Extraction of Narratives from Podcast Transcripts

Anonymous ACL submission

Abstract

As one of the oldest forms of human communication, narratives appear across a variety of genres and media. Computational methods have been applied to study narrativity in novels, social media, and patient records, leading to new approaches and insights. However, other types of media are growing in popularity, like podcasts. Podcasts contain a multitude of spoken narratives that can provide a meaningful glimpse into how people share stories with one another. In this paper, we outline and apply methods to process English-language podcast transcripts and extract narrative content from conversations within each episode. We provide an initial analysis of the types of narrative content that exists within a wide range of podcasts, and compare our results to other established narrative analysis tools. Our annotations for narrativity and pretrained models can help to enable future research into narrativity within a large corpus of approximately 100,000 podcast episodes.

1 Introduction

Storytelling is an intricate and culturally rich psychological phenomenon. When storytellers share a narrative with an audience, they are doing more than just telling a story (Piper et al., 2021). They are taking their audience on a shared journey, navigating through emotions, insights, and cultural reflections. Our understanding of the complex psychological framework underpinning narrative structures is still in its early stages (Piper et al., 2021).

Previous work in Natural Language Processing (NLP) has examined narratives in novels (Gio, 2023; Han, 2023), social media sites such as Reddit(Yan et al., 2019), Twitter (Ganti et al., 2023) and Facebook (Ganti et al., 2022) and medical records (Tange et al., 1997). Narrative analysis in these studies has explored aspects such as feature analysis in online Health communities (Ganti et al., 2022) or the spread of health misinformation on Twitter (Ganti et al., 2023), contributing to a deeper understanding of how narratives are constructed and communicated in diverse textual sources.

In recent years, podcasts have emerged as a significant medium, rich in linguistic variety and style. Their diverse topics, ranging from formal news journalism to conversational chats and spanning both fiction and non-fiction, allow researchers to delve into language use across various emotional and thematic contexts. Once transcribed, podcast datasets can bridge the gap between formal and informal language, serving as a crucial resource for uncovering various insights and patterns from modern language. One important feature of podcasts that has received little attention, however, is narrativity. Many podcast episodes contain examples of people sharing stories, either in the form of personal experiences or storytelling involving external characters and events. Given the large number of often lengthy podcast episodes, automatically extracting and analyzing this narrative content from podcasts may help to explore the potential for new avenues in research, content creation, recommendation systems, and other applications.

In this study, we build upon the previous work in this domain and introduce a novel model for extracting narratives from podcasts. The extraction of narratives from podcast data poses several unique challenges. Unlike written texts, podcasts rely on oral communication, which follows a different style and structure (Yang et al., 2019), and additional noise may be introduced due to imperfect transcription tools. Podcasts span a broad range of topics and formats, which makes it more difficult to apply narrative analysis and detection tools that are tailored to particular genres or media. Podcasts often follow a conversational format with multiple speakers, making the identification and separation of narrative threads more intricate. Often, the main content of the podcast is interspersed with extraneous content such as advertisements, which should be ignored when identifying narratives.

In this work, we make the following contributions: (1) we develop a podcast transcript processing workflow to remove non-English and extraneous content; (2) we annotate a set of podcast episodes for sentence-level narrativity and fine-tune language models for the task of narrative detection; (3) we define a simple yet effective method for characterizing the overall narrativity of a podcast and compare it to an existing measure of narrativity. We find that we are able to accurately filter out extraneous content from podcast transcripts given only the text, and our narrative detection methods provide a meaningful way to measure podcast narrativity that does not rely on narrative arc features which, unlike narratives in other media, are not always present within a given podcast episode. Our results suggest that categories such as fiction, true crime, and daily news contain a high degree of narrative content and should be useful types of podcasts to explore in future work on narrative analysis. We release[1] our annotations and pretrained models that can be used for both extraneous content removal and narrative detection, and, at the time of writing, access to the dataset may be freely requested[2] for non-commercial research purposes.

2 Related works

2.1 NLP for Narratives

A long line of work in NLP has focused on narrative analysis. A range of narrative elements that have been studied already within NLP, from the extraction of characters and their relations (Massey et al., 2015) to studies of language models' ability to represent time in books (Kim et al., 2014). Among other work, Antoniak et al. (2019) performed a computational analysis of birth stories on social media, Levi et al. (2022) developed data and models for the extraction of narrative elements from news text, and Gala et al. (2020) explored gender bias in narrative tropes. However, it is important to note that much of this previous work begins with a dataset that is known beforehand to contain narrative-style text, and therefore researchers can directly begin to analysis specific aspects of narratives. In our own work with podcasts, we cannot make this assumption, since not all podcasts follow a narrative-type format, and therefore an important first stage is to extract narratives from the episodes in which they may or may not occur.

Existing work on narrative detection, while mostly successful has mostly focused on specific domains such as online patient communities (Dirkson et al., 2019), Facebook posts related to breast cancer support (Ganti et al., 2022), or tweets about the COVID-19 pandemic (Ganti et al., 2023). Given that these are dramatically different media and rely on written text rather than transcripts of spoken conversations, we cannot directly use the data or models from previous work, and therefore focus on building a narrative extraction pipeline that is specifically tailored for podcast transcripts.

2.2 NLP for Podcast Analysis

Podcasts are emerging mediums, rich in linguistic variety and style. Once transcribed, podcast datasets can bridge the gap between formal and informal language, serving as a crucial resource for uncovering various insights and patterns from modern language. The Spotify Podcast Dataset (Clifton et al., 2020a) is one such dataset that facilitated a wide range of research in areas such as summarization (Kashyapi and Dietz, 2020; Song et al., 2020), recommender systems (Kashyapi and Dietz, 2020; Nazari et al., 2020), search and information retrieval (Alexander et al., 2021). The dataset was used as a part of the TREC 2020 Podcast Track for (1) retrieval and (2) summarization (Jones et al., 2021). Abstractive techniques, with the BART transformer model (Lewis et al., 2020) trained on news summarization and fine-tuned using the creator's descriptions as targets, were the most predominant summarization models (Song et al., 2020; Manakul and Gales, 2020; Karlbom and Clifton, 2020; Rezapour et al., 2021; Zheng et al., 2020) in TREC 2020 summarization track (Rezapour et al., 2022). Podcasts were also analyzed for user engagement and popularity. Reddy et al. (2021b) analyzed podcasts through quantitative analysis and found stylistic features having stronger correlations with engagement in less popular podcasts. Yang et al. (2019) employed iTunes to compile a podcast dataset consisting of 88,728 episodes, using 10 minutes from each episode to predict their popularity, seriousness, and energy levels through acoustic features.

Podcasts were also analyzed in the fields of healthcare and science. MacKenzie (2019) extracted and studied 952 English science podcasts from public websites dedicated to podcast promotion and found exponential growth in the number

[1] url will be added upon publication
[2] https://podcastsdataset.byspotify.com/

of series from 2010 to 2018, with 65% of them hosted by scientists and 77% targeting a general audience. Furthermore, (Dumbach et al., 2023) extracted 29 healthcare podcasts, totaling 3,449 episodes, through web mining. They tracked AI trends using 102 buzzwords in these podcasts, identifying 14 distinct topic clusters. Additionally, they assessed sentiment to detect trends, finding that the speakers expressed a more positive sentiment toward these trends.

Our study builds on previous research in podcast analysis, providing a novel perspective and method for examining narrativity. Our proposed approach enriches our understanding of podcast content and paves the way for future investigations into the nuances of storytelling within this medium.

3 Data

Dataset description. The Spotify Podcast Dataset consists of 105,360 podcast episodes, mostly in English (Clifton et al., 2020a). Each episode comes with an automatically generated transcript, using Google's Cloud Speech-to-Text API, its audio, an RSS header, and a short description written by the podcast creators. The automatic speech recognition system displayed stability, with an 18.1% word error rate and 81.8% accuracy in named entity recognition across a varied dataset (Clifton et al., 2020a). The dataset consists of approximately 18,000 distinct shows spanning a range of topics such as news, science, and sports.

Filtering ads and promotions. We are primarily focused on the transcripts of podcasts to detect narrativity. As shown in previous work (Reddy et al., 2021a), podcasts often include advertisements and promotions that carry non-relevant information to the main themes of the discussion. This presence of extraneous content can result in distorted analysis outcomes or misleading representations of the podcast's core narrative. To detect and remove boilerplate and noise from transcripts, we followed Reddy et al.'s approach (Reddy et al., 2021a). We first created three sets of labeled sentences, each representing ads and promotions in podcasts. The first set included only sentences taken from the episode descriptions. The second set comprised sentences from the transcript dataset, while the third set consisted of a combination of sentences from both the descriptions and the transcripts. Sentences were randomly selected from a diverse range of podcast episodes to ensure representation across various genres and topics and were annotated as either extraneous (ads and promotions) or non-extraneous.

	Test		
	Description	**Transcript**	**Combination**
Description	89%	76%	85%
Transcript	82%	93%	86%
Combination	89%	**94%**	91%

Table 1: Extraneous sentence classification using BERT. Models are trained and tested on sentences from podcasts' episode descriptions, transcripts, and both.

We used these annotated sets to train a binary classifier to detect whether a sentence is extraneous or not. We fine-tuned BERT (Devlin et al., 2019) using our labeled dataset and evaluated the performance using three separate test sets similar to the training datasets. Our results (Table 1) show that the best performance, in terms of F1 score, was achieved when the model was trained on the combined dataset and tested on transcripts only.

Additionally, to further evaluate the generalizability of our model, we performed an additional test on data obtained from (Vaiani et al., 2022). This dataset consists of 2,203 manually annotated data taken from episode descriptions from the same dataset provided by Spotify. Our best-performing model, trained on the combination of descriptions and transcripts, was tested on this new data, achieving an F1-score of 89% on this dataset, which matches the results presented by the authors of that dataset of podcast descriptions, while we only trained on our own annotated data. While we aim to remove extraneous content from the transcripts rather than the descriptions, this result confirms that our trained model is in-line with previous work on this task.

Finally, we employed our best-performing model to automatically label the remaining sentences in our dataset. A total of 1,623,451 sentences, constituting 0.45% of the sentences, were labeled as extraneous and subsequently removed from our dataset. Manual evaluation of the removed content confirmed that they predominantly focused on product promotion.

Non-English transcripts. The Spotify Podcasts dataset was transcribed using the Google API (Clifton et al., 2020a). Consequently, podcasts that were initially in languages other than English were transcribed into English, resulting in the generation of incoherent and noisy texts, i.e., while the transcripts for non-English episodes appear in English, they might not convey any meaningful content. As a result, using any language detection model on these transcripts would be misleading. To address this issue, we used the episode descriptions of the podcasts. Since these descriptions are typically written by the podcast creators in the original language, they offer a more reliable indicator of the actual language. We utilized the Langdetect library for language detection[3]. which resulted in identifying 1,420 episodes as non-English.

After removing extraneous content and non-English transcripts, the total number of transcripts decreased from 105,361 to 103,934.

Podcast categories. The narrative structure of a podcast can vary based on its genre and the topics discussed. For instance, crime podcasts might use words with a negative connotation, whereas self-improvement or motivational podcasts often convey a positive tone. The metadata files included in the podcast dataset do not specify the categories (i.e., genres). However, the categories can be obtained from the RSS headers of each podcast. For each episode, we extracted its category labels to conduct a more in-depth narrativity analysis.

Upon reviewing the categories and comparing them with a sample of transcripts, we found some categories ambiguous and not well-defined (e.g., 'Leisure' mainly includes gaming podcasts but also general leisure topics, 'Kids and Family' includes podcasts for kids as well as parenting podcasts). Therefore, in addition to iTunes categories, we created a new set of categories using topic modeling. In line with previous research (Reddy et al., 2021b; Clifton et al., 2020b; Yang et al., 2019), we use Latent Dirichlet Allocation (LDA) topic modeling (Blei et al., 2003) to extract 100 distinct topics from our corpus of 103,933 podcasts. We then manually assigned distinct categories to each topic for better interpretation. Table 2 shows a sample of the extracted topics.

Genre	Words
Identity terms	woman, men, female, man, male, gay, black, also, girl, like
Finance	year, number, million, percent, hundred, price, dollar, think, rate, market
Races	race, run, running, mile, marathon, bike, year, runner, really, time
Cryptocurrency	bitcoin, coin, crypto, people, like, nt, money, exchange, lightning, network
Drugs and Alcohol	drink, cigar, drinking, drug, beer, alcohol, bar, wine, smoke, smoking
Filler 1	nt, think, get, would, really, gun, damage, going, like, character
Filler 2	nt, like, got, man, know, right, saying, na, get, yall
Filler 3	going, one, really, get, kind, little, pretty, bit, lot, actually
Films	star, movie, war, think, nt, character, like, trek, going, one
Medicine	injury, bone, joint, nerve, pain, tissue, spinal, fracture, question, patient
Professional Wrestling	match, wrestling, fight, show, think, nt, ring, guy, wrestler, see
Stories	would, fire, king, one, man, death, could, men, stone, dead
United States	country, people, English, world, also, American, U, America, language
Crime	police, nt, murder, would, case, crime, found, year, could, death
Net sports	team, think, player, year, coach, guy, sport, league, going, like
Clothing	shoe, wear, store, wearing, brand, fashion, shirt, look, clothes, buy
American Football	defensive, back, going, receiver, guy, team, player, game, offensive, really
Football	think, player, nt, season, league, club, week, goal, football, going
Nutrition	body, weight, fat, eating, food, calorie, diet, eat, going, lose
Beauty	hair, look, skin, makeup, beauty, face, really, love, dress, product
Career	new, job, people, get, city, York, got, work, go, said
Education	teacher, student, learning, teaching, teach, learn, language, education, skill
Gaming	card, dog, deck, one, play, magic, think, board, turn, amber
Psychology	behavior, relationship, person, brain, people, child, human, control, u, often

Table 2: High probability words from examples of LDA topics for podcast transcripts along with manually assigned labels.

4 Narrative Extraction Methodology

In this section we describe the baseline method from LIWC, which can assign narrativity scores to podcast transcripts, and our approach to building text classification models that we evaluate and use later for the extraction of narrative sentences from podcast transcripts.

4.1 LIWC Narrative Arc

The Linguistic Inquiry and Word Count (LIWC) narrative arc analysis feature (Boyd et al., 2022) identifies and quantifies words and phrases associated with three key narrative components: staging, plot progression, and cognitive tension.

- **Staging** refers to the introduction of characters, setting and plot in the early stages of a narrative.

[3]https://github.com/shuyo/language-detection

- **Plot progression** refers to the sequence of events that unfold in a narrative, including rising action, climax, and falling action.
- **Cognitive tension** refers to the uncertainty, suspense, or conflict that keeps readers engaged in a narrative.

To calculate staging, plot progression, and cognitive tension, LIWC counts the number of words belonging to each category that appear in the text. Each input text (in our case, podcast transcript) is broken into five equally-sized segments, and each of the three scores is computed for each segment. The results are then normalized to account for the length of the segment, meaning that the scores are expressed as a percentage of the total number of words in the segment. Then, for each score, the "arc" comprised of the scores for each of the five segments is compared to a reference that was computed over a set of documents know to follow a traditional narrative structure, and the correlation between the computed arcs and the reference arcs is provided as a score for staging, plot progression, and cognitive tension. The overall narrativity score is an average of the three.

4.2 Narrativity Annotation

The LIWC Narrative Arc tool provides a transcript-level narrativity score, but does not allow for a more fine-grained analysis of narratives within podcasts. To explore this level of granularity further and evaluate models for sentence-level narrative extraction, we annotated individual sentences from podcast transcripts for their narrativity.

Data selection. We selected and annotated the transcripts on the sentence level as sentences are fundamental building blocks of text, and this will allow us to assess and annotate if a given sentence is a part of a narrative or not regardless of the narrative arc of the podcast. To ensure diversity in our selection of podcasts, we adopted a multi-step approach. In our datset, the overall narrativity score of LIWC ranges from -59.91 to 97.81, with the former indicating the lowest narrativity and the latter indicating the highest that we observed. To evenly distribute our selection across this range, we categorized the episodes into five separate groups based on the LIWC narrativity overall ranges, each comprising 20,000 episodes. From each group, we chose the top 20 episodes based on their narrativity scores, resulting in 100 selected episodes. Within each selected episode, we randomly sample twelve consecutive sentences for annotation. Since narrativity is context dependent, we included one sentence before and one sentence after each target sentence to account for context. A total of 1200 sentences were selected for the training and 304 sentences were chosen for testing from a total of 100 distinct podcast episodes.

Data annotation. We first developed our annotation guidelines through a series of pilot phases. During each of these, we selected 100 random sentences in each phase (which were not part of the training set described in the previous section), to develop a comprehensive annotation guideline to label narrativity of sentences. Three annotators independently applied the guidelines iteratively, evaluating if a sentence is narrative or not. After each round of annotation, the annotators met to discuss the results and collectively refined the annotation guidelines based on their observations. Following 3 iterations, all annotators reached a consensus on the final annotation guidelines (Appendix B.2.) The guidelines were then used to label the full training dataset. Two annotators labeled each sentence, and if a consensus was reached, the agreed-upon label was used. Otherwise, a third annotator intervened to break the tie. After a tiebreak process, the Krippendorf's alpha score was 0.534.

4.3 Classification Model

Given the annotated dataset, we then explored several approaches for building text classifiers that would be able to automatically label the rest of the dataset for narrativity at the sentence level. For encoder transformer based models, we utilized BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), and RoBERTa (Liu et al., 2019) to build our narrativity classifiers, using models accessible through HuggingFace (Wolf et al., 2020): with bert-base-uncased, distilbert-base-uncased, and roberta-base configurations. In each case, we used the default tokenizers, and the `[CLS]` input token served as input to a trainable classification layer. For autoregressive generative models, we experimented with GPT-3.5-turbo and GPT-4 models accessed via the OpenAI API[4]. Our experimental approach involves presenting these models with either an instruction or a prompt as input, to which they generate responses as completions. We explore both zero-shot and few-shot learning, and also considered several prompt variations for the models.

[4] https://openai.com/blog/openai-api

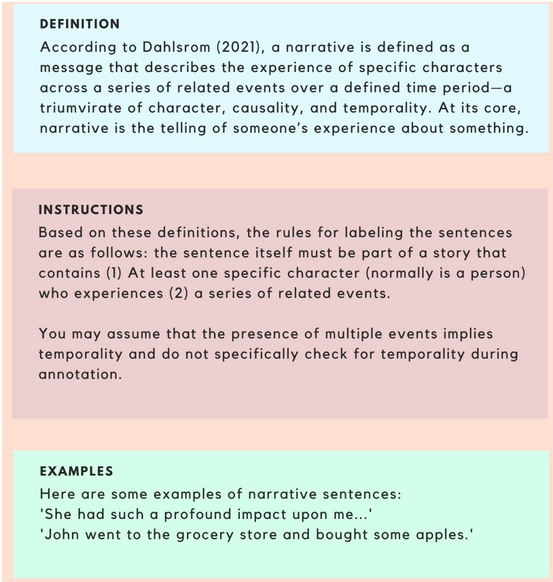


Figure 1: Prompt components for GPT Models. From top to bottom, the blocks display the definition (blue), instructions (pink), few-shot examples (green).

These prompts include not only examples of sentences from the dataset but also the inclusion of narrative definitions and additional instructions. The components that were included are outlined in Figure 1.

5 Podcast Narrative Analysis

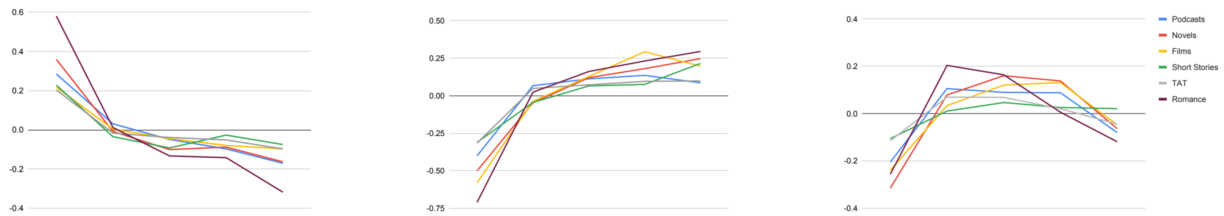
In this section, we use our narrative extraction models to estimate the overall narrativity of each podcast to explore the topics that are most associated with narrativity. We compare our results with another popular method for automatically quantifying narrativity and find that in the domain of podcasts, our method appears to better identify texts that have a high degree of narrativity.

LIWC Narrative Arc. Figure 2 shows the arc of the narrative graphs in podcasts vs. the other types of texts. As shown in Boyd et al. (2020), the most significant disparity between the non-fiction texts and the traditional stories was evident in the cognitive tension dimension.In our case, the curves are quite similar to the standard "arc of narrative" showed in Figure 2c. Note that this captures the average trend and individual podcasts' narrativity scores varied.

Furthermore, we used LIWC's overall narrativity score to extract categories of podcasts with the highest and lowest narrativity. Table 3 presents the top 10 categories with the highest and lowest average scores. Several sports-related podcasts exhibit higher narrativity than those in other categories. Although we anticipated Fiction to rank among the categories with the most narrativity, it was among those with the lowest overall narrativity scores. This suggests that the narrativity analysis of LIWC may not be directly applicable to podcast data, as the structure and format of spoken content can differ from written text. Further, podcasts from the fiction category often tell as single story that is broken up across multiple distinct episodes, making the narrative arc of each individual episode incompatible with the expected arc that is needed in order to achieve a high LIWC narrativity score.

Table 5 shows the top 10 LDA topics (as described in section 3) with the highest and lowest narrative scores using the LIWC overall narrativity score. Here we can see that several sports-related categories again had high overall narrativity scores (with the exception of the American Football topic), while podcasts with topics related to religion and medicine had lower scores.

Narrativity Detection. Table 7 shows the result of our narrativity detection using the transformer-based models. Both BERT-base and DistilBERT achieved high performance in terms of accuracy and F1 score. RoBERTa models, both base and large, seem to perform less effectively on this specific narrative detection task. Our results show that encoder-based models like BERT and DistilBERT can be very competitive to autoregressive models at detecting narratives from transcript data, though the latter only required a small number of training examples compared to the fine-tuned models. Although BERT performed slightly better than DistilBERT overall, we opted to use our fine-tuned DistilBERT model due to computational efficiency purposes, since it is a much smaller model. For the generative models, shown in Table 8, GPT-4 outperformed GPT 3.5-turbo in nearly all zero-shot and few-shot experiments. GPT-4 with few-shot learning and instructions outperformed the other models. Overall, we noticed that the few-shot prompts typically led to better results than zero-shot counterparts. GPT-3.5 was more sensitive to the specific prompting approach, showing a much higher range of F1-scores across the various configurations, while GPT-4 achieved similarly high results regardless of the configuration. While we do not use these models to annotate the full dataset, we find the results of these models promising for future exploration given the limited amount of training



(a) Staging (b) Plot Progression (c) Cognitive Tension

Figure 2: The narrative arcs of podcasts (blue lines) compared to narrative arcs from the genres of text presented by Boyd et al. (2020).

Category	LN
Tennis	51.69
Tech News	46.90
Podcasting	46.63
After Shows	42.67
Hinduism	42.27
Gov. & Org	40.32
Management	38.23
Running	37.77
Wrestling	37.22
Sports & Rec.	36.49
History	17.44
Astronomy	17.27
Language Learning	17.02
Fiction	15.82
Science Fiction	14.69
Outdoor	13.61
Mathematics	13.27
Amateur	0.55
News Commentary	-3.36
Physics	-17.55

Table 3: Categories with the highest and lowest LIWC Narrativity (LN) Scores

Category	AS
Fiction	0.73
Gov. & Org	0.67
True Crime	0.66
History	0.64
Daily News	0.62
Film History	0.62
News	0.59
Kids & Family	0.55
Personal Journals	0.54
College / School	0.54
Medicine	0.29
Investing	0.27
Marketing	0.26
Management	0.25
Language Learning	0.25
Science	0.18
Tech News	0.18
Astronomy	0.17
Mathematics	0.15
Physics	0.00

Table 4: Categories with highest, lowest Average Narrativity Scores (AS, ours).

Topic	LN
Investing	46.81
Wrestling	45.01
Basketball	40.07
Health & Nutrition	40.06
Working Out	37.04
Animals	36.54
Mental Health	36.35
Filler 3	36.22
Arts	35.93
Well-being	35.02
Gaming	15.76
Relationships	14.92
Podcast Start	14.17
Med. & Diseases	13.24
Filler 2	13.57
Filler 1	12.73
Celebrations	11.87
Christianity 1	5.04
American Football	-1.55
Medicine	-15.69

Table 5: Topics with the highest and lowest LIWC Narrativity (LN) Scores

Topic	AS
Routine	0.77
Effusiveness	0.70
Music	0.69
Mystery	0.66
Love Relationship	0.59
Astrology	0.59
History	0.58
Med. & Diseases	0.56
Filler 2	0.55
Wrestling	0.55
Net Sports	0.24
Medicine	0.23
Football	0.21
Business	0.21
Christianity 1	0.16
Wars	0.14
Podcast Start	0.06
Investing	0.05
Christianity 2	0.04
American Football	0.00

Table 6: Topics with highest and lowest average narrativity scores (AS, ours)

Model	F1	Accuracy	Precision	Recall
BERT base	**0.812**	**0.803**	0.794	0.833
BERT large	0.738	0.675	0.619	**0.917**
RoBERTa base	0.598	0.625	0.701	0.633
RoBERTa large	0.526	0.500	0.517	0.600
DistilBERT base	0.799	0.800	**0.802**	0.800

Table 7: Narrative classification using transformer encoder models. The best results for each metric are listed in **bold**.

data required.

Analysis of our results Based on the results from the classifiers, we chose to employ DistilBERT for annotating the rest of the sentences in our transcripts, as it not only demonstrated the highest precision among all the models but is also a smaller version of BERT, designed for computational efficiency. After annotating every sentence in our transcripts, we **calculated our own narrativity scores** for each transcript by dividing the number of narrative sentences by the overall sentence count in that transcript.

To compare to the LIWC narrative arc scores, we first used iTunes podcast categories to better understand the narrativity characteristics of the podcasts. Table 4 presents the top ten categories with the highest and lowest average narrativity scores. As shown in the table, unlike the results given by LIWC's narrativity, categories like Fiction, True Crime, and History have a high score. In fact, based on Spotify [5], a fictional audio podcast is a type of podcast that presents fictional stories, or dramas through the audio medium, therefore, expected to be more narrative compared to other genres.

Based on the narrativity definition adapted from Dahlstrom, narrative texts consist of characters who are involved in a series of related events. Film history or Fiction often encompass a greater abundance of these narrative elements compared to genres such as marketing podcasts. When compar-

[5] https://www.masterclass.com/articles/types-of-podcasts-explained

Model	*-shot	definition	Instruction	F1	Accuracy	Precision	Recall
GPT3.5	z			0.67	0.72	0.74	0.72
	z	✓		0.68	0.72	0.73	0.72
	z		✓	0.64	0.70	0.68	0.70
	z	✓	✓	0.66	0.71	0.70	0.71
	f			0.68	0.71	0.69	0.71
	f	✓		**0.78**	**0.78**	**0.78**	**0.78**
	f		✓	0.63	0.69	0.67	0.69
	f	✓	✓	0.65	0.72	0.76	0.72
GPT4	z			0.72	0.72	0.75	0.72
	z	✓		0.73	0.73	0.73	0.73
	z		✓	0.76	0.76	0.76	0.76
	z	✓	✓	0.71	0.72	0.71	0.72
	f			0.72	0.72	0.73	0.72
	f	✓		0.74	0.73	0.75	0.73
	f		✓	**0.78**	**0.79**	**0.79**	**0.79**
	f	✓	✓	0.75	0.76	0.76	0.76

Table 8: Narrative classification using GPT-3.5 and GPT-4 models under different configurations. "z" denotes zero-shot learning, and "f" signifies few-shot learning. A check mark indicates the presence of the instruction or the definition in that configuration.

ing our narrativity scores to those from LIWC, we identified more categories that shared the lowest average narrativity between the two sets of results. Specifically, podcasts related to scientific disciplines, such as Physics, Mathematics, and Astronomy, tend to have lower narrativity scores when using either method. This can be attributed to these genres typically featuring content with few characters and events, which explains their consistently low narrativity across different models.

Correlation Analysis. We conducted a Pearson correlation analysis to assess the relationship between our narrativity scores and LIWC's narrativity score. The correlation coefficient between the two results was 0.05, showing a divergence in the conceptualization of narrativity between the two methods. In addition to narrativity components, we also used LIWC's psycho-linguistic features (Boyd et al., 2022) in the analysis of correlation. Our results showed a strong correlation (~ 0.7) between 'focuspast' and narrativity. This strong correlation can explain why the highest narrativity scores are associated with podcasts in storytelling genres as shows table 4 where the frequent use of past tense verbs is a common narrative technique (Piper et al., 2021). The remaining correlation results are presented in Appendix A.

Narrativity of podcasts based on extracted topics. Table 6 shows the top 10 LDA topics (as described in section 3) with the highest and lowest narrative scores using our proposed model. These results show that topics related to things like routines, which clearly describe sequences of actions, had high narrativity scores. This is likely because these routines are often told in a first-person narrative style. Topics related to religion, business and investing had lower narrativity scores. These results again stand in contrast with those obtained when using the LIWC narrativity scores.

Comparing narrativity measures. We believe that although the results are different when comparing between the LIWC overall narrativity scores and the scores we computed using the output of our model, each method can serve its own purpose. The LIWC narrative arc score is able to determine if the overall progress matches a standard narrative arc, while our proposed supervised-learning based approach is able to accurately detect narrative sentences even within podcast episodes that do not follow this standard arc. This allows us to identify types of podcast that have a high frequency of narrative content even when the podcasts don't follow a typical narrative structure overall.

6 Conclusion

In this work, we studied narrativity within podcasts, which have grown in popularity in recent years. In order to clean the dataset, we implemented an extraneous content detection system and demonstrated competitive results with existing works. Our classifier can work on both episodes description and transcripts at the same time. We then annotated a dataset and trained text classification models for the task of narrative sentence detection. We use one of our best models to annotate the entire Spotify podcasts dataset for narrativity, and then compare the types of podcast that had a high proportion of narrative sentences with those that had high narrative scores based on other tools such as LIWC narrative arc. We found that our method was able to identify high narrativity in fiction and true crime podcasts, which are expected examples of categories that should contain narrative content. We aspire for this research to serve as a starting point for future investigations into podcast narrativity, and we believe that the tools and annotations that we have created will facilitate future analyses in this area.

7 Limitations

The transcriptions for this study were generated in 2020. While they served the purpose at the time, it's worth acknowledging that there have been advancements in automatic transcription technology. The use of an updated transcription model could potentially lead to more accurate transcriptions, which may be considered for future research to enhance the quality of data analysis.

Even after participating in three rounds of training sessions, the annotators still encountered several disagreements among themselves. With further training, it might be possible to improve the reliability of annotations.

Furthermore, the narrative labels applied to the complete dataset are derived from predictions made by a transformer-based encoder model that possesses imperfect predictive capabilities, leading to some additional noise in the analyses based on these labels.

8 Ethical Considerations and Impact

The podcast data used in this research have been provided by Spotify and are available exclusively for research purposes. The data used have been obtained through authorized channels and are used in compliance with Spotify's terms and conditions for research. We are committed to promoting open and collaborative research practices. The annotations associated with the sentences derived from this study will be made publicly available for future research endeavors. We believe that sharing this resource will contribute to the advancement of knowledge and foster innovation in the field of computational social science.

References

2023. Author as character and narrator: Deconstructing personal narratives from the r/amitheasshole reddit community. 17:233–244.

2023. Happenstance: Utilizing semantic search to track russian state media narratives about the russo-ukrainian war on reddit. 17:327–338.

Abigail Alexander, Matthijs Mars, Josh C Tingey, Haoyue Yu, Chris Backhouse, Sravana Reddy, and Jussi Karlgren. 2021. Audio features, precomputed for podcast retrieval and information access experiments. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 3–14. Springer.

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Ryan L. Boyd, Kate G. Blackburn, and James W. Pennebaker. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, 6(32):eaba2196.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020a. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskovich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020b. 100,000 podcasts: A spoken english document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Michael Dahlstrom. 2021. The narrative truth about scientific misinformation. *Proceedings of the National Academy of Sciences*, 118:e1914085117.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Anne Dirkson, Suzan Verberne, and Wessel Kraaij. 2019. Narrative detection in online patient communities.

Philipp Dumbach, Leo Schwinn, Tim Löhr, Phi Long Do, and Bjoern M Eskofier. 2023. Artificial intelligence trend analysis on healthcare podcasts using topic modeling and sentiment analysis: a data-driven approach. *Evolutionary Intelligence*, pages 1–22.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 212–217, Online. Association for Computational Linguistics.

Achyutarama Ganti, Eslam Hussein, Steven Wilson, Zexin Ma, and Xinyan Zhao. 2023. Narrative style and the spread of health misinformation on twitter. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Achyutarama Ganti, Steven Wilson, Zexin Ma, Xinyan Zhao, and Rong Ma. 2022. Narrative detection and feature analysis in online health communities. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 57–65, Seattle, United States. Association for Computational Linguistics.

Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth JF Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2021. Trec 2020 podcasts track overview. *arXiv preprint arXiv:2103.15953*.

Hannes Karlbom and Ann Clifton. 2020. Abstractive podcast summarization using BART with longformer attention. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

Sumanta Kashyapi and Laura Dietz. 2020. TREMA-UNH at TREC 2020. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65.

Effi Levi, Guy Mor, Tamir Sheafer, and Shaul Shenhav. 2022. Detecting narrative elements in informational text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1755–1765, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lewis E MacKenzie. 2019. Science podcasts: analysis of global production and output from 2004 to 2018. *Royal Society open science*, 6(1):180932.

Potsawee Manakul and Mark Gales. 2020. CUED_speech at TREC 2020 podcast summarisation track. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

Philip Massey, Patrick Xia, David Bamman, and Noah A Smith. 2015. Annotating character relationships in literary texts. *arXiv preprint arXiv:1512.00728*.

Zahra Nazari, Christophe Charbuillet, Johan Pages, Martin Laurent, Denis Charrier, Briana Vecchione, and Ben Carterette. 2020. Recommending podcasts for cold-start users based on music listening and taste. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1041–1050.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

S. Reddy et al. 2021a. Detecting extraneous content in podcasts. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Sravana Reddy, Mariya Lazarova, Yongze Yu, and Rosie Jones. 2021b. Modeling language usage and listener engagement in podcasts. pages 632–643.

Rezvaneh Rezapour, Sravana Reddy, Ann Clifton, and Rosie Jones. 2021. Spotify at TREC 2020: Genre-aware abstractive podcast summarization. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

Rezvaneh Rezapour, Sravana Reddy, Rosie Jones, and Ian Soboroff. 2022. What makes a good podcast summary? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2039–2046.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2020. Automatic summarization of open-domain podcast episodes. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

Huibert J Tange, Arie Hasman, Pieter F de Vries Robbé, and Harry C Schouten. 1997. Medical narratives in electronic medical records. *International Journal of Medical Informatics*, 46(1):7–29.

Lorenzo Vaiani, Moreno La Quatra, Luca Cagliero, and Paolo Garza. 2022. Leveraging multimodal content for podcast summarization. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 863–870, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Xinru Yan, Aakanksha Naik, Yohan Jo, and Carolyn Rose. 2019. Using functional schemas to understand social media narratives. In *Proceedings of the Second Workshop on Storytelling*, pages 22–33, Florence, Italy. Association for Computational Linguistics.

Longqi Yang, Yu Wang, Drew Dunne, Michael Sobolev, Mor Naaman, and Deborah Estrin. 2019. More than just words: Modeling non-textual characteristics of podcasts. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*.

Chujie Zheng, Kunpeng Zhang, Harry Jiannan Wang, and Ling Fan. 2020. A two-phase approach for abstractive podcast summarization. In *The 29th Text Retrieval Conference (TREC) notebook. NIST*.

Appendix

A Correlation between scores

We compare our proposed narrativity score with the LIWC narrative arc scores across various dimensions, and the results are presented in Figure A2. Further, we used the LIWC dictionary-based category counting functionality to compute the scores for various LIWC categories, leading to the results presented in Figure A3.

B Annotation guideline

In this section we will define what is narrative and introduce the guideline to narrative annotation.

B.1 What is a narrative?

As per Dahlsrom's definition in 2021 (Dahlstrom, 2021), a narrative can be described as a communication that recounts the journey of particular characters through a sequence of interconnected events within a specified timeframe. This concept fundamentally revolves around conveying someone's personal experience or perspective on a subject.

B.2 Annotating sentences from podcasts for narratives

Based on these definitions, the rules for labeling the sentences are as follows: The sentence itself must be part of a story that contains

1. At least one specific character (normally is a person) who experiences.
2. A series of related events.

You may assume that the presence of multiple events implies temporality and do not specifically check for temporality during annotation.

B.2.1 Characters

1. Character/characters need to refer to specific individuals.
2. Characters can be the speaker (1st person), but can also be someone else who is mentioned in the text (2nd or 3rd person).

B.2.2 Events

An event, can be characterized as a notable occurrence that takes place at a particular moment and location, and it typically leads to significant outcomes. In the tangible world, this could encompass incidents such as an explosion triggered by a bomb, the birth of a successor, or the passing of a renowned individual.
Example: But actually she embodies so much wisdom in her teaching.

B.2.3 Context

In this case, we should use context sentences; a sentence can be a part of a narrative context.
Example:
Sentence 1: I just I don't know.
Sentence 2: It doesn't feel I can do it if I'm really really tired and if I'm not I'm like I should be doing something more than this for at least a few poses, and it's strange because the feedback I've had from students.
Sentence 1 can be seen as not a narrative sentence, but while reading the next sentence, we can see that it's a part of narrative context. So both of them can be narrative.

B.2.4 Clarifications

Emotions, thoughts, or other non-observable actions can be considered an event. The characters involved don't necessarily need to take any actions but should be involved in or experiencing the events somehow. Events can be fictional, false, or occurring in the future. They don't need to be actual things that have definitely happened.

B.2.5 Examples

Example 1: So Ruth and I started working together last year. **Label:** 1
Example 2: It doesn't work like that. **Label:** 0

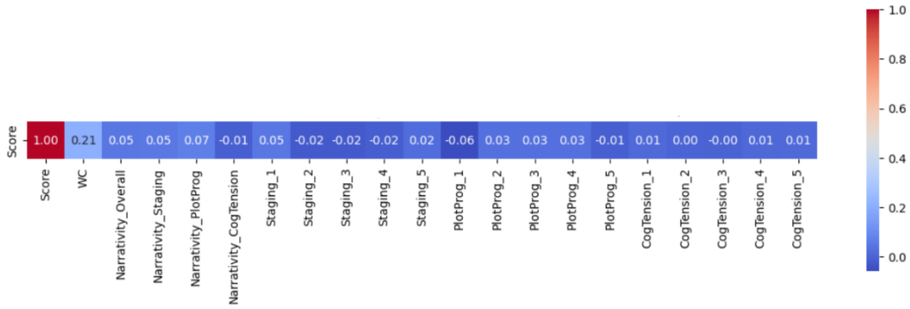


Figure A2: Correlation between LIWC narrative arc score and our score.

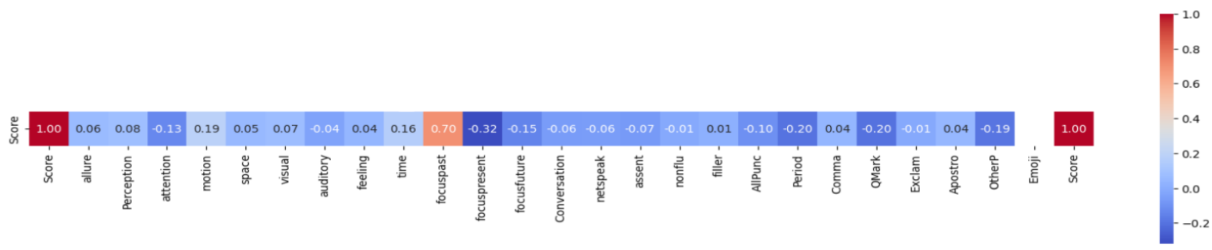


Figure A3: Correlation between sample LIWC categories and our narrativity score.