Dynamics of Toxicity in Political Podcasts

Naquee Rizwan
nrizwan@kgpian.iitkgp.ac.in
Indian Institute of Technology
Kharagpur, West Bengal, India

Nayandeep Deb
nayandeepdeb125@kgpian.iitkgp.ac.in
Indian Institute of Technology
Kharagpur, West Bengal, India

Sarthak Roy
sarthak.cse22@kgpian.iitkgp.ac.in
Indian Institute of Technology
Kharagpur, West Bengal, India

Vishwajeet Singh Solanki
vsinghsolanki@kgpian.iitkgp.ac.in
Indian Institute of Technology
Kharagpur, West Bengal, India

Kiran Garimella
kg766@comminfo.rutgers.edu
Rutgers School of Communication
and Information
USA

Animesh Mukherjee
animeshm@cse.iitkgp.ac.in
Indian Institute of Technology
Kharagpur, West Bengal, India

Abstract

Toxicity in digital media poses significant challenges, yet little attention has been given to its dynamics within the rapidly growing medium of podcasts. This paper addresses this gap by analyzing political podcast data to study the emergence and propagation of toxicity, focusing on conversation chains—structured reply patterns within podcast transcripts. Leveraging state-of-the-art transcription models and advanced conversational analysis techniques, we systematically examine toxic discourse in over 30 popular political podcasts in the United States.

Our key contributions include: (1) creating a comprehensive dataset of transcribed and diarized political podcasts, identifying thousands of toxic instances using Google's Perspective API, (2) uncovering concerning trends where a majority of episodes contain at least one toxic instance, (3) introducing toxic conversation chains and analyzing their structural and linguistic properties, revealing characteristics such as longer durations, repetitive patterns, figurative language, and emotional cues tied to anger and annoyance, (4) identifying demand-related words like 'want,' 'like,' and 'know' as precursors to toxicity, and (5) developing predictive models to anticipate toxicity shifts based on annotated change points.

Our findings provide critical insights into podcast toxicity and establish a foundation for future research on real-time monitoring and intervention mechanisms to foster healthier discourse in this influential medium.

Warning: Contains potentially abusive/toxic contents.

1 Introduction

Understanding and addressing toxicity in digital media is an ongoing challenge for researchers and practitioners alike. While much attention has been paid to platforms such as social media and online forums, comparatively less is known about the nature and dynamics of toxicity within the rapidly growing medium of podcasts. In this study, we aim to bridge this gap by collecting and analyzing political podcast data, focusing specifically on conversation chains – structured reply patterns within podcast transcripts – to study the emergence and propagation of toxicity in this medium.

Podcasting has seen remarkable growth in recent years, emerging as a mainstream medium for information and entertainment. As of 2024, there are over 500 million podcast listeners worldwide, representing 23.5% of all internet users. In the United States alone, 47% of adults have listened to a podcast within the last month, a figure that has more than tripled over the past decade. Politics and government rank as the third most popular podcast topic, with 41% of listeners regularly tuning in to content in this category [8, 25]. This growth underscores the importance of podcasts as a platform for political discourse, but it also highlights their potential as vehicles for misinformation, conspiracies, and hate speech [30]. Unlike traditional forms of media, podcasts remain largely unmoderated, further exacerbating the risks of unchecked toxicity. Understanding toxicity in podcasts is therefore not only academically significant but also crucial for ensuring healthy democratic discourse.

Studying toxicity in podcasts presents unique challenges. Unlike text-based content, the audio format of podcasts makes them inherently difficult to monitor and analyze using traditional strategies. Transcribing large volumes of audio content incurs significant computational and financial costs, making large-scale studies impractical without significant resources. Further, audience interaction in podcasts is limited compared to social media platforms, where users can directly respond to or fact-check content. This lack of interactivity complicates efforts to understand how toxic narratives resonate with audiences. Moreover, analyzing toxicity within podcast transcripts requires not only identifying toxic language but also tracking how it ebbs and flows over time – examining what triggers toxic conversations and how they evolve within conversational structures. These complexities highlight why naïve approaches to studying podcast toxicity often fail to capture its multifaceted nature.

Existing research provides valuable starting points, but fails to address the problem in a comprehensive way. Advances in automatic transcription tools, such as Whisper, now allow for large-scale analysis of audio data. However, as previous work on datasets like RadioTalk [3] demonstrates, transcription errors and content ambiguity remain significant barriers, particularly for conversational and emotive formats such as podcasts. In addition, though the study of toxicity online is an active research area [21], there is little work on understanding conversational structures of toxicity and how it evolves [32].

In this paper, we introduce the concept of toxic conversation chains in podcasts. Using state-of-the-art transcription models and conversational analysis techniques, we explore how toxicity emerges and spreads in political podcasts. Our approach combines scalable transcription workflows with innovative analysis of conversational dynamics, providing novel insights into the structure and flow of toxic discourse in this medium.

Our work represents a critical step towards systematically analyzing toxicity in podcasts and lays the foundation for future research in this area.

Our **key contributions and observations** are as follows:

- We collect, transcribe and diarize a dataset of over 30 popular political podcasts from the US, reaching tens of millions of people, and identify thousands of instances of toxicity in these podcasts (as defined by Google's Perspective API).
- We identify concerning trends of a majority of episodes from many popular podcasts containing at least one instance of toxicity.
- We build toxic conversation chains from the transcriptions where each chain has a highly toxic part which we call the *anchor* and a series of preceding and following parts to express the context of the anchor.
- Analysis of these message chains reveals that the anchor text is (a) longer in duration, (b) more repetitive, (c) incorporates figurative language such as metaphors and hyperboles, and (d) is closely tied to emotions like anger, fury, and annoyance.
- Linguistic analysis reveals an intriguing observation: preceding words like 'want,' 'like,' and 'know,' which often express demands, can escalate into toxic conversations.
- We annotated the top 100 toxic conversation chains to identify change points where toxicity levels shift. Using these annotations, we developed a model to predict the temporal trajectory of these conversations. Our findings suggest that such automated methods hold significant potential for real-time monitoring and the development of effective intervention mechanisms in the future.

2 Related works

Toxicity on social media: Toxicity, broadly characterized as disrespectful, harmful, or unreasonable communication, has been a significant focus in social media research [28]. It manifests itself through harassment, hate speech, and polarized discourse, often perpetuated by bots and trolls that amplify its spread. Studies highlight that toxicity often combines with misinformation and polarization, forming a "cycle of online extremism" that exacerbates its societal impact [27].

Detection of toxicity: Defining and detecting toxicity requires nuanced approaches due to its context-sensitive nature. Techniques like BERT-based sentiment analysis have proven effective in detecting toxic content on platforms such as Twitter, leveraging large labeled datasets to improve classification accuracy [19]. Further, context-aware models are necessary to differentiate between benign negativity and harmful toxicity, particularly in multi-dimensional interactions [9]. Perspective API [16] is a valuable tool for researchers to analyze toxicity of online content, allowing them to study the spread of hate speech, cyberbullying, and other harmful online behaviors in different languages and cultures.

Toxicity in podcasts: Unlike traditional social media, podcasts present unique challenges due to their audio form. Transcription is often necessary for textual analysis, but introduces errors and challenges such as speaker identification and role prediction [6] that can skew results. Podcasts also allow for prolonged discussions, creating more opportunities for toxicity to escalate. Recent work has curated datasets to study toxicity in political podcasts, identifying patterns in conversational structures that promote or mitigate toxicity [17, 29]. Similarly, the authors in [2] analyze the political biases and content strategies of YouTube and Rumble podcasts, revealing a right-wing orientation in Rumble's content and a more diverse perspective on YouTube. As a mitigation strategy, the authors in [20] propose a novel approach to combat misinformation in podcasts using auditory alerts to notify listeners of potential misinformation in real-time.

Given this space, to the best of our knowledge, we are the first to look at toxicity on podcasts and to make use of the conversational aspects in podcasts. Our contributions are two fold: first is to show the prevalence of toxicity in popular political podcasts and the development of methods to look at toxicity in conversations rather than considering them as a one-off instance. Our methods and analysis extend beyond just podcasts and can be applicable in any conversational setting, which is prominent but highly under explored in NLP [32].

3 Dataset

In this section, we discuss the details of curating, transcribing and diarizing the dataset being used for this work.

We construct the dataset based on channels list curated and shared by [30]. They identify popular political podcasts by analyzing *Apple Podcasts'* top hundred lists from two specific time periods and selecting talk shows that discuss politics, policy, current events, or news. The process is expanded by including related political punditry podcasts from the 'you might also like' recommendations, creating a comprehensive sample for analysis. *RSS feeds* are used to download the wav files and the podcasts span around the years 2022-2023. The full list of 31 podcast shows we used in this paper are shown in Table 3 in the Appendix. We acknowledge that these channels might have a bias with over representation from right leaning sources since the original objective of the dataset curated by [30] was to study conspiracies and misinformation. Having said that, care should be taken in interpreting the results particularly in the context of prevalence of toxicity, since our dataset only represents a small, biased subset of political podcasts.

That said, even though the set of podcasts we consider is small, these shows might have disproportionate impacts and people who are in power might disproportionately listen to these. For e.g., a lot of January 6, 2021 US Capitol attack [1] coordination and top down messages iterate through the podcasts of personalities like Steve Bannon. Hence, the claims made on these podcasts may have particularly consequential implications for broader public opinion and political discourse. So, as a part of this work, we try to understand the toxicity being driven through conversational chains so that such content can be auto-moderated and avoided. A detailed summary of the dataset over thirty-one podcast channels, and some metadata associated with them is shown in Appendix 3.

We use Whisper [22] to transcribe and Pyannote[2] to diarize and identify individual speakers [4]. We filter out overlapping speech

[1] https://www.britannica.com/event/January-6-U-S-Capitol-attack
[2] https://pyannote.github.io/

segments to avoid performance degradation in speaker clustering during diarization, which is accomplished using the same Pyannote model. The diarization process typically achieves high-quality results with over 90% accuracy, though some misclassifications do occur. For a detailed analysis of diarization quality, please refer to [4]. Since our study focuses on conversation chains (see Section 4), it is crucial to maintain continuous speaker conversations without interruptions. In this context, the 90% accuracy is sufficient for our analysis, as we are examining conversation chains rather than individual speaker behaviors. When necessary, we outline the specific cleaning approaches used in subsequent sections.

After processing, each episode in the dataset is made up of a number of speaker turn chunks, as returned by Pyannote library. Each chunk contains (a) start time (b) end time (c) unique speaker id, and, (d) diarized text. All these metadata are corresponding to these uneven chunks. As a next step, we further organize our data in the Section 4 for making conversational chains which are uniformly distributed to carry out our analysis.

4 Toxic conversation chains

In this section, we discuss the process employed to accurately calculate the toxicity scores and the subsequent formulation of toxic conversation chains using these toxicity scores.

4.1 Toxic score calculation

To compute the toxicity of a speaker's speech we pass the diarized text corresponding to the speech to Perspective API.[3] We use the benchmark Toxicity attribute, available with the API to obtain the toxicity scores. Since the Perspective API can handle a limited number of tokens we split the diarized conversation text corresponding to a speaker turn into *chunks* of 17 seconds since this is the median conversation time of a turn in our dataset (see Figure 1). We club 1-4 chunks (approximately *one* minute) into a *segment*. Each segment corresponds to a specific speaker identified by a speaker ID, a set of (at most) four chunks, the start and end timestamps and the diarized text corresponding to the chunks in the segment. The toxicity score of a segment is set to the maximum of the toxicity score across these chunks within the segment. If a speaker turn exceeds the duration limit of a segment, it is broken into the required number of contiguous segments. We perform all our analysis at the level of these one minute segments so that the observations are statistically meaningful and not over dominated by a speaker speaking over a long stretch.

Figure 2 shows the top 10 podcasts with the highest percentage of episodes containing at least one instance of toxic conversations. Certain shows such as *Get Off My Lawn Podcast w/ Gavin Mcinnes*, *The New Abnormal*, and *Louder with Crowder* have at least one instance of a toxic conversation in all of their episodes. Table 3 shows the full statistics for all the shows in our dataset. It is really surprising to see instances of toxicity in hundreds of episodes even on shows with tens of millions of listeners (e.g. *The Dan Bongino Show*).

[3]https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

4.2 Toxic chain formulation

A toxic chain constitutes an *anchor* segment, the *preceding ten* and the *following ten* segments of the anchor segment (see Figure 1). A segment is identified as an anchor segment if its toxicity score is 0.7 or higher. This threshold is set as per the guidelines noted in documentation of the Perspective API.[4] We take the previous and the next *ten* segments to understand what leads to the rise in toxicity and what happens after a peak toxic threshold has been reached. Anchoring on this threshold, we perform our analysis on a total of 8,634 chains. A set of representative examples of toxic conversation chains are provided in Figure 8 & Appendix 11.

4.3 Distribution of the toxic chains across podcasts

We compute the distribution of the toxic chains across the different podcast channels. Figure 3 shows the distribution of top ten channels. The top *five* contributing podcast channels are *Louder with Crowder*, *Mark Levin Podcast*, *The New Abnormal*, *The Dan Bongino Show* and *Get Off My Lawn Podcast w/ Gavin McInnes* in decreasing order of contribution aggregating to a total of *63.6%*. Together, these podcasts potentially reach tens of millions of listeners every week.[5] Many of these channels have been reported to be hosted and attended by people making homophobic, and racist remarks.[6,7,8]

5 Analysis of the toxic conversation chains

In this section we present an in-depth analysis of the segments constructed in the previous section.

5.1 Time coverage

The duration of each segment in a conversation chain is the difference between the `start` and the `end` time obtained during diarization process. We calculate the mean of the segment's duration across all the chains and plot them in Figure 4. Error bars indicate 95% confidence intervals. Interestingly, the anchor segment has the highest mean implying that speakers tend to speak for longer duration when their speech is most toxic.

Further, the rise in the mean duration is not sudden and neither is the fall. While the ascend starts from `Segment -1` itself, the decay is more gradual and flattens after `Segment 3`.

5.2 Basic textual properties

We now study the basic textual properties of the diarized text corresponding to the anchor segments and compare them with the other segments in a chain. For all our textual analysis, we use the `PyNLPl`[9] library.

(i) Token count: We calculate the total number of tokens in the diarized text for each segment in a conversation chain. To break the words into tokens, we use the *word_tokenize* from NLTK library.

[4]https://support.perspectiveapi.com/s/about-the-api-score?language=en_US
[5]e.g. Just Dan Bongino and Mark Levin podcasts are extremely popular with over 20 million weekly listeners. See https://bit.ly/3ZEArcS
[6]https://tinyurl.com/louder-crowder
[7]https://tinyurl.com/dan-bingo-show1
[8]https://tinyurl.com/dan-bingo-show2
[9]https://pypi.org/project/PyNLPl/

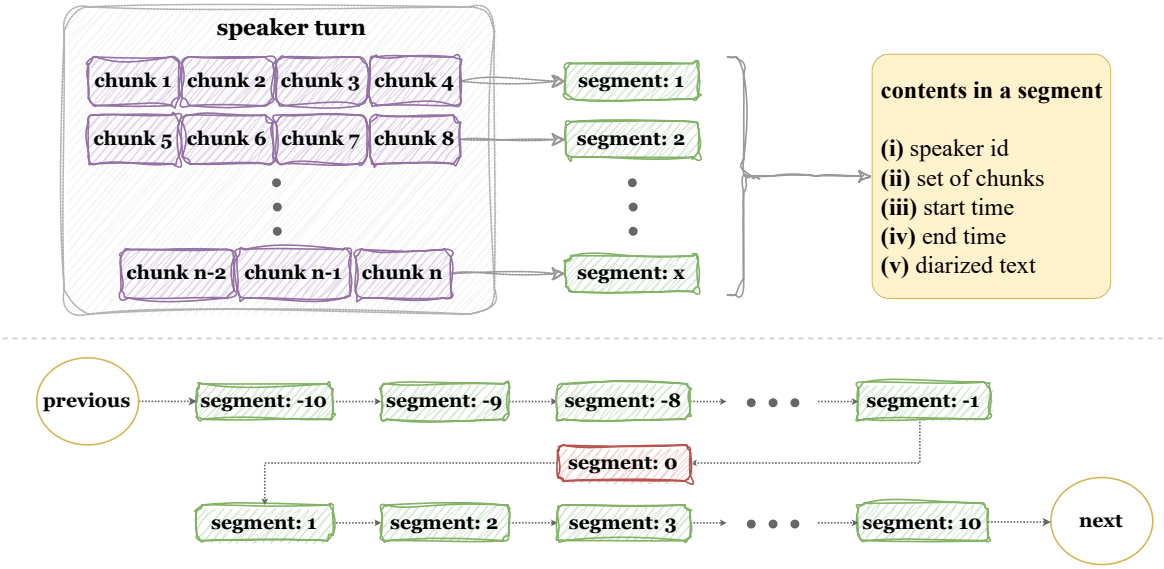

Figure 1: Top: Computation of segments from chunks and their contents. Bottom: Schema for toxic conversation chains. The segment marked in red color represents the anchored segment with toxicity above a threshold of 0.7.

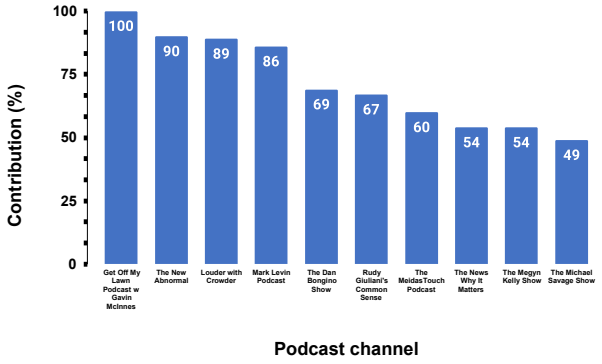

Figure 2: Top 10 podcast shows with most amount of toxic content. The bars show the percentage of episodes containing at least one toxic conversation.

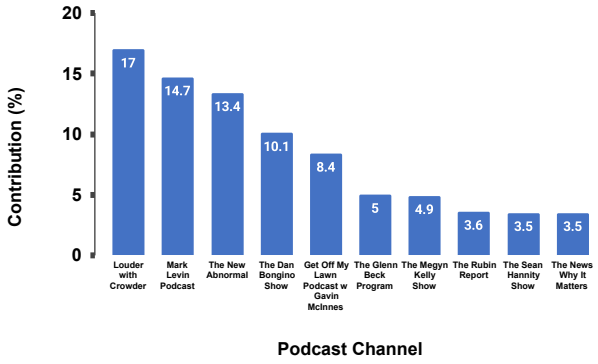

Figure 3: Distribution of toxic conversation chains across the podcast channels. Percentage contribution for top 10 podcast channels are shown.

Figure 5 shows that the mean token count is the largest for the anchor segment which naturally follows from the earlier observation that these segments have the longest mean duration.

(ii) Type token ratio (TTR): This metric is used to evaluate the diversity of vocabulary in a text. It is given by the formula: $\frac{\#types}{\#tokens}$, where *types* are unique *tokens*. Higher values indicate greater lexical diversity which means that the text has less redundancy. Lower values suggest repetitive textual content. Figure 5 shows that the mean TTR for anchor segment is lower than all other segments in the chain indicating that there is more repetition (possibly of the same hateful remark) in the anchor segment.

(iii) Token entropy: Using the unigram probability values of the tokens we compute the entropy of the diarized text. Figure 5 shows that the anchor segment has a higher average entropy compared to the other segments indicating that the text in the anchor segment is more random and less well-formed.

(iv) Perplexity: Perplexity measures the prediction ability of a language model. Formally, if a language model has a perplexity P, then it means that the model is as uncertain as if it is selecting P equally likely choices which is mathematically represented as: $P = Pr(w_1, w_2, ..., w_n)^{-1}$ where $w_1, w_2, ..., w_n$ are a sequence of tokens and Pr is the language model. For our use case, Pr is the unigram model where the unigram probabilites are obtained from the diarized text. Figure 5 shows that the mean perplexity is the

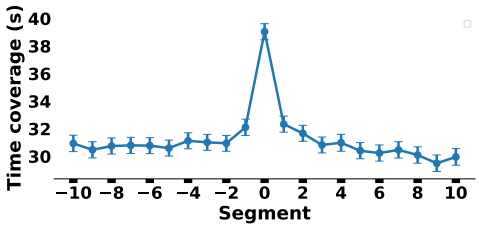

Figure 4: Mean of time coverage with 95% confidence interval across segments in seconds.

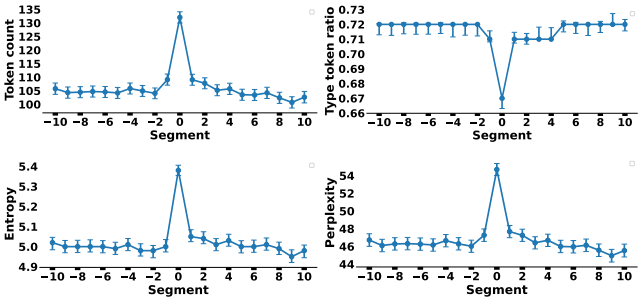

Figure 5: Mean at 95% confidence intervals for token count, type to token ration (TTR), entropy & perplexity across segments.

highest for the anchor segment indicating that the conversation is least organized/coherent in this segment.

5.3 Figurative language

Human conversation, especially in a public discourse like podcasts are often strewn with figurative language like hyperboles and metaphors [5]. Hyperboles incorporate exaggeration for emphasis and metaphor is used to make implied comparison. In this section, we report the extent of such figurative language present in the conversation chains. We use the `bert-large-uncased` model fine-tuned on the STL task [1] to infer the hyperboles and metaphors from the diarized text. From Figure 6, we observe that, as expected, the usage of metaphors in podcasts is generally high ranging between 69.58% – 73.53% for the non-anchor segments. Remarkably, it reaches to an all high of 81.91% for the anchor segment. Further, we observe that the anchor segment has nearly double the percentage of hyperboles (18.03%) compared to the highest among non-anchor segments (9.19%). Thus, in summary, the toxic segment has the highest exaggeration level and more frequently invokes implied comparison. Both of these work as possible means to emphasize on the embedded toxic remark making it intense as well as subtle at the same time.

5.4 Empath features

Empath is a metric which is frequently used in text analytics to assess the emotional content of the text. We also use empath as a metric to understand the underlying emotions and behaviourial

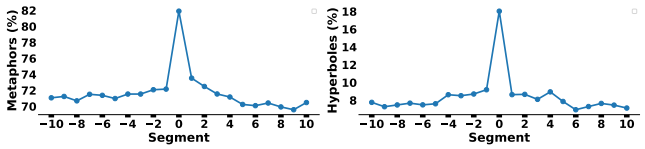

Figure 6: Metaphors (left), and, Hyperboles (right) across segments. The numbers are percentages (%).

patterns latent in the conversation chains. Since our inputs are conversation chains and not flat text, we finetune the DistilBERT [24] model over the dataset presented in [23] to infer the 32 empath features from these chains. In Figure 7 we show the top eight most frequent emotions that have occurred at least 5% of times within a segment. We observe considerably high intensity for the following empath features – 'angry', 'furious', 'annoyed', 'disguted' in the anchor segment. On the other hand, there is a reduced intensity of the following empath features – 'surprised', 'hopeful', 'afraid', 'anticipating' in the anchor segment. Thus toxic conversations are laden with aggression resulting in the loss of decorum in the conversation.

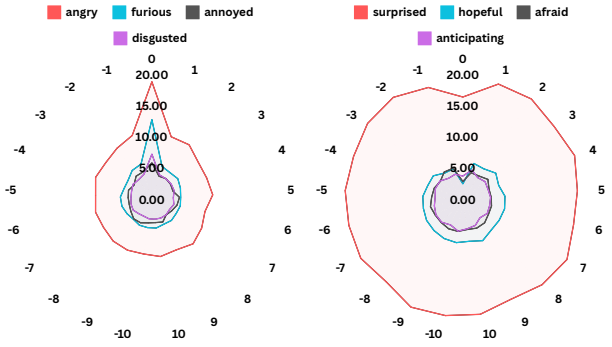

Figure 7: The plot on the left presents the Empath features; those which increased for anchor segment. The plot on the right illustrates the ones with decreased value. In each of these plots, anchor point is '0'; left & right semicircles represent previous & next segments respectively.

5.5 Keywords

In order to identify and compare the keywords present in the segments we extract the top ten toxic key-phrases from each segment in every conversation chain using KeyBERT [10]. Each phrase is limited to a maximum of five-grams. The word embeddings for KeyBERT algorithm are obtained from the Detoxify [12] library.[10] To ensure diversity in our extracted key-phrases, we use the maximal margin relevance utility of the KeyBERT algorithm enforcing a diversity score of 0.75 . In addition, we eliminate stop words & punctuations and only retain strings containing characters from the English alphabet. We combine all the keywords obtained from the preceding ten segments and represent them as a word cloud;

[10]https://huggingface.co/unitary/toxic-bert

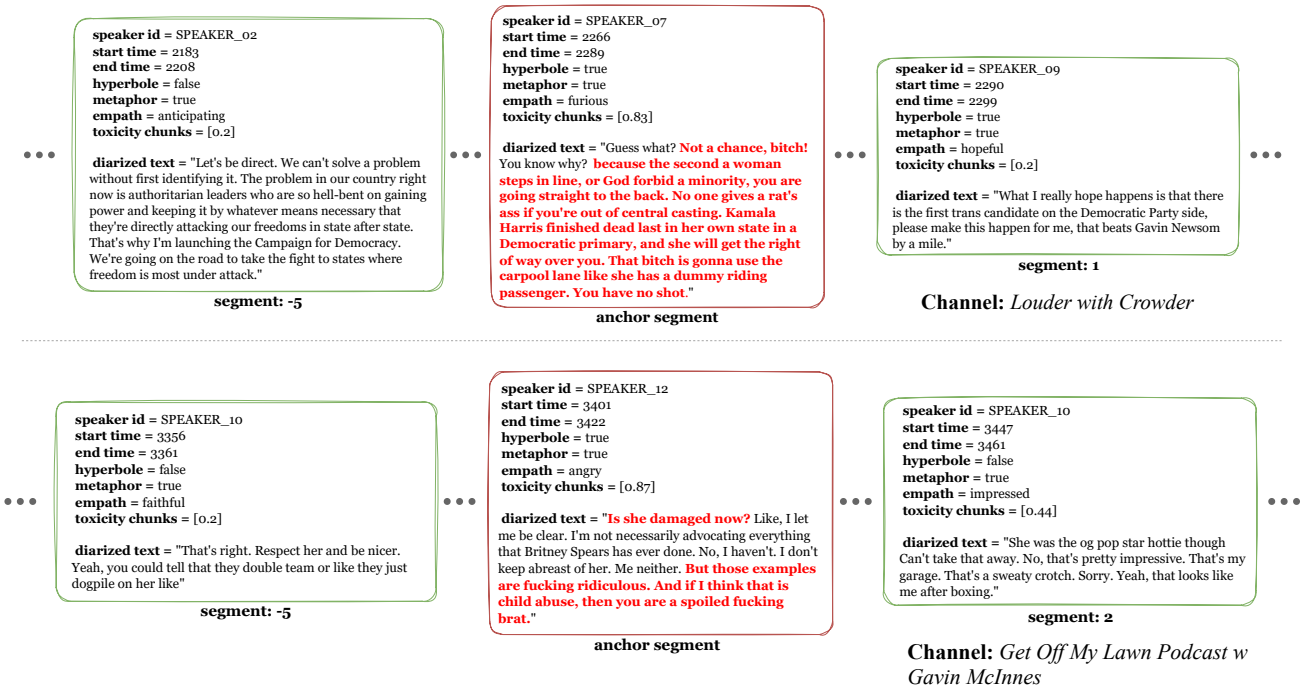

Figure 8: Examples: Since it is not feasible to illustrate all segments, one among previous and next segments are shown along with anchor segment from the toxic conversation chain. Toxic texts in the anchor segment are marked in red. Note: start and end times are in seconds.

similarly, we combine all the keywords obtained from the following ten segments and represent them as another word cloud. We also obtain the word cloud for the anchor segment separately. These word clouds in series are illustrated in Figure 9. We observe that most of the words are linked to political and controversial themes like 'right', 'republican', 'democrat', 'women', 'biden', 'american' & 'liberal' which naturally follows from the choice of our dataset. The anchor segment has a number of toxic keywords including 'idiot', 'stupid', 'f*cking/f*ck', 'sh*t', 'a*s' & 'moron'. The high similarity between the previous and next word clouds indicate that the anchor segment introduces a disruption in the flow of the main conversation which rewinds back to normal only at the end of the anchor segment. Finally, in both the previous and the next word clouds, words like 'want', 'know', 'people' & 'yeah' appear which reflect an expression of demand. Thus hostility in the speech of a particular speaker seem to be fueled by words of demand from either the anchor or other speakers/participants in podcast conversations. .

Figure 9: Word clouds for previous, anchor and next segments.

5.6 Topical shifts

The keyword analysis in the previous section indicates that there is a significant change in the conversation content during the transition from the previous to the anchor segment. This hints to the fact that there is a possible topical/thematic shift during such a transition. In order to establish this we perform topic modeling using the BERTopic [11] model. We extract the topics considering the previous ten aggregated segments a one document, the next ten aggregated segments as a second document and the anchor segment as the third document. We set the number of topics to three and report the top ten most representative words for that topic which has the highest probability of association with a document. The results are noted in Table 1, which reveal significant shifts in thematic focus and toxicity level during the transition from the previous to the anchor segment and the anchor segment to the next segment. Precisely the anchor topic is highly toxic while the preceding and the following topics are more related to demands

thus offering insights into the progression and contextual drivers of toxic conversations.

Order in chain	Induced topics
Preceding	*like, know, people, go, right, yeah, think, get, say, going*
Anchor	*bitch, stupid, son, fuck, shit, fucking, idiot, damn, shut, guy*
Following	*like, know, people, go, right, yeah, think, get, say, going*

Table 1: Topic transitions in the conversation chains induced by BERTopic.

6 Change points in toxic chains

While linguistic properties studied in the previous section enables us to characterize the conversation chains, it is not possible to formulate actionable insights for designing effective interventions. In order to achieve this goal one has to acknowledge that toxic conversation chains are inherently dynamic, with evolving interactions that may involve multiple participants. Understanding the trajectory of these conversations requires identifying key points where significant shifts occur. Automatic change point detection can enable researchers to: (i) **Isolate key moments**: Automatically identify segments of conversations where toxicity sharply increases/decreases or new topics are introduced; and, (ii) **Understand contextual triggers**: Track down the patterns latent in events or statements that precede toxicity spikes or topic changes and thus reveal actionable insights for designing intervention strategies.

6.1 Automatic change point detection

We utilize standard change point detection (CPD) algorithms to automatically identify change points within conversation chains. Specifically, we consider various search methods – Pelt, KernelCPD, Window, BottomUp & Binseg alongside different cost functions including `rbf`, `cosine`, `l2` & `linear`. All implementations are done through `ruptures` [26] library.[11] According to the documentation, twelve combinations over these search methods and cost functions are possible; however only `rbf` as the cost function coupled with Pelt, KernelCPD, BottomUp & Binseg search methods successfully identifies at least one change point. As a result, we conduct our experiments and identify change points using these four algorithms.

6.2 Manual annotation of change points

To evaluate the predictive performance of the baseline CPD algorithms we conduct a manual annotation exercise of the change points. We consider as many as top 100 toxic conversation chains based on the toxicity scores of the anchor segments. We then get all the change points in each of these chains annotated by three annotators. All the three annotators are experts in hate speech analysis research. To ensure consistency and reliability in the annotation process, we ask the annotators to base their judgments on the following factors – (a) *tone*: shifts in sentiment or intensity, such as transitions from neutral or mildly toxic statements to overt hostility, (b) *topical shift*: the emergence of new discussion themes or the cessation of previously dominant topics and (c) *change in toxicity*: escalations or reductions in the degree of harmful language or expressions. Further since this is a sensitive task we instruct the annotators to (a) maintain confidentiality and handle the conversational data responsibly, (b) focus on objective evaluation without imposing personal biases and (c) annotate at most 15 data points per day to ensure annotation quality and mental well being of annotators. We appropriately remunerate the annotators with Amazon gift vouchers.

Overall, the annotators marked a median of 3 change points for most of the chains with the minimum and maximum being 1 and 7 respectively. The total number of change points marked by them across all the 100 chains is 984. We believe that this is a unique and a first-of-its-kind dataset that can be used in future to train/evaluate other predictive algorithms.

6.3 Performance evaluation

Evaluation metrics: In order to measure the correspondence between the manually annotated change points and those generated by the CPD algorithms we use four standard metrics – *Hausdorff distance*,[12] *rand index*, *precision*, and *recall*. While calculating the metrics, if a particular change point is marked by a majority of the annotators it is considered as a valid change point. While the total number of unique points annotated by all 3 annotators is 615, after the majority voting we arrive at 257 points. The predictions of the algorithms are evaluated against these majority voted change points for evaluation.

Key results: The main results are noted in Table 2. We find that the KernelCPD method stands out as the best-performing algorithm across multiple metrics, particularly in terms of Hausdorff distance, rand index, and precision, where it consistently outperforms the other methods. It achieves the highest precision and a high recall value with the margin of two, making it very reliable in detecting change points with both high accuracy and minimal error. Pelt performs decently in most metrics, but lags behind KernelCPD in terms of Hausdorff distance, rand index and precision, indicating that it may miss some important change points. BottomUp provides moderate performance but is generally less reliable than KernelCPD, particularly in terms of precision. Finally, Binseg shows poor results across all metrics, suggesting that it is the least effective algorithm for detecting change points in toxic conversation chains. We present two representative examples of the change points across two different chains in Figure 10.

7 Discussion

Social media has been extensively studied with respect to intervention and demarcation strategies [13, 14, 31]. Rigorous research projects have lead to the development of various strategies like counter speech [18], text detoxification [7], and meme intervention [13], among many others, including some controversial strategies [15]. Podcasts are comparatively different, since the interaction of the audience with guests/hosts is nearly negligible. Also, unlike

[11] https://centre-borelli.github.io/ruptures-docs/

[12] https://en.wikipedia.org/wiki/Hausdorff_distance

Metrics		Aggregation	Pelt	KernelCPD	BottomUp	Binseg
Hausdorff		avg	6.89	**4.66**	6.89	7.51
		med	7	**4**	7	7
rand index		avg	0.75	**0.84**	0.72	0.53
		med	0.76	**0.86**	0.74	0.54
precision	1	avg	0.3	0.53	0.25	0.15
		med	0.33	0.5	0.25	0
	2	avg	0.37	**0.59**	0.34	0.22
		med	0.33	**0.57**	0.29	0
	4	avg	0.45	0.54	0.35	0.3
		med	0.4	0.5	0.33	0.33
recall	1	avg	0.62	0.76	0.72	0.15
		med	0.6	**1**	0.75	0
	2	avg	0.74	0.85	0.93	0.22
		med	0.75	**1**	**1**	0
	4	avg	0.88	0.79	**0.97**	0.29
		med	**1**	**1**	**1**	0.25

Table 2: Performance comparison of the CPD algorithms across different metrics and two different ways of aggregating the metrics (mean and median). The number 1, 2 and 4 linked with precision and recall are the margins of errors allowed. The best results are highlighted .

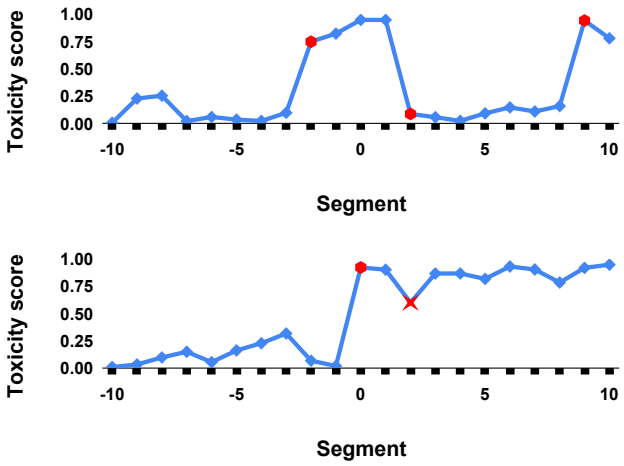

Figure 10: Plots representing two samples comparing human annotation with KernelCPD method's detected change points. Correctly predicted points are marked with a red hexagon and incorrect predictions are marked with red cross.

social media content, they are generally characterized by long conversations and figuring out exact toxic segments in real time is difficult. Thus, for current scenarios, intervention is limited to only beeping or stripping away such contents which is not real time and is also often ignored by the podcast owners and streaming platforms due to the lack of automation strategies. Further, unlike the access of social media contents, podcasts can easily be made available in offline mode and can create a much bigger indirect impact. In addition, the toxic clips can be spread on social media without any further context thus spicing up the whole thing and leading to even worse polarization.

In this paper we introduced a dataset of transcribed political podcasts and formulated toxic conversation chains from the transcribed text. Firstly, we make the surprising observation on the high prevalence of toxicity (detected with high confidence) in podcasts which reach tens of millions of users. Next, we made various interesting observations regarding the linguistic characteristics of these chains. We noted that the anchor segment corresponding to the most toxic part of the chain has a lot of repetition, is less well-formed and has emotions and sentiments related to anger and fury. Expressions of demand in conversations can quickly lead the discourse to higher levels of toxicity. Subsequently, we manually annotate 100 toxic chains with change points for training and evaluation of automatic methods. We plan to release this dataset upon acceptance. Based on this ground-truth dataset we established that CPD algorithms can be effectively used to monitor the toxicity levels of conversation chains in real time.

The CPD algorithm as demonstrated here can be used to automatically monitor the toxicity levels in real time. Such real time monitoring can be very useful in controlling toxicity by making the participating hosts/guests aware of the rising toxicity in their conversation through a recommendation/alert system. Based on adherence to such recommendations, points, badges and other incentives may be awarded to the participants to promote healthy interaction and deescalate toxicity. In cases of extreme violation automatic termination of the recording of the conversation can also be implemented.

As future work, we plan to extend our study to incorporate audio signals as well accompanied by complete and thorough analysis over different audio features. We also plan to develop intervention strategies for this complex scenario. We also plan to expand our dataset and scalable tooling to other political podcasts to get clear estimates on the prevalence of toxicity in podcasts. Finally, an important aspect that our paper does not answer is the impact of such toxicity on the listeners. Our findings lead us to the question on whether toxicity might become normalized if coming from popular podcast hosts who reach tens of millions of people and the consequences it might have on our political discourse. Some of the tools we developed could be used by social scientists to study and answer such questions.

Ethics statement. Our dataset comprises popular podcasts curated from publicly available RSS feeds, which are widely accessible and listened to by tens of millions of people. These podcasts are analyzed using automated transcription and diarization techniques, and while individual speakers are segmented to enable conversational analysis, their identities are neither inferred nor used in the analysis to ensure privacy. We recognize the potential for biases in various stages of our pipeline, including transcription, diarization, toxicity detection, and change point detection. To mitigate these biases, we employed a rigorous evaluation process, comparing multiple algorithms at each stage and selecting the best-performing ones based on empirical results. While our analysis yielded promising findings, we emphasize the importance of conducting a comprehensive audit of the pipeline to further evaluate its reliability and fairness, particularly in production or high-stakes contexts. Our study adheres to ethical guidelines for research on publicly available data and prioritizes transparency, privacy, and the minimization of harm. We aim to contribute constructively to the discourse on toxicity in podcasts while acknowledging the limitations of our methods and the need for ongoing scrutiny and improvement.

References

[1] Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A Match Made in Heaven: A Multi-task Framework for Hyperbole and Metaphor Detection. arXiv:2305.17480 [cs.CL] https://arxiv.org/abs/2305.17480

[2] Utkucan Balci, Jay Patel, Berkan Balci, and Jeremy Blackburn. 2024. Podcast Outcasts: Understanding Rumble's Podcast Dynamics. arXiv:2406.14460 [cs.SI] https://arxiv.org/abs/2406.14460

[3] Doug Beeferman, William Brannon, and Deb Roy. 2019. Radiotalk: A large-scale corpus of talk radio transcripts. *arXiv preprint arXiv:1907.07073* (2019).

[4] Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. *arXiv preprint arXiv:2104.04045* (2021).

[5] Christian Burgers, Elly A. Konijn, and Gerard J. Steen. 2016. Figurative Framing: Shaping Public Discourse Through Metaphor, Hyperbole, and Irony. *Communication Theory* 26, 4 (04 2016), 410–430. https://doi.org/10.1111/comt.12096 arXiv:https://academic.oup.com/ct/article-pdf/26/4/410/21973874/jcomthe0410.pdf

[6] Ann Clifton, 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*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 5903–5917. https://doi.org/10.18653/v1/2020.coling-main.519

[7] Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024. Overview of the Multilingual Text Detoxification Task at PAN 2024. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, Guglielmo Faggioli, Nicola Ferro, Petra Galuščáková, and Alba García Seco de Herrera (Eds.). CEUR-WS.org.

[8] Edison Research. 2024. *The Podcast Consumer 2024.* Technical Report. Edison Research. https://www.edisonresearch.com/the-podcast-consumer-2024-by-edison-research/ Accessed: 2024-12-02.

[9] Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Varna, Bulgaria, 260–266. https://doi.org/10.26615/978-954-452-049-6_036

[10] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265

[11] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794 [cs.CL] https://arxiv.org/abs/2203.05794

[12] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

[13] Prince Jha, Raghav Jain, Konika Mandal, Aman Chadha, Sriparna Saha, and Pushpak Bhattacharyya. 2024. MemeGuard: An LLM and VLM-based Framework for Advancing Content Moderation via Meme Intervention. arXiv:2406.05344 [cs.CL] https://arxiv.org/abs/2406.05344

[14] Kaylee Payne Kruzan, Kofoworola D.A. Williams, Jonah Meyerhoff, Dong Whi Yoo, Linda C. O'Dwyer, Munmun De Choudhury, and David C. Mohr. 2022. Social media-based interventions for adolescent and young adult mental health: A scoping review. *Internet Interventions* 30 (2022), 100578. https://doi.org/10.1016/j.invent.2022.100578

[15] Paddy Leerssen. 2023. An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Computer Law & Security Review* 48 (2023), 105790.

[16] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. arXiv:2202.11176 [cs.CL] https://arxiv.org/abs/2202.11176

[17] Benjamin Litterer, David Jurgens, and Dallas Card. 2024. Mapping the Podcast Ecosystem with the Structured Podcast Research Corpus. *arXiv preprint arXiv:2411.07892* (2024).

[18] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou Shalt Not Hate: Countering Online Hate Speech. *Proceedings of the International AAAI Conference on Web and Social Media* 13, 01 (Jul. 2019), 369–380. https://doi.org/10.1609/icwsm.v13i01.3237

[19] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2022. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. arXiv:2012.10289 [cs.CL] https://arxiv.org/abs/2012.10289

[20] Sachin Pathiyan Cherumanal, Ujwal Gadiraju, and Damiano Spina. 2024. Everything We Hear: Towards Tackling Misinformation in Podcasts. In *International Conference on Multimodel Interaction (ICMI '24)*. ACM, 596–601. https://doi.org/10.1145/3678957.3678959

[21] María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open* 10, 4 (2020), 2158244020973022.

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[23] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. arXiv:1811.00207 [cs.CL] https://arxiv.org/abs/1811.00207

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL] https://arxiv.org/abs/1910.01108

[25] Statista. 2024. Audio podcast consumption in the U.S. 2024. https://www.statista.com/statistics/270365/audio-podcast-consumption-in-the-us/ Accessed: 2024-12-02.

[26] Charles Truong, Laurent Oudre, and Nicolas Vayatis. 2020. Selective review of offline change point detection methods. *Signal Processing* 167 (2020), 107299. https://doi.org/10.1016/j.sigpro.2019.107299

[27] Janikke Solstad Vedeler, Terje Olsen, and John Eriksen. 2019. Hate speech harms: A social justice discussion of disabled Norwegians? experiences. *Disability & Society* 34, 3 (2019), 368–383.

[28] Emily A Vogels. 2021. The State of Online Harassment | Pew Research Center. https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/. (Accessed on 05/02/2021).

[29] Valerie Wirtschafter. 2021. The challenge of detecting misinformation in podcasting. *Brookings* (25 August 2021). https://www.brookings.edu/articles/the-challenge-of-detecting-misinformation-in-podcasting/ Accessed: 2024-12-02.

[30] Valerie Wirtschafter. 2023. Audible reckoning: How top political podcasters spread unsubstantiated and false claims. *Brookings* (March 2023). https://www.brookings.edu/articles/audible-reckoning-how-top-political-podcasters-spread-unsubstantiated-and-false-claims/ Accessed: 2024-12-02.

[31] Seid Muhie Yimam, Daryna Dementieva, Tim Fischer, Daniil Moskovskiy, Naquee Rizwan, Punyajoy Saha, Sarthak Roy, Martin Semmann, Alexander Panchenko, Chris Biemann, and Animesh Mukherjee. 2024. Demarked: A Strategy for Enhanced Abusive Speech Moderation through Counterspeech, Detoxification, and Message Management. arXiv:2406.19543 [cs.CL] https://arxiv.org/abs/2406.19543

[32] Xinghua Zhang, Haiyang Yu, Yongbin Li, Minzheng Wang, Longze Chen, and Fei Huang. 2024. The Imperative of Conversation Analysis in the Era of LLMs: A Survey of Tasks, Techniques, and Trends. *arXiv preprint arXiv:2409.14195* (2024).

Podcast channel name	Number of episodes	Average episode duration (min)	Average tokens count	Toxic episodes	Percentage toxic episodes (%)
Bannon's War Room	1184	54 (2)	9339 (662)	172	15
Bill O Reilly's No Spin News and Analysis	1117	18 (17)	2553 (2460)	72	6
The Sean Hannity Show	990	37 (5)	6375 (983)	208	21
The Glenn Beck Program	725	81 (37)	12619 (5703)	275	38
The Dan Bongino Show	448	48 (15)	8483 (2706)	309	69
Mark Levin Podcast	440	103 (18)	14280 (3095)	380	86
Human Events Daily with Jack Posobiec	403	27 (8)	4614 (1602)	34	8
Conservative Review with Daniel Horowitz	398	62 (6)	9673 (1210)	95	24
The Rubin Report	372	43 (13)	8280 (2513)	161	43
The News Why It Matters	334	44 (0)	8313 (423)	181	54
The Megyn Kelly Show	332	95 (10)	17626 (2117)	180	54
Tim Pool Daily Show	321	86 (11)	15472 (2295)	114	36
The Ben Shapiro Show	229	48 (17)	10133 (3664)	62	27
The New Abnormal	220	50 (16)	9132 (2919)	198	90
Louder with Crowder	210	62 (33)	12588 (6900)	186	89
The Charlie Kirk Show	200	40 (15)	7127 (2903)	30	15
The Michael Savage Show	196	54 (19)	9253 (3384)	96	49
The Jordan B Peterson Podcast	145	99 (21)	16236 (3772)	27	19
The Michael Knowles Show	133	46 (17)	8112 (3176)	43	32
Verdict with Ted Cruz	133	43 (11)	7142 (1833)	25	19
Hold These Truths with Dan Crenshaw	115	53 (20)	9699 (3478)	13	11
Bret Weinstein DarkHorse Podcast	94	106 (21)	17414 (3496)	39	41
Pseudo Intellectual with Lauren Chen	88	11 (2)	2202 (379)	10	11
Fireside Chat with Dennis Prager	87	32 (7)	4309 (1168)	13	15
Conversations With Coleman	82	71 (19)	12920 (3681)	22	27
The One w/ Greg Gutfeld	71	15 (3)	2690 (645)	34	48
Candace Owens	47	29 (14)	5620 (2782)	13	28
Get Off My Lawn Podcast w/ Gavin McInnes	24	61 (20)	10012 (3391)	24	100
The MeidasTouch Podcast	15	52 (40)	9017 (6961)	9	60
The Matt Walsh Show	10	46 (24)	8081 (4228)	3	30
Rudy Giuliani's Common Sense	3	36 (4)	5414 (750)	2	67

Table 3: We crawl a total of 31 podcast channels encompassing a total of of 9,166 episodes. Statistics for the crawled dataset sorted according to the number of episodes is presented. Columns for duration (in minutes) and number of tokens present corresponding mean values. In these columns, numbers in parenthesis specify standard deviation. Columns for number of toxic episodes (those containing atleast one toxic conversation chain) and the corresponding percentage distribution are also provided.

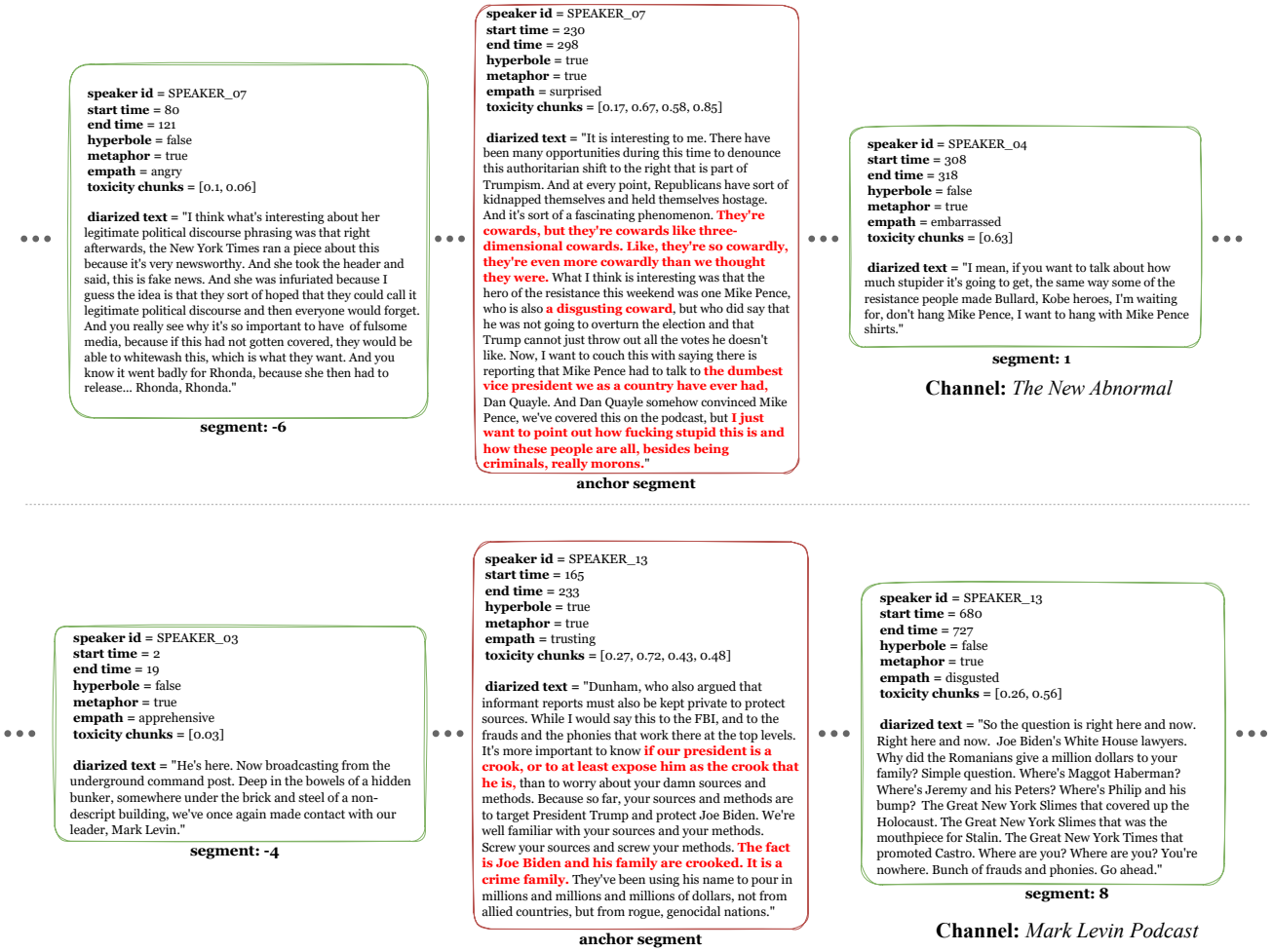

Figure 11: Conversation Examples: One among previous and next segments are plotted along with anchor segment since it is not feasible to plot all segments. Toxic contents in the anchor segment are marked in red color. Note: start and end times are in seconds.