Mid Roll Advertisement Placement Using Multi Modal Emotion Analysis

Sumanu Rawat, Aman Chopra, Siddhartha Singh, and Shobhit Sinha(✉)

Manipal Institute of Technology, Manipal 576104, Karnataka, India
sumanurawat12@gmail.com, amanchopra64@gmail.com, singh.siddhartha23@gmail.com, shobhit.sinha19@gmail.com

Abstract. In recent years, owing to the ever-increasing consumer base of video content over the internet, promoting business via advertising between the videos has become a powerful strategy. Mid roll ads are the video ads that are played between the content of a video being watched by the user. While a lot of research has already been done in the field of analyzing the context of the video to suggest relevant ads, little has been done in the field of effective placement of the ads so that it does not deteriorate users' experience. In this paper, we are proposing a new model to suggest at which particular spot in a video, an advertisement should be placed such that most people will watch more of the ad. This is done using emotion, text, action, audio and video analysis of different scenes of a video under consideration.

Keywords: Advertisement placement · Audio analysis · Emotion analysis · Sentiment analysis · Video processing · LSTM · CNN · Artificial Neural Network

1 Introduction

Videos have an intuitive way of fostering attention. YouTube and Facebook are two famous platforms where video traffic is very high. While YouTube's mobile video consumption rises nearly by 100% every year, Facebook generates an average of 8 billion views every day. Considering the rate at which internet video traffic is increasing, mid roll advertisements are becoming a popular way to propagate business. Apart from the context and quality of the ad, it's placement also plays a crucial role in whether the user will watch and be interested in the ad's content. According to [5], mid roll ads could generate negative attitudinal responses and higher ad avoidance because consumers would likely find such ads more intrusive and irritating if they are not placed appropriately. Hence, the ad placement becomes very crucial.

Automatic mid roll ads might be placed in between scenes, or during a climax scene which severely deteriorates users' experience which in turn affects users' sentiments towards an ad or the brand. Generally, all videos consist of scenes

I. V. Tetko et al. (Eds.): ICANN 2019, LNCS 11730, pp. 159–171, 2019.
https://doi.org/10.1007/978-3-030-30490-4_14

which most of the users are particularly interested in watching. Placing ads near those scenes might increase the chances of that ad being watched by most of the users. It is also of extreme importance to place ads during scene transitions and not between a particular scene to not affect users' experience.

In this paper, we have predicted prospective points inside a video where an ad may be placed to increase the number of users that watch the ads without deteriorating their experience. This is done by analyzing underlying sentiments of different scenes of a video. In this paper, the classification models are trained on text sentiment, text emotion (extracted from speech), video emotion, audio, and the action of different scenes of a video to predict salient positions in an unseen video where a mid roll ad can be placed effectively to achieve the desired results. The paper has been laid out in the following manner. Section 2 deals with the primary objective of this paper. Section 3 talks about the key researches being already done in this field. Section 4 lays down the methodology the paper has followed to solve this problem. Section 5 discusses the case study. Sections 6 and 7 mention the results and conclusion.

2 Objective

This paper has proposed a design for a model which suggests effective markers in a video where an advertisement may be placed. Most of the times identifying where an advertisement should be placed turns out to be a difficult task as manual placement of an ad does not take into consideration the viewing patterns of the target audience. Placing an ad before a particular scene depends on a lot of factors like scene transition, emotion, and sentiment of the upcoming scene. Hence, analyzing these factors and training a model based on these extracted features can suggest places in the video where a suitable contextual ad can be placed without deteriorating the users' experience and encouraging maximum views.

3 Related Work

There have been studies related to object level and contextual Video Advertising. In work [12], the authors have proposed an approach to automatically detect objects that are continuously occurring and then select ads based on their relevance. According to the paper, the selected ads can be inserted at the time when the related objects appear. The drawback is that the related object might appear in between an engrossing scene and placing the ad between the scene might severely affect the users' experience. In work [7] the authors have identified the ads based on the contextual relevance of the video. For ad placement, they have identified insertion points which only depends on frame change and contextual relevance. While these two are significant factors, placing an ad does not just depend on these two factors, it also depends on the significance of the scene in a video which can be predicted using audio, text and frame analysis. In [11], the authors have proposed a dynamic framework to detect the climax of a

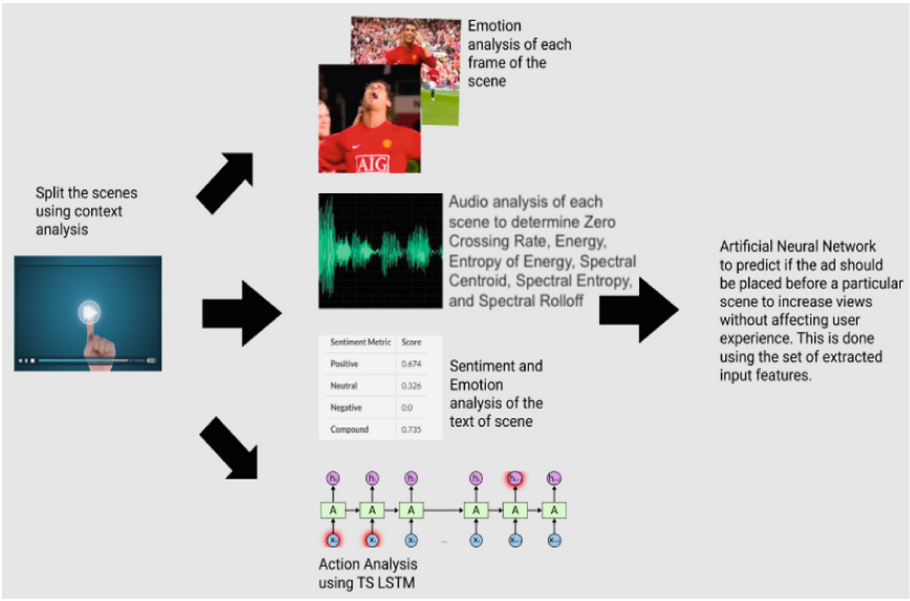

Fig. 1. Proposed methodology

scene for understanding the story in Video Advertisement. In [10], the authors have proposed a Computational effective Video-in-Video Advertising strategy where advertisements are inserted according to established psychological rules, by assessing the emotional impact of program content. This paper, apart from considering emotion analysis in the spatial domain, also considers action analysis in the temporal domain to better train the ANN model for improved results.

4 Methodology

Figure 1 shows the workflow diagram. The video is first divided into different scenes using content analysis. Multiple analysis like audio, video and text are done to generate data as these values play a major role in predicting the importance of each scene. They might indicate whether a scene carries high importance in the video or not. The individual scenes are then passed to an LSTM model for early action detection. All these extracted features are then fed to our classification model which suggests whether an ad should be placed before a particular scene or not.

4.1 Data Collection

A database of 56,435 scenes from 800 videos of different genres like 'Games', 'Art & Entertainment', 'Food & Drink', 'Business and Industrial' and 'Science' was created from the standard YouTube-8M Dataset [1].

4.2 Data Preprocessing

Machine learning datasets usually provide us with massive amounts of data scraped off the Internet. This data needs to be processed and constructed differently according to the needs of a case study. Multiple scenes were extracted from each video using standard python library *pyscenedetect* which splits the video into separate clips using Content-Aware detection. This paper experiments with its parameters like the threshold value to come up with fast, effective and relevant scene splitting. The optimal threshold value was 40. The start and end time of each scene in a video are saved. The start time values of different scenes are the prospective points where an advertisement may be inserted.

4.3 Model Generation

Each scene goes through a series of analysis, the output of which is fed to the classification model which decides ad placement before a particular scene.

Text Analysis. The video clip is first converted into an audio file using FFmpeg. The text is extracted from each audio clip using Google Cloud speech API. VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based open source sentiment analysis tool [4]. Sentiment Analysis is a subfield of Natural Language Processing which classifies a sentence based on its emotional value. Each text phrase is given a sentiment score which ranges from [−1 to 1]. The negative sign in a sentiment score indicates a negative sentiment (sad, angry, gloomy), while the positive scores depict cheerful sentiments. The magnitude of this score measures the intensity of the given sentiment. To extract the context from text both sentiment analysis and text emotion analysis is necessary. While sentiment scores are just an indication of whether a phrase is positive, negative or neutral, the emotion analysis actually helps in categorizing the phrase into different categories of emotions. A CNN Model was trained on the dataset obtained from [8] to categorize the emotions of text in one of the four categories comprising of anger, joy, sadness, and fear. The CNN Model trained gives an accuracy of 93%. Figure 2 shows sentiment scores of some text extracted from an audio file.

Audio Analysis. A standard python library *pyaudioanalysis* is used to extract features from the audio files. We have used the short term feature extraction functionality of *pyaudioanalysis* for audio feature extraction. Short term feature extraction for an audio signal works by splitting each audio clip into frames of 50 ms with a 50% overlap of 25 ms. The library provides us with 6 audio features for each frame of an audio clip namely:

1. **Zero Crossing Rate** - The rate of sign-changes of the signal during the duration of a particular frame [3].
2. **Energy** - The sum of squares of the signal values, normalized by the respective frame length.

	Text	Sentiment score
0	My name is Aman	0.0000
1	Hundreds were killed in tribal violence last m...	-0.8625
2	Tonight's gonna be a good night	0.4404
3	Maria's level of happiness rose to ecstatic wh...	0.7845
4	The lakers lost the game	-0.3182

Fig. 2. Vader sentiment analysis

3. **Entropy of Energy** - The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4. **Spectral Centroid** - The center of gravity of the spectrum.
5. **Spectral Entropy** - Entropy of the normalized spectral energies for a set of sub-frames.
6. **Spectral Rolloff** - The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.

For each audio clip, we calculate the difference between the maximum value of a feature and the minimum value of a feature occurring in any of the 50 ms frames. The change in the energy level of a clip gives us a sense of high randomness or a 'burst' of excitement within a particular scene. The user experience would deteriorate if we would try to place an advertisement during this scene. This delta value is then scaled between 0 and 1 using the minmaxscaler of the scikit-learn preprocessing library. Scaling is done on per video basis as opposed to scaling on the whole dataset at once. Since the absolute value of these features may heavily vary with each video, our per video normalization provides a degree of evenness to the dataset. Every video contains scenes with a relatively high point and a relatively low point, which makes this a significant decision in our research and hence showed a good improvement in the results. Figure 3 shows Zero Crossing Rate (zcr) and Energy of different frames of a particular audio file.

Video Analysis. Emotion analysis of the video segment is an important attribute. To calculate the combined emotion of a scene extracted by *pyscenedetect*, we first split each scene into multiple frames which is essentially an image

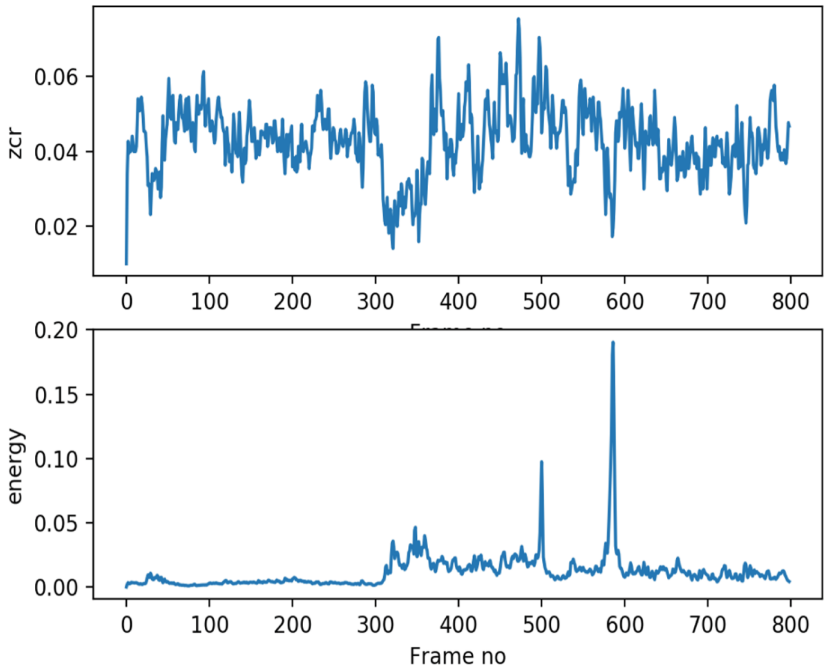

Fig. 3. ZCR, energy graph

using OpenCV. For emotion analysis, the paper has made use of Microsoft Azure cognitive emotion API. It identifies the face in the image and detects the emotion. The emotions detected are [anger, contempt, disgust, happiness, fear, neutral, sadness, and surprise]. These emotions are understood to be universally communicated with particular facial expressions. The emotion is neutral if the face is not detected in a particular frame. Now, to calculate the combined emotion of a particular scene, we take the emotion category with the highest confidence from each frame and label the scene with the category that occurred the most number of times in that scene. In case of a tie, our framework considers positive emotion. Since the interval of a particular scene is very small (7–12 s), one small scene will depict one emotion in most of the cases.

Action Analysis. The paper has taken a lot of spatial features into account using text, audio and image analysis. These features treat each scene as independent entities and have no relation to what happened before the scene took place. To detect the importance of a particular scene, features in the temporal domain are as important. Hence, the paper suggests action analysis which is used as an input attribute to the classification model. The extracted scenes are very small and therefore early detection of action is very important. The paper follows the methodology suggested in [2] to use a multi-stage LSTM Model for early action detection. This model gives an accuracy of 80.1% given only the first

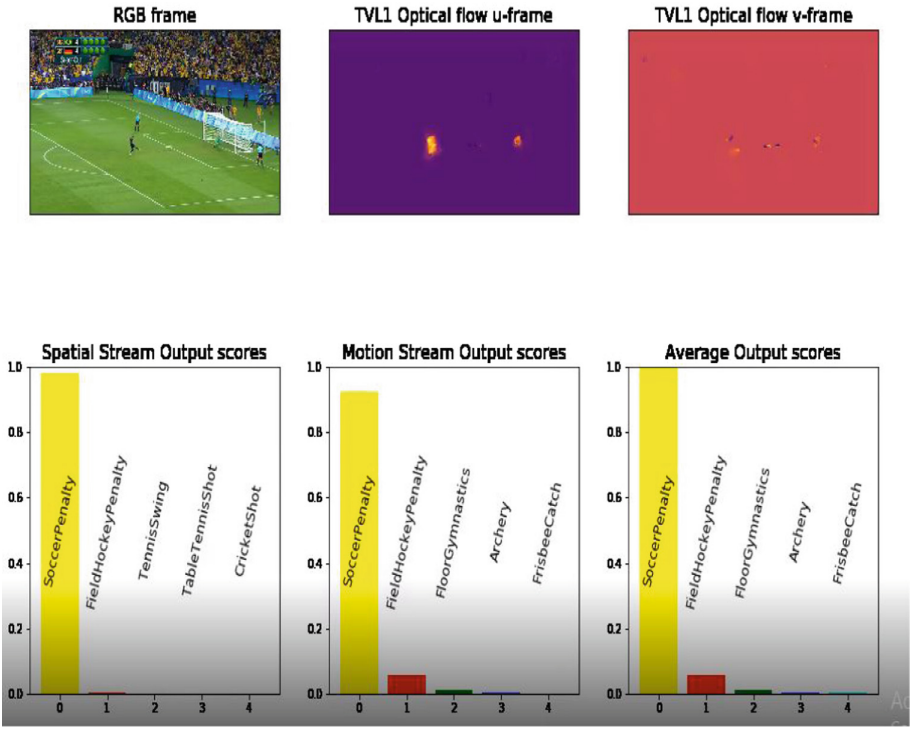

Fig. 4. LSTM Model detecting a soccer penalty

1% of the video. The LSTM model under consideration is trained over UCF101 action Recognition dataset which is a collection of 101 actions and it predicts one of these actions when the clip is fed into multi-stage recurrent architecture based on LSTMs. If the action is not present in the 101 actions on which the model is initially trained, it outputs the nearest action. These actions are then categorized into 5 categories comprising of [Human-Object Interaction, Body-Motion Only, Human-Human Interaction, Playing Musical Instruments, Sports] which are labeled from 1 to 5. This categorical numerical attribute of action is used as an input feature to the classification model.

Figure 4 shows the LSTM Model detecting a soccer penalty.

Min-Max Scaling. Before feeding the data into a classification algorithm, it is important to normalize the numeric features to prevent unnecessary imbalance in the data while training. The paper has used Min-max scaling as shown in Eq. 1, which scales a range of values down, such as their values vary from 0 to 1. The paper has applied min max scaling formula on the features namely text sentiment, Zero Crossing Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Entropy, Spectral Rolloff, Text_Emotion, Scene_Emotion, and Action.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

One Hot Encoding. Since the problem is a classification problem, the number of output columns should be equal to the desired number of output classes which is 2 as it is a binary classification problem. One hot encoding the output label column gives an array of 2 columns. Each of these columns denotes an output class. Whichever class a particular row belongs to, is assigned a value of 1 and the others are kept as 0.

Annotations. After extracting 56,435 scenes from the YouTube-8M dataset, it was necessary to annotate the videos with prospect ad insertion points to get the output in order to train and validate the classification model. Using *pyscenedetect*, we already got the start time of each clip of a video where an ad might be inserted, but to train the classification model, we had to find the points where the ads will not deteriorate the users' experience. We made use of the Amazon Mechanical Turk platform for this purpose. User psychology and their attitude towards advertisement are subjective and it also depends upon the background of the user, but using Amazon Mechanical Turk platform, we were able to give the task to annotators of different age groups and diverse location. We encouraged them to use VATIC (Video Annotation Tool from Irvine, California) [9]. We restricted participation on our tasks to annotators with at least 90% approval rate who submitted a minimum of 500 approved tasks in the past. We also added some special qualifications for the annotators, for instance, annotators should have some amount of knowledge and interest in the genre of the video they are annotating like 'Games', 'Science', 'Art & Entertainment' etc. We submitted each video for annotation to three workers satisfying the qualification. Using these qualifications, we tried to make sure that the model is generalized. A total of 56 annotators helped the study. Each was asked to watch the video and annotate the clips of the particular video where they won't mind the ad to be inserted. To ensure quality, annotators were also asked to describe what happens at the end of the video. We manually inspected a subset of them and found that the timestamps were reasonable.

Table 1 shows the accuracy of different algorithms used for model generation

The dataset was then shuffled with a random seed value and 25% of the dataset was used as the testing set. We split the remaining videos into one fifth (15%) for validation and four-fifths (60%) for training.

Table 1. Accuracy of algorithms

Algorithms	Accuracy
VADER Sentiment Analysis	96%
CNN Text Emotion Analysis	93%
Early Action detection using LSTM	80.1%

4.4 Classification Algorithms:

For this problem statement, the classification models predict whether an ad should be inserted before a scene of the video or not. The model is trained with 10 features i.e., Zero Crossing Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Entropy, Spectral Rolloff, Text_Sentiment, Text_Emotion, Scene_Emotion, Action and then the trained model is used for classification. We have used a comparative analysis of different classification models to come up with a model with the highest accuracy.

Logistic Regression. Logistic regression is used when the result expected is categorical in nature. In this paper, a special case of Logistic Regression, which is used to classify the result into two classes is used. This is called Binomial Logistic Regression. The main idea behind the Logistic Regression used is to establish a relationship between the attributes of the dataset and the probable outcome of the classes. In this paper, Binomial Logistic Regression uses a series of Bernoulli trials (pass, fail) to classify the start time of the clips as prospective ad insertion points. Our model of logistic regression makes use of L1 regularization in an attempt to introduce additional information that proves to be useful in the prevention of over-fitting to some extent. The equation to be solved by L1 regularized logistic regression is shown in Eq. 2.

The simplicity of logistic regression sets a bar, to begin with, and to compare the results of other classification algorithms used in the paper, namely, Support Vector Machine and Artificial Neural Networks.

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} log(exp(-y_i(X_i^T w + c)) + 1) \tag{2}$$

Support Vector Machine. Due to a limited set of points in many dimensions, Support Vector Machine tends to be very good because it is able to find the linear separation that should exist. It also performs well with outliers as it will only use the most relevant points to find a linear separation (support vectors). In this project, we have used SVM with Radial Basis Function (RBF) kernel to find the prospect ad insertion points and compare its accuracy with baseline models like Logistic Regression and Complex models like Artificial Neural Networks.

Artificial Neural Networks. For this problem statement, the artificial neural network is predicting whether an ad should be inserted before a scene of the video or not. The artificial neural network is trained with 10 features i.e., Zero Crossing Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Entropy, Spectral Rolloff, Text_Sentiment, Text_Emotion, Scene_Emotion, Action and then the trained model is used for classification. Apart from the input and output layers, this model incorporates 2 hidden layers of 10 and 8 nodes respectively. The back-propagation algorithm is used to train the feedforward Multi-Layer Perceptrons to minimize the value of Categorical Cross Entropy loss function. As mentioned

above, the dataset was divided into three sets. 60% of the data was used for training, 25% was used for testing and the remaining 15% was used for validation. After some experimentation, the number of iterations till convergence was found to be 47.

Transfer Function: Perceptrons perform the sigmoid function on output data as an activation function before passing the result on to the next layer. The Sigmoid function was chosen over the popularly used Rectified Linear Unit (ReLU) to maintain the relevance of negative sentiments. The output layer is subjected to the Softmax activation function because it strengthens the winning class and weighs down the losing one. It selects the output class with a higher probability.

Optimizing Algorithm: The optimization algorithm that showed promising results for the problem under consideration was the incorporation of Nesterov Momentum into Adam optimizing algorithm. The paper uses Keras's implementation of the 'nadam' optimizer for this purpose.

Regularization: It penalizes the weight matrices of the nodes of Artificial Neural Network. Here, we have used L1 Regularization to add a regularization term to update the general cost function. A drop out layer with $p = 0.45$ was also added. ANN are complex models and hence these steps become necessary to avoid over-fitting.

5 Case Study

The dataset used in this study consists of 56,435 scenes extracted from 800 videos of different genres from the standard YouTube-8M Dataset. The videos are split into scenes using *pyscenedetect* based on content analysis, fade in/out and fast cut detection of each shot. The Conditional attributes are extracted for each scene and are shown in Table 2. The Zero Crossing Rate, Energy, Entropy of Energy, Spectral Centroid, Spectral Entropy, Spectral Rolloff are calculated using audio analysis of each scene. The text sentiment analysis is done after extracting the text from speech and applying Vader sentiment analysis to get the sentiment score. The text emotion is calculated using a Convolution Neural Network model pre-trained on the dataset obtained from [8]. The action of the clip is identified using LSTM and the video emotion is categorized after splitting the scene into frames and passing each frame to Microsoft Azure's cognitive services to detect the emotion of a scene after combining the emotions of all the frames of that scene. Every scene also contains a video identifier from which it has been extracted along with start and end time of the scene. The advertisement will be placed at the start time of a particular scene predicted by the model.

Classification models are then trained on these features after feature normalization. After training, the model does a binary classification as to whether the advertisement should be placed before a particular scene or not with the aim of maximizing views without disturbing the users' experience.

Table 2. Dataset description

Conditional attributes	
Attributes	Description
Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame
Energy	The sum of squares of the signal values, normalized by the respective frame length
Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes
Spectral Centroid	The center of gravity of the spectrum
Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames
Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated
Text_Sentiment	Vader sentiment score of the text of the scene
Text_Emotion	Classifies the text of the scene into four predefined emotions
Scene_Emotion	Classifies the frames of scene into eight predefined emotions
Action	Early Actions predicted by LSTM which are segregated into five categories

Table 3. Accuracy of classification algorithms

Algorithms	Accuracy (Rounded off to first decimal place)
Logistic Regression	72.8%
Support Vector Machine	82.2%
ANN	86.5%

6 Results

To evaluate the proposed model, the research collects 56,435 scenes from 800 videos of YouTube-8M Dataset. After training the models on 33,861 scenes from 480 videos of varied genres, the models were tested and validated. Table 3 shows the accuracy of different algorithms used for binary classification.

After comparative analysis, it was found that ANN with 2 hidden layers of 10 and 8 nodes respectively, sigmoid transfer function, 'nadam' optimizer, L1 regularisation and a dropout layer with $p = 0.45$ performed the best with an accuracy of 86.5%. Since it is a binary classification problem, to further validate our model, we also calculated the F1 score which is the harmonic mean of the precision and recall. The F1 score was calculated to be 0.885.

After carefully examining the insertion points suggested by the model on random videos, it was found that most of the points were after some climax positions

and not towards the end of the videos which is in sync with a recent survey conducted by Facebook in which most people lose interest towards the end of the video, hence making it a not so effective position for ad insertion. Hence, the prospect ad insertion points to maintain a balance between ad completion rate and users' experience is after climatic positions (the highest dramatic tension or a major turning point in the video) near the middle of the videos.

In [10], the authors use CAVVA (Computational Affective Video-in-Video Advertising) to show a subjective user-study according to theories from marketing and consumer psychology and our results of prospect ad insertion points are in sync with the study. We also compared our results with the research done in the field of the Effectiveness of Video Ads [6] and our prospect insertion points in mid roll ads justify the result in that study. Most of the work done in this field is survey based and theoretical. We have proposed a new model to automatically place ads in a video which would not disturb users' experience as well as increase the ad completion rates.

7 Conclusion and Future Work

This paper has proposed a framework using audio, video and text processing along with early action analysis using LSTM, and Artificial Neural Network to automatically suggest mid roll advertisement insertion points. Considering the amount of traffic that online videos attract, this framework will help in establishing the right balance between video marketing and users' experience while watching the video. Inappropriate placement of ads can deteriorate a users' experience as well as their sentiments towards the ad even though the quality of the ad is up to the mark. Hence the placement of an ad is equally or even more important than the contextual relevance of an ad. This framework will hence promote automatic mid roll ad placement rather than manually analyzing and placing them.

In the future, the authors want to apply BiLSTM for early action detection and analysis as it might provide more accurate results. The research will try to incorporate more features from the temporal domain to increase the applicability of the model. Due to computational limitations, the amount of data is less but in future, an attempt will also be made to increase the dataset further and to validate the model on a larger dataset of videos with more varying genres.

References

1. Abu-El-Haija, S., et al.: YouTube-8M: a large-scale video classification benchmark. CoRR abs/1609.08675 (2016). http://arxiv.org/abs/1609.08675
2. Akbarian, M.S.A., Saleh, F., Salzmann, M., Fernando, B., Petersson, L., Andersson, L.: Encouraging LSTMs to anticipate actions very early. CoRR abs/1703.07023 (2017). http://arxiv.org/abs/1703.07023
3. Giannakopoulos, T.: pyAudioAnalysis: an open-source Python library for audio signal analysis. PLoS ONE **10**, e0144610 (2015). https://doi.org/10.1371/journal.pone.0144610

4. Hutto, C., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text (2014). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109
5. Kim, S.: Effects of ad-video similarity, ad location, and user control option on ad avoidance and advertiser-intended outcomes of online video ads (2015). http://hdl.handle.net/11299/175210
6. Krishnan, S.S., Sitaraman, R.K.: Understanding the effectiveness of video ads: a measurement study. In: Proceedings of the 2013 Conference on Internet Measurement Conference, IMC 2013, pp. 149–162. ACM, New York (2013). https://doi.org/10.1145/2504730.2504748
7. Madhok, R., Mujumdar, S., Gupta, N., Mehta, S: Semantic understanding for contextual in-video advertising, April 2018
8. Mohammad, S.M., Bravo-Marquez, F.: WASSA-2017 shared task on emotion intensity. CoRR abs/1708.03700 (2017). http://arxiv.org/abs/1708.03700
9. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. Int. J. Comput. Vis. 1–21. https://doi.org/10.1007/s11263-012-0564-1
10. Yadati, K., Katti, H., Kankanhalli, M.S.: CAVVA: computational affective video-in-video advertising. IEEE Trans. Multimedia **16**, 15–23 (2014)
11. Ye, K., Buettner, K., Kovashka, A.: Story understanding in video advertisements, September 2018
12. Zhang, H., Cao, X., Ho, J.K.L., Chow, T.W.S.: Object-level video advertising: an optimization framework. IEEE Trans. Ind. Inform. **13**(2), 520–531 (2017). https://doi.org/10.1109/TII.2016.2605629