A Study on Multi-modal Ensemble Models for Commercial Video Memorability Prediction

Olufela Fieulleteau[1,†], Conner Perriello[1,†] and Sejong Yoon[1,*]

[1]*The College of New Jersey, USA*

Abstract

This working notes paper reports our results in participating in the 2026 edition of the Predicting Movie and Commercial Memorability Benchmark. Among multiple subtasks, we focused on Subtask 2, Commercial Memorability. Specifically, we tackled Challenge 2.1, the commercial video memorability prediction task. We investigated several non-visual features capturing emotional, semantic, and audio cues. With experiments, we sought the best combination of the features, including the provided features. Using an ensemble of six models, the best Spearman's Rank Correlation on the test set was 0.225.

1. Introduction

This working notes paper summarizes our approach and results of our participation in the MediaEval 2026 edition of the movie and commercial memorability prediction task. We participated in one of the challenges under the **Subtask 2** category, the commercial video memorability prediction task. Specifically, we focused on **Challenge 2.1**, which aims to predict the memorability score of a given commercial video directly. A detailed description of the task, as well as the dataset used (VIDEM) [1] can be found in the overview paper [2].

In this work, we investigate the utility of non-visual features in combination with popular visual features. We consider three additional categories of features: acoustic, semantic, and emotional. Most of these features, either in identical or similar forms, have been investigated in previous editions of the challenge. For example, for the short-term video memorability task, a combination of visual, aesthetic, emotional, and textural information was found as the best-performing method [3]. Audio features [4] were also considered in the dataset based on TRECVid.

On the other hand, for the commercial video prediction task, not many prior participants attempted acoustic and emotion features. In the 2025 edition of the challenge, the best-performing method utilized visual features and metadata including information that is only obtainable after publication (e.g., engagementRate) or that may introduce auxiliary bias not directly related to the content (e.g., channelName) [5]. Other methods utilized visual and text-based information, or relied heavily on vision-text encoding (e.g., CLIP) [6, 7, 8]. There was one attempt utilizing audio features [9], based on the built-in spectral feature extractors of the Librosa library. However, the effectiveness of those features was not entirely clear. A study [10] on advertisement effectiveness using muted commercial videos reported that (a) there was little difference in effectiveness for muted ads on the mobile platform, (b) there was reduced effectiveness measured by brand recall, but improved liking, and (c) the negative effect of muting on brand recall was significant only if the audio contained substantial information that is essential for recall.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online

*Corresponding author.

†These authors contributed equally.

fieullo1@tcnj.edu (O. Fieulleteau); perriec1@tcnj.edu (C. Perriello); yoons@tcnj.edu (S. Yoon)

To this end, we saw a need for studying the utility of audio features for the challenge task. We designed audio features considering the characteristics of commercial videos. In the same spirit, we also considered semantic and emotional features that can be extracted from the caption.

2. Approach

In this section, we describe each feature and the prediction model briefly. The code to reproduce the work can be found in https://github.com/yoonsejong/ME2026_Memorability/.

2.1. Acoustic Features

Acoustic features are extracted using Librosa [11]. First, audio tracks are extracted from all videos. Next, we estimate the fundamental frequency (f_0) using the Probabilistic YIN (PYIN) algorithm [12]. Using this, we can compute the mean, variance, and range of pitch. Next, using the residual of f_0 over frames, we can compute the pitch velocity v_t, and the statistics of velocity, i.e., the mean and variance are also computed. Based on a raw audio sample y_t, we compute the root mean square (RMS) feature, and using the RMS, we compute the energy mean, variance, and range, defined as the statistics of RMS.

In addition to these standard statistical features, we also consider audio characteristics that may be important in commercial videos. We adopt the variance of pause duration where the pause duration is defined as the interval between speech normalized by the sampling rate. In addition, we consider the pitch change over a voiced segment of the audio track as advertisements often place emphasis on later parts of the commercial videos. We also compute the signal-to-noise ratio (SNR) in decibels where the noise is defined as the bottom 35th percentile of the entire RMS energy. Typically, values larger than this threshold indicate voiced audio, likely including meaningful information. Using this definition, we also compute the speech ratio, which is the ratio of audio frames classified as containing active speech. Finally, we compute the speaking rate, which is the estimated number of syllables per second of active speech. Syllable counts are approximated from the provided transcript.

2.2. Semantic Features

Semantic features are extracted using text structural information (part of speech) that may correlate with the effectiveness of commercial videos. Specifically, we computed the ratios of nouns, verbs, and adjectives. Further, we computed the lexical diversity defined as the number of unique words. Each of these four features is normalized by the total number of words. We also considered the average word length, average sentence length, and rarity defined as the negative log of the frequency of a used word in a corpus.

2.3. Emotion Features

For the emotion features, we used the NRCLex [13] to extract a numerical assessment from the word usage in the transcript. Ten different emotions (joy, anger, fear, sadness, positive, negative, anticipation, disgust, surprise, trust) are extracted from the transcript and the sum of emotions is added as the intensity of the emotion features. In addition, one dominant emotion is determined from the numerical magnitude, and if the identified dominant emotion is one of the four major emotions (joy, anger, fear, sadness), that information is encoded as a one-hot vector. If none of the four is dominant, the vector is all-zeros indicating neutral emotion.

2.4. Overall Model Design

We designed our prediction model as an ensemble of multiple, classic regression models. We considered five models: Linear regression, Ridge regression, a simple Multi-layer Perceptron (MLP), Support Vector Regression (SVR), and Gradient Boosting (GB). Model predictions are fused by rank-averaging, i.e., first, the predictions of each model are normalized into the $[0, 1]$ range and then the average of each sample's predictions across models is taken.

For feature selection (FS), we tried PCA with a varying number of components to keep (30 and 50). We also tried a target-oriented FS method. Specifically, we selected features (either 30 or 50) with those values that best correlated with the target memorability score, using Spearman's rank correlation. Therefore, we tested five different FS strategies (None, Spearman30, Spearman50, PCA30, PCA50). We also considered visual features to see the synergistic impact of the non-visual features. Since the visual features are often in high dimensions, we also tested computing the Radial Basis Function (RBF) kernel of the original features. The features are combined using equal weights. Like in the case of FS, the kernels are computed based on the training split. We also tested vanilla concatenation instead of kernels for comparison.

3. Results

To find the best combination of features, models, feature selection, and the feature preprocessing method, we conducted experiments considering combinations of various features, grouping them based on modalities. Specifically, we considered the following:

- Feature Sets (Individual and Group: 33 sets): HandCrafted (HSVHistogram, RGBHistogram, LBP), DeepVisual (AlexNet, DenseNet121, EfficientNetB3, ResNet50, VGG, ViT), Audio, Temporal (R3D), EmotionSemantic (Emotion, Semantic), etc.

For the full description of feature sets, we refer the readers to the code. A total of 1,650 (33×5 models $\times$ 5 FS methods $\times$ 2 preprocessing methods) combinations were tested. We used 5-fold cross-validation (CV) and applied default parameters of the Scikit-Learn package for all models tested. For MLP, we used a fully connected network with two hidden layers, 256 and 128 dimensions each.

Table 1 shows Spearman's rank correlation, comparing tested combinations with vs. without using each proposed feature. The ablation study reports the mean and standard deviation of all combinations that meet the condition. The results clearly indicate that the audio features are useful in improving the memorability prediction. On the other hand, the emotion features' impact is not significant and the proposed semantic features may even harm the prediction.

Table 2 shows the models we selected for the submission. Models to build each ensemble were selected based on the CV experiments. For Run 1, following the task rule, we only used features that were extractable from the provided devset. Table 3 shows the final results of our participation in this round. The best model was Run 4, which used all DeepVisual features.

Table 1
Effectiveness of Proposed Features (Validation)

Feature	Mean Spearman (w/o)	Mean Spearman (w/)	Diff.	p-value
Audio	0.0053 $\pm$ 0.0558	0.0388 $\pm$ 0.0493	0.0335	0.0000 ***
Emotion	0.0138 $\pm$ 0.0576	0.0159 $\pm$ 0.0525	0.0022	0.4584
Semantic	0.0165 $\pm$ 0.0559	0.0095 $\pm$ 0.0563	-0.0070	0.0203 *
Emotion+Semantic	0.0135 $\pm$ 0.0571	0.0170 $\pm$ 0.0533	0.0035	0.2386

Table 2
Ensemble models used in submissions

Run#	Model#	Features	Model	Selector	Preprocessing
1	1	R3D	SVR	PCA50	Kernel
	2	DenseNet121	GB	None	Concatenation
	3	Emotion+Semantic	GB	Spearman50	Kernel
2	1	R3D	SVR	PCA50	Kernel
	2	DenseNet121	GB	None	Concatenation
	3	Audio	MLP	Spearman50	Concatenation
3	1	All	SVR	None	Kernel
	2	HandCrafted+Audio+Emotion+Semantic	SVR	Spearman50	Kernel
	3	R3D	SVR	PCA50	Kernel
	4	Audio	MLP	Spearman50	Concatenation
	5	HandCrafted+Audio+Emotion+Semantic	Ridge	None	Concatenation
	6	DenseNet121	GB	None	Concatenation
4	1	All	SVR	None	Kernel
	2	HandCrafted+Audio+Emotion+Semantic	SVR	Spearman50	Kernel
	3	R3D	SVR	PCA50	Kernel
	4	Audio	MLP	Spearman50	Concatenation
	5	HandCrafted+Audio+Emotion+Semantic	Ridge	None	Concatenation
	6	DeepVisual+Audio	SVR	Spearman30	Kernel
5	1	All	SVR	None	Kernel
	2	HandCrafted+Audio+Emotion+Semantic	SVR	Spearman50	Kernel
	3	R3D	SVR	PCA50	Kernel
	4	Audio	MLP	Spearman50	Concatenation
	5	HandCrafted+Audio+Emotion+Semantic	Ridge	None	Concatenation
	6	DenseNet121	GB	None	Concatenation
	7	DeepVisual+Audio	SVR	Spearman30	Kernel

Table 3
Final results

	Validation			Test		
Run#	Spearman	Pearson	MSE	Spearman	Pearson	MSE
1	0.186	0.190	0.093	0.073	0.124	0.095
2	0.210	0.208	0.095	0.08	0.096	0.098
3	**0.283**	0.279	0.081	0.16	0.239	0.078
4	0.257	0.265	0.084	**0.225**	0.283	0.078
5	0.278	0.275	0.082	0.161	0.249	0.078

4. Discussion

We found four takeaways from this year's participation: (a) the proposed audio features are useful, particularly when combined with other visual features, (b) emotion features may be useful in combination but the evidence is insignificant, (c) simple textual semantic features are not useful, and (d) precomputed kernels are useful for high-dimensional feature fusion.

Declaration on Generative AI

During the preparation of this working notes paper, the authors used Claude and Gemini in order to: grammar and spelling check.

Acknowledgment

The authors thank Rukiye Savran Kiziltepe for assisting in extracting the proposed audio features from the VIDEM dataset.

References

[1] R. S. Kiziltepe, S. Sahab, R. V. Santana, F. Doctor, K. Paterson, D. Hunstone, A. G. Seco de Herrera, VIDEM: Video effectiveness and memorability dataset, in: I. Rojas, G. Joya, A. Catala (Eds.), Advances in Computational Intelligence, Springer Nature Switzerland, Cham, 2026, pp. 41–54.

[2] I. Martín-Fernández, A. Ganesh, M. G. Constantin, C.-H. Demarty, M. Gil-Martín, S. Halder, B. Ionescu, A. Matran-Fernandez, R. Savran Kiziltepe, A. García Seco de Herrera, Overview of the mediaeval 2026 predicting movie and commercial memorability task, in: Proceedings of the MediaEval 2026 Workshop, Amsterdam, The Netherlands and Online, 2026.

[3] M. G. Constantin, C.-H. Demarty, C. Fosco, S. Halder, G. Healy, B. Ionescu, S. V. Luncanu, I. Martín-Fernández, A. Matran-Fernandez, R. Savran Kiziltepe, A. F. Smeaton, L.-D. Stefan, L. Sweeney, A. García Seco de Herrera, A review of computational memorability: A benchmark framework, International Journal of Computer Vision 134 (2026) 298. doi:10.1007/s11263-026-02880-6.

[4] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, K. Wilson, Cnn architectures for large-scale audio classification, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Press, 2017, p. 131–135. URL: https://doi.org/10.1109/ICASSP.2017.7952132. doi:10.1109/ICASSP.2017.7952132.

[5] S. M. T. Mariappan, M. Ramasamy, B. Arul, Mediaeval 2025 : A multimodal approach for predicting movie and commercial memorability using stacking and gradient boosting, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[6] E. N. Tekay, I. A. Isleyen, M. Karakus, R. S. Kiziltepe, Multimodal feature fusion for video and brand memorability, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[7] A. Pramov, Llm-based fusion of multi-modal features for commercial memorability prediction, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[8] M. Adeel, K. Fatima, M. I. Ayoubi, M. Usmani, M. A. Tahir, Exploring visual, textual, and engagement features for memorability predictions, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[9] P. Balasundaram, D. Parthasarathy, T. Shaw, N. Jahnavi1, P. M, Memorability: Predicting movie and commercial memorability using visual and audio features, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[10] S. Bellman, S. Arismendez, D. Varan, Can muted video advertising be as effective as video advertising with sound?, SN Business & Economics 1 (2021) 1–27. doi:10.1007/s43546-020-00030-9.

[11] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, SciPy 2015 (2015). URL: https://doi.org/10.25080/Majora-7b98e3ed-003. doi:10.25080/Majora-7b98e3ed-003.

[12] M. Mauch, S. Dixon, Pyin: A fundamental frequency estimator using probabilistic threshold distributions, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 659–663. doi:10.1109/ICASSP.2014.6853678.

[13] S. M. Mohammad, P. D. Turney, Crowdsourcing a word-emotion association lexicon, Computational Intelligence 29 (2013) 436–465.