as vectors of numbers. Distances between words are “informative about the semantic similarity of underlying concepts they connote for the corpus on which they were built” (Rodriguez and Spirling 2022, 101), allowing researchers to probe and measure meanings and associations between words and concepts. Some practical uses include helping translation algorithms detect from context whether *avocat* in a French text should be translated into English as *avocado* or *lawyer*. Research has also used these features to identify biases that are encoded in the everyday discourse which forms the corpora of natural language processing, for instance in quantitatively tracing changing associations between occupations and gender over time (Durrheim et al. 2023; Petreski and Hashim 2023). Using a similar logic, this corpus of podcasts produced in Canada can identify how political topics were discussed through the semantic associations and relationships between words and concepts.

We examine here the topic of housing, which was a major political issue in Canada during the data collection period in 2024. The high cost of renting or buying a home has long been a hot topic in Toronto and Vancouver, which saw a disconnection between local labour income and housing prices before 2020 arguably driven by housing investment and financialization (Gordon 2020, Reisenbichler 2025). However, in the post-pandemic period, rising rents and sale prices became a wider problem even in cities that had been relatively insulated from this dynamic. During and after the Covid-19 pandemic, homelessness also became increasingly visible in Canada, with many cities seeing tent encampments (Flynn et al. 2022). By 2024, the topic was increasingly prioritized in nationwide political discussion, with one poll finding that 88 per cent of Canadians were concerned with housing accessibility and affordability (Sheppard and Coletto 2024). The global relevance of housing as a political issue should also be noted, as housing unaffordability and its economic implications significantly contribute to polarization in Europe (Ansell & Cansunar 2021). The salience of housing as a policy issue makes it a suitable proof of concept for this technique.

Using the concordance tools in NLTK, specifically the Word2Vec model, allows us to produce a two-dimensional map of the word embeddings relevant to “housing” in the corpus (figure 6). The relative distance of different words to the centre can be interpreted as depicting how closely they were associated with housing in the corpus. Words further away from the centre or not appearing on the diagram at all are thus less connected to housing within the discourse of this sample of politics podcasts in Canada. The axes have no analytical significance, rather it is relative distances from the centre and between words that matter. This is unlike a word cloud based on frequency of words, instead, the distance of words from the centre and relative positions or clusters of words indicate their contextual and semantic proximity within the corpus of politics-focused podcast transcripts.

Notably, “investment” is closer to the centre than “affordable” and “affordability”, suggesting that this framing is more central than housing as a social issue even outside of real estate-specific discussions, which we further explore in the following section. Other words close to the centre include “infrastructure”, “plan”, “funding”, and a cluster of “government”, “policy”, and “solution”, suggesting discussions also framed housing as an infrastructural issue that demanded more or improved government policy interventions. Towards the outside of the diagram are words that also suggest the other issues that housing was frequently linked to, notably health and immigration. The presence of “healthcare” likely captures housing being frequently mentioned alongside other major policy moves in 2024, when the Canadian government was implementing universal dental care and pharmaceutical insurance. “Immigration” highlights that elevated immigration levels after 2021 were often claimed to be a contributing factor in rising housing prices.

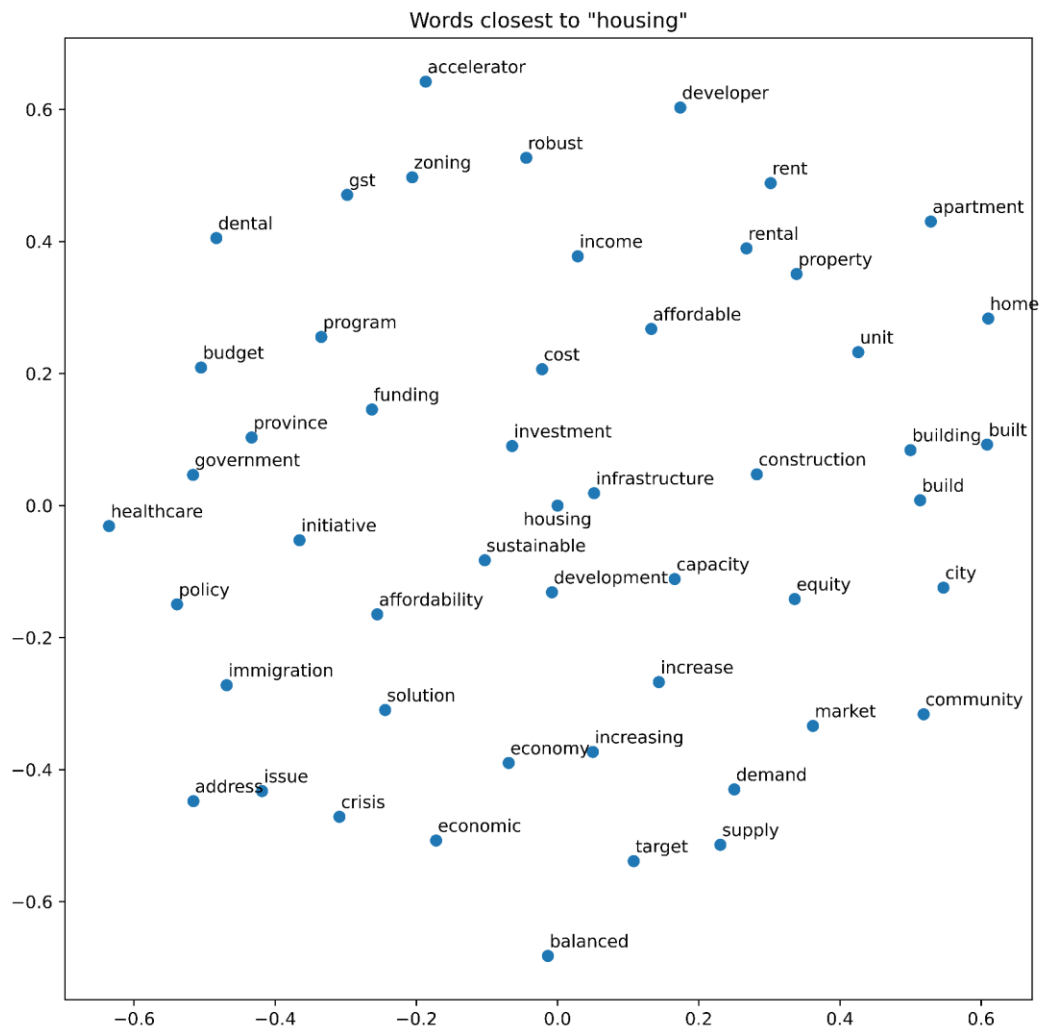


Figure 6. Word embeddings analysis, centred on “housing”.

These insights, while based on prior knowledge of the political discourse at the time, can also be verified with a keywords-in-context approach. The word “housing” appeared in the corpus 19,501 times. A random sample of 25 occurrences of the word together with the 200 characters coming before and after it is presented in Table 11. “Affordable”, “affordability”, and “cost” are found in close proximity, highlighting that this was a major concern whenever housing was mentioned. The situation is frequently described as a “crisis” and linked closely to both pocketbook and macroeconomic trends. In the sample, a pundit notes that politicians are connecting the housing issue to inflation and grocery prices.

Housing is also frequently presented as a policy failure, which governments are responsible for correcting. One occurrence in the sample frames housing as “infrastructure,” and the close proximity of this word to housing in the overall corpus suggests the government’s proper role was often presented as being similar to that in roads and utilities. At the same time, words like investment, equity, and market are slightly further away but reflect that finance and its effects on housing supply are also important parts of the discussion. Overall, the analysis suggests that podcast content tended to discuss active government intervention more frequently than market-oriented policy approaches.⁷

Housing as investment or social issue in the wider podcast ecosystem

We now further demonstrate possibilities of these techniques by returning to the wider sample to examine the broader discussion of housing issues in the wider Canadian podcast ecosystem. We identify housing discourse in the Canadian podcast ecosystem through a two step LLM-based classification pipeline. In the first step, we use an LLM as a binary classifier to determine if podcast descriptions and episode descriptions explicitly discuss housing as a theme.⁸ Then, for positive cases, we use an additional classifier to determine if these variables discuss housing through the lens of *real estate* or through the lens of *social issues*. In doing so, we are able to identify housing discourse in the podcast ecosystem and to gauge which perspective is more prominently represented.⁹ In line with the emerging consensus around the use of LLMs in social science research (Barrie, Palmer, and Spirling 2025), we adopted the open weight Cohere R7B (Cohere 2024) and Cohere Command-A (Cohere 2025) models.

After completing the classification and conducting multiple validation checks, we can segment housing podcasts based on if and how they discuss the housing theme. The first group comprises podcasts that mention housing as a *primary focus* in their podcast description. The second group comprises podcasts that do not mention housing in their podcast description but occasionally release housing-themed episodes. Finally, within each group, we then distinguish between episodes that adopt the “real estate” and “social issues” frames.

The figure below illustrates the size and composition of these two groups. The x axis refers to the cumulative percentage of housing-focused podcasts. The y axis refers to the cumulative share of housing-focused episodes. The two coloured and shaded Lorenz-style curves indicate the respective results for the two groups. Darker red shading indicates a higher proportion of “real estate” focused episodes.

⁷ Please see Appendix E for a more granular view of housing discourse.

⁸ For clarity, the analysis in this section is based on podcast and episode description variables, not on transcripts.

⁹ Additional information about our LLM procedures and quality control measures are included in appendix 4.

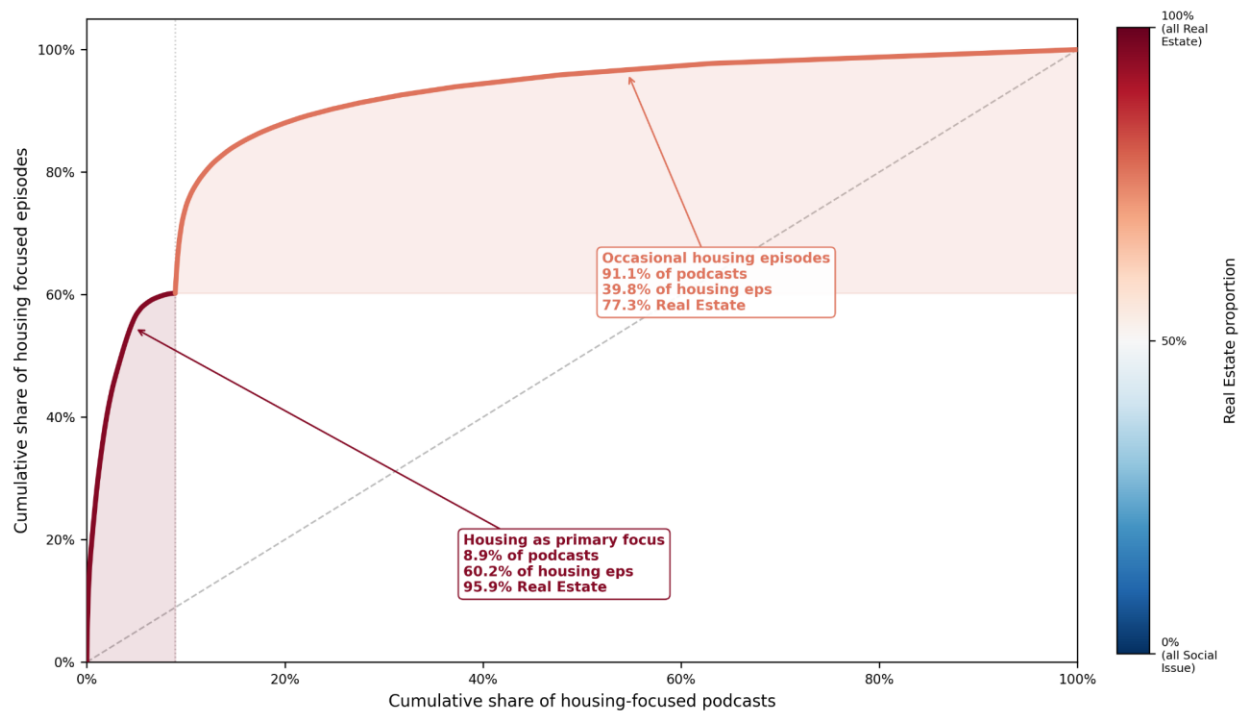


Figure 7. Distribution and substantive themes in housing podcasts.

Figure 7 illustrates several structural features of housing discourse in the Canadian podcast ecosystem. First, the long-tailed distribution pattern indicates the presence of a small group of prolific housing podcasters and a much larger group of occasional housing podcasters. The podcasts with an explicit housing focus are responsible for about 60 per cent of all housing episodes, while the remaining 40 per cent of housing focused episodes are spread over more than 1,200 podcasts. These groups are not perfectly segmented, in that several prolific housing podcasters do not mention housing in their podcast description. Some examples that fall into this group include general investing podcasts that often discuss real estate. Nevertheless, the overall pattern is clear.

Strikingly, real estate podcasts dominate housing discourse in the Canadian podcast ecosystem. Among housing-focused podcasts, almost 96 per cent of the episodes adopt a real estate framing. In the secondary group, the percentage drops, but only to 77 per cent. Taken together, the data suggest that housing discourse in the Canadian podcast ecosystem is concentrated among a small number of producers and overwhelmingly filtered through the lens of real estate investing.

It is important to note that “real estate” and “social issues” are not mutually exclusive categories. For example, a social issue focused podcast could release an episode on abusive investment practices or a real estate podcast could release an episode on strategies for first-time homebuyers.

To probe the finding above, we display podcasts based on the proportion of episodes that engage with real estate versus social issue themes.

The figure below illustrates the ratio of “real estate” to “social issue” framed podcast episodes produced in the dataset. The x -axis illustrates the number of real estate episodes, while the y -axis illustrates the number of social issue episodes. Both are presented on a log scale. The colour is mapped to the proportion, such that darker red podcasts have a higher ratio of real estate episodes, darker blue indicates more social issue episodes, and transparent indicates an even match. Finally, podcasts with housing as a primary focus are labelled with stars, while the secondary group is labelled with dots.¹⁰

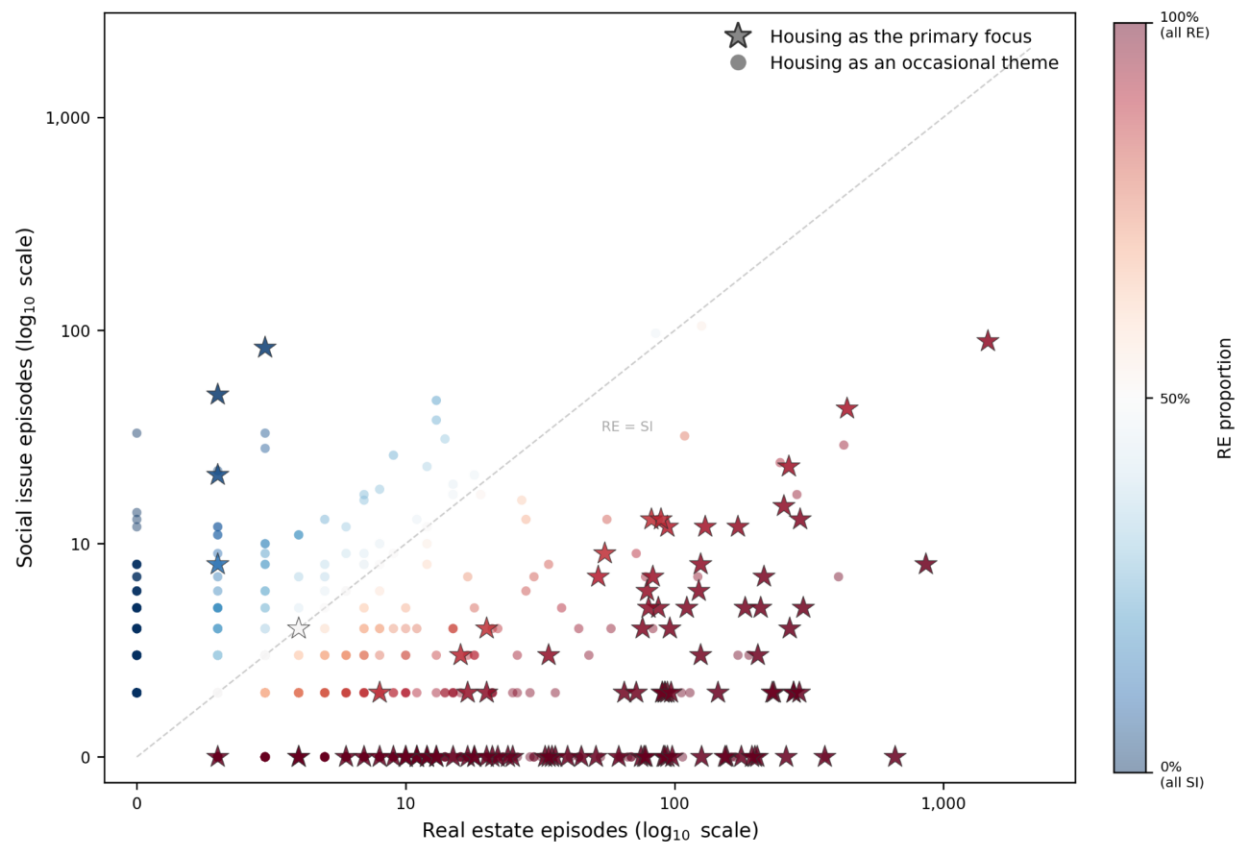


Figure 8. Thematic focus among housing podcasts.

The figure underscores the dominance of real estate perspectives in housing podcasts. There are only four dedicated housing podcasts that focus on social issues, and one has fewer than 10 episodes. The most prolific of these released fewer than 100 episodes. In this sense, the social issue frame is more prominent in podcasts with broader topical coverage that do not focus

¹⁰ Note: 69% of podcasts ($n = 975$) have fewer than five housing-focused episodes and cluster near the origin; overplotting in this region means density is underrepresented visually.

exclusively on housing issues. This contrasts with the many housing-focused podcasts on the real estate side, including the many on the lower horizontal axis that never engage with social issues. Many real estate podcasts have released between 100 and 1,000 episodes, with one alone producing more than 1,400 episodes. Interestingly, many real estate focused podcasts appear to engage with social issues, including many with 10 to 100 episodes. Still, it is clear that real estate investing is a far more prominent framing.

This vignette illustrates an effective technique for locating substantive themes and mapping their structure within a podcast ecosystem. In addition to generating substantive findings, this technique could also be useful for researchers who need to determine which podcast episodes to transcribe given their topical interests, given the costs associated with podcast transcription.

Conclusion

This study demonstrates that podcasts are both salient and accessible as sources of data about political media and communications. Our analysis proceeded through podcast-level metadata, episode-level metadata, and transcribed content, revealing both the structural characteristics of the Canadian podcast ecosystem and the discourse patterns within audio programming focused on politics, news, and current affairs. The metadata analysis establishes that Canadian podcasting is substantively diverse, geographically concentrated in major urban centers, predominantly English-language, and characterized by varied publication schedules with significant output concentration among high-frequency producers. Additionally, we demonstrated how podcast and episode level metadata can be used in tandem to track substantive themes through the ecosystem. The content analysis shows that online-first podcasts exhibit greater sentiment variation than broadcast formats, that discussions of economic topics during 2023–2024 were more sentimentally negative when focused on inflation, and demonstrated how word embedding methods can be used to map how contested issues like housing are discussed across the ecosystem.

However, the study also exhibits several important limitations. One limitation concerns data coverage. Since our seed dataset was purchased in 2024, we cannot claim to have full coverage of the Canadian podcast ecosystem. Moreover, our analysis is limited to freely available podcasts, which excludes much of the premium content behind paywalls. Future research would require us to probe these bounds for additional content.

For the metadata analysis, inconsistency and vagueness of raw data limit are limiting factors. As the housing analysis demonstrates, these limitations can be navigated by curating and bolstering additional variables, but this remains a labour-intensive endeavor. Moreover, our transcribed corpus represents only a narrow slice of the podcast landscape—overtly political podcasts produced in Canada during a specific time window. This coverage is thus illustrative rather than comprehensive. As in other forms of media research, questions of inclusion and exclusion shape

what can be inferred from the data. Expanding the boundaries of analysis will require more computational resources, since converting large volumes of audio into text is time and energy intensive and best performed on accelerator hardware such as GPUs. Future work could experiment with lighter transcription models, distributed processing, or shared research infrastructure to extend coverage across time and topic without prohibitive cost.

This study also focuses on what is said within podcasts, not on how these messages circulate. The Canadian podcast ecosystem is deeply intertwined with global flows of media. Canadians listen to shows produced abroad, and Canadian podcasts reach international audiences. Understanding the political role of podcasts therefore requires attention to patterns of distribution as well as content. Future work could integrate download data, cross-platform link sharing, or social media traces to capture how audiences encounter and spread podcast material. Studying these circulation dynamics would clarify not only what is said, but to whom and with what reach. From the perspective of politics and media, Canadian politics has been characterized by governments and parties communicating with the public using intense, repetitive, and tightly controlled messaging, adapting branding strategies from consumer industries (Marland 2016). Future research could examine if the recent popularity of the podcast format will challenge this centralized, monolithic approach to communications by political parties, or only create new channels through which to impose it.

The metadata-first approach adopted here reveals production patterns without requiring transcription of the entire corpus. This two-stage method of describing the ecosystem through RSS feeds before analyzing transcribed content offers a template for future work. Researchers can map podcast landscapes at scale through metadata analysis, then strategically select subsets for intensive content analysis based on research questions. This approach balances analytical opportunities while managing computational costs.

Finally, transcription inevitably reduces podcasts' complexity. Converting speech to text omits features such as tone, rhythm, and emphasis that carry political meaning. Prior research shows that nonverbal cues in voice can influence perceptions of authority, empathy, and credibility. While text-based methods have clear advantages—they are computationally efficient and compatible with established natural language processing tools—future research could analyze audio directly. As Dietrich et al. (2019) have shown, acoustic features like tone can be politically informative, even when decoupled from the content of a message. Examining acoustic features alongside text would restore some of the dimensionality lost in transcription and open new avenues for studying affect and performance in political communication.

This study shows that podcasts can and should be treated as data. Our analysis of Canadian political podcasts demonstrates the kinds of observations that become possible once audio and metadata are organized into a corpus. At the same time, the work highlights how much of the medium has yet to be studied. Podcasts reach sizable audiences, cover a wide range of political

topics and operate across many parts of the media environment. Building a broader understanding of this space will require larger datasets, more computing resources, and shared tools that expedite collection and processing. This creates an opening for a sustained research agenda that incorporates podcasts into routine empirical study and treats them as a regular part of political communication.

Acknowledgements

A version of this paper was presented at the *Aligning Computational Tools for the Political Science Research Lifecycle* meeting for the 2026 APSA Virtual Research Group. We would like to thank Patrick Brandt for providing comments on this draft.

Funding information

Initial funding for this project was provided from the Media Ecosystem Observatory of the Max Bell School of Public Policy at McGill University. Funding was provided through a doctoral fellowship to Scott Patterson.

This work was supported by compute credits from a Cohere Labs Research Grant.

References

- Adler Berg, Freja Sørine. 2025. "Analysing podcast intimacy: four parameters", *Convergence*, 31(4): 1423–38.
- Angus Reid. 2024. "Canadians feeling the financial heat this summer as housing pain intensifies; few see prices for essentials falling", Angus Reid Institute, July 8. <https://angusreid.org/economic-outlook -canada-summer-2024/>
- Ansell, Ben, and Asli Cansumar. 2021. "The political consequences of housing (un)affordability", *Journal of European Social Policy*, 31(5): 597–613.
- Anson, Ian G. 2016. "Just the Facts? Partisan Media and the Political Conditioning of Economic Perceptions." *Political Research Quarterly* 69(3): 444–56. doi:10.1177/1065912916648016.
- Araci, Dogu. 2019. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models" (MSc diss., University of Amsterdam), <https://arxiv.org/abs/1908.10063>
- Asker, David, and Elias Dinas. 2019. "Thinking Fast and Furious: Emotional Intensity and Opinion Polarization in Online Media", *Public Opinion Quarterly*, 83(3): 487–509.
- Barrie, Christopher, Alexis Palmer, and Arthur Spirling. 2025. "Replication for Language Models Problems, Principles, and Best Practices for Political Science." Preprint, November 4. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf
- Best, Jacqueline. 2024. "Central Banks' Knowledge Controversies", *New Political Economy*, 29(6): 857–71.
- Brandt, Jessica, Bret Schafer, Valerie Wirtschafter, and Peter Benzoni. "Echoes Across the Airwaves: How Kremlin Narratives about Ukraine Spread (or Don't) on US Political Podcasts." Brookings Institute, September 2023. <https://www.brookings.edu/articles/echoes-across-the-airwaves/>.
- Bratcher, Tegan R. 2022. "Toward a Deeper Discussion: A Survey Analysis of Podcasts and Personalized Politics." *Atlantic Journal of Communication* 30(2): 188–99. doi:10.1080/15456870.2020.1862846.
- Bridgman, Aengus, Alexei Abrahams, Thomas Bergeron, et al. 2024. "The Canadian Information Ecosystem." Preprint, January 5. <https://doi.org/10.31219/osf.io/b29q8>.

- Burnham, Michael, Kayla Kahn, Ryan Yang Wang, and Rachel X. Peng. 2026. "Political DEBATE: Efficient Zero-Shot and Few-Shot Classifiers for Political Text", *Political Analysis*, 34(3): 329–43.
- Burn-Murdoch, John. 2023. "Should we believe Americans when they say the economy is bad?" *Financial Times*, December 1.
- Burst, Tobias, Pola Lehmann, Simon Franzmann, Denise Al-Gaddooa, Christoph Ivanusch, Sven Regel, Felicia Riethmüller, Bernhard Weßels, and Lisa Zehnter. 2024. manifestoberta. Version 56topics.sentence.2024.1.1. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB) / Göttingen: Institut für Demokratieforschung (IfDem). doi:10.25522/manifesto.manifestoberta.56topics.sentence.2024.1.1
- Chan-Olmsted, Sylvia, and Rang Wang. 2022. "Understanding podcast users: Consumption motives and behaviors", *New Media & Society*, 24(3): 684–704.
- Clifton, Ann, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. "100,000 podcasts: A spoken English document corpus." In *Proceedings of the 28th International Conference on Computational Linguistics*.
- Cohere. 2004. "Introducing Command R7B: Fast and efficient generative AI." Cohere Blog. December 13, 2024. <https://cohere.com/blog/introducing-command-r7b>.
- Cohere. 2025. "Command A: An Enterprise-Ready Large Language Model." *Arxiv*. Preprint, April 7. <https://arxiv.org/pdf/2504.00698>
- DeMets, Sydney A., and Emma S. Spiro. 2025. Podcasts in the periphery: Tracing guest trajectories in political podcasts. *Social Networks*, 82: 65-79.
- Dietrich, Bryce J., Matthew Hayes, and Diana Z. O'Brien. 2019. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech," *American Political Science Review*, 113(4): 941–62.
- Dowling, David. 2024. *Podcast Journalism: The Promise and Perils of Audio Reporting*. Columbia University Press.
- Durrheim, Kevin, Maria Schuld, Martin Mafunda, and Sindisiwe Mazibuko. 2023. "Using Word Embeddings to Investigate Cultural Biases", *British Journal of Social Psychology* 62(1): 617–29. doi:10.1111/bjso.12560

- Elmer, Greg, and Anthony Burton. 2022. "Rebel personalities: Canada's far-right media", *First Monday* 27(9). doi:10.5210/fm.v27i9.12546
- Florini, Sarah. 2015. "The Podcast 'Chitlin' Circuit': Black Podcasters, Alternative Media, and Audio Enclaves." *Journal of Radio & Audio Media* 22(2): 209–19. doi:10.1080/19376529.2015.1083373.
- Flynn, Alexandra, Joe Hermer, Caroline Leblanc, Sue-Ann MacDonald, Kaitlin Schwan, and Estair Van Wagner. 2022. *Overview of Encampment across Canada: A Right to Housing Approach*. Canadian Human Rights Commission.
- Goel, Natasha, and Eric Merkley. 2025. "The nature of online talk: Incivility of opposing views and affective polarization," *Political Behavior*, 47: 1483–1500.
- Gordon, Joshua C. 2020. "Reconnecting the Housing Market to the Labour Market: Foreign Ownership and Housing Affordability in Urban Canada." *Canadian Public Policy* 46(1): 1–22.
- Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.
- Hammersley, Ben. 2004. "Audible Revolution", *The Guardian*, February 12. <https://www.theguardian.com/media/2004/feb/12/broadcasting.digitalmedia>
- Hutto, Clayton, and Eric Gilbert. 2014. "Vader: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 216–25. <https://ojs.aaai.org/index.php/icwsm/article/view/14550>.
- Jiang, John, and Michael Shen. 2023. "Traditional Media, Twitter, and Four Business Scandals", *Journal of Quantitative Description: Digital Media*, 3, doi:10.51685/jqd.2023.016
- Johnson, Kristine, and Michael McCall. 2025. "Trust in Pod: Listener Trust of News Content Heard on Different Genre Podcasts", *Media & Communication*, 13, doi:10.17645/mac.9182
- Kayser, Mark Andreas, and Michael Peress. 2026. "Do Voters Respond to the Economy or to News Reporting on the Economy? A Mediation Analysis", *Journal of Politics*, 88(1): 267-81.
- Litterer, Benjamin, David Jurgens, Dallas Card. 2025. "Mapping the Podcast Ecosystem with the Structured Podcast Research Corpus." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

- Lotfian, Reza and Carlos Busso. 2019. "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings." *IEEE Transactions on Affective Computing*.
- Marland, Alexander J. 2016. *Brand Command: Canadian Politics and Democracy in the Age of Message Control*. UBC Press.
- McLuhan, Marshall. 1964. "Radio: The Tribal Drum", *AV Communication Review*, 12(2): 133-145.
- Morris, Jeremy Wade, Samuel Hansen, and Eric Hoyt. 2019. "The PodcastRE project: Curating and preserving podcasts (and their data)." *Journal of Radio & Audio Media*: 26(1): 8-20.
- Moschella, Manuela. 2024. *Unexpected Revolutionaries: How Central Banks Made and Unmade Economic Orthodoxy*. Cornell University Press.
- OECD. 2025. *OECD Economic Surveys: Canada 2025*. Organisation for Economic Co-operation and Development. <https://doi.org/10.1787/28f9e02c-en>
- Pehlivan, Zeynep, Saewon Park, Alexei Sisulu Abrahams, Mika Desblancs-Patel, Benjamin David Steel, and Aengus Bridgman. 2025. "Building a Media Ecosystem Observatory from Scratch: Infrastructure, Methodology, and Insights." arXiv:2506.10942. Preprint, arXiv, June 12. <https://doi.org/10.48550/arXiv.2506.10942>.
- Perks, Lisa Glebatis, and Jacob S. Turner. 2019. "Podcasts and Productivity: A Qualitative Uses and Gratifications Study", *Mass Communication & Society*, 22(1): 96-116.
- Petreski, Davor, and Ibrahim C. Hashim. 2023. "Word Embeddings Are Biased. But Whose Bias Are They Reflecting?" *AI & Society* 38(2): 975–82. doi:10.1007/s00146-022-01443-w.
- Prior, Markus. 2013. "Media and Political Polarization", *Annual Review of Political Science*, 16: 101-127.
- Prior, Markus, Gaurav Sood, and Kabir Khanna. 2015. "You Cannot Be Serious: The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions." *Quarterly Journal of Political Science* 10: 489–518.
- Rae, Maria. 2023. "Podcasts and Political Listening: Sound, Voice and Intimacy in the Joe Rogan Experience." *Continuum* 37(2): 182–93. doi:10.1080/10304312.2023.2198682.
- Reisenbichler, Alexander. 2025. *Through the Roof: Housing, Capitalism, and the State in America and Germany*. Cambridge University Press.

- Rodriguez, Pedro L., and Arthur Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *The Journal of Politics* 84(1): 101–15. doi:10.1086/715162.
- Sheppard, Eddie, and David Coletto. 2024. "Housing concerns grow as over half of Canadians fear losing their home due to financial changes", Abacus Data, December 13, <https://abacusdata.ca/housing-concerns-grow-as-over-half-of-canadians-fear-losing-their-home/>
- Sherrill, Lindsey A. 2022. "The 'Serial Effect' and the True Crime Podcast Ecosystem", *Journalism Practice*, 16(7): 1479-1494.
- Statistics Canada. 2022. "While English and French are still the main languages spoken in Canada, the country's linguistic diversity continues to grow", *The Daily*, August 17, <https://www150.statcan.gc.ca/n1/daily-quotidien/220817/dq220817a-eng.htm>
- Steppat, Desiree, Laia Castro Herrero, and Frank Esser. 2022. "Selective exposure in different political information environments – How media fragmentation and polarization shape congruent news use", *European Journal of Communication*, 37(1): 82-102.
- Swiatek, Lukasz. 2018. "The Podcast as an Intimate Bridging Medium," in *Podcasting: New Aural Cultures and Digital Media*, edited by Dario Llinares, Neil Fox, and Richard Berry, 173-187. Springer.
- Tan, Kenneth Paul. 2025. "Podcasting Politics in Singapore: Hegemony, Resistance, and Digital Media", *Critical Asian Studies*, 57(2): 211-235.
- Triton Digital. 2024. *Canadian Podcast Report*. www.tritondigital.com.
- Vilceanu, M. Olguta. 2025. "Parasocial Intimacy, Change, and Nostalgia in Podcast Listener Reviews", *Media & Communication*, 13, <https://doi.org/10.17645/mac.9059>.
- Waisbord, Silvio. 2018. "Populism as media and communication phenomenon", in *Routledge Handbook of Global Populism*, 221-34. Routledge.
- Werner, Annika, Onawa Lacewell, Andrea Volkens, Theres Matthieß, Lisa Zehnter, Leila van Rinsum. 2021. *Manifesto Coding Instructions*, 5th re-revised edition. Comparative Manifesto Project, https://manifesto-project.wzb.eu/down/papers/handbook_2021_version_5.pdf
- Yilmaz, Ihsan, Ana-Maria Bliuc, Robb Norrie, Chloe M. Smith, James Smith, Cerys Evans, Daniel S. Courtney and Daniel Barnett. 2026. "Populism in the age of social media: A systematic

review of recent digital populism research”, *New Media & Society*, advance.
doi:10.1177/14614448261420218

Zhu, Ying, and Yali Yang. 2025. “The ‘renaissance’ of auditory medium in China: a study on the motives of Chinese podcasting users and corresponding development strategies of Chinese podcasting”, *Media International Australia*, advance. doi:10.1177/1329878X241302499.

Appendices

Appendix A: Podcast-level Variable Distributions

The table below depicts key variables that we collect at the podcast level of analysis. While we aggregated most of these variables from podcast RSS feeds, our variable list is supplemented with descriptive features that were included in our purchased dataset, such as location and social media handles.

Table A1. Podcast-level metadata.

Variable	Description
Site Name	Title of the podcast
Category	Topical or genre classification
Podcast Description	Textual summary of the show
Author Name	Creator or host attribution
Location	Geographic label
Facebook, Twitter/X, and Instagram URL	Social Media Links

This dataset forms the basis for describing the Canadian podcast ecosystem and for selecting the political subset used in subsequent analysis.

The following tables provide detailed breakdowns of the podcast-level metadata variables listed in Table A1.

Table A2. Top podcast categories.

Rank	Name	Percentage
1	Canada Podcasts	30.0%
2	Society & Culture	3.8%
3	Music	2.6%
4	Comedy	1.9%
5	Christianity	1.8%
6	News	1.5%
7	Sports	1.3%

8	Business	1.1%
9	Self-improvement	1.0%
10	Canadian Books	1.0%

The "Canada Podcasts" label dominates the classification scheme, accounting for nearly one-third of all categorized podcasts.¹¹ This catch-all designation appears to function as a default geographic marker rather than a meaningful topical classification. Beyond this label, the distribution reveals standard podcast genres—Society & Culture, Music, Comedy, News—alongside niche categories like Christianity and Canadian Book Podcasts.

What are the geographic features of the Canadian podcast ecosystem? The **Location** variable contains **335 unique labels**. As with the Category label, the results point to both informative and vague features. The top 10 most frequent locations are displayed in the table below.

Table A3. Top locations.

Location Tag	Percentage
Canada	52.9%
Toronto, Ontario	12.1%
Vancouver, British Columbia	4.3%
Ottawa, Ontario	2.9%
Ontario	2.4%
Montréal, Québec	1.9%
Edmonton, Alberta	1.4%
Winnipeg, Manitoba	1.2%
328 other locations (<1% each)	18.6%

Inconsistent granularity makes deeper geographic attribution challenging. The most confounding feature is that from the “Canada” label, we are unable to tell if the podcast is national in scope or underspecified. While less concerning, the presence of province-level labels province-level labels, like Ontario, pose an additional attribution challenge.

¹¹ While only 30% of podcasts included the “Canada Podcasts,” all other podcasts in our dataset included locations tags with a Canadian location.

The results illustrate that Canada’s urban centres are among the most well represented specific locations in the Canadian podcast ecosystem. Toronto alone accounts for 12.1% of located podcasts, followed by Vancouver, Ottawa, Calgary, and Montreal. This pattern aligns with the concentration of media production infrastructure in major metropolitan areas. However, with raw data alone, we are unable to tell if podcast production is primarily urban or if small towns are also highly represented. Future research is needed to make this determination.

How is the Canadian podcast ecosystem segmented in terms of language? To address this, we classified the languages of the **podcast description** variable.¹²

Table A4. Language demographics.

Language	Percentage
English	88.2%
French	11.4%
Other	0.4%

English dominates Canada’s bilingual podcast ecosystem, approaching 90 per cent of the classified podcasts. This is greater than the 75 per cent of Canadians who speak English, and conversely the proportion of French podcasts is about half the 21 per cent of Canadians who speak that language (Statistics Canada 2022). The “other” category comprises 21 podcasts across twelve additional languages. These podcasts tend to target prominent diaspora communities.

Are Canadian podcasters active on social media? We addressed this question using the **Facebook, Twitter/X, and Instagram URL variables**. The results are displayed below.

¹² We did so using the *langid* python library.

Table A5. Social media platforms.

Platform	Percentage
Facebook	52.2%
Twitter/X	49.6%
Instagram	46.9%
At least 1 platform	66.3%

Appendix B: Episode-level Variable Distributions

The episode level of analysis contains a larger and more heterogeneous set of variables. We present these variables in groups according to their prevalence across the dataset.

Table B1. Episode-level metadata.

Level of Coverage	Variable Name
Universal (>95% coverage)	Episode Title
	Episode Description
	Publication Date
	Duration
Common (50-94% coverage)	Explicit Content
	Image Link
Uncommon (<50% coverage)	Keywords
	Season

The variation in field completeness reflects the decentralized nature of podcast distribution. While core RSS elements (title, publication date, audio enclosure) are nearly universal, additional metadata fields vary across creators and distribution platforms.

Appendix C: Politics-focused podcast subset

List of politics-focused podcasts included in the analysis, with format categories as used. Only podcasts with more than 10 episodes captured are included; this criteria notably omits “The Blueprint”, an official Conservative Party of Canada podcast, from the analysis.

Podcast name	Affiliation/notes	Format
Uncolonized	Independent	Podcast
Pierre Nantel	QUB	Broadcast
GAC Files Dossiers d’AMC	Global Affairs Canada	Podcast
Information Morning Fredericton	CBC	Broadcast
Benoît Dutrizac	QUB	Broadcast
Antoine Robitaille	QUB	Broadcast
Ontario Morning	CBC	Broadcast
CPAC Today in Politics	CPAC	Broadcast
CTV News Toronto at Six	CTV	Broadcast
Philippe-Vincent Foisy	QUB radio	Broadcast
Mainstreet Halifax	CBC	Broadcast
Scarborough Centrists	Independent	Podcast
Morning North	CBC	Broadcast
Fight Back with Libby Anaimer	Zoomer Radio	Broadcast
The Real Canadian Podcast	Independent	Podcast
iPolitics	iPolitics	Media
The Agenda with Steve Paikin (Audio)	TVO	Broadcast
The CGAI Podcast Network	Canadian Global Affairs Institute (think tank)	Podcast

Uncommons: Canadian Politics with Nathaniel Erskine-Smith	Independent (host is Liberal MP)	Podcast
Beyond the Headlines	CIUT (University of Toronto student radio)	Broadcast
Canadian Taxpayers Podcast	Advocacy group	Podcast
The Richard Syrett Show	NewsTalk Sauga AM960	Broadcast
Real Talk Ryan Jespersen	Independent (host is ex-broadcaster)	Podcast
At Issue	CBC	Broadcast
Front Burner	CBC	Media
Rabble Radio	Independent	Podcast
10/3 Canada Covered	National Post	Media
Canadian Politics with a twist	Independent	Podcast
inFocus with David Coletto	Abacus Data	Podcast
Metro Morning	CBC	Radio
Information Morning Saint John	CBC	Radio
Canadian Politics is Boring	Independent	Podcast
CANADALAND	Canadaland	Podcast
This Matters Daily News Podcast	Toronto Star	Media
PrimeTime Politics with Peter Van Dusen	CPAC	Broadcast
OPPO	Canadaland	Podcast
West of Centre	CBC	Media
The Backbench	Canadaland	Podcast

Green Majority Radio	Independent	Podcast
The Bridge with Peter Mansbridge	Independent (host is ex-broadcaster)	Podcast
The House	CBC	Broadcast
PolitiCoast The BC Politics Podcast	Independent	Podcast
The Breakdown with Nate Pike	Independent	Podcast
Canadian Common Sense	Independent	Podcast
Policy Options podcast	Think tank	Podcast
NB Poli Podcast	Independent (hosts are municipal politicians)	Podcast
Power and Politics	CBC	Broadcast
CTV Power Play with Vassy Kapelos Podcast	CTV	Broadcast
CFRC Daily News Briefs	Queen's University student radio	Broadcast
#onpoli, a TVO podcast	TVO	Media
The Political Paradox Podcast	Independent	Podcast
Blue Skies Podcast with Erin O'Toole	Independent (host is former Conservative Party leader)	Podcast
All in a Day	CBC	Broadcast

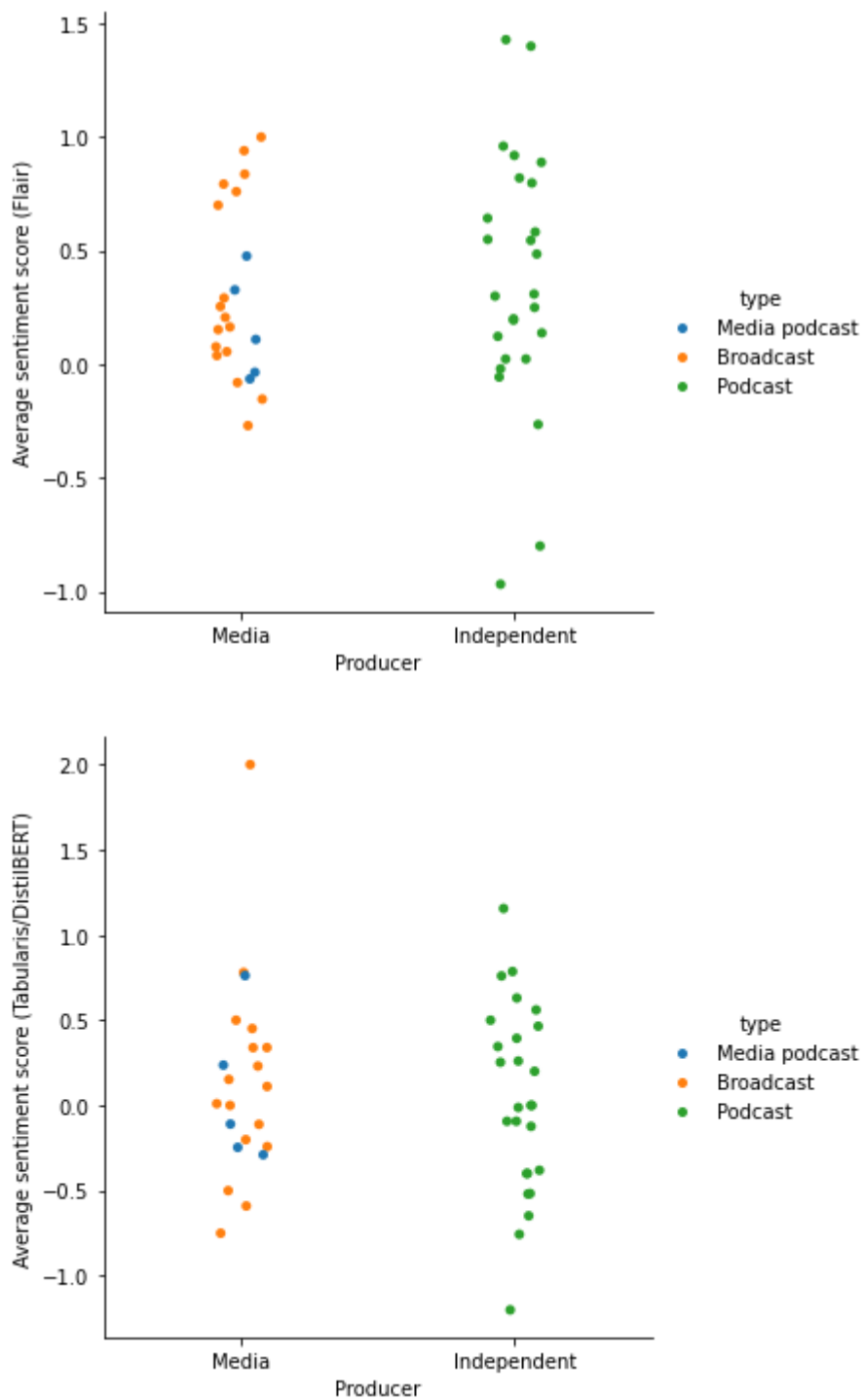


Figure A1. Average sentiment scores for politics-focused podcasts, using different models.

Appendix D: Podcast Tool

We include the following application to assist researchers with analysis of podcasts. This tool is particularly helpful for finding podcast RSS feeds and facilitating the downloading of files. Additionally, we include a module that runs the OpenAI Whisper transcription model. However, this tool is not optimized to run on the hosting platform for the tool. Users who wish to conduct batch transcription can download the repository and run the transcription model locally.

The first tool identifies the RSS feeds associated with a given podcast. Most listening platforms present user-friendly URLs that do not reveal the underlying feed. Our tool accepts a podcast name or a platform link and returns the corresponding RSS feed when one is available. It does this by checking public search interfaces and validating that the retrieved link points to a functioning RSS document. This helps researchers assemble feed lists without manually inspecting multiple sources.

The second tool provides an interface for working with automatic speech recognition. It allows users to upload an audio file and obtain a transcript generated by the OpenAI Whisper model. This tool is useful for testing transcription settings and ensuring the output quality is acceptable before scaling up. It provides a simple way to explore the characteristics of Whisper without writing code.

The tools we release are intended to make the underlying process transparent. The source code for the tool is available in the project's Github repository. They show how to identify podcasts, extract metadata and produce transcripts. Researchers with access to more powerful computing environments can adapt these components to meet the needs of larger projects.

The tools can be accessed here: https://github.com/patterson-s/podcast_rss

Appendix E: Supplemental view of housing discourse in transcribed episodes

The table below contains a random sample of concordances for “housing” in the transcribed corpus subset .

Table E1. Random sample of concordances for “housing.”

But again, those things don’t necessarily always get enforced. We know that employment and	housing	discrimination still happen regularly. So having legislation in and of itself or having sort
that made it into the announcement and the work that [Miramichi] is doing with their regional	housing	authority, that they’re developing up there, with us, we’re trying to stand up a municipal
that are sort of perceived to be progressive, but they also have, you know, huge issues with	housing,	homelessness and addiction. And what I’m seeing is a really sort of mobilized nimby movement
intended for newcomer students in case they arrive in Canada and are having trouble finding	housing	so they can attend the school and rent an apartment from basically the school itself as well
who come from modest means. Look, if you come from a wealthy family, you can get into the	housing	market. You come from modest means and offer a family of homeowners. You’re new to this
community housing sector dominated by small-sized organizations that each have a narrow	housing	focus. According to Steve, Canada’s community housing system has evolved on the project by
us purchasing the building, or that through rent supplement, I want to make a dent on the	housing	front. All right. That is all the time we have for today. And I think those are promising
place here, but we had the answers in 2019, right? And I wasn’t. Answers to what? The	housing	problems that we were having. Why was, you know, people say there’s a housing crisis now.
They felt that was what Canadians wanted to see. So, they’re going to cut the GST on new	housing	units. They made this pitch that they were going to bring together all the grocery company CEOs

the exit polls show. They show two interesting things really. One housing crisis, big issue	housing	prices on the cost of living, and they're really focused on the next government doing a
couple other announcements out of the government before we get into our main discussion of the	housing	debates. For those working in gig work and online platforms, the government has like pre
who's on social assistance. In Ontario, it's a little over \$500. So deeply affordable	housing	would geared to those people and it would be quality housing. So we just don't have that
policy statement, planning, planning authorities shall provide an appropriate range and mix of	housing	options and different densities to meet projected market based and affordable housing needs
a clear example of this, where it's like, we have one problem, which is we need way more	housing.	And then we have another problem, which is like, we need good jobs for people. Why can we
has been a big issue here as it as been in so many parts of the country. There's a real	housing	crisis in Whitehorse specifically because the population's been growing, things are not
critical, because if you don't have the right infrastructure to support it, whether it's	housing,	whether it's roads, whether it's power, you're going to struggle in that area, I think
services. Crime and safety is really important for a lot of people top of mind transit and	housing.	All that stuff is here in the budget. And with all the uncertainty right now, we're not
their files. And certainly not in the short term, and we can go into why, in particular in	housing,	those files aren't going to advance in the short term for other reasons, but Chantal is
Facebook ad library and they did a, you know, a PowerPoint style ad, you know, attacking the	housing	policy and it's just bullet point, bullet point, bullet point. And it's you know, it

his frozen moment in time at least, Canadians are saying that they care more about housing and	housing	affordability, which you might call Canada's housing crisis than they do about climate change
a value in the idea of and the government is already looking at this having more affordable	housing	built by the government you know we see this in the hints that they are going to combine the
listing competitive in a credit constraint market. So right now, you know, we talk about how	housing	affordability was so bad this time last year and just because of price, because rates were
in now over the next few months is going to make a huge difference for people, affordable	housing	that the NDP has forced the Liberals to finally invest in. We have a disaster under Pierre
only slightly better that's because well weaker population growth does translate into weaker	housing	demand they also expected to result in fewer new listings now over time it is limiting our
in China. How much it gets out of control is yet to be seen but we're already seeing that	housing	prices and so for example developers who are having trouble meeting their debt obligations

Appendix F: NLP pipelines and evaluation protocols for “Housing as investment or social issue in the wider podcast ecosystem”

This appendix details the natural language processing (NLP) pipelines and evaluation protocols used for section 4.3. In total, we include details for 6 pipelines and quality control evaluations.

Definitions

For shorthand, we refer to the following groups to distinguish housing-focused podcasts.

- **G1:** Podcasts whose primary editorial focus is housing.
- **G2:** Podcasts that do not mention housing in their description but that have occasional housing-focused episodes.
- **G1.5:** An intermediate group — podcasts that mention housing alongside other issues (for example, an urbanism podcast with housing as one theme, or a personal finance podcast with a real estate theme). G1.5 was ultimately folded into G1.

Tools

The following tools were used throughout the classification pipeline.

- **Cohere Command-R7B** (command-r7b-12-2024): a small, open-weight generative language model used for binary classification tasks across the main housing detection funnel.
- **Cohere Command-A** (command-a-03-2025): a larger, open-weight generative language model used occasionally where the classification task required more nuanced judgment.
- **Cohere embed-v4.0** (embed-v4.0): a word embedding model used to generate dense vector representations of text.
- **Cohere rerank-v4.0-pro** (rerank-v4.0-pro): a cross-encoder reranking model used to re-score candidate texts by semantic relevance to a query.

Our decision to use open-weight LLMs is in line with emerging practices for LLM use in social science research (Barrie, Palmer, and Spirling 2025). While we ran our pipeline using the Cohere API service, the Command-R7B and Command-A models can be run offline or on a private cloud using downloaded model weights. Both models are licensed under a Creative Commons Attribution Non-Commercial 4.0 International licence, making them suitable for research purposes.

Evaluation Considerations

After each NLP run, we evaluated results through human review. Our primary considerations were:

- **Precision** expresses the proportion of true positives among the positively classified group: of all podcasts or episodes the model labeled as housing-related, how many genuinely are?

- **False negative rate (FNR)** expresses the proportion of actual housing-positive items that were missed by the classifier and incorrectly placed in the negative group.

A primary challenge is that we lack the capacity to evaluate the complete dataset. More than 6,000 podcast descriptions and more than 600,000 episode descriptions is prohibitively large for manual review. We addressed this through a stratified evaluation approach. For the negative pool, we defined rerank score bands and sampled within each band, concentrating evaluation effort on the highest-risk strata. This allows estimation of false negative rates without exhaustive labeling.

As we proceeded through evaluations, we accumulated high quality evaluated data. In part 5, we attempted to reincorporate this training data into an additional trained classifier, but ultimately abandoned the approach.

Part I — First-pass classification of G1 and G2

In part 1, we created our initial list of housing-focused podcasts and episodes. Our podcast and episode pipelines followed similar procedures: embed, filter by cosine similarity, rerank, classify with Command-R7B, and then stop after a series of consecutive negative classifications.

Pipeline 1 — Podcast description classification (G1)

- 1) **Embed.** All podcast descriptions were embedded using embed-v4.0 (1024-dimensional).
- 2) **Cosine similarity filter.** Descriptions were ranked by cosine similarity to a fixed housing reference vector. The reference was produced by embedding the following query:
 - *“housing affordability crisis real estate market rental housing homeownership housing supply zoning development tenant landlord mortgage rent prices housing policy homelessness social housing housing crisis tenant rights affordable housing”*
- 3) **Rerank.** Top candidates were re-scored using rerank-v4.0-pro against the query:
 - *“podcast covering housing real estate rental market mortgage affordability housing crisis homelessness tenant rights zoning housing policy social housing.”*
- 4) **LLM classify.** Candidates were passed in descending rerank order to Command-R7B for binary classification. A sequential early-stop strategy halted the pass after a fixed number of consecutive negative decisions. Temperature: 0.0.

Prompt

System:

You are a housing content classifier for a research project on Canadian media.
You MUST respond with ONLY a raw JSON array – no explanation, no markdown, no code fences.
Start your response with [and end with].

User:

Classify each podcast description below.

Answer `is_housing=true` if this podcast regularly covers ANY aspect of housing, broadly defined. This includes – but is not limited to:

Real estate markets, buying/selling, investing, property management, mortgages

Rental markets, tenant rights, landlord issues, rent control

Housing affordability, the housing crisis, homelessness, shelters

Affordable housing, social housing, public housing policy

Zoning, urban planning, development, NIMBYism

Housing as a political/social issue (e.g. progressive housing policy debates)

Answer `is_housing=false` only if housing is clearly absent or purely incidental (e.g. a general politics show that rarely touches housing).

Return a JSON array, one object per item (0-indexed):

```
[{"index":0,"is_housing":true,"confidence":0.95,"reasoning":"one sentence"}]
```

Items:

```
[0] <podcast name>
```

```
<podcast description>
```

```
...
```

Evaluation

Two rounds of human evaluation were conducted. Round 1 exhaustively reviewed all 63 positives produced and sampled negatives across four rerank bands. The approach yielded zero false positives, but considerable false negatives, especially in Band A and B. In round 2, we extended our evaluation of false negatives, while oversampling from Band B.

Metric	Round 1	Round 2	Combined
Positives evaluated	63 / 63	—	63 / 63
Precision	63/63 = 100%	—	63/63 = 100%
Band A FNR	9/17 = 52.9%	2/3 = 66.7%	11/20 = 55.0%
Band B FNR	5/20 = 25.0%	37/94 = 39.4%	42/114 = 36.8%
Band C FNR	0/10 = 0.0%	0/20 = 0.0%	0/30 = 0.0%
Band D FNR	0/5 = 0.0%	0/20 = 0.0%	0/25 = 0.0%
False negatives found	14	39	53
Total labeled items	115	137	252

The results suggest that the prompt can effectively recognize housing content, but that the restrictive embedding and rerank pipeline was producing false negatives. After completing

evaluation, we added 53 false negatives to G1, arriving at an n=116. Additionally, we accumulated a dataset of human-evaluated true negatives. We noted that we would need to return for further false negatives robustness checks and proceeded to G2.

Pipeline 2 — Episode description classification (G2)

- 1) **Pre-computed embeddings.** Episode descriptions were embedded in an earlier pass and stored in `podcast_data.episode_description_embeddings` (1024-dimensional). G1 podcasts were excluded.
- 2) **Cosine similarity filter.** Episodes ranked by cosine similarity to the same housing reference vector as Pipeline 1.
- 3) **Rerank.** Top candidates re-scored using `rerank-v4.0-pro` against:
 - *“episode about housing real estate rental market mortgage affordability housing crisis homelessness tenant landlord zoning housing policy.”*
- 4) **LLM classify.** Command-R7B binary classification with sequential early-stop. Same batch size and temperature as Pipeline 1.

Prompt

System: Same as Pipeline 1.

User:

Classify each episode description below.

Answer `is_housing=true` if this episode substantially covers ANY aspect of housing, broadly defined. This includes:

Real estate markets, buying/selling, investing, mortgages
 Rental markets, tenant rights, landlord issues, rent prices
 Housing affordability, the housing crisis, homelessness
 Affordable/social/public housing, housing policy
 Zoning, urban development, gentrification
 Housing as a political or social justice issue

Answer `is_housing=false` only if housing is a brief passing mention or clearly off-topic.

Return a JSON array, one object per item (0-indexed):

```
[{"index":0,"is_housing":true,"confidence":0.95,"reasoning":"one sentence"}]
```

Items:

```
[0] <podcast> – <episode title>
    <episode description>
    ...
```

Evaluation

Sample. 50 positives randomly drawn from 7,809 classified positives. Negatives stratified into five bands by rerank score.

Metric	Value
Positives evaluated	50 / 50

Precision	43/50 = 86%
Band A FNR	6/30 = 20.0%
Band B FNR	0/20 = 0.0%
Band C FNR	1/10 = 10.0%
Band D FNR	0/5 = 0.0%
Band E FNR	0/5 = 0.0%

The model again demonstrated high precision. However, unlike the previous pipeline, the false-negative rates were lower, even in the threshold Band A. At this stage, we determined that we had an effective episode-classification prompt. However, since we did not classify the entire dataset, we noted that false negative may in fact be under-reported.

Part II — Typology classification

After seeding our initial G1 and G2 lists, we attempted to classify them by thematic framing: housing as real estate or housing as a social issue. That these were important themes became clear after reading several hundred podcast descriptions and episode descriptions.

Two variations on the classification schemes were conducted in parallel as a robustness. The first distinguished between “real estate” and “housing policy” framing. The second distinguished between “investor operations” and “crisis analysis” framings.

Pipeline 3 — Typology classification

- 1) **Rerank (two passes).** Each episode was scored by rerank-v4.0-pro against two queries simultaneously — one oriented toward real estate framing, one toward policy framing — to generate a prior signal.
- 2) **LLM classify.** Command-R7B applied a two-step prompt: first confirm housing relevance, then assign one type from the schema plus tag dimensions. Note, the tagging was ultimately abandoned.
- 3) **Two runs.** Run 1 used types `real_estate / housing_policy`. Run 2 used `investor_operations / crisis_analysis` with different rerank queries, as an alternative framing for comparison.

Prompt

System (both runs):

You are a qualitative coder analyzing Canadian podcast episodes for housing content.

You MUST respond with ONLY a raw JSON array – no explanation, no markdown, no code fences.
Start your response with [and end with].

User (Run 1):

For each podcast episode below, do the following:

STEP 1 - Confirm housing relevance:

Set `housing_confirmed=true` only if housing is a SUBSTANTIAL topic of this episode (not just a passing mention). Set `housing_confirmed=false` if housing is incidental.

Provide a one-sentence `housing_reasoning`.

STEP 2 - If `housing_confirmed=true`, apply the typology:

TYPES (assign exactly one - mutually exclusive):

`real_estate`: The ideal type of a podcast oriented toward property as a commodity – concerned with markets, transactions, investment returns, industry practice, and the financial mechanics of housing. A pure `real_estate` podcast treats housing primarily as an asset class.

`housing_policy`: The ideal type of a podcast oriented toward housing as a social good – concerned with affordability, rights, homelessness, planning, advocacy, and the political economy of shelter. A pure `housing_policy` podcast treats housing primarily as a public issue.

TAG DIMENSIONS (assign all that apply – may be empty list):

`geography_level` (Geographic Scope): national, provincial, municipal

`geography_region` (Named Region): Canada, Toronto/GTA, Vancouver/Metro Vancouver, British Columbia, Ottawa, Calgary, Saskatchewan, Manitoba, Quebec, Okanagan, Nova Scotia, Edmonton, Kingston

`audience` (Target Audience): general_public, investors, industry_professionals, homebuyers, advocates_researchers

`sub_topic` (Sub-Topics & Perspectives): affordability, investing, market_trends, policy, homelessness, urban_planning, mortgages, social_justice, personal_finance, industry_practice, economic_analysis, tenant_rights, rental_market, commercial, zoning, property_management

Valid type names: ["real_estate", "housing_policy"]

Each item includes a [Rerank signal] showing how strongly its description matches each type query. Use this as a prior but let the description content override it.

Return a JSON array, one object per item (0-indexed):

```
[{"index":0,"housing_confirmed":true,"housing_reasoning":"...", "type":"real_estate",
  "geography_level":["national"],"geography_region":["Toronto/GTA"],
  "audience":["investors"],"sub_topic":["market_trends","investing"]}]
```

If `housing_confirmed=false`, set type to null and all tag arrays to [].

Episodes:

[0] <podcast> – <episode title>

[Rerank signal: policy=0.32, real_estate=0.71]

<episode description>

...

Rerank queries (Run 1):

Policy: “housing affordability crisis tenant rights homelessness social housing public housing advocacy zoning urban planning housing policy rental regulation NIMBY.”

Real estate: “real estate market investing buying selling property mortgage investment returns rental income REIT property management landlord capital gains market analysis.”

Typology evaluation

150 episodes were sampled from the 7,809 G2 positives. Stratification was by cross-run agreement, so that evaluation effort concentrated on cases where the two competing schemas disagreed.

Stratum	real_estate	social_issue	ambiguous	not_housing
both_agree_RE	10	—	—	—
both_agree_SI	4	14	—	2
RE_vs_CA	37	6	2	5
HP_vs_IO	15	8	2	5
run1_only	10	—	3	2
run2_only	8	1	2	4
neither	—	1	—	9

The classifiers were functional but the evaluation surfaced several ambiguities in the coding scheme. First, some episodes address the housing crisis from a real estate investor's viewpoint. Second, construction activity may warrant its own sub-category. Third, commercial real estate episodes were often mistakenly included. These limitations motivated a follow-up quality-control round.

G2 quality control

Two follow-up rounds validated typology labels on G2 episodes.

We conducted two additional validation rounds to validate typology labels for G2. However, in the process, we noticed that many positive examples were taken from podcasts that had been mistakenly excluded from G1. Basically, this round of evaluation was more useful for surfacing G1 false negatives than for G2 quality control.

Round	RE pool	SI pool	Total	G1 candidates found
Round 1	50	50	100	36 episodes (18 podcasts)
Round 2	40	40	80	1 episode

The 30 g1-labeled items in Round 1's RE pool (out of 50) indicated that the original G1 funnel had missed an entire pocket of real-estate-focused podcasts. The 18 newly-identified podcasts were added to G1. Then, we added an additional keyword detector that expanded G1 even further, to 144 podcasts.

Part III — Auditing the expanded G1

With G1 at 144 podcasts, we worried that we may have included some false negatives. We audited G1 with the following process.

Pipeline 4 — G1 corpus audit

All 144 current G1 podcasts were queried directly from `housing_results.podcast_level` and reviewed exhaustively using the G1 Audit tab in an evaluation app. Each podcast's full description was shown alongside its housing similarity score. Label options: housing focused / housing mentioned / drop. No LLM was involved — this was a manual review step.

Evaluation

Label	Count	Share
housing focused	118	82%
housing mentioned	22	15%
drop	4	3%

The four podcasts labeled drop were exclusively about commercial real estate or construction services with no residential housing content:

- The Industrial Real Estate Podcast
- Commercial Real Estate Podcast
- Commercial Real Estate Library
- Freeman's Construction Inc

While auditing G1, we considered forming an additional group (G1.5) to account for cases where housing was mentioned in the podcast description, but among other themes. The 22 “housing mentioned” podcasts retained from the G1 audit served as seeds. However, we ultimately folded these into G1. At this point, we considered G1 to be complete.

Part IV — Full-corpus episode classification

Having completed G1, we returned to G2. In pipeline 2, we determined that we had an effective prompt but were at risk of false negatives. To address this, we decided to classify all episode descriptions. We attempted to do so with two pipelines. In the first, we used our accumulated evaluation data to train a spaCy TextCat classifier – which is essentially a less sophisticated by computationally cheap neural classifier. After poor results, we then classified the full corpus with Command-R7B.

Pipeline 5 — spaCy TextCat classifier

We first tried a trained spaCy TextCat classifier for three reasons. First, we already had human-labeled training data accumulated through the evaluation rounds of Pipelines 1–3, so the supervised setup was inexpensive to bootstrap. Second, spaCy models are cheap to apply at the 600k-episode scale. Third, the architecture supports sequential retraining, meaning the model could be improved through sequential evaluations and training rounds.

The classifier applied a three-stage triage pipeline to each episode description.

- 1) **Keyword auto-positive.** Descriptions matching any of 24 explicit housing phrases (e.g., “mortgage,” “real estate,” “landlord,” “housing crisis,” “homelessness”) were immediately classified as housing-positive with a score of 1.0 and triage label `keyword_auto_positive`. No model call was made.
- 2) **spaCy binary classification.** Descriptions containing weak housing signals (broader terms like “rent,” “construction,” “zoning”) that did not match the strong-keyword list were scored by a trained spaCy TextCat model (`housing_desc_binary_v1`). Score threshold: 0.5. Descriptions below all keyword thresholds received a `no_signal` label and were not passed to the model.
- 3) **spaCy typology classification.** Descriptions classified as positive by the binary model were passed to a second trained TextCat model (`housing_desc_typology_v1`) to assign `real_estate` or `social_issue` framing.

Evaluation

Three successive evaluation rounds were conducted, each sampling 20 items per positive band. After each round, the newly-labeled items were folded back into the training set and the model was retrained before the next round. Round 1 also included negative bands to estimate false negative rates.

Band	Triage method	Score range
K	Keyword auto-positive	Strong keyword → score 1.0
MH	ML positive, high	spaCy score ≥ 0.80

ML	ML positive, medium	spaCy score 0.50–0.80
N1	Near-miss ML negative	spaCy score 0.35–0.50
N2	Clear ML negative	spaCy score < 0.35
N3	No-signal	Keyword-filtered out

Round	Positive bands	Negative bands	Total	Precision
1	K:20, MH:20, ML:20	N1:20, N2:10, N3:10	100	22/60 = 36.7%
2	K:20, MH:20, ML:20	—	60	24/60 = 40.0%
3	K:20, MH:20, ML:20	—	60	29/60 = 48.3%

Negative band results (Round 1 only):

Band	FNR	N evaluated
N1 (near-miss)	1/20 = 5.0%	20
N2 (clear negative)	2/10 = 20.0%	10
N3 (no-signal)	0/10 = 0.0%	10

The model was terrible. Precision did not exceed 48% across three rounds of retraining, meaning more than half of episodes flagged as housing-positive by the model were false positives, despite the iterative retraining strategy. The spaCy approach was abandoned.

Pipeline 6 — Full-corpus Command-R7B

Command-R7B was applied directly to all episode descriptions using the same classification prompt as Pipeline 2. A random sample of 70 episodes was for human evaluation.

Prompt

Same system and user prompts as Pipeline 2 (episode_focus variant). See Pipeline 2 above.

Evaluation

Metric	Value
R7B-positive evaluated	35 / 35
R7B-positive precision	31/35 = 88.6%
R7B-negative evaluated	35 / 35
R7B false rejection rate	2/35 = 5.7%

Command-R7B substantially outperforms the spaCy classifier (88.6% vs. 37–48% precision) and produces a low false-rejection rate of 5.7%, indicating the model rarely misclassifies genuine housing content as negative. R7B is retained in the database as the primary G2 signal.