It's What You Say and How You Say It: Exploring Textual and Audio Features for Podcast Data

Anonymous ACL submission

Abstract

Podcasts are relatively new media in the form of spoken documents or conversations with a wide range of topics, genres, and styles. With a massive increase in the number of podcasts and their listener base, it is beneficial to understand podcasts better, to derive insights into questions such as what makes certain podcasts more popular than others or which tags help in characterizing a podcast. In this work, we provide a comprehensive analysis of hand-crafted features from two modalities, i.e., text and audio. We explore multiple feature combinations considering podcast popularity prediction and multi-label tag assignment as proxy downstream tasks. In our experiments, we use document embeddings, affective features, named entities, tags, and topics as the *textual features*, while multi-band modulation and traditional speech processing features constitute the *audio features*. We find the audio feature *prosody* and textual affective features, *sentiment* and *emotions* are significant for both the downstream tasks. We observe that the combination of textual and audio features helps in improving performance in the popularity prediction task.

1 Introduction

Podcasts have emerged as an exciting medium for entertainment, advertising, news, and information dissemination. According to Nielsen (Nielsen, September 2021), the total number of podcast titles is reaching 2 million with a steady increase in listeners across all demographics. The Interactive Advertising Bureau (IAB) believes that the US-Podcast revenue will see a big jump from $842 million in 2020 to $2 billion by 2023. Content creator apps like Anchor (Anchor, 2022), and Riverside (Riverside.fm, 2022) provide an easy framework to record, edit, and publish a podcast on the media platforms. With a steady increase in the listener base and podcast content, several open problems in handling and accessing this information have emerged. Jones et al. (2021) highlights some of the unique challenges and future directions in the domain of podcast information access. They point out that the existing technologies addressing tasks such as *Search*, *Recommendation*, *Summarization*, and *User experience* are inadequate to handle the multi-genre, multi-style and multi-format composition of podcasts. As this opens up an exciting landscape for future research, we believe that a comprehensive feature analysis of podcasts can serve as an important groundwork in tackling some of these problems. These features need to address *what* podcasts contain and *how* they are delivered.

Podcasts represent inherently heterogeneous data consisting of music and speech in different spoken and written styles and formats. With this work, we investigate the efficacy of individual traditional features and their combinations in understanding *whats* and *hows* in the context of podcast data. Textual and audio features such as tags, sentiment, emotions, and modulation-based features are traditionally considered as a basis of downstream applications. We formulate an application-oriented framework to evaluate and understand the interplay among these features. We consider two separate applications, i.e., *Podcast Popularity Prediction*, and *Podcast Tag Assignment*, each targeting a specific set of features and modality. For example, we hypothesize that textual features may be more informative in the tag assignment task, whereas audio features may be more relevant in the popularity prediction task, as audio captures the style and listening experience important for popularity and tags are more dependent on the content. We also hypothesize that combining features from different modalities can be more informative in characterizing a podcast.

To summarize, in this work, we evaluate various hand-crafted features, from both podcast audio and its textual transcript, that are necessary to understand podcasts. We briefly describe each of these features and how they contribute to representing podcasts (Section 2). We consider the aforementioned proxy downstream tasks to study and evaluate these features. We compile and curate the dataset introduced in (Yang et al., 2019) (Section 3). We experiment with multiple combinations of text and audio features to evaluate their effect and usefulness in the context of the tasks mentioned above (Section 4). We further provide an in-depth analysis on these features and their efficacy in the context of existing experiment setup (Section 5). Our contribution in this work is a detailed feature analysis useful for future research addressing podcast data.

2 Features

As a podcast is essentially an audio experience with focused textual content, we consider both text and audio modalities in this study. Specifically, we focus on widely used hand-crafted features for our experiments. Next, we present a brief overview of features and their characteristics in the context of a podcast.

2.1 Audio Features

Audio features we consider need to represent *how* a podcast is delivered. For audio modeling, we use short-time modulation features derived through multi-band filtering and energy operators and the other traditional speech processing features such as Mel-frequency Cepstrum Coefficients (MFCC), prosody features, and emotions to extract salient properties.

Modulation: Short-time features like energy and absolute amplitude may assign less importance score to *content-rich* podcasts spoken with low amplitude. For example, the spectrogram and magnitude-based features typically used for podcast analysis may ignore the low energy and high-frequency sound (Dimitriadis et al., 2005). We believe such sound may contribute to predicting the popularity of a podcast. Though the magnitude information correlates well with the understanding of speech, the spectral content can also convey comparable speech intelligibility and also provides information about the speaker characteristics that we believe can distinguish popular podcasts (Boashash, 1992). Here, we introduce the features that equally consider the energy, amplitude, and spectral content. For any multi-component audio signal like the podcast, we need to break it down into its Amplitude, and Frequency modulation (AM-FM) components (Boashash, 1992), since there could be multiple frequencies varying as a function of time. (Zlatintsi et al., 2012) use these AM-FM components to assign a measure of interest (importance) to audio frames for audio event detection and summarization task. For this work, we believe that these AM-FM components can prove beneficial in predicting the popularity of a podcast since they can capture the dynamic nature and preserve the subtle harmonic structures present in audio (Dimitriadis et al., 2005).

Here, we compute the AM-FM features by multiband filtering the audio signal using 40 Gabor filters (Evangelopoulos and Maragos, 2006). A nonlinear energy tracking operator, the Teager Energy Operator (TEO), estimates the squared product of the instant amplitude and frequency for every multiband filtered signal. To extract these individual features, we rely on the Energy Separation Algorithm (ESA) (Evangelopoulos and Maragos, 2006; Maragos et al., 1993). The ESA tracks a filter that records the Maximum average Teager Energy (MTE) and also computes their corresponding values of Mean Instantaneous Amplitude (MIA) and Mean Instantaneous Frequency (MIF). We extract 1198-dimensional MTE, MIA, and MIF representations using $12s$ snippets of the leading 10 minutes of every podcast (Yang et al., 2019). To reduce the computational requirements, we resample the podcast at $16kHz$ and use a window size of $25ms$ and a window shift of $10ms$.

MFCC: MFCC is based on the human auditory system and employs a nonlinear scale to correlate with the human perception of the frequency contents of a sound. We consider 39-dimensional delta and double-delta MFCC representation using the similar parameters adopted for extracting modulation-based features.

Prosody: Prosody (non-verbal) features effectively capture the speaker characteristics, their speaking style, emotional state, and approximately identify the listener's interest towards a section of an audio (Adell et al., 2005). We use PRAAT to extract 15-dimensional prosody features on the entire podcast (Boersma and Van Heuven, 2001). The prosody features include $F0$-median, $F0$-mean, $F0$-standard deviation, $F0$-minimum, $F0$-maximum, number of pulses, number of periods, their mean and standard deviation, number of unvoiced frames, number and degree of voice breaks, mean autocorrelation, mean noise-to-harmonic and mean harmonic-to-noise ratio.

Affective Audio Features: The emotional state in speech and audio is one of the most important paralinguistic messages captured during human interactions. We use a Wav2Vec2 fine-tuned model on the IEMOCAP database to extract 4-dimensional *audio emotions* features (neutral, happy, sad, and angry) of an entire podcast (Baevski et al., 2020; Yang et al., 2021).

2.2 Textual Features

According to a Nielsen report (Insights, 2020), podcast engagement has steadily seen growth in the number of heavy as well as light listeners owing to quality content being offered to the listeners. As an exercise to understand the characteristics of such content, we evaluate various textual features. We consider a set of textual features that can help understand *what* a podcast talks about.

Tags: It is observed that every podcast is assigned a set of tags representing its categorization such as *Arts*, *Society*, *Sports*, *Business*, etc. This assignment is often user-defined and captures the scope of the podcast. These tags are a mix of fine-grained and coarse-grained categories. We use an average 100-dimensional Glove embedding vector (Pennington et al., 2014) representing the assigned set of tags to the podcast.

Topics: Even though there exist a set of tags describing a podcast, topics give a detailed list of concepts covered. The topics along with the pre-assigned tags give a much wider representation of the podcast content. We use the unsupervised topic detection algorithm Top2Vec (Angelov, 2020) to generate a list of topics for a given podcast. The topic words are represented using an average 100-dimensional Glove embedding vector for a given podcast.

Affective Textual Features: Emotions play an important role in human cognition, including perception, attention, learning, and reasoning (Tyng et al., 2017). Affective content is more engaging to users than neutral content (Xu et al., 2014). Accordingly, we consider two perspectives of affective features given below.

1. **Sentiment**: Sentiment of a podcast refers to the inclination of the podcast content towards positive or negative polarity. We use the NLTK sentiment analyzer to extract the sentiment of the content. We create a three-dimensional sentiment representation vector for each podcast with scores corresponding to negative, neutral, and positive sentiment that sum up to 1.

2. **Text Emotions**: We consider six basic human emotions, i.e., anger, sad, happy, fear, disgust, and neutral, assigned to sentences in podcast transcripts. We use a zero-shot sentence classification setup with task-aware sentence representations (Halder et al., 2020)[1] to predict the emotion and the corresponding confidence for a sentence, which we further utilize to get an emotion representation for a given podcast.

   For each sentence of the podcast, the model predicts one of the above six emotions denoted by $e_1, e_2, e_3, e_4, e_5, e_6$ and gives their respective confidence scores denoted by $s_1, s_2, s_3, s_4, s_5, s_6$. We further denote the probability of an emotion e_i by $P(e_i) = c_i/n$, where c_i is the corresponding count of emotion e_i and n denotes the count of all sentences with detected emotions in the podcast. The confidence score of a model m for a given emotion e_i is denoted by s_i where $P(m|e_i) = s_i$. For the final representation of our emotion probability vector, we calculate posterior probability of each emotion e_i for a given model m as

$$P(e_i|m) = \frac{P(e_i).P(m|e_i)}{\sum_{i=1}^{6} P(e_i).P(m|e_i)}$$

Named Entities: Named entities like '*person*' and '*organization*' can play a significant role in attracting audience attention. These are generally not covered in tags or topics features explained above. We consider named entities in the podcast title as one of the features. We predict named entities and the corresponding confidence score using the TARS (Halder et al., 2020) zero-shot sentence classification setup. We predict named entities with four dimensions: '*Person*', '*Organization*', '*Location*', and '*Others*' represented by n_1, n_2, n_3, and n_4 respectively. To obtain a four-dimensional named

[1] https://github.com/flairNLP/flair/blob/master/resources/docs/TUTORIAL_10_TRAINING_ZERO_SHOT_MODEL.md

entities representation for a podcast, we use a similar formulation used in the text emotion feature representation.

Document Embeddings: Pre-trained language models like BERT (Devlin et al., 2019) and DistilBERT (Sanh et al., 2019) are often used to extract content representations. Specifically, we use contextual embeddings to get vector representations for the text content. Considering the longer length of textual transcripts, we use Longformer embeddings (Beltagy et al., 2020) to extract podcast content representations.

3 Dataset

Yang et al. (2019) introduced a dataset of 6,511 English language podcasts for the podcast popularity prediction task from various categories like *Arts*, *Society*, *Sports*, *Business*, etc. They scraped the publicly available iTunes podcast directory to get 46,358 episodes from 18,433 channels active from July 2016 to July 2017. They considered the ranking of channels on the iTunes chart as the basis of channel (and episodes) popularity. Top-200 podcast channels were considered popular, and all the episodes from the popular channels were labeled as *popular*. The remaining episodes were labeled as *unpopular*. At most, one episode published in the most recent two weeks from a channel was considered in the dataset[2]. Joshi et al. (2020) use textual transcripts from this dataset to predict podcast popularity.

For our experiments, we enhance this data further first by scraping the audio files of the podcasts using the links provided along with the dataset. We downloaded $3,526$ audio files out of the total $6,511$ podcast episodes, as the remaining podcast links were broken or information was missing. In our experiments, we use the text transcripts as provided in the dataset. Since the transcript had no punctuation, we predict the punctuation in the transcripts using a bidirectional RNN and attention-based punctuation restoration technique (Tilk and Alumäe, 2016). Table 1 shows the distribution of the *popular* and *unpopular* episodes in the dataset. The percentage of *popular* and *unpopular* episodes are roughly the same in both modalities, though we have fewer audio files as compared to the text transcripts.

For the analysis of text and audio features on multi-label tag assignment, we extract tags from the RSS feed of the podcasts. Out of 3,526 podcasts with both the text and audio data, the RSS feed (and tags) is available for 3,306 podcasts. Table 2 shows the distribution of tags in 3,306 podcasts. Originally, a podcast can be assigned one or more tags from 105 fine-grained tags. Since we have a small and highly imbalanced dataset, we manually merge these 105 fine-grained tags into 19 coarse-grained tags as described by Apple podcasts[3]. This allows us to map fine-grained tags such as {'*Music*', '*Music Commentary*', '*Music History*', and '*Music Interviews*'} under one coarse-grained tag {'*Music*'}. Table 3 shows the dataset distribution after the dataset was split into 80:20 for train and test set.

	Text	Audio
Popular	837 (12.86%)	454 (12.87%)
Unpopular	5674 (87.14%)	3072 (87.13%)
Total	6511	3526

Table 1: Distribution of the *popular* and *unpopular* podcasts in the dataset.

Tag	#Podcasts	#Popular Podcasts	%Sentiment Difference
True Crime	18	13	3.14
Fiction	28	6	4.74
Government	41	10	5.90
History	71	9	4.12
Kids & Family	100	23	9.37
Science	126	25	6.21
Music	147	28	9.97
Technology	206	31	9.17
TV & Film	283	28	9.80
Religion & Spirituality	287	38	8.45
Comedy	313	50	8.39
Arts	359	38	9.19
Education	359	65	8.48
Sports	361	20	8.72
News	370	55	6.24
Health & Fitness	371	59	8.63
Leisure	376	38	9.05
Business	486	60	9.39
Society & Culture	514	51	8.12

Table 2: Distribution of tags and their popularity in the dataset. Here, we consider 3,306 podcasts with available data from both text and audio modalities. In %Sentiment Difference column, we show the difference in average %positive and %negative sentiment for each tag.

[2] https://github.com/ylongqi/podcast-data-modeling

[3] https://podcasts.apple.com/us/genre/podcasts/id26

Split	#Podcasts	#Tags	#Tags per Podcast
Train	2671	19	1.46
Test	635	19	1.43

Table 3: Dataset distribution for the task of multi-label tag prediction.

4 Experimental Setup

The podcast data is observed to be multi-modal and heterogeneous, motivating us to experiment with combinations of diverse feature sets. Our experiments consist of studies with individual features as well as their combinations. We conduct our experiments using a 20 core CPU with $64GB$ RAM. To analyze and evaluate the efficacy of various features mentioned in Section 2 to characterize and understand podcast data, we use two formulations, i.e. *podcast popularity prediction* and *podcast tag assignment*.

4.1 Podcast Popularity Prediction

In this task, we seek to investigate the factors influencing the popularity of a podcast. We posit that multi-modal features capture important information to predict podcast popularity. We experiment with hand-crafted multi-modal features and analyze their efficiency using a fine-tuned XGBoost and bagging classifier-based architecture. We perform the grid search on hyperparameter space resulting in the best found combination for XGBoost as: *gamma* = 0.2, *maximum depth* = 14, *estimators* = 120, *reg_alpha* = 0.8, and *reg_lambda* = 1.2. Figure 1 shows our binary podcast popularity classification model. We concatenate the features and use a 5-fold cross-validation as an initial step. We reduce the feature dimension using PCA with 256 components in each fold. Since the data is highly imbalanced (as shown in Table 1), we upsample the minority class (i.e., *popular*) using the SMOTE algorithm (Chawla et al., 2002). Next, we train the XGBoost classifier on the upsampled training data. We then use the hyperparameter tuned bagging classifier with 80 estimators to further address the challenge of data imbalance and synthetic data, as it trains the base XGBoost classifier on the random subset of the original dataset and aggregates the predictions. We use the test data (with PCA and without upsampling) to predict the popularity of the podcasts. We report the results using the macro-F1 score.

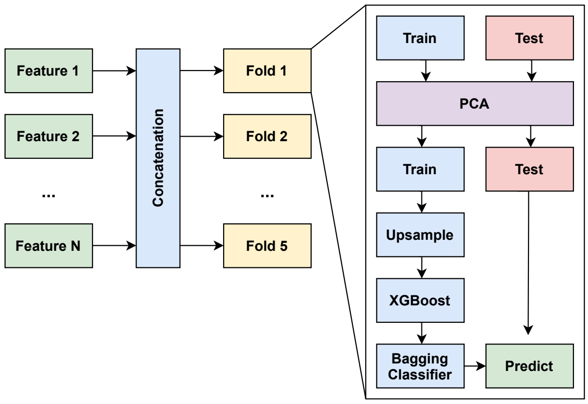

Figure 1: XGBoost and bagging based architecture for podcast popularity prediction. For every fold, the classification setup involves PCA, upsampling, XGBoost, and a bagging classifier.

4.2 Podcast Tag Assignment

In the novel multi-label podcast tag assignment task, we seek to automatically assign appropriate tags to a given podcast. We consider tags available from the RSS feed of the podcast as the ground truth, since these tags are assigned by the hosts of the podcast. As mentioned in Section 3, due to the small and imbalanced nature of the dataset, we manually merge the available 105 fine-grained tags into 19 coarse-grained tags. A podcast can have one or more coarse-grained tags associated with it, similar to a multi-class, multi-label setup. We want to observe the effect of various text and audio features in only text, only audio, and text+audio modalities. One approach to solve multi-label classification is using binary relevance, where we transform the problem into multiple binary models with a one-vs-rest setup. However, since we aim to do feature analysis in different modalities, we create a unified model for multi-label tag assignment by using a simple two-layer perceptron network. For this, we concatenate different combinations of text and audio features and feed them through a fully connected layer, followed by a dropout layer with a dropout probability of 0.3 and another fully connected layer. We use binary cross-entropy loss function with class_weights to handle dataset imbalance. We calculate class_weights for each tag l as N/N_l, where N is the total number of podcasts in the training dataset, and N_l denotes the numbers of podcasts with tag l. Since the dataset is highly imbalanced, we use the weighted-F1 score as the primary evaluation metric. Similar to (Kar et al., 2018), we also evaluate the models on tag recall

Feature(s)	Macro-F1
All Text	0.545
(−) Tags	0.500
(−) Text Emotions	0.537
(−) Sentiment	0.539
(−) Topics	0.547
(−) Named Entities	0.549
All Audio	0.802
(−) MIF	0.603
(−) Prosody	0.784
(−) MFCC	0.801
(−) MTE	0.801
(−) MIA	0.801
(−) Audio Emotions	0.802
All Audio + All Text	0.807

Table 4: Podcast popularity prediction results using multi-modal features. The tags and MIF report the individual best representation, whereas the combination of all audio and text features proves to be more informative in characterizing a podcast.

(TR) and unique tags learned (TL) by the model over weighted-F1 scores. Tag recall is the average recall per tag. Tag recall is calculated as follows:

$$TR = \frac{\sum_{i=1}^{T} R_i}{|T|}$$

Here, R_i is the recall of the i^{th} tag, and $|T|$ is the total number of tags. We train every setup for a maximum of 75 epochs with early stopping criteria. For all experiments, we use a learning rate of 0.0001, batch size of 4, and Adam optimizer.

5 Results and Analysis

In this section, we report our results and analysis on the performance of hand-crafted features in different modalities for the two aforementioned downstream tasks.

5.1 Podcast Popularity Prediction

In Table 4, we present the ablation study of macro-F1 scores using the text and audio features. The tags and MIF are the most significant features in predicting podcast popularity. Their exclusion results in a performance drop of 9% and 33% w.r.t. the all text and audio features, respectively. This may be due to MIF's ability to preserve the spectral content information (which otherwise is ignored in only energy-based computations) for the popularity prediction task. The affective features (sentiment and text emotions) contribute equally well to the score. However, the topics and named entities negatively contribute to the prediction results. On their exclusion from all text features, we observe a marginal rise of 0.36% (topics) and 0.73% (named entities) in the macro-F1 score. We also observe an insignificant contribution from the audio emotions to the overall performance. Since the combinations of all audio and text representations report the highest macro-F1 score of 0.807, we further experiment with different combinations of these features to identify the best performing multi-modal feature combination.

Feature(s)	Macro-F1
Modulation + Prosody + Tags	0.820
(−) Modulation	0.520
(−) Prosody	0.803
(−) Tags	0.811
Modulation + Prosody + All Text	0.810
(−) Modulation	0.506
(−) Prosody	0.791
(−) Topics	0.798
(−) Sentiment	0.801
(−) Named Entities	0.802
(−) Tags	0.809
(−) Text Emotions	0.812

Table 5: Podcast popularity prediction results using the top two combinations of multi-modal features with the highest macro-F1 score. The modulation is the most dominant representation of all multi-modal features.

Table 5 shows the top two multi-modal feature combinations (i.e., modulation + prosody + tags and modulation + prosody + all text) with the highest macro-F1 score and their ablation results. As can be seen from both the combinations, the modulation features are the most significant ones in capturing the intricacies of a podcast for this task. We believe that this performance improvement is solely due to the ability of modulation-based features in modeling the dynamic and non-linear aspects of an audio (Evangelopoulos and Maragos, 2006).

5.2 Podcast Tag Assignment

In Table 6, we present our multi-label tag assignment results for text, audio, and multi-modal frameworks with their ablation analysis. We observe that the text modality significantly outperforms audio and multi-modal frameworks with over 135% increase in weighted-F1 score. Specifically, the '*topics*' emerge as the most prominent feature since its

Feature(s)	w-F1	TR	TL
All Text	44.02	78.22	19
(−) Topics	35.31	76.34	19
(−) LF Embeddings	41.89	71.85	19
(−) Named Entities	43.37	79.44	19
(−) Sentiment	43.42	77.07	19
(−) Text Emotions	43.83	79.34	19
All Audio	18.79	80.40	18
(−) Prosody	16.26	57.89	14
(−) MIA	17.55	51.20	15
(−) MTE	17.87	40.78	15
(−) Audio Emotions	18.68	71.84	15
(−) MFCC	19.30	63.84	17
(−) MIF	22.54	70.39	17
All Text + Audio	18.62	69.36	17
(−) LF Embeddings	17.57	70.23	16
(−) Prosody	17.71	63.86	15
(−) MTE	17.90	53.89	15
(−) Audio Emotions	17.91	54.09	12
(−) Topics	18.36	71.92	16
(−) Sentiment	18.49	52.59	15
(−) Named Entities	18.60	59.59	17
(−) MIA	18.63	58.46	15
(−) Text Emotions	18.98	65.40	17
(−) MFCC	19.26	55.45	15
(−) MIF	31.14	55.47	16

Table 6: Ablation analysis of text and audio features for multi-label tag assignment task. We use weighted-F1 (w-F1), tag recall (TR) and unique tags learned (TL) as the evaluation metrics. LF Embeddings represent the Longformer embeddings of the podcast transcript.

removal results in a 19.78% drop in the weighted-F1 score. Textual features like '*topics*' provide a detailed list of concepts covered in the podcast. This is important in understanding *what* the podcast is about and, in turn, assigning relevant tags.

Longformer Embeddings are the next most important textual feature, followed by Named Entities, Sentiments, and Emotions. Overall, all the textual features play an essential role in tag assignment as the weighted-F1 scores drop after removing any of these features.

The audio features as standalone representations do not perform well in tag assignment tasks. Even with all the audio features, the model fails to learn all the tags. After removing MFCC and MIF, the weighted-F1 scores increase by 2.71% and 19.95%, respectively. Even though these features can capture human speech very well for the task of short-form audio classification (Bergstra et al., 2006), they fail to provide desirable performance in the case of long-form content like podcasts. This falls in line with our hypothesis that textual features competently capture the complex nature of podcasts for the task of tag assignment. However, prosody features are the most dominant in the audio modality as with its exclusion, the model fails to learn five tags altogether while also producing a significant drop of 13.46% and 27.99% in the weighted-F1 score and tag recall, respectively. Similarly, after excluding MTE features, we see a drop of 04.89% and 49.27% in weighted-F1 and average tag recall metrics, respectively. This may be due to MTE features' ability to retain the signal envelope variations where a speech activity is detected.

The model also does not benefits from multi-modal setup. Similar to the audio modality, the MFCC, and MIF features fail to capture the tags appropriately. We note an increase of 67.23% in the weighted-F1 score with the removal of MIF features. In a multi-modal framework, longformer embeddings perform the best, followed by audio emotions, prosody features, and sentiments. To understand why sentiments perform well for tag assignment in both text and multi-modal scenarios, we take the average percentage of negative, neutral, and positive sentiment across all podcasts under each tag. On average, 82.6% of every podcast has neutral sentiment irrespective of the tags. In Table 2 We report the difference between the average %positive and %negative sentiment per tag. We believe the model uses these differences to learn mappings to corresponding tags. We can see that the model fails to learn {'*True Crime*', '*Fiction*', '*Government*', and '*Science*'} tags completely in the absence of sentiments. These amount to four out of five tags with the lowest difference between average positive and negative sentiment.

5.3 Results Summary

For the popularity prediction task, the *tags* and *MIF* features are most effective for text modality and audio modality experiments, respectively. The MIF representation identifies the important events in a podcast that may impact its popularity. We observe that the model benefits from the inclusion of the tags feature. We can identify from Table 2 that some tags such as '*True Crime*' in general are much more popular over tags like '*Society & Culture*'. The combination of all audio and text features provides a more informative representation in predicting the popularity of a podcast (Table 4). In particular, the combination of modulation, prosody, and tags yields the highest macro-F1 score for popularity prediction (Table 5).

For tag assignment, topics and Longformer embeddings are the most prominent features from text modality as they effectively capture the content within the podcast. The combination of all text features gives us the highest weighted-F1 score. The standalone audio and multi-modal frameworks give a mediocre performance for multi-label tag assignment. Prosody and affective text features are found to be essential in multi-modal setups irrespective of the downstream tasks.

6 Related Work

One of the first works introducing large-scale podcast data and relevant tasks is by Clifton et al. (Clifton et al., 2020) from *Spotify*. They compiled a corpus of 100,000 podcast episodes comprising nearly 60,000 hours of speech along with transcriptions. Recent work in Alexander et al. (2021) further enriches this dataset with precomputed audio features based on prosody and MFCCs. They demonstrate how these features can be used in podcast segment categorization based on delivery(e.g., entertaining, subjective, or discussion). We extract and use similar features in our analysis but on a different dataset. Earlier works based on this dataset, such as abstractive summarization in (Zheng et al., 2020), and PodSumm in (Vartakavi and Garg, 2020) consider pre-trained models such as BART, BERT, and T5. These works do not specifically consider hand-crafted audio and textual features, and their efficacy remains relatively unexplored.

Another similar dataset of note was compiled by Yang (Yang et al., 2019) consisting of data from nearly 88,728 podcast episodes on Apple iTunes. Along with data, they also introduce an Adversarial Learning-based Podcast Representation (ALPR) that captures non-textual aspects of podcasts. They evaluate these representations in the context of podcast popularity prediction and prediction of seriousness-energy in podcasts reporting state-of-the-art results. We enhance this data and use their insights to formulate our experimental framework. Joshi et al. (Joshi et al., 2020) consider DistilBERT based embeddings as textual features with the triplet loss to address the popularity prediction task for the data introduced in (Yang et al., 2019) with state-of-the-art results. They note that polarity and subjectivity of features remain similar with no marked difference, thus not very informative for the popularity prediction. We seek to investigate this further by considering various hand-crafted features from both podcast audio and its text transcript, along with enhanced data. While the DistilBERT embeddings with triplet loss act as a black box and are difficult to explain, we focus on hand-crafted features for greater explainability about how different features contribute towards popularity.

(Dhanaraj and Logan, 2005) studied audio and text modalities for popularity prediction of songs using Support Vector Machine and boosting classifiers. A lot of work has been done in genre prediction on short-form audio content using MFCCs (Mandel et al., 2006)(Bergstra et al., 2006). Recently, (Wilkes et al., 2021) performed feature analysis in text, audio, and video modalities for the task of music genre prediction using machine learning classifiers.(Cascante-Bonilla et al., 2019) use audio, text, and video modalities from movie trailers, posters, plots, and other metadata to predict movie genre. However, to the best of our knowledge, no one has explored multi-modal feature analysis on podcasts on the task of multi-label tag assignment. Inspired by these works, we seek to understand how features from different modalities perform (separately or combined) in the context of proxy applications.

7 Concluding Remarks

Podcasts are spoken-documents ranging across a wide variety of genres, topics, and styles. Owing to the rapid growth in popularity and global reach, there is a definite need to explore and investigate this new engagement medium and relevant research landscape. In this work, we study different hand-crafted features and their combinations based on podcast audio and its textual transcript to characterize the podcast data. As can be seen, features capturing distinct qualities like speaker style, content affect, and subject content coverage (i.e., prosody, sentiment, emotions, and topics) are significant irrespective of the downstream task. We believe that the analysis can be helpful in several other downstream tasks such as podcast summarization, retrieval, and recommendation.

References

Jordi Adell, Antonio Bonafonte, and David Escudero. 2005. Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech. *Procesamiento del lenguaje natural*, (35):277–283.

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.

Anchor. 2022. The easiest way to make a podcast. *accessed 11th January 2022*.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

James Bergstra, Norman Casagrande, Dumitru Erhan, Douglas Eck, and Balázs Kégl. 2006. Aggregate features and a da b oost for music classification. *Machine learning*, 65(2-3):473–484.

Boualem Boashash. 1992. Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals. *Proceedings of the IEEE*, 80(4):520–538.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. 2019. Moviescope: Large-scale analysis of movies using multiple modalities. *arXiv preprint arXiv:1908.03180*.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, et al. 2020. 100,000 podcasts: A spoken english document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ruth Dhanaraj and Beth Logan. 2005. Automatic prediction of hit songs. In *ISMIR*, pages 488–491. Citeseer.

Dimitrios Dimitriadis, Petros Maragos, and Alexandros Potamianos. 2005. Robust am-fm features for speech recognition. *IEEE signal processing letters*, 12(9):621–624.

Georgios Evangelopoulos and Petros Maragos. 2006. Multiband modulation energy tracking for noisy speech detection. *IEEE Transactions on audio, speech, and language processing*, 14(6):2024–2038.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213.

Nielsen Insights. 2020. Podcast content is growing audio engagement. *accessed 11th January 2022*.

Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, et al. 2021. Current challenges and future directions in podcast information access. *arXiv preprint arXiv:2106.09227*.

Brihi Joshi, Shravika Mittal, and Aditya Chetan. 2020. Did you “read” the next episode? using textual cues for predicting podcast popularity. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 13–17.

Sudipta Kar, Suraj Maharjan, and Thamar Solorio. 2018. Folksonomication: Predicting tags for movies from plot synopses using emotion flow encoded neural network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2879–2891.

Michael I Mandel, Graham E Poliner, and Daniel PW Ellis. 2006. Support vector machine active learning for music retrieval. *Multimedia systems*, 12(1):3–13.

Petros Maragos, James F Kaiser, and Thomas F Quatieri. 1993. Energy separation in signal modulations with application to speech analysis. *IEEE transactions on signal processing*, 41(10):3024–3051.

Nielsen. September 2021. Insights for podcast advertisers. *Podcasting Today*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Riverside.fm. 2022. Record podcasts and videos from anywhere. *accessed 11th January 2022*.

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*.

Ottokar Tilk and Tanel Alumäe. 2016. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech 2016*.

Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik. 2017. The influences of emotion on learning and memory. *Frontiers in psychology*, 8:1454.

Aneesh Vartakavi and Amanmeet Garg. 2020. Podsumm–podcast audio summarization. *arXiv preprint arXiv:2009.10315*.

Ben Wilkes, Igor Vatolkin, and Heinrich Müller. 2021. Statistical and visual analysis of audio, text, and image features for multi-modal music genre recognition. *Entropy*, 23(11):1502.

Min Xu, Jinqiao Wang, Xiangjian He, Jesse S Jin, Suhuai Luo, and Hanqing Lu. 2014. A three-level framework for affective content analysis and its case studies. *Multimedia tools and applications*, 70(2):757–779.

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*, pages 276–284.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Chujie Zheng, Harry Jiannan Wang, Kunpeng Zhang, and Ling Fan. 2020. A baseline analysis for podcast abstractive summarization. *arXiv preprint arXiv:2008.10648*.

Athanasia Zlatintsi, Petros Maragos, Alexandros Potamianos, and Georgios Evangelopoulos. 2012. A saliency-based approach to audio event detection and summarization. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1294–1298. IEEE.