OCanada-TVNews: A Two-Year Dataset for Gender and Topic Diversity in Canadian TV News Broadcasts

Zeynep Pehlivan
McGill University
zeynep.pehlivan@mcgill.ca

Abstract

This paper presents OCanada-TVNews, a two-year dataset of Canadian TV news broadcasts, from four major channels: CBC The National, CTV The National, Global News, and RadioCanada Info, retrieved from channels' YouTube playlists. The dataset consists of 3982 videos, in total 999 hours. Each video is processed through a systematic pipeline that involves audio extraction, speaker diarization, gender detection, and topic classification. In addition, for each video, metadata containing user engagements is also provided. By merging speaker segments with gender labels and topic categories, this dataset offers a detailed view of who speaks and what is discussed on a daily basis. An exploratory analysis highlights persistent disparities in gender representation and provides insights into the topics that dominate coverage. OCanada-TVNews serves as a basis for a further study of media diversity, content trends, and potential biases in broadcast journalism.

Code — https://github.com/ZeynepP/ocanada-tvnews
Datasets — https://figshare.com/s/3cd465d04ff4dda93390

Introduction

Social media platforms are increasingly popular for news consumption, but television broadcasts remain a primary source of information for much of the population (online or on linear television). Although television news audiences are decreasing in most countries, (Newman et al. 2021) shows that it has remained consistent in Canada around 40% since 2020. This enduring trust in TV news highlights its continued influence in shaping public perceptions, particularly with regard to social issues such as gender. For example, media has the power to challenge or reinforce traditional gender stereotypes, directly affecting how people perceive gender roles, but also how the topics are presented. However, gaining comprehensive access to televised news broadcasts can be challenging due to limited public data. By focusing on Canadian TV news content shared on YouTube, our dataset bridges this gap, capturing both the traditional influence of TV journalism and the growing role of social media in modern news distribution.

We have collected daily TV news for two years (2023 and 2024) from four Canadian TV news channels on YouTube: CBC The National, CTV The National, Global News, and RadioCanada Info (in French). This dataset comprises 3982 videos (999 hours of content) and includes transcripts, speaker diarization, gender labels as female and male, and topic labels such as politics, economy, science, etc., enabling researchers to track how coverage shifts or remains consistent over time.

A major motivation for developing this dataset is to shed light on gender and topic representation in broadcast news. We recognized the need for empirical data illustrating where women's voices are heard and where they remain underrepresented. This dataset opens the door to several important research questions, such as:

- How do TV news allocate airtime between male and female speakers on different topics, such as politics, science, and human interest stories?
- Which topics dominate coverage, and do certain channels focus more on specific issues (e.g., conflict, environment, or economy)?
- As coverage is daily, can we identify shifts in topic emphasis or speaker demographics that correlate with major events and beyond?

In addition to the dataset itself, we release the full source code used to generate OCanada-TVNews. This includes modules for collecting YouTube metadata, performing speaker diarization, gender recognition, topic classification, and data analysis. Researchers can reuse and adapt this pipeline to create similar datasets on other topics or channels, facilitating broader studies of audiovisual media using publicly available online content. Finally, by highlighting gender representation, in particular, the imbalance of female speakers in key topic areas, this data set underscores important questions about media coverage and offers a foundation for future work in social science, communication studies, and computational social science.

Related Work

Numerous studies (Rao and Taboada 2021; Asr et al. 2021; Soumah et al. 2023; Pelloin et al. 2024) have explored gender representation and topic analysis in news media and on-

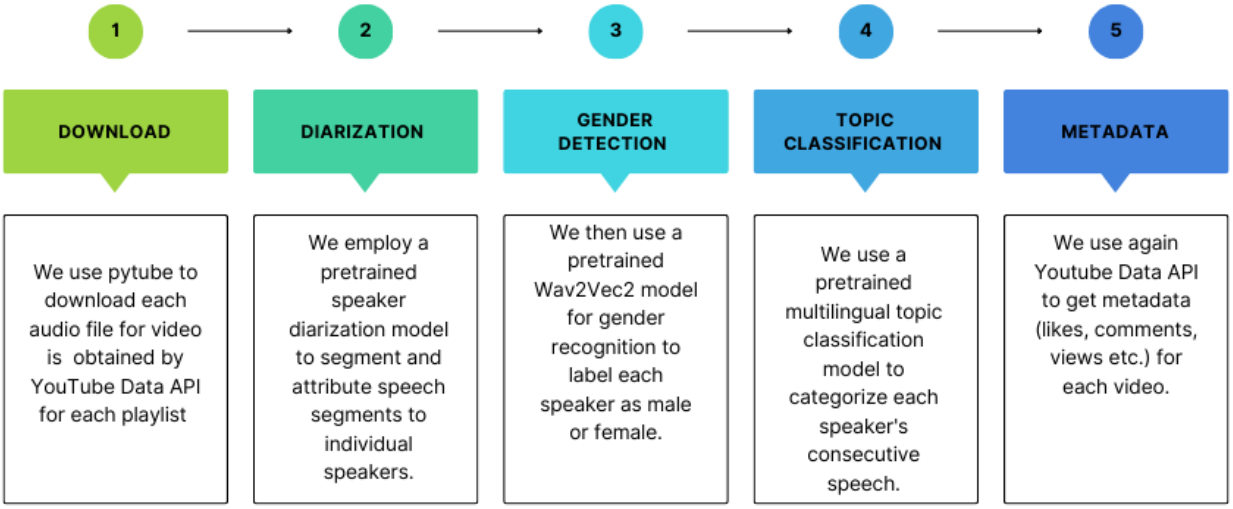

Figure 1: Processing Steps in the OCanada-TVNews Pipeline

line platforms, using datasets that vary in scope, methodology, and focus.

The Gender Gap Tracker (Asr et al. 2021) and The Radar de Parité (Soumah et al. 2023) are automated systems that measure men and women's presence in mainstream Canadian news outlets. They collect and analyze articles from Canadian news websites and apply Natural Language Processing to identify speaker quotes and assign gender. While they provide valuable insights into gender representation in French/English news in Canada, they are primarily focused on text-based media.

The Global Media Monitoring Project (GMMP) (Macharia 2020) provides comprehensive reports on gender representation in media globally, offering valuable insights into the presence and roles of women in news stories. However, GMMP's reports may lack sufficient hours of data for in-depth, continuous analysis. Another relevant study (Pelloin et al. 2024) utilizes machine learning techniques to classify news subjects and assess gender biases in French TV news broadcasts. Their work relies on data from the National Institute of Audiovisual (INA), which provides comprehensive 24/7 TV channel distributions, enabling large-scale analysis of traditional broadcast media. Unfortunately, Canada lacks an equivalent institution that archives such extensive audiovisual data. To address this, we source our data from YouTube, which not only provides access to TV news broadcasts but also includes valuable metadata such as user engagement metrics, offering a richer perspective on audience interactions across bilingual Canadian news media. The study presented in (Pelloin et al. 2024) serves as an excellent example of how our dataset can be utilized for comparable or expanded research.

In summary, while previous studies have advanced our understanding of gender and topic disparities in traditional news media in Canada, our dataset offers a novel contribution by providing a broadcast TV news dataset. It combines bilingual content, speaker-level gender classifications, topic modeling, and YouTube metadata, creating a versatile resource for analyzing the interplay between media representation and audience reactions.

Methodology

This section explains how the dataset was collected and prepared for analysis. It is divided into two main parts: Dataset Construction and Data Description. The first part shows the steps used to retrieve and process daily TV news content from YouTube playlists, while the second part summarizes the dataset's structure.

Dataset Construction

This dataset is sourced from YouTube News playlists affiliated with four major Canadian broadcast channels: CBC The National [1], CTV The National [2], Global News [3], and Radio Canada Info [4]. The collection period covers the years 2023 and 2024. Our dataset construction process involves five key steps, as illustrated in Figure 1. This pipeline ensures a systematic approach to collecting, processing, and analyzing two years of daily TV news videos from YouTube playlists for four major TV channels in Canada:

- **Download** By using the YouTube Data API [5], we extracted the unique video IDs, for each playlist created by content providers specially to rebroadcast their TV broadcast news, for two years 2023 and 2024. We then

[1] PLvntPLkd9IMcbAHH-x19G85v_RE-ScYjk
[2] PLLzHOgGvydCmI3CNmzqtLofFKCufwOAHz
[3] PLA0c-X5PdUCXybJu-Jk-lJTb4s6VCEnwk
[4] PLZr1y64TPtN843GY096K8Nyzx-lZWpP9U
[5] https://developers.google.com/youtube/v3

Channel Name	# Videos	Duration (hours)	Female	Male	Female (without presenter)	Male (without)
CBC News	624	423.80	235.08	188.71	178.51	149.42
CTV News	677	234.11	111.49	122.61	80.52	85.95
Global News	735	242.32	138.91	103.41	68.15	94.50
Radio Canada Info	1946	99.57	53.45	46.12	21.82	28.07
Total	**3982**	**999.81**	**538.94**	**460.87**	**349.00**	**357.96**

Table 1: Summary of Playlists by Channel.

used pytubefix (JuanBindez 2023) library to download each video file (only audio). During this stage, some videos were skipped if they were restricted by region, or if their metadata was incomplete or they are no longer available (less than 1% of the dataset).

- **Diarization** This step segments the audio track into different speaker turns, enabling us to identify when different individuals spoke during the broadcast. We use PyAnnote Speaker Diarization 3.1 [6] (Plaquet and Bredin 2023; Bredin 2023) to process the downloaded audio files. This tool performs a frame-level multiclass classification over 5-second windows for local diarization. The local diarizations are then refined using Agglomerative Hierarchical Clustering, leveraging ECAPA-TDNN embeddings (Desplanques, Thienpondt, and Demuynck 2020) to group speaker segments effectively. The results are saved in the RTTM format which is specially designed for speaker diarization tasks, identifying who speaks when (Fiscus et al. 2006). Within this format, each line details the start time and duration of the segment,as well as the speaker's ID. Those files are also provided as part of the dataset.
- **Gender Detection** We applied a pre-trained gender recognition model, specifically wav2vec2-large-xlsr-53-gender-recognition-librispeech (Fiury 2023), to classify each speaker turn as female or male based on only audio. This model is a fine-tuned version of Facebook's Wav2Vec2-XLS-R-300M (Baevski et al. 2020), tailored for gender recognition tasks using the Librispeech-clean-100 dataset. It achieves a loss of 0.0061 and an F1 score of 0.9993 on the evaluation set. This gender detection model operated on the segmented audio produced during the diarization phase. We also transcribed each speaker's speech to use in the topic modeling step by using WhisperX multilingual model (Bain et al. 2023)
- **Topic Classification** We used the multilingual news topic classification model (CLASSLA 2023) to classify the content of each speaker's consecutive speech into relevant topics. The classifier uses the top level of the IPTC Media Topic NewsCodes schema, consisting of 17 labels. We also annotated first speaker as presenter and add a "presenter" field on data to facilitate filtering.
- **Metadata** To enrich the dataset and make it more usable for different research questions, we also provide the metadata for each video obtained by using YouTube Data API. It contains video details, user engagements.

Below is an example of the data stored in a json file for each channel. This file combines the output of the diarization, gender detection, and topic classification steps.

```
{
video_id : 0DiuxWWnvYw,
channel:CBC,
analysis :[ {
    video: 0DiuxWWnvYw,
    speaker: SPEAKER_24,
    start: 1.316,
    end: 6.036,
    duration: 4.72,
    gender: 0,
    gender_probability: 0.998506844,
    text: Tonight, ...,
    date: 2023-12-22 08:39:29,
    channel:CBCTheNationalNewsPlaylist,
    presenter: 1,
    topic: conflict, war and peace,
    topic_probability: 0.9051486254
    } ...],
video_metadata : {video: { ...
  snippet: {
    publishedAt: 2023-12-22T08:39:29Z,
    channelId: UCK...,
    title: CBC News: ...,
    description: Dec. 21, 2023...,
    thumbnails: {...},
    channelTitle: CBC News...,
    tags: [Prague shooting,...],
    categoryId: 25,
    liveBroadcastContent: none,
    defaultLanguage: en},
    defaultAudioLanguage: en-CA
  },
  contentDetails: {duration:...},
  statistics: { viewCount: 57380,...},
  }
}
```

[6] https://huggingface.co/pyannote/speaker-diarization-3.1

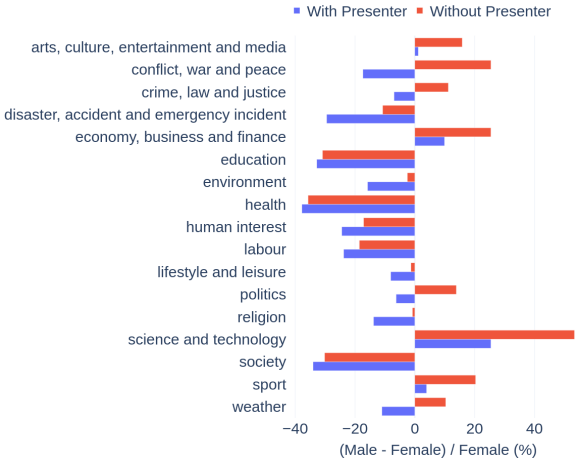

Figure 2: Female vs Male distribution over topics

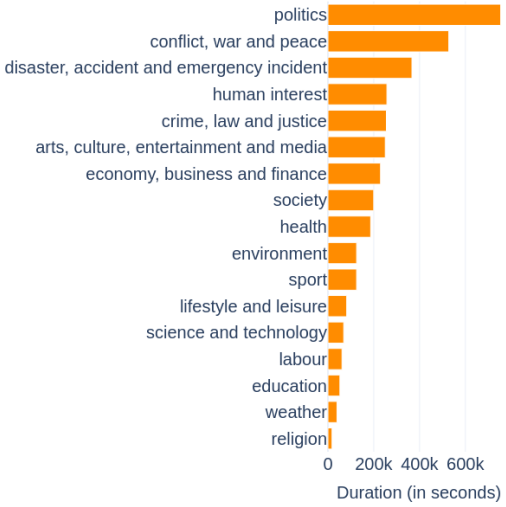

Figure 3: Overall Topic Distribution

Data Description

Here is a brief description of each field:

speaker Unique identifier for each speaker in one audio track.

start, end The time offsets (in seconds) from the beginning of the audio track where the speaker segment begins/ends.

duration The duration of the speaker segment.

gender, gender_probability Predicted speaker gender (0 for female, 1 for male) and the associated confidence score.

text Transcribed speech for the current speaker turn.

date YouTube video publication date in UTC.

id YouTube unique video ID.

channel Playlist name (e.g., CBC, RadioCanada).

presenter Indicating whether the speaker is a presenter.

topic, topic_probability Topic classification label and its associated confidence score.

video_metadata The snippet section includes basic details such as the publication date, title, description, language settings, and thumbnails. The statistics section captures engagement data, including view counts and other metrics, while contentDetails may provide specifics like video duration or format.

In other words, each file merges identification (channel, date, and video ID) with audio segmentation results (speaker IDs, start/end times, transcripts), semantic analyses (gender, topic classification) and metadata (view count, like count, user info etc.) for each channel. This structure facilitates a deeper exploration of not only the topics being discussed but also the participants involved, their audience interactions, and how these conversations and engagement trends evolve over time.

Exploratory Data Analysis

Table 1 provides an overview of the YouTube playlists used for the dataset by news channel. For each channel, the table includes the total number of videos, the approximate duration of all videos, and female and male speakers' distribution in hours.

Except for Radio Canada, which publishes multiple extractions from TV news broadcasts daily, the other channels typically publish a single news video per day. The daily average number of videos published is 2.88 for Radio Canada Info and 1.01 for CBC News, CTV News, and Global News. This disparity reflects Radio Canada's unique approach to content dissemination, focusing on extracting and publishing shorter segments from its broadcasts, in contrast to the other channels' consolidated podcast-style uploads.

Figure 2 compares the gender representation gap across news topics, measured as the percentage difference in speaking duration between male and female speakers. Positive values indicate greater male speaking time, while negative values indicate greater female speaking time. Each topic displays two bars: one that includes presenter speech (blue), and one that excludes it (red). This visual highlights how presenter airtime can substantially affect gender balance metrics—particularly in categories such as *health*, *society*, and *science and technology*, where the disparity either increases or reverses depending on whether presenters are included during analysis.

Figure 3 illustrates the overall distribution of speaking time across topics in the OCanada-TVNews dataset. The prominence of categories like *politics*, *conflict, war and peace*, and *disaster, accident and emergency incident* reflects the dominant themes in Canadian TV news coverage over the two-year period. This distribution provides a useful lens for examining editorial priorities and recurring narrative frames in mainstream news broadcasts.

Temporal distribution of top 5 most covered topics (extracted from Figure 3) is illustrated in Figure 4, and high-

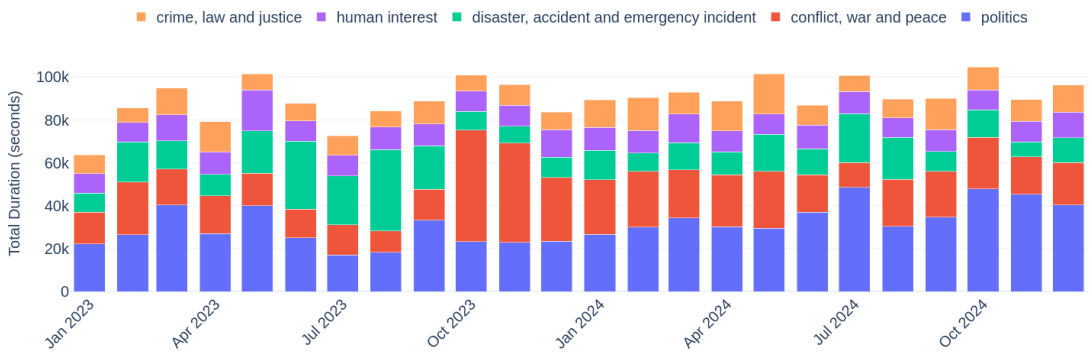

Figure 4: Temporal Distribution of Top 5 News Topics

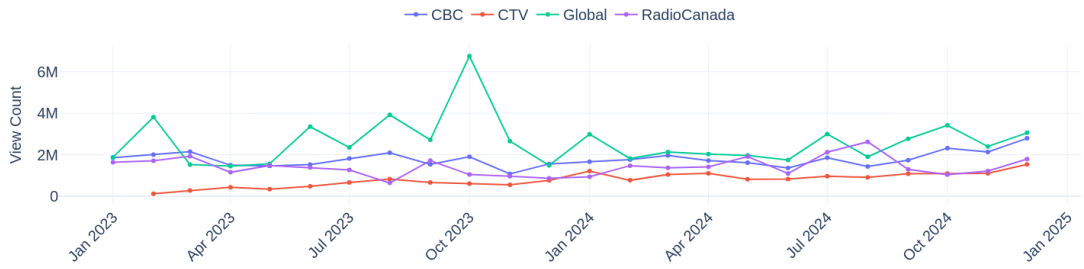

Figure 5: View Count Trends Across Channels

lights significant shifts in coverage during major events. For instance, October 2023 shows a substantial spike in the "conflict, war, and peace" category, aligning with the onset of the Gaza-Israel conflict. Similarly, August 2023 exhibits a notable rise in the "disaster, accident, and emergency incident" category, coinciding with the Canadian wildfires. These trends demonstrate the dataset's ability to capture temporal variations in news topics, enabling researchers to focus on specific periods and analyze how media coverage responds to major events.

In addition, Figure 5 shows the view counts for videos across channels over time, reflecting audience engagement in response to significant events. Peaks in viewership often correlate with high-interest periods, such as those associated with the Gaza-Israel conflict or the Canadian wildfires. This highlights the dataset's utility not only in analyzing the content of news coverage but also in understanding how audience attention shifts in response to global and national events. Together, these insights provide a comprehensive perspective on the dynamic interplay between news topics and audience engagement.

Discussion and Conclusion

This work highlights the interplay between gender representation and news coverage topics in Canadian television broadcasts uploaded to YouTube. The findings reveal differences in speaking duration between male and female speakers, demonstrating that news channels allocate airtime in ways that may reinforce or reduce existing gender disparities. By focusing on both presenters and non-presenters, the analysis shows that on-screen hosts can significantly influence overall impressions of gender balance. Additionally, the topic breakdown points to varying editorial priorities across channels, suggesting that some may emphasize areas such as politics or conflict more heavily than others.

A key strength of this dataset is its daily scope: analyzing news over two years allows researchers to capture fluctuations tied to major events, policy changes, and societal shifts. However, the dataset currently labels speakers as male or female, a dichotomy that does not capture nuances such as non-binary identities. Other limitations include the potential for transcription errors and the reliance on YouTube's availability for content. Future research can build on this work by refining speaker identification methods, incorporating more inclusive gender labeling, and expanding the range of topics assessed. The dataset also provides opportunities for cross-lingual studies given Radio-Canada's French-language content.

In summary, OCanada-TVNews offers a comprehensive, time-based view of Canadian TV news, including valuable metadata on speaker identity, gender, and topic classification. It not only sheds light on how gender dynamics unfold in the news but also underscores how editorial priorities differ across major Canadian channels. Researchers, media practitioners, and policy makers can use this resource to track shifts in public discourse and investigate whether news coverage reflects the diversity of voices in society. By making the dataset openly available, this paper invites further exploration and discussion of gender equity in the media,

laying the groundwork for additional studies into the social impact of broadcast journalism.

Ethical Statement The proposed dataset adheres to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability. It is hosted on FigShare, assigned a unique DOI to ensure findability. The dataset is provided in a standardized, well-documented format with detailed metadata and schema descriptions, enabling seamless interoperability across research tools. To support reusability, we will also include code for processing and analysis, facilitating cross-platform research on gender representation and topical trends.

We only collected publicly available YouTube data, adhering to platform policies and respecting user privacy. By maintaining compliance with legal and ethical guidelines, we ensure the dataset promotes responsible research while protecting the privacy and rights of individuals and communities.

References

Asr, F. T.; Mazraeh, M.; Lopes, A.; Gautam, V.; Gonzales, J.; Rao, P.; and Taboada, M. 2021. The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media. 16(1): e0245533. Publisher: Public Library of Science.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, 12449–12460. Curran Associates, Inc.

Bain, M.; Huh, J.; Han, T.; and Zisserman, A. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. arXiv:2303.00747.

Bredin, H. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *INTERSPEECH 2023*, 1983–1987. ISCA.

CLASSLA. 2023. multilingual-IPTC-news-topic-classifier. Accessed: [2025-01-10].

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, 3830–3834.

Fiscus, J.; Ajot, J.; Michel, M.; and Garofolo, J. 2006. The Rich Transcription 2006 Spring Meeting Recognition Evaluation. 309–322. ISBN 978-3-540-32549-9.

Fiury, A. 2023. wav2vec2-large-xlsr-53-gender-recognition-librispeech. Accessed: [2025-01-10].

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

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

JuanBindez. 2023. pytubefix: A Python library for downloading YouTube videos. Accessed: [2025-01-10].

Macharia, S. 2020. Global Media Monitoring Project (GMMP). In *The International Encyclopedia of Gender, Media, and Communication*, 1–6. John Wiley & Sons, Ltd. ISBN 978-1-119-42912-8.

Newman, N.; Fletcher, R.; Schulz, A.; Andi, S.; Robertson, C. T.; and Nielsen, R. K. 2021. Reuters Institute digital news report 2021. *Reuters Institute for the study of Journalism*.

Pelloin, V.; Dodson, L.; Émile Chapuis; Hervé, N.; and Doukhan, D. 2024. Automatic Classification of News Subjects in Broadcast News: Application to a Gender Bias Representation Analysis. arXiv:2407.14180.

Plaquet, A.; and Bredin, H. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *INTERSPEECH 2023*, 3222–3226.

Rao, P.; and Taboada, M. 2021. Gender Bias in the News: A Scalable Topic Modelling and Visualization Framework. 4. Publisher: Frontiers.

Soumah, V.-G.; Rao, P.; Eibl, P.; and Taboada, M. 2023. Radar de Parité: An NLP system to measure gender representation in French news stories. arXiv:2304.09982.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Answer: Yes, the research question focuses on improving understanding of gender and topic representation in news media without violating privacy norms or societal ethics. The dataset only includes publicly available data and adheres to platform policies.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Answer: Yes, the abstract and introduction clearly align with the paper's contributions to dataset creation, topic and gender analysis, and the exploration of media content trends.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Answer: The methodological pipeline (diarization, gender detection, topic classification) is designed to align with the dataset's scope and objectives. Together, these methods are robust, reproducible, and validated for their suitability to address the research questions, ensuring the claims about temporal and topical trends, as well as gender disparities, are supported by the dataset.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Answer: Yes, the potential artifacts are highlighted, including biases in gender detection models and topic classifiers due to imbalanced training data and regional content distributions.

   (e) Did you describe the limitations of your work? Answer: Yes, limitations such as the reliance on YouTube data, potential biases in speaker diarization and gender detection are explained.

(f) Did you discuss any potential negative societal impacts of your work? Answer

(g) Did you discuss any potential misuse of your work? Answer: No becasue we do not think there can be any potential misuse of our work

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Answer: Yes and we removed the comments from Youtube metadata to prevent these kind of issues.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Answer: Yes

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Answer: NA

(b) Have you provided justifications for all theoretical results? Answer: NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Answer: NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Answer: NA

(e) Did you address potential biases or limitations in your theoretical framework? Answer: NA

(f) Have you related your theoretical results to the existing literature in social science? Answer: NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Answer: NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? Answer: NA

(b) Did you include complete proofs of all theoretical results? Answer: NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Answer: NA

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Answer: NA

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Answer: NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Answer: NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Answer: NA

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Answer: NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Answer: Yes, all existing models and tools (e.g., PyAnnote, wav2vec2, Hugging Face models) are properly cited.

(b) Did you mention the license of the assets? Answer: No, but shared the link where license information is available.

(c) Did you include any new assets in the supplemental material or as a URL? Answer: No.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Answer: Yes, only publicly available YouTube data from Youtube Data API is used, in compliance with platform policies.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Answer: No because it only contains publicly available data and we removed comments data.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Answer: Yes, the dataset is FAIR-compliant, with documentation, accessibility, and a persistent DOI for findability.

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

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

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

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

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

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