Did You “Read” the Next Episode?
Using Textual Cues for Predicting Podcast Popularity

Brihi Joshi* **Shravika Mittal*** **Aditya Chetan***
Indraprastha Institute of Information Technology, Delhi
{brihi16142, shravika16093, aditya16217}@iiitd.ac.in

Abstract

Podcasts are an easily accessible medium of entertainment and information, often covering content from a variety of domains. However, only a few of them garner enough attention to be deemed ‘popular’. In this work, we investigate the textual cues that assist in differing popular podcasts from unpopular ones. Despite having very similar polarity and subjectivity, the lexical cues contained in the podcasts are significantly different. Thus, we employ a triplet-based training method, to learn a text-based representation of a podcast, which is then used for a downstream task of “popularity prediction”. Our best model received an F1 score of 0.82, achieving a relative improvement over the best baseline by 12.3%.

1 Introduction

Predicting the popularity of media content, such as songs, podcasts, etc., before its release can have significant implications for the producers, artists, etc. Traditionally, this task has been attempted with hand-crafted feature sets (Tsagkias et al., 2008), and utilising various audio features (Dhanaraj and Logan, 2005). However, hand-crafted feature sets are often not scalable, while audio-based features ignore the textual cues that are present in the data. Recently, with the rise in popularity and efficacy of Deep Learning, Neural network-based models (Yang et al., 2017; Zangerle et al., 2019) have also been proposed for hit-song prediction. There have also been some attempts (Yang et al., 2019) to learn a general representation for media content, but only based on the audio of the content, not from the textual cues.

In this work, we attempt to study the following: *How does the textual content of popular podcasts differ from that of unpopular ones?* First, we conduct experiments to assess the polarity of popular podcasts, and observe that it is quite similar to that of unpopular podcasts. This observation is also prevalent while studying the subjectivity of the transcripts. Furthermore, there is little to no variation when polarity and subjectivity are studied over time. We then analyse the differences in the keywords and the general topical categories interspersed between popular and unpopular podcasts. It is observed that content generally centered around ‘Politics’, ‘Crime’ or ‘Media’ is more popular than others. Keeping this in mind, we design a triplet-training method, that leverages similarities between the popular and unpopular podcast samples to create representations that are useful in the downstream podcast popularity prediction task.

2 Related Work

The problem of “popularity prediction” has been explored for different types of media content, in a variety of ways. For instance, Hit song prediction has been an active area of research. Dhanaraj and Logan (2005) used spectral features like MFCCs to train an SVM for predicting whether a song would be a hit or not. Yang et al. (2017) proposed a Convolutional Neural Network based architecture for predicting the popularity of a song, using audio-based features. More recently, Zangerle et al. (2019) employed a combination of low-level and high-level audio descriptors for training Neural Networks on a regression task. However, these works have not taken textual cues into account when predicting the popularity of a song. Sanghi and Brown (2014) made an attempt to use lyric-based features that incorporated the rhyming quality of the song. However, they did not learn a representation based on the lyrics.

For podcasts, Tsagkias et al. (2008) gave a framework for assessing the credibility of pod-

*Equal contribution. Ordered randomly.

casts. Their notion of credibility included preference of the listeners. The framework was also shown to be reasonably effective in predicting popular podcasts (Tsagkias et al., 2009). This framework included highly refined hand-crafted features, based on both audio, textual and content describing the podcast on its platform. Recently, Yang et al. (2019) proposed a GAN-based model, for learning representations of podcasts, based on non-textual features, and showed its applications in downstream tasks like music retrieval and popularity prediction.

Finally, popularity prediction is also challenging because of the class imbalance that is inherent in the problem definition itself. Popular podcasts or songs would always be in a minority in a corpus. This makes the task of learning a good representation for them difficult. To overcome this, we exploit the triplet-based training procedure (Hoffer and Ailon, 2015) for generating a balanced distribution of both popular and unpopular podcasts as the "anchor" podcast. (See Section 5.1)

3 Dataset

In our study, we use the dataset collected by Yang et al. (2019) as a part of their podcast popularity prediction task. The dataset consists of 6511 episodes among which, there are 837 popular and 5674 unpopular (*long-tail*) podcasts. Based on the iTunes chart ranking, channels corresponding to the top 200 podcasts were treated as "top channels" and episodes from these top channels were then labelled as popular. Yang et al. (2019) provide a random 60-40 split of the dataset as a training and testing set. The average duration of the podcasts is 9.83 minutes.

In this work, we only use the transcripts that are provided with the podcast audio. Each transcript contains the start and end timestamps (in milliseconds) along with every spoken token in a new line. We remove the timestamps and stop words for all transcripts. We also do not consider non-verbal vocalisations in the transcript (for example, "*ooooo*", "*ahhh*", etc.) for our analysis. After pre-processing, the podcast transcriptions contain 1557 tokens on an average.

4 Data Analysis

4.1 Polarity Analysis

In order to understand the general polarity and sentiment across popular and unpopular podcasts, we extract the polarity scores of each podcast using TEXTBLOB[1], which is calculated by averaging the polarity of pre-defined lexicons, inferred from the words in the podcast. The polarity values range between -1 to 1, where anything above 0 is considered to be 'positive'.

We average the obtained polarity scores for all the podcasts, for each of the popular and unpopular categories. It was observed that the overall polarity of popular and unpopular podcasts is roughly **the same** – as the average polarity score for the popular class was 0.14 and for the unpopular class was 0.15.

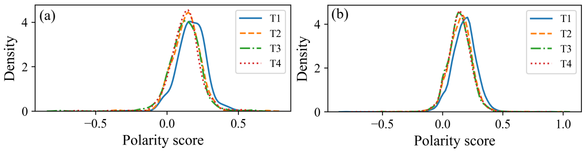

Figure 1: Density distribution of raw polarity scores for (a) Popular and (b) Unpopular podcasts over four time intervals.

In order to understand how polarity varies *over time*, we split each podcast into four time-chunks based on the three quartiles ($Q1$, $Q2$ and $Q3$), which we call $T1$, $T2$, $T3$ and $T4$, in order, with the help of the timestamps provided with the podcast transcripts.

Figure 1 shows the density distributions for raw polarity scores over the four splits (based on timestamps) for the two categories. It is observed that **both** popular and unpopular podcasts start-off with a positive tone, slowly transitioning into neutral content. However, there is limited observable distinction between popular and unpopular podcasts based on polarity.

4.2 Subjectivity Analysis

Similar to Polarity analysis, we looked into subjectivity scores for each podcast using TEXTBLOB, which is calculated by averaging the subjectivity of pre-defined lexicons, inferred from the words in the podcast. The values vary between 0 and 1 such that, the higher the score the more 'opinion based' (subjective) the text is.

As was observed for polarity, the overall subjectivity of popular and unpopular podcasts is exactly the **same** – as the average subjectivity score obtained across all podcasts was 0.48 for both popular and unpopular classes.

[1] https://textblob.readthedocs.io/en/dev/

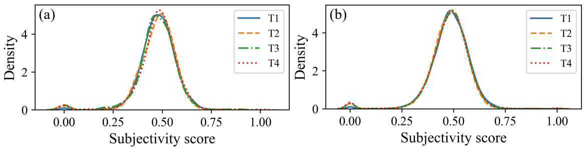

Figure 2: Density distribution of raw subjectivity scores for (a) Popular and (b) Unpopular podcasts over four time intervals.

To capture how subjectivity varies over time we used the same four timestamp based podcast chunks as was used for Polarity analysis. Figure 2 shows the density distributions for raw subjectivity scores over the four splits for the two categories.

It can again be observed that **both** popular and unpopular podcasts maintain their subjectivity over time with no significant differences across categories.

4.3 Lexical Analysis

We use EMPATH (Fast et al., 2016) to analyse the topical signals with the help of 194 pre-defined lexicons (for example – 'social media', 'war', 'violence', 'money', 'alcohol', 'crime' to name a few) that highly correlate with LIWC (Tausczik and Pennebaker, 2010).

We extract the scores from EMPATH for each category, for each podcast. The most and the least relevant lexical categories for popular podcasts, ordered by their significance values are given in Table 1.

Rank	Lexical Categories
1	**Government**
2	Crime
3	Politics
4	Money
5	Law
190	Hygiene
191	Social Media
192	Urban
193	Worship
194	Swimming

Table 1: Lexical Categories that are more likely to be present in popular podcasts, than unpopular podcasts: We run a Welch's two sample t-test on the category scores for each podcast. Top-5 lexical categories shown are more significantly ($p < 0.05$) present in popular podcasts, than unpopular ones. Bottom-5 categories are ordered according to least significance ($p > 0.95$).

4.4 Keyword co-occurrence

We also study what kind of keywords are present in popular and unpopular podcasts. We rank bigrams based on their Pointwise Mutual Information (PMI) scores and report the top 10 in Table 2.

It can be observed that in podcasts belonging to the popular class, keyword pairs like *'Hillary Clinton'*, *'Donald Trump'*, or *'Gordon Hayward'* outshine highlighting the possibility of domain areas such as 'Politics', 'Sports', or 'Celebrities' to be responsible for making a podcast popular. This can also be seen in Section 4.3, which shows that 'Government' related topics are widely present in popular podcasts.

On the other hand the top keyword pairs extracted from unpopular podcasts belong to more generic domains like 'Cities', 'Lifestyle', etc., to name a few.

Popular		Unpopular	
Bi-gram	PMI	Bi-gram	PMI
Los Angeles	42.26	Web Site	85.62
United States	37.95	New York	80.63
New York	26.73	E Mail	80.50
Gordon Hayward	15.11	Fourth July	61.28
North Korea	14.56	Two Thousand	60.50
Blue Apron	13.87	High School	52.82
Hillary Clinton	12.40	Las Vegas	43.00
Donald Trump	9.02	Hong Kong	41.90
Fourth July	8.19	Real Estate	37.44
San Francisco	8.01	Wal Mart	34.71

Table 2: Top 10 bi-grams (ranked by their PMI values) for Popular vs. Unpopular podcasts: The keyword bi-grams in bold are encompassed by topics that are shown to be highly relevant for popular podcasts in Section 4.3.

5 Podcast Popularity Prediction

5.1 Proposed Method

Owing to the lack of a balanced dataset for popularity prediction, we use the Triplet Training strategy. In this method, instead of having class labels like 'popular' or 'unpopular' for the podcasts, we group the podcasts into triplets – each triplet has an anchor a podcast, which is often the reference for comparison, a positive podcast p which belongs to the same class as a, and a negative podcast n which belongs to the other class. The intuition is to reduce the distance between the representations of podcasts belonging to the same class and vice versa. After extracting the representation

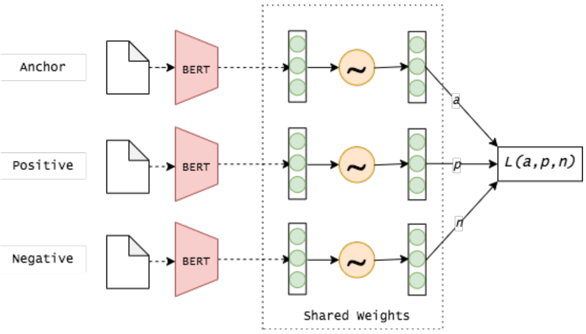

Figure 3: Triplet Training Architecture: The podcast triples are first passed through a DISTILBERT model, followed by a 2-layer Neural Network, with a RELU non-linearity in between. The weights are shared across the triplet during training.

of all the three podcasts in a triplet from a network with shared weights, we use the Triplet loss given below, as introduced by Schroff et al. (2015).

$$\mathcal{L}(a,p,n) = \sum_{i=1}^{N}\left[\|f(a_i)-f(p_i)\|_2^2 - \|f(a_i)-f(n_i)\|_2^2+\alpha\right]$$

where a_i, p_i and n_i are the anchor, positive and negative podcast samples in the i^{th} triplet, f is a function that outputs an embedding for the podcasts and α is the margin between the positive and negative podcast samples.

We use a pre-trained DISTILBERT (Sanh et al., 2019) model[2] to create initial representations for the podcasts, followed by two fully connected layers, which shared weights during the triplet training phase. The architecture can be seen in Figure 3. The output of the final layer is a 128-dimensional vector, that is used as an embedding for the downstream popularity prediction task.

5.2 Evaluation and Results

The following methods are used to extract the representations of podcasts to predict their popularity:

- **TF-IDF:** TF-IDF weights (Ramos et al., 2003) corresponding to each word in a podcast are used to fill a vector, the size of which equals the size of training set's vocabulary.

- **WORD2VEC (WV):** WORD2VEC (Mikolov et al., 2013) embeddings for each word in a podcast are averaged to create a single embedding representing the podcast.

Method	Macro-Avg F1
TF-IDF	0.61
WV	0.59
DB	0.73
DB-T	**0.82**

Table 3: Popularity Prediction: Macro-average F1 score for the baselines and the proposed Triplet training strategy for the popularity prediction task.

- **DISTILBERT (DB):** The embedding corresponding to the [CLS] token in a pre-trained DISTILBERT (Sanh et al., 2019) is taken as an embedding for a podcast.

- **DISTILBERT-Triplet (DB-T):** The embedding corresponding to the [CLS] token in a pre-trained DISTILBERT is trained in a Triplet manner as shown in the proposed method (Figure 3), and the output of the final neural network is a 128-dimensional embedding for the podcast.

For each of the methods listed above, embeddings corresponding to every podcast are extracted. We use a supervised classifier like XGBOOST (Chen and Guestrin, 2016) with binary labels for popularity. Results for the various methods are given in Table 3. Appropriate hyperparameter tuning is done over 5-fold cross validation, including adding penalties for misclassifying the minority (Popular) class. It can be seen that our proposed method (DB-T) significantly outperforms the others, achieving a relative improvement over the best baseline (DB) by 12.3%.[3]

6 Conclusion

In this work, we explore how textual cues like polarity, subjectivity, lexicons and keywords differ in popular and unpopular podcasts. We then employ a triplet-based training procedure to counter the class imbalance problem in our data, which yields a relative improvement of 12.3% over the best performing baseline. In future work, we plan to explore this problem in a multi-modal setting, by constructing multi-modal embeddings that leverage both audio and textual data. We also plan to leverage temporal information associated with the transcripts, in the form of timestamps of the spoken words, for the task of popularity prediction.

[2] We use the DISTILBERT BASE model provided by huggingface's transformers library (Wolf et al., 2019)

[3] Code and saved models are available at: https://github.com/brihijoshi/podpop-nlp4musa-2020/

References

Tianqi Chen and Carlos Guestrin. 2016. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Ruth Dhanaraj and Beth Logan. 2005. Automatic prediction of hit songs. In *ISMIR*.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition*, pages 84–92, Cham. Springer International Publishing.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. New Jersey, USA.

Abhishek Sanghi and Daniel G. Brown. 2014. Hit song detection using lyric features alone. In *Proceedings of the 15th International Society for Music Information Retrieval Conference 2014 (ISMIR 2014)*.

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

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Manos Tsagkias, Martha Larson, and Maarten de Rijke. 2009. Exploiting surface features for the prediction of podcast preference. In *Advances in Information Retrieval*, pages 473–484, Berlin, Heidelberg. Springer Berlin Heidelberg.

Manos Tsagkias, Martha Larson, Wouter Weerkamp, and Maarten de Rijke. 2008. Podcred: A framework for analyzing podcast preference. In *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web*, WICOW '08, page 67–74, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

L. Yang, S. Chou, J. Liu, Y. Yang, and Y. Chen. 2017. Revisiting the problem of audio-based hit song prediction using convolutional neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 621–625.

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*, WSDM '19, page 276–284, New York, NY, USA. Association for Computing Machinery.

Eva Zangerle, Ramona Huber, and Michael Vötter Yi-Hsuan Yang. 2019. Hit song prediction: Leveraging low- and high-level audio features. In *Proceedings of the 20th International Society for Music Information Retrieval Conference 2019 (ISMIR 2019)*, pages 319–326.