IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 January 2023, accepted 8 February 2023, date of publication 15 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3245982

RESEARCH ARTICLE

CLEFT: Contextualised Unified Learning of User Engagement in Video Lectures With Feedback

SUJIT ROY[1], VISHAL GAUR[1], HAIDER RAZA[2], (Senior Member, IEEE), AND SHOAIB JAMEEL[3]

[1]Brainalive Research Pvt. Ltd., Kanpur 208001, India
[2]School of Computer Science and Electronics Engineering, University of Essex, CO4 3SQ Colchester, U.K.
[3]School of Electronics and Computer Science, University of Southampton, SO17 1BJ Southampton, U.K.

Corresponding author: Sujit Roy (sujit@braina.live)

This work was supported by Brainalive Research Pvt. Ltd., Kanpur, India (http://braina.live/). The work of Haider Raza was supported by the Economic and Social Research Council (ESRC) funded by the Business and Local Government Data Research Centre under Grant ES/S007156/1.

ABSTRACT Predicting contextualised engagement in videos is a long-standing problem that has been popularly attempted by exploiting the number of views or *likes* using different computational methods. The recent decade has seen a boom in online learning resources, and during the pandemic, there has been an exponential rise of online teaching videos without much quality control. As a result, we are faced with two key challenges. First, how to decide which lecture videos are engaging to intrigue the listener and increase productivity, and second, how to automatically provide feedback to the content creator using which they could improve the content. The quality of the content could be improved if the creators could automatically get constructive feedback on their content. On the other hand, there has been a steep rise in developing computational methods to predict a user engagement score. In this paper, we have proposed a new unified model, CLEFT, that means ''Contextualised unified Learning of user Engagement in video lectures with Feedback'' that learns from the features extracted from freely available public online teaching videos and provides feedback on the video along with the user engagement score. Given the complexity of the task, our unified framework employs different pre-trained models working together as an ensemble of classifiers. Our model exploits a range of multi-modal features to model the complexity of language, context agnostic information, textual emotion of the delivered content, animation, speaker's pitch, and speech emotions. Our results support hypothesis that proposed model can detect engagement reliably and the feedback component gives useful insights to the content creator to further help improve the content.

INDEX TERMS NLP, emotions, video engagement, contextual language models, text-based emotions, BERT.

I. INTRODUCTION

The ongoing pandemic has resulted in various teaching and research organizations resorting to online lectures. What is predominantly seen today is that different academicians across the globe are creating teaching materials and have started to share them online with students as Open Educational Resources (OERs) [1] such as Massive Online Open Courses (MOOCs) to boost online learning. Users are now overwhelmed by the amount of data, for instance, on YouTube alone, searching for ''Deep Learning Lectures'' retrieves hundreds of results ranging from content created by various individuals to organisations worldwide. The ideal option for any user is to select the one that they could engage with reliably. While beliefs and biases [2] do surround the choice of the videos, e.g., videos from a popular academician in a well-known organisation could be regarded as more interesting than others, this may not always be true. Simply relying on user ratings or modelling text extracted from the video recordings alone might not lead to desirable results because these features do not capture the overall inherent quality of the content, e.g., whether the creator is

The associate editor coordinating the review of this manuscript and approving it for publication was Sunil Karamchandani.

explaining the concepts via videos or animations or even role-plays.

Utilising user feedback (e.g., the video is too one-sided or monotonous) is becoming an important and time-sensitive challenge for successfully leveraging intelligent and user-centric systems in various applications. In the domain of online educational resources, it is imperative to provide timely feedback of user engagement over a population. The feedback can not only make it easier for content creators to create suitable videos as per the target audience, but also it will be more effective in the online teaching tools. There has been growing interest in the domain of contextualised engagement in OER [3]. We argue that only a context agnostic model [3] cannot provide true user engagement, instead, an ideal model is which utilises the features of a video lecture extensively and gives feedback to the content creator. Additionally, we develop an approach, that can assimilate the information to improve the engagement with the target population by providing automatic feedback to the content creator. This is a crucial component because such automatic feedback can help the creator understand the key shortcomings in the content, e.g., whether the tone is too monotonous. As a result, the creator can improve the content via an engaging voice to keep the users engaged.

Following [3], we define engagement as how much intriguing a resource is with respect to the context of the learner. Here context could mean learning needs, goals of the learner, among others.

The problem of automatically studying engagement is important because educators can create content that will optimize the engagement levels on their content. Besides, manual techniques, such as employing an army of domain-expert volunteers to view long videos and providing feedback is too time-consuming. In our model, the automated system provides feedback to an educator that the engagement quality of the teaching content is not faithful. While it can be argued that personalisation is more appropriate in such problem scenarios, personalisation exploits users' historical data and many users might not be willing to share their data such as their location, their click-through patterns, among others. Therefore, works such as [3] have largely focused on non-personalised prediction problems. What is interesting in our work is that we provide feedback to the content creator to help understand which areas they need to improve, e.g., should they add more animations. One key advantage of our model is that it is crucial to a user who might end up spending plenty of time searching for ideal content suited to their learning behaviour, e.g., some users prefer more animations in their videos.

In recent years, there has been a growing interest in the contextualised engagement [4]. In this modelling paradigm, we model the extracted text from the videos and likes associated with videos in a unified way. We argue that there are other key features in the videos, which can increase or decrease user engagement. The categorical model assumes that there are discrete emotional categories such as Ekman's six basic emotions – "anger," "disgust," "fear," "joy," "sadness," and "surprise" [5]. Emotion recognition has been widely studied [6], [7], [8], [9], [10], [11]. In our model, we have modelled the emotional variation of the speaker to time. There have been previous attempts at predicting emotions from speech [12], [13], [14], [15], [16], [17] where the average classification accuracy of speaker-independent speech emotion recognition systems is less than 70% in most of these proposed techniques. For example, it is 50% in [14], 67% in [15], 80% in [16], and 65% in [17]. We improve upon these existing methods and develop our speaker-independent emotion recognition model unified with an ensemble of learners to model the user engagement problem reliably. Another key advantage is that our model helps automatically give feedback to the content creator.

Given the extensive and discrete number of educational materials, new and automated ways are being devised, where mostly cascaded models are used to predict the engagement score [3]. Regarding the OERs, this means to scale down the learners' efforts of finding the material without compromising its quality. Such objectives are usually accomplished after studying the personalization factor [18] that is defined as contextual engagement, which determines the extent of learners' context about a particular learning source. Automatically learning user engagement has several direct benefits, for instance, online search systems can exploit them to retrieve content that is not only relevant but also engaging. This can have an impact on the overall productivity of the user base because they can quickly find something that can intrigue them.

We also hypothesise that engagement cannot be directly exploited using text alone in lecture videos as it has been done in the past work [3]. There is various complementary information that we could exploit from online lecture videos such as speaker intonation, speaker's use of animations, emotions, among others that have not been explored in prior works. We develop a novel unified framework that goes beyond the current techniques that measure user engagement. We expect that over time, our research would have a significant impact in the education domain, where users could find a plethora of engaging free materials online. It will lead to a significant positive step towards achieving the global development goals because if the quality of teaching and learning is improved without increasing the cost of delivering them, we will see several people being educated especially in the developing world. To the best of our knowledge, this is the first work in the area of contextualized engagement, where we exploit other features beyond just text extracted from video lectures. We also publicly share our dataset to further the research in this domain because previous methods have not made their data publicly available.

This work answers the following key research questions (RQ):

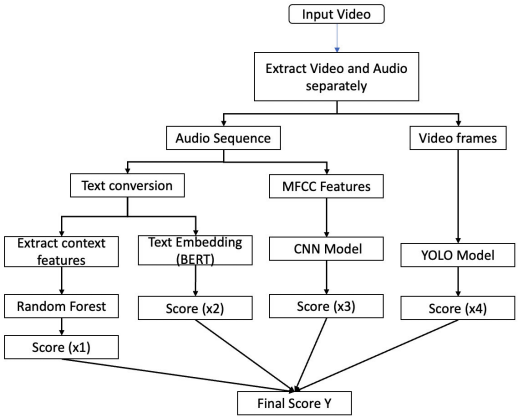

FIGURE 1. Flowchart of our unified CLEFT model for making a prediction over a leacture video.

- RQ1: What are different features that play a key role in automatically modelling user engagement in video lectures?
- RQ2: Can we develop a model that could exploit different pre-trained machine learning models to measure user engagement and which work end-to-end?
- RQ3: Given that videos occupy a large amount of disk space and can produce a large amount of (sometimes redundant) features, can we develop a model that is capable of reliably predicting when the number of videos is small?

Our key contributions include:

- A novel unified model that learns to fine-tune its parameters by exploiting the predictive error of several pre-trained models in an ensemble setup.
- The model not only predicts an engagement score but also provides feedback to the content creator.
- We have conducted experiments on publicly available free videos and the dataset will be shared with the community given the lack of freely available public benchmark datasets in this problem domain.
- Our model can reliably predict under settings when we have fewer data which is very crucial when we model a range of videos in different languages.

There are several key innovations in our work that make our model novel. For instance, unlike traditional machine learning models, our architecture allows for automatic fine-tuning the model as training is underway. The fine-tuned model to automatically adjust the parameters towards a more reliable trained model. we have used different existing machine learning models in our framework that aid towards the overall predictive performance. These independent models are trained in a unified way where the key challenge lies in the way they are jointly trained. For instance, our model automatically ensures that we do not over-train the contextual language model so that its pre-trained parameters are corrupted. In our framework, it is because of the different machine learning framework, we are able to make our model handle multi-modality that is present in the data.

II. RELATED WORK

Bulathwela et al. in [3] developed a technique to learn text feature vectors from video lectures. They obtained a proprietary dataset from VideoLecture.net, which is not publicly available at the time of writing this manuscript. They extracted different features from the videos in this dataset and trained a regression model to determine the overall engagement score. They have defined engagement as a loaded concept that can have different definitions to different communities. For example, engagement is measured using different metrics depending on the modality of the educational resource. While the work is closely related to ours, there are a few key differences. We have developed a novel unified approach to model various multi-modalities present in the videos unlike [3]. We have argued that only textual features might not suffice when predicting the overall engagement score and we demonstrate that using a range of multi-modal features helps improve the performance of the model. The authors in [3] focus on modelling educational engagement to model engagement as a function of the context of the learner. Our work and [3] model context-agnostic engagement through several content-based features. Zhou et al. in [19] developed a novel unsupervised model sequence mining and information retrieval coupled with a clustering algorithm to extract engagement patterns of learners. Their goal is to mainly extract consistent patterns in learning behaviour. As a result, this work is fundamentally different from ours, where our goal is not to learn consistency levels.

Chen et al. in [20] have developed a technique to model automated disengagement that detects learners' maladaptive behaviours, e.g. mind-wandering and impetuous responding. While their work does not develop a novel computational model as ours, our framework is different from this work where our goal is to model user engagement and provide feedback to the user. Recently, Bulathwela et al. in [18] developed a novel recommendation framework [21] that considers several background information of a user, for instance, learner's knowledge of the topic. While they extend their prior work in [3], their main focus in this work is to develop a recommender system for predicting the engagement score. In Bulathwela et al. [22], the authors have created a new dataset of video lectures, which is not publicly available yet. The authors have also developed a novel tool [23] that aids users in finding appropriate videos for learning.

We have exploited different features, which could help to improve the engagement score of our model, for instance, we have exploited emotions from text and speech to model engagement. We have also exploited object detection techniques to further improve our results. As a result, our model can learn from various key representative features which are crucial towards determining the overall engagement score including modelling explainability. For instance, by modelling different objects in the videos during teaching, we can model that the teacher is using different teaching

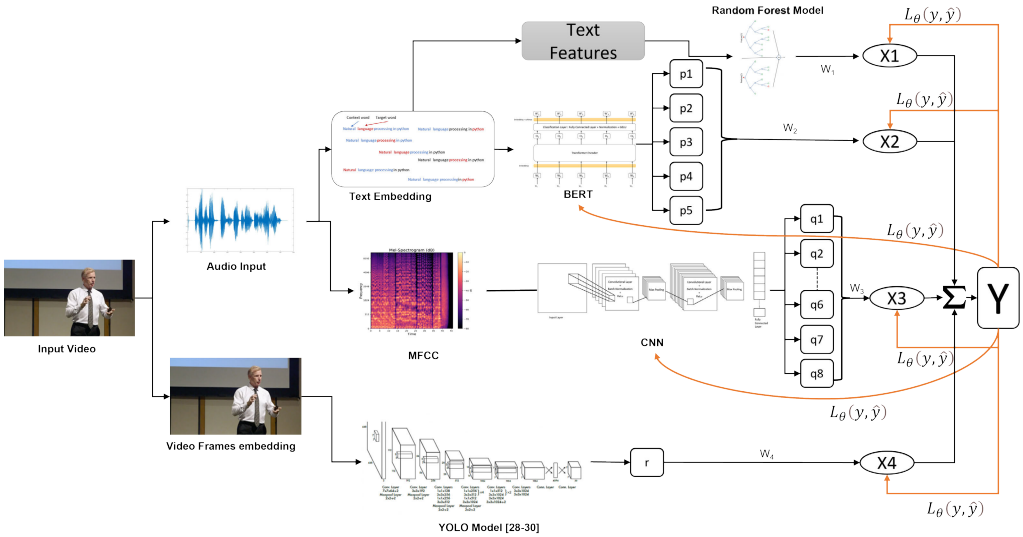

FIGURE 2. Architecture diagram of our unified CLEFT model (RQ2). Orange lines indicate that the model is fine-tuned iteratively while training is underway.

methods. The importance of emotions has been highlighted in many other works, for instance, Nias et al. studied the emotional aspects of teachers in the UK. The authors conclude that the emotions of the instructor play a vital role in teaching since it helps improve engagement of the subject. Recently, in [24], Jimenez et al. studied the important question, "What is the nature of preservice teachers' emotions throughout their engagement in the sequence?" They concluded that "emotions in science education as they illustrate the importance of providing preservice teachers with opportunities to explore their emotions especially about self-regulation when engaging in teaching sequences in teacher preparation."

III. CLEFT: OUR UNIFIED ENGAGEMENT SCORE MODEL

A. MODEL OVERVIEW

Our full end-to-end prediction framework is presented in Figure 1. To address RQ1 mentioned above, we study the role played by different features in our model. We first begin with publicly available lecture videos that are obtained from various sources. Further, we process the video and extract the audio segment and video segment of the same length from the video. Post extraction of audio, we used the same audio to extract text features as well as MFCC features. For extracting text-based features we have used IBM Watson-based speech-to-text method to create a corpus of text words from the video. Further, we did pre-processing of text to extract features listed in Table 1. After the extracted features from the text, we pass it through the trained random forest model for the prediction of the score. Based on the text features we also used BERT for emotion classification to produce the probability of each emotion. After extraction of MFCC features, we passed the input of MFCC features through our trained CNN model for the probability score of emotions based on speech. The extracted video is processed frame by frame by giving it as an input to the YOLO network for the detection of objects. We cumulate all the scores and produce one final score as a video engagement score. A detailed description and training of each model involved in this process are discussed in section III-B followed by each individual model in C, D, E, and F. The section also covers the training mechanism as well as the process for optimizing the errors.

B. DETAILED MODEL DESIGN

In this section, we describe our complete framework, which models the engagement score along with providing feedback to help the content creator improve their videos. The engagement score measures how likely the user will engage with the content. The feedback provided by the model will help the content creator in understanding the key insights about the content. The design principle of our model is to exploit the advantages of different existing pre-trained complementary models. These models are combined, as an ensemble of models, working as a unified machine learning model where they make predictions jointly. This design paradigm gives us a direct advantage that we can model multi-modal features using the most suitable computational model for that feature type, for instance, frame representations can most ideally be learned using a Convolutional Neural Network (CNN) model than a random forest model which is most ideally designed to model text in our problem setup, and subsequently making predictions in a unified way. While all individual pre-trained models play a key role in the overall predictive performance, some can be further fine-tuned based on the data characteristic leading to more faithful results, for instance, the pre-trained BERT [25] model can be fine-tuned given the data than using the original fine-tuned BERT alone. Similarly, we can fine-tune the CNN

model on our data in an iterative way. As a result, we exploit the key advantage of transfer learning.

We model emotion-based features from text and speech followed by object detection. Emotion in teaching has received increased attention in literature [26]. Emotions in the classroom are not only a private matter but also a political space in which students and teachers interact with implications in larger political and cultural struggles [27]. Besides, automatic object detection can help the model understand what else a teacher uses while teaching, e.g., are there some classroom activities organised. In our model, these complementary models contribute towards our overall goal of engagement prediction and feedback. To extract and learn different features from publicly available datasets, we have used a variety of state-of-the-art models which we apply in the predictive analysis. Our overall framework comprises of, 1) text-based context agnostics, 2) text-based emotions, 3) speech-based emotions, and 4) object detection. There are certain key tasks that we need to do, for instance, extraction of features, learning those features using relevant models and using our unified machinery on these individual complementary pre-trained models to derive their weights leading to a prediction score.

In our novel modelling architecture depicted in Figure 2, from the text transcript, we predict $X1$ by random forest and $X2$ by BERT model, where $p1$, $p2$, $p3$, $p4$, $p5$ is the output of text-based emotion. Using the audio feature we predict $X3$ based on the probability of $q1, q2, \ldots, q8$ which represents speech-based emotion. Finally, $X4$ is the count of objects that appeared in the video. We then combine these outputs from models and predict Y where the parameter training of our model and the fine-tuning of the pre-trained BERT and CNN models take place simultaneously based on $\hat{Y}$.

We have used four different pre-trained models, giving complementary advantages, to learn from different features. In our end-to-end framework, CNN and BERT models can be simultaneously fine-tuned iteratively while the unified model training is underway. In fine-tuning, we freeze relevant layers and fine-tune only specific layers which are needed for our task, for instance, in the pre-trained text language model, we only fine-tune the contextual layers, mainly, layer 12. This has been commonly done in the literature. Our framework is depicted in Figure 2 where we extract audio from video, and audio extraction of the speech to text is performed using the IBM Watson speech to text platform. After a speech to text, we have extracted 13 features based on their continued use in studies [3], [28], [29], [30], [31].

Table 1 depicts these extracted features, and apart from these features, we also trained the model to extract emotions from the text data. For training, the Emotion ISEAR, DAILYDIALOG, and KAGGLE Datasets were used. The model was trained on 27,261 sentences considering five classes, namely, "Joy," "Sad," "Fear," "Anger," "Neutral." For emotion classification, the BERT model has been used. Mathematically, our model formulation is shown in Equation 1 where we linearly combine (convex combination) different models using four parameters. While these parameters could be arbitrarily assigned or given the same weights, we have trained these model parameters based on the data. To this end, we have used the backpropagation model to update these model parameters in each iteration, and simultaneously fine-tuned the individual models.

C. TEXT BASED CONTEXT AGNOSTICS

We have used VideoLecturesNet (VLN) dataset [3] which has been categorized into 21 different subjects, such as Computer Science, Physics, Philosophy, etc. We extracted the context engagement model, which are, 'duration', 'conjugate rate', 'normalization rate', 'tobe verb rate', 'auxiliary rate', 'preposition rate', 'pronoun rate', 'document entropy', 'easiness', 'fraction stopword coverage', 'fraction stopword presence', 'title word count', 'word count', 'speaker speed', 'median engagement rate'. The subset of cross-modal and language-based features were selected from the VLN dataset. The 14 extracted features can be seen in Table 1. The description of the features is also shown in table 1 which are our novel features to suit our problem task.

We split the data in 67% for training and 33% for the test. The random forest regressor was used to train as the model with median engagement rate as a prediction variable ranging from 0-1 and was stored as X1 in CLEFT. This model gave us better results in our case than the support vector-based model. Mean squared error was calculated from the predicted variable. The median engagement rate was calculated based on user feedback, star rating of videos, number of views, and likes for the videos.

D. TEXT BASED EMOTIONS

For evaluating emotions in a text we have used the International Survey on Emotion Antecedents and Reactions (ISEAR) and Dailydialogue publicly available datasets. The ISEAR dataset [32] contains the emotional statement that helped us to further train the model using the textual data. It contains 7666 sentences which are further divided into 7 emotional categories, where we have only considered 5, i.e. "joy," "sadness," "anger," "fear," and "neutral". We have utilized the broadly classified emotion which can be identified by the text as well as speech for addressing its influence over attention. Several studies suggest that emotional events are remembered clearly and accurately for a longer period of time [33], [34], [35], [36], [37]. Additionally, emotional context is of higher importance when we try to tell a story.

The dailydialogue dataset [38] is a high-quality multi-turn dialog dataset. The dataset is constituted of human written statements which makes them less noisy. The dataset contains information that reflects our daily conversations and consists of a variety of topics about our daily lives. The dataset is also manually annotated in similar emotional categories to the aforementioned datasets. This dataset contains a total of 13,118 dialogues which are then split into 11,118 training dialogues and 1000 dialogues of validation and test set each.

TABLE 1. Extracted features from the VLN data-set.

Feature	
Conjugate_rate	Count of conjunctions used by the speaker in a particular video
Pronoun_rate	Count of pronouns used by the speaker in a particular video.
preposition_rate	Count of prepositions used by the speaker in a particular video.
tobe_verb_rate	Count of times the speaker used (“be”, “being”, “was”, “were”, “been”, “are”, “is”) in a video lecture.
auxiliary_rate	Count of times speakers used auxiliary verbs in a video.
normalization_rate	Count of words used by the speaker which were ended with suffixes (“tion”, “ment”, “ence”, “ance”)
fraction_stopword_coverage	It is the ratio of stopwords used by the speaker to the total stopwords
Fraction_stopword_presence	It is the ratio of stopwords used by the speaker to the total number of words used by the speaker in a particular video
Easiness	It will give the readability level of the text of a video.
Document_entrophy	Entropy of words for a particular video.
word_count	Total count of words used by a speaker in a particular video.
title_word_count	Count of title words for a particular video.
Duration	Length (Time) of video file.
Speaker speed	Count of words used by a speaker per minute.

TABLE 2. Parameters for speech based emotion detection model. B is referred to as batch size.

Layers	Output Shape	Param	Activation
Input	[B, 180]		
conv1D_1	[B, 161, 128]	2688	ReLU
batch normalisation	[B, 161, 128]	512	
conv1D_2	[B, 152, 64]	81984	ReLU
batch normalisation	[B, 152, 64]	256	
flatten	[B,4864]		
dense_1	[B,520]	5059080	
dense_2	[B,8]	4168	Softmax
Total params : 5,148,688			
Trainable params : 5,148,304			
Non-trainable params : 384			

The emotions were characterised into 5 categories, i.e., “joy,” “sadness,” “anger,” “fear,” and “neutral”. The model used for training the dataset was K-train based text BERT model. The maximum length of the unigrams is 35000 and the tokens are 350. We run the model for 5 epochs with a batch size equal to 12. The learning rate was kept $2e^{-5}$. The output of the model was treated as an array with the probability of all the emotions, shown as p1, p2, p3, p4, p5 and was stored as X2 in CLEFT.

E. SPEECH BASED EMOTIONS

Emotions play an important role in teaching [39]. A monotonous video without any emotions will be relatively less engaging than those videos where the teacher exploits emotions. Besides emotion detection in the text, we also conduct emotion detection in speech which would lead to a more reliable understanding of the engagement factors in videos.

We describe speech-based emotions with three features, i.e., Mel-frequency cepstral coefficients (MFCC), chroma, and Mel spectrogram frequency which are extracted from the speech waveform from the RAVDESS dataset. Ryerson Audio-Visual Database for Emotional Speech and Song (RAVDESS) dataset [40] consists of speech and song, audio and video files. For our analysis, we focused on the emotional speech and song files. There are 1440 files in the RAVDESS dataset that assist in analyzing the emotions from the speech. The RAVDESS dataset contains 24 actors (12 male and 12 female), who record the speech in lexically similar statements in a neutral North American accent. The speech dataset is further categorized into seven different emotions namely, “Calm,” “Happy,” “Sad,” “Angry,” “Fearful,” “Surprise” and “Disgust”. There are mainly two different levels of emotional intensities (Normal and Strong). There is also an additional third neutral expression.

The speech was classified into 8 categories, which are, “Angry,” “Sad,” “Happy,” “Neutral,” “Fear,” “Disgust,” “Surprise” and “Calm”. MFCC: Mel Frequency Cepstrum (MFC) is a representation of linear cosine transform of a short-term log power spectrum of a speech signal on a non-linear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are together make up an MFC. MFCC extraction is of the type where all the characteristics of the speech signal are concentrated in the first few coefficients [41]. Chroma-based features are used for identifying pitch-based information or they can also be referred to as pitch class profiles. The chroma representation is used for intensities of 12 distinct musical chromas of the octave at each time frame. By using chroma we generated a chromagram based on 12 pitch classes. The pitch classes in the particular order are as follows C, C#, D, D#, E, F, F#, G, G#, A, A#, B.

The total number of training samples was 1152 and validation samples were 288. We used 1D CNN for training the data. For the input layer, we extracted all the features and combined them horizontally, where the length of the input vector was 180. The length of the MFCC feature was 40, Mel was 128 and chroma was 12. After the input layer, we used 1D CNN layer with 128 filter sizes of 20 and a stride of 1. Post the CNN layer we used the batch normalization layer and the activation function was ReLU. We used another layer of 1D CNN with 64 filter sizes of 10. Another batch normalisation layer was used and the activation function was ReLU. We further used a dense layer of size 520 and then another dense layer of size 8 with softmax as an activation function. The training batch size was 16, the learning rate

was $10^{-}4$, the loss was categorical cross-entropy and Adam was used as a training optimizer. Table 2 describes the model architecture and parameters. The output of the model was treated as an array with the probability of all the speech-based emotions, shown as $q1, q2, \cdots, q8$ and was stored as $X3$ in CLEFT.

F. OBJECT DETECTION

Our motivation to include object detection is primarily to capture different objects in the videos [42], [43], for instance, if there are animations in the videos, it would result in more objects than just the teacher and the students, i.e., human types. Other use cases include if the teacher uses a variety of objects in a class to teach students in addition to traditional objects already found in the classrooms, for instance, a Physics teacher using a range of real-world objects to explain a concept. YOLO v3 uses logistic regression to compute the target score. It gives the score for all targets in each boundary box. YOLO v3 can give the multilabel classification because it uses a logistic classifier for each class in place of the softmax layer used in YOLO v2. YOLO v3 uses darknet 53. It has fifty-three layers of convolution. These layers are more in-depth compared to darknet 19 used in YOLO v2. Darknet-53 contains mainly 3×3 and 1×1 filters along with bypass links [44], [45]. The output of the model is where it counted the objects and animations that appeared in the video with respect to time and was stored as X4 in CLEFT.

In Equation 1, X_1 is based on contextual engagement provided with contextual engagement score ranging from 0-1 based on 14 textual features. This is the regression-based model. X_2 provides the overall emotional distribution over 5 classes for the lecture transcript. This is the text-based model which exploits representation vectors obtained from the BERT language model. X_3 provides the emotion feedback with reference to time, based on the speaker's tone and delivery speech which mainly exploits the speech data. X_4 detects the number of animation and objects with reference to time, which is the object detection model.

$$y = \alpha X_1 + \beta X_2 + \gamma X_3 + \delta X_4 \tag{1}$$

where y is the prediction score. The individual weight parameters $\alpha, \beta, \gamma, \delta$ are the coefficients. Initially, $\alpha, \beta, \gamma, \delta$ will be initialised with random weights. After every watched video, we are collecting the user rating for the video (out of 5) and positive and negative comments. This user feedback is the real truth and is denoted by $\hat{y}$. To minimise the error we use the Huber loss [46] given by:

$$L_\theta(y, \hat{y}) = \begin{cases} 0.5 * (y - \hat{y})^2, & |y - \hat{y}| \leq \theta \\ \theta * (|y - \hat{y}| - 0.5 * \theta), & \text{otherwise} \end{cases} \tag{2}$$

where $\hat{y}$ is the normalised user feedback truth, y is the predicted output and θ is the hyperparameter for a large or small error. Our objective is to minimize $L_\theta(y, \hat{y})$ based on $\alpha, \beta, \gamma, \delta$. The loss will be backpropagated for the individual model as well as the coefficients. Since it is a linear equation,

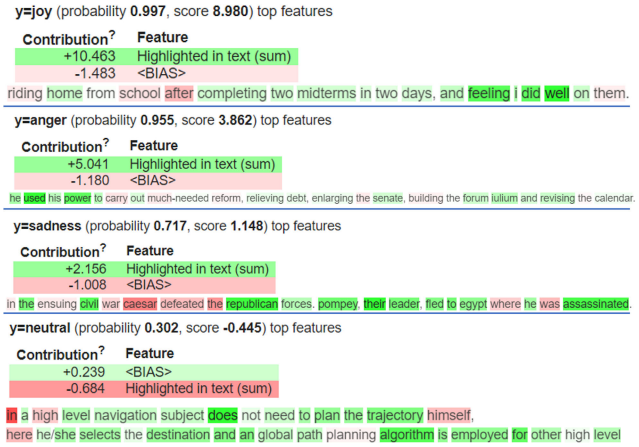

FIGURE 3. Feedback provided by our CLEFT model, where green indicates the positive emotion and red negative emotion.

the coefficients can also be optimised by the Gaussian process, but it would be highly unlikely that there will be a unique solution. Therefore, by using backpropagation we are aiming to optimise the model as well as the coefficients for the prediction of engagements. We run the backpropagation model for a certain number of iterations until the model parameters converge. In the end, we obtain the converged weight parameters followed by fine-tuned models learned in a unified way. The advantage that we get is that these parameters are trained in a consolidated parameter space which leads to more reliable results than approaches that use cascaded techniques where the output of one or more is fed as input to the next model. We found that, in our case, the results obtained from cascading models were too poor and that it was difficult to engineer the pipeline framework because there is no such pre-defined sequence rule for the models that we could engineer, e.g., should X_1 come before X_2 or vice-versa. In contrast, in our CLEFT, all models work simultaneously. For initialising, the coefficients are $\alpha = 0.5$, $\beta = 0.1$, $\gamma = 0.2$ and $\delta = 0.2$.

IV. EXPERIMENTS AND RESULTS

Our main experimental results are presented first where we compare our full model with variations of our model. We remove certain key components from our model which becomes a baseline model. This will help demonstrate that our full model outperforms these baseline methods. Given that there is no other published baseline to compare our full model, we thus adopt this quantitative comparison strategy to demonstrate the effectiveness of our full model. As a qualitative study, through our results, we demonstrate how our model provides reliable feedback to the content creator. As an ablation study, we have demonstrated that the individual pre-trained components of our model learn reliably from the data which help in the overall predictive performance of our model.

Dataset: We have applied our proposed architecture CLEFT on our new dataset, where we used the user ratings

based on engagement from the video as the target variable and benchmark. For collecting this dataset, we have deployed the model on the cloud server, and users are shown a video from open-source videos lectures according to their interest. After completing the video, users are asked to provide a rating for their engagement on a scale of 1 to 10. We have, until now, 50 videos in our collection dataset. The average video length was 29.5 minutes, and the topics cover domains of school level physics, literature and history. Videos were obtained from open source lectures, such as NPTEL, byjus, and Unacademy. Note that even with this dataset size, we can obtain reliable features from these full-length videos to train our model. We expect that with the large-scale datasets such as those used in [3], we can further help improve our model performance.

Deep learning has demonstrated that it can scale to large datasets and reliable parameter learning can be done when a large number of instances are used. However, sometimes storing and processing large datasets incurs a huge amount of cost that is very relevant to some datasets especially videos. The problem is even more acute if these videos are lecture videos that are longer in length and occupy a large amount of space. We chose 50 videos in our setting because we wanted to experiment our model in case where using commodity hardware one can replicate our results reliably. We also noted that these 50 videos generated enough features for us to model the problem reliably as lecture videos are much longer in length.

For every video processing, it was separated in two segments, one with video frame embeddings and the other with audio. We extracted the features from the audio and video discussed above and calculated the engagement score by leveraging our model, CLEFT. The ground truth mean engagement score as reported by users, after normalisation, is 0.88 (88%). The predicted mean engagement score by CLEFT is 0.85 (85%). Figure 4 depicts the predicted engagement score by all the models including the baseline methods. Predicted engagement score $X_1 + X_2 + X_3 + X_4 = 0.85, X_2 + X_3 + X_4 = 0.63, X_1 + X_3 + X_4 = 0.72, X_1 + X_2 + X_4 = 0.76, X1 + X_2 + X_3 = 0.72$, where $+$ denotes the model components x_i used in our framework. The engagement score shows the impact of individual model on the final output.

Figure 3 depicts our qualitative results based on shapely values. The text associated with feedback has words which are assigned different weights based on the overall engagement score. The overall engagment score is computed by different components of the model. As demonstrated in our qualitative results, our CLEFT model provides useful feedback to the content creator after they provide the video to our CLEFT model. This characteristic of our model is particularly used in cases where the creator intends to understand the shortcomings of the video automatically before sharing the content with users, whereas, manual even getting this amount of reliable manual feedback will be very time consuming. The model feedback behind the emotion ''joy,'' ''anger,'' ''sadness,'' and ''neutral'' is presented. In the

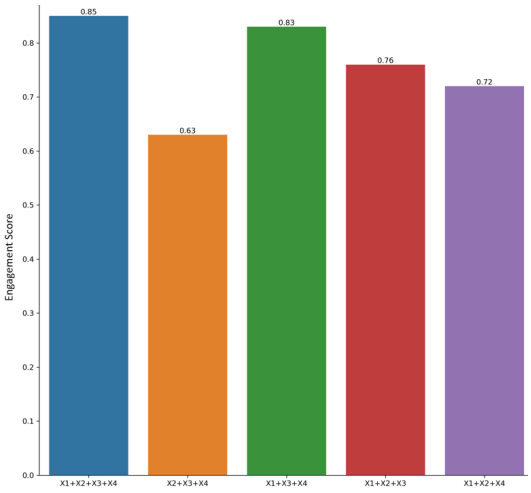

FIGURE 4. Variation in predicted mean engagement score by leaving one model out and our full CLEFT model. We can notice that compared to individual baseline models mentioned above and leaving out certain models out, our full model obtains the best result quantitatively.

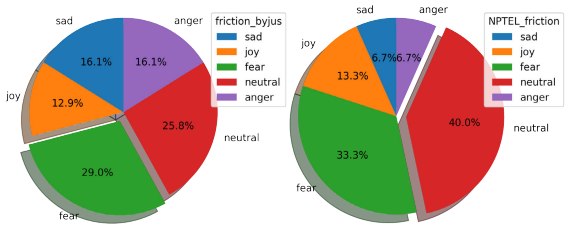

FIGURE 5. Emotion distribution of two different videos by different open content providers on the same topic.

figure, green depicts the positive contribution and red/pink shows negative contribution or detracts the model from the prediction. The shade of the colour represents the strength of the coefficients in the inferred model. We can clearly observe that model is able to predict relative emotion with reference to the text. For example feeling of exam going well is joy, wheres Julius Ceaser's assassination is sad and a topic of trajectory planning is considered to be neutral. The model still further needs to be tuned with more data with related to context to provide better accuracy with multiple emotions. Figure 5 shows the overall distribution of emotions in the text data. It was observed that the content with more varying emotions has higher engagement than the content with one emotion.

A. ABLATION ANALYSIS

In this section, we present our ablation study where we demonstrate that individual models do play a role in predicting the overall engagement score. As a result, these components are crucial to our model.

Figure 6 shows the variation of speech-based emotion over the video length where ''Happy,'' ''Surprised,'' ''Neutral,'' ''Fear'' are dominant. To generate this figure, we extracted 10 secs of speech with a moving window of 10 secs and a

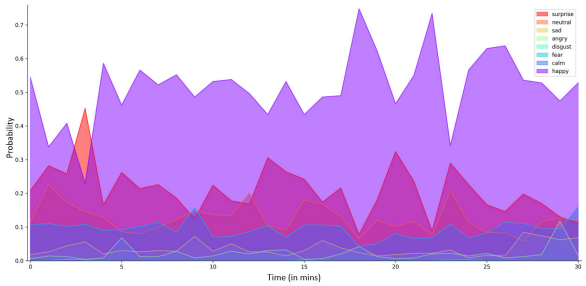

FIGURE 6. Speech emotion response of video 1 with respect to video time stamp.

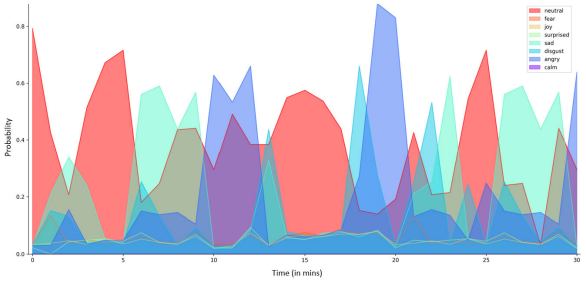

FIGURE 7. Speech emotion response of video 2 with respect to video time stamp.

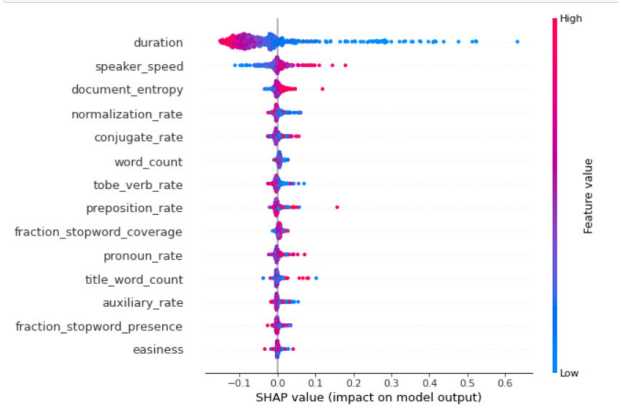

FIGURE 8. Shap values for VLN dataset for context engagement. The duration of video is a very important feature and it is showing that longer length of video shows drop in engagement.

hop of 10 secs as well. Subsequently, the model prediction probability for every emotion was used. Similarly, Figure 7 shows the variation of speech-based emotion over the other video where "Anger," "Sad," "Disgust," "Neutral" are dominant.

Using random forest regressor for contextualised engagement we obtained a mean squared error of 0.0173. Figure 8 shows the shap summary plot of different features by random forest regressor. The summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value. The colour represents the

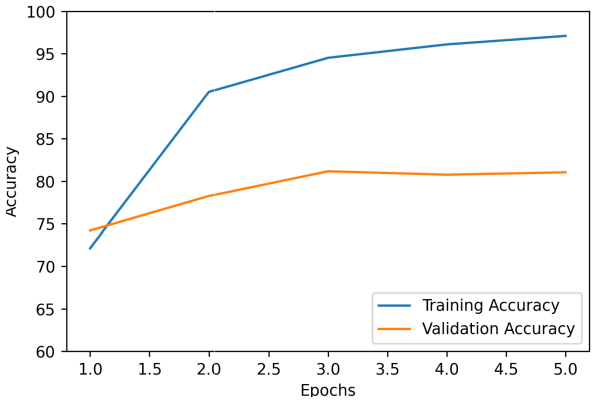

FIGURE 9. Training and test accuracy for decoding text based emotion using K-train BERT architecture.

value of the feature from low to high. The features are ordered according to their importance. We can see that the length of the video is the most important feature and longer lengths of videos have less impact on the model. Preposition_rate has a low affect on the model. Similarly, tobe_verb_rate has major effect on model but the used rate should be <0.03% ("be," "being," "was," "were," "been," "are," "is"). Auxilary_rate ("will," "shall," "cannot," "may," "need to," "would," "should," "could," "might," "must," "ought," "ought to","can't," "can") will provide with positive effect if the used rate is <0.025%. Speaker speed has more importance to the model and speed contributing positively to engagement is 115-120 words per minute. It should not be very low or very high. The easiness level determined should be more than 83 to have a positive impact on the model. A normalisation rate greater than 0.1 has a positive effect on the engagement score. Preposition_rate has a very minor effect on the model.

Figure 9 shows the training and validation accuracy for the k-train based BERT model for decoding text-based emotion. We used a pre-trained k-train model and further fine-tuned it on our dataset. The model was provided with the text extracted from speech to text. The model runs for 5 epochs beyond which no improvement was observed. The model achieved a test accuracy of 81% with 5 classes. Table 3 shows the precision, recall, and F1 score of individual emotions. Precision for "Joy" and "Fear" is the same 0.85%, anger is 0.82%, neutral is 0.79% and sadness is 0.77%. Macro-average values is calculated by:

$$P = \frac{P_1 + P_2 + \cdot + P_n}{N} \tag{3}$$

where P can be precision or recall or F1 score. P_i is precision/recall/f1 score for class i and N is the total number of class.

Figure 10 shows the training and test accuracy for speech-based emotion detection. The model was trained for 200 epochs and achieved an overall accuracy of 70.83% with 8 classes.

TABLE 3. Results for emotion classification using Ktrain-BERT model for text.

Emotion	Precision	Recall	F1-score
Joy	0.85	0.84	0.84
sadness	0.77	0.83	0.80
fear	0.85	0.83	0.84
anger	0.82	0.75	0.78
neutral	0.79	0.83	0.81
macro avg	0.82	0.82	0.81
accuracy : 0.81			

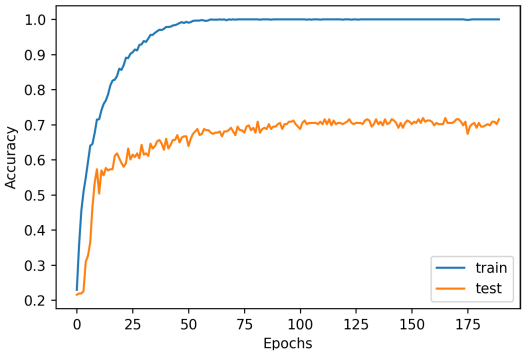

FIGURE 10. Training and test accuracy for decoding speech based emotion from Ravdess dataset.

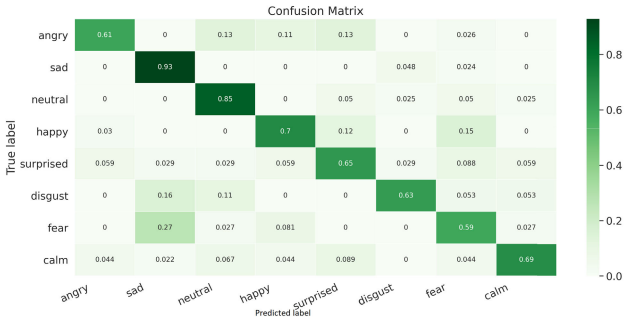

FIGURE 11. Confusion matrix for speech based emotion detection.

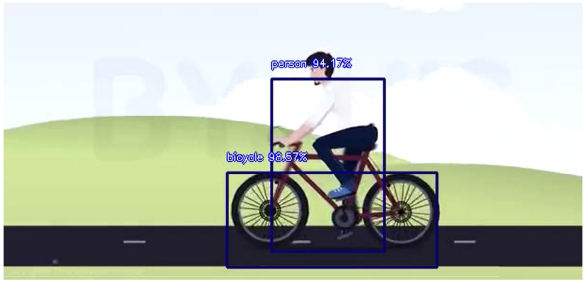

FIGURE 12. Yolov3 object detection in online video tutorials.

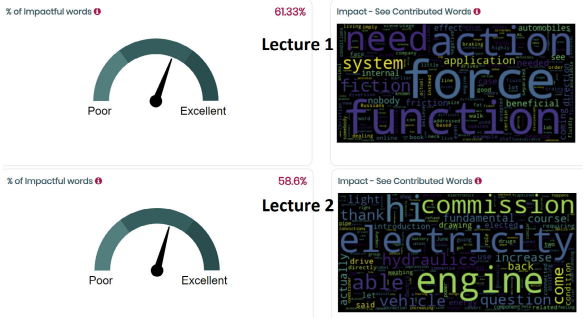

FIGURE 13. The two lectures which were measured in their of the impact that they can create.

Figure 11 shows the confusion matrix of emotions from speech data. The precision for angry is 0.82 %, calm is 0.72%, disgust is 0.74%, fearful is 0.68%, happy is 0.59%, neutral is 0.75%, sad is 0.59% ans surprised is 0.86%. Macro average precision, recall and f1 score is 0.72%, 0.71% and 0.71% respectively.

Figure 12 shows the detection of objects inside the specific video frame with 3fps and it keeps the count of the objects shown with reference to time. As a result, we can automatically detect the activities with different objects in the video. We are extracting the information of objects and matching it with the pretext context of that object for the relevancy of the topic. In this way, we can gauge the role of various objects towards the overall engagement score.

V. DEPLOYMENT IN A PRODUCTION SETTING

Our model was deployed online in a production environment and was provided to students. Learning Management System (LMS) organisation helped in assessing online lectures and evaluating them in real-time with their users. The deployed model evaluated the two sample lectures from NPTEL, one a lecture on Friction and the other on eVehicles, based on the design and delivery of the content. The design of the lecture was evaluated on Impact, Complexity, Content Richness and Segmentation whereas the delivery was evaluated based on Vividness, Facial Expression, Speech Expression, and Speech Speed Performance. The goal of this user study is to measure whether our model can learn reliably from the data. The data used in this study including the number of users in the study can be obtained by contacting the first author which not only includes the videos but also the associated user study data. Detailed parameter settings of the deployed model can also be obtained by contacting the authors.

LMS company requested tutors to upload their online content on the server, where the model evaluated every video and provided feedback to the creators. Post analysis LMS company requested students of grade 11 or older to score the video on a scale of 1 to 10 in terms of engagement. Post lecture, content creators asked questions related to the topic in form of a checkbox to gauge the understanding of content. We describe the key terminologies below along with the results associated with these terminologies.

1) Impact (Figure 13) – Content impact measures the proportion of meaningful words to that of total words used in the content. The briefer the content is the higher is the learner engagement. Impact measure helps

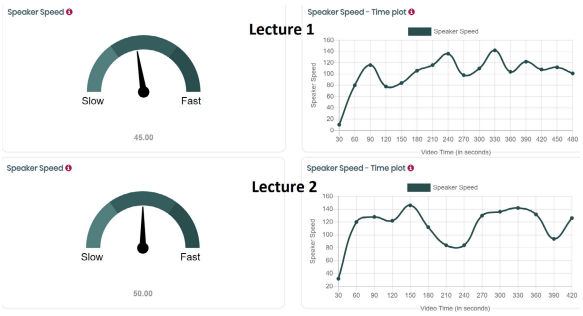

FIGURE 14. Comparison between two videos on the basis of speaker speed.

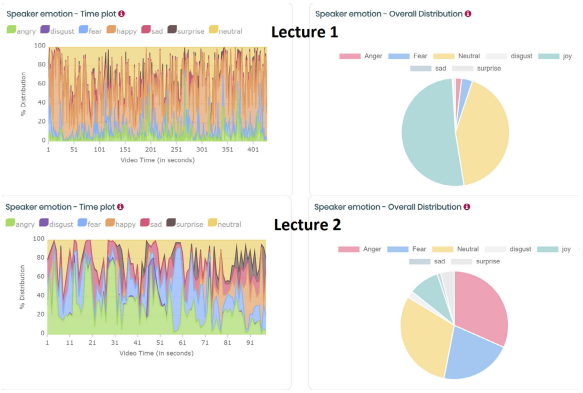

FIGURE 15. Comparison between two videos on the basis of speaker emotion.

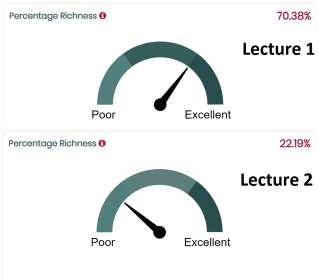

FIGURE 16. Comparison between two videos on the basis of richness of content.

maintain brevity by creating maximum impact through the lowest possible content.

2) Complexity (Figure 17) – Content complexity measures the depth and diversity of the words used in design and delivery. Content complexity feedback helps optimise the usage of rare and unique words to keep the content simple and thus maximise learner engagement.
3) Content Richness (Figure 16) – Content richness provides feedback on optimal usage of educational media in the video. A video lecture gets the highest possible visual engagement when it has more multimedia in it. However, creating such a video is complex and highly time-consuming. Content richness helps instructors identify the optimal amount of educational media to be used without worrying about the engagement level of the video.
4) Segmentation (Figure 17) – Segmentation provides feedback on the optimal duration of the video for better learner attention.
5) Vividness (Figure 17) – Content vividness measures the contextual word usage in speaker communication. The usage of striking words is proven to better engage listeners over bland word choices. Appropriate word choices eliminate the overuse of words and reduce the monotonicity in the narration.
6) Facial Expression (Figure 15) – Speaker facial decoding reports the emotions expressed in the speaker's face during content delivery. Major emotions that are reported and proven to impact the learner mood are happy, sad, surprise, angry, disgust and fear.
7) Speech Expression (Figure 17) – Delivery tone and rhythm feedback report the emotions conveyed in the speaker's audio pitch. Speech tone and rhythm are found to influence the listener's mental state. The feedback helps the speaker achieve a tone that expresses joy and surprise that is proven to have better engagement among the listeners.
8) Speech Speed Performance (Figure 14) – Speaker speed measures the words-per-minute in speech. Speaker speed influences the processing of the learner's acquired information. While a lower speed is proven to be perceived as less challenging among the listeners, speed in the upper range is proven to be equally dangerous in losing learner attention. Speaker speed feedback helps optimize the speed based on the context of the content.

The model deployed was used by instructors to generate the automatic feedback on their videos and modify the content accordingly. As per them, they received positive feedback from the audience after they showed the revised videos to the users. To further improve the model feedback quality, the instructors provided their insights into the automatic feedback quality generated by our model. As a result, we further tuned the model parameters which helped improve the automatic feedback on the production system. We thus found evidence from real users that the automatic feedback component of our model is useful to real users that helps improve the quality of the tutoring videos.

A. LIMITATIONS

We have presented promising results obtained from our model, CLEFT. However, we must highlight certain limitations of the model. We have shown that our model can be reliably applied under settings where we have less data, e.g., 50 or less videos (RQ3). We have not yet ascertained that our model can reliably scale to millions of videos. While deep learning methods have shown to perform well on large-scale datasets, we are confident that our model can generalise well in large-datasets too including scaling to such collections. The key issue is not modelling a large number of videos.

	NPTEL_Friction	NPTEL_evehicles
DESIGN	64.65%	46.47%
Impact	61.33%	58.60%
Complexity	68.40%	58.67%
Content Richness	70.38%	22.19%
Segmentation	70.47%	70.45%
DELIVERY	61.91%	76.28%
Vividness	48.24%	65.22%
Facial Expression	57.71%	69.39%
Speech Expression	100.00%	100.00%
Speech Speed Performance	60.00%	80.00%

FIGURE 17. Comparison between two video content on the basis of design and delivery from real users. Percentages are computed with reference to the full video lectures: (Total correct hit/Total hit + miss) * 100.

What remains a challenge is downloading and storing a large number of videos and processing them that requires specialised and expensive hardware resources.

While it can be argued that our model can underperform in situations where the setting is not in the teaching and learning environment, for instance, if someone records voice in a busy area where there are plenty of movements. One of the key features that we use is domain-dependent textual features, for instance, the vocabulary used in the lecture belong to a certain domain that could be Physics, Mathematics. Our assumption is that a video recorded in a busy mall will not have words that come from such domain-dependent distributions. While the model can be further improved, for instance, using latent topics to constrain the model to remain within a particular topical theme, we will further improve our current model by considering latent topics as a future work. We will also model background noise to alleviate certain shortcomings in the model.

VI. DISCUSSION AND CONCLUSION

We have developed a new framework CLEFT for contextualised adaptive engagement. In [3], the authors have focused on the textual information only for establishing contextualised engagement which is insufficient to provide feedback. As a result, we have extracted additional discriminative features, which are, textual emotion, speech emotion variability with time, animation and object detection from the video lectures, and unified them to create a prediction variable and update the vector-based on user feedback. Our novel model unifies the individual pre-trained models and learns their weight parameters in a completely unsupervised way. Our results show that our model can reliably provide engagement scores followed by feedback which existing models cannot do. We encourage content creators to use the automated feedback and reflect upon the created content. We hope that our automated feedback will help improve the content quality over time.

The individual models in our CLEFT framework were first trained on publicly available datasets, and the training performances were reported on those datasets as baseline results and to establish that each one of them can learn from their respective data. Subsequently, we unified these models to develop our novel CLEFT framework and compared the performance of each model against the CLEFT model. As evidenced in the results in Figure 4 the major impact on the prediction is based on X_1 which was trained on the VLN dataset with subset features, used in [3]. Removing X_2, from the overall model, did not impact the prediction significantly which models the emotions based on the textual content. Likely, textual emotions are not very crucial in domain-specific videos. As a result, the emotion of text can have the least impact on domain-specific predictions. Removing X_3, which is based on emotion decoding over speech reduced the predicted engagement score significantly. It is also observed that variation of positive emotion increases engagement compared to negative emotion. Figure 6 shows the variation in emotion of speech over time where "Happy," "Surprised," "Neutral," "Fear" are dominant and Figure 7 shows the variation in emotion of speech over time where "Anger," "Sad," "Disgust," "Neutral" are dominant. Engagement score of video, Figure 6, were significantly better than the video in Figure 7. Variation in the emotion of speech over time helps to increase the engagement score than having a single emotion tone for a longer period. Removing X_4 which accounts for object count over the video also had a significant drop in the engagement score as the animation plays a key role in engagement. However, the impact of this model would be greater if we could account for more than 80 objects as fixed in Yolo. Based on textual data the length of the video, tobe_verb_rate, Auxilary_rate, speaker speed, and the easiness level of the text is important for engagement prediction. For video to be engaging the length of the video should be short, the use of tobe_verb_rate should be $<0.03\%$, use of Auxilary_rate should be $<0.025\%$, speaker speed should be in the range of 115-120 wpm, easiness level should be more than 83.

In the future, we will design an estimator and policy based on the observed truth y and ground truth $\hat{y}$ to make the prediction stronger and user-centric. To this end, we will

develop a new reinforcement learning framework where user feedback is incorporated into the model.

ACKNOWLEDGMENT

The dataset for the study is made available by Brainalive Research Pvt. Ltd. https://forms.gle/TgmmefJPxnwJVroG6. The authors would also like to thank Gnaneswara Rao Gorle from Brainalive Research for his support on some results.

REFERENCES

[1] M. Ehlers, R. Schuwer, and B. Janssen, "OER in TVET: Open educational resources for skills development," *UNESCO-UNEVOC Int. Centre Tech. Vocational Educ. Training*, 2018.

[2] R. White, "Beliefs and biases in web search," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2013, pp. 3–12.

[3] S. Bulathwela, M. Pérez-Ortiz, A. Lipani, E. Yilmaz, and J. Shawe-Taylor, "Predicting engagement in video lectures," 2020, *arXiv:2006.00592*.

[4] S. Bulathwela, M. Pérez-Ortiz, R. Mehrotra, D. Orlic, C. De La Higuera, J. Shawe-Taylor, and E. Yilmaz, "SUM20: State-based user modelling," in *Proc. 13th Int. Conf. Web Search Data Mining*, Jan. 2020, pp. 899–900.

[5] P. Ekman, "Are there basic emotions?" *Psycol. Rev.*, vol. 99, no. 3, pp. 550–553, 1992.

[6] X. Li, J. Pang, B. Mo, and Y. Rao, "Hybrid neural networks for social emotion detection over short text," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 537–544.

[7] Y. Wang, S. Feng, D. Wang, G. Yu, and Y. Zhang, "Multi-label Chinese microblog emotion classification via convolutional neural network," in *Proc. Asia–Pacific Web Conf.* Cham, Switzerland: Springer, 2016, pp. 567–580.

[8] C. Baziotis, N. Athanasiou, A. Chronopoulou, A. Kolovou, G. Paraskevopoulos, N. Ellinas, S. Narayanan, and A. Potamianos, "NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning," 2018, *arXiv:1804.06658*.

[9] H. Meisheri and L. Dey, "TCS research at SemEval-2018 task 1: Learning robust representations using multi-attention architecture," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 291–299.

[10] P. Du and J.-Y. Nie, "Mutux at SemEval-2018 task 1: Exploring impacts of context information on emotion detection," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 345–349.

[11] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowl. Inf. Syst.*, vol. 62, no. 8, pp. 2937–2987, Aug. 2020.

[12] M. Lugger, M.-E. Janoir, and B. Yang, "Combining classifiers with diverse feature sets for robust speaker independent emotion recognition," in *Proc. 17th Eur. Signal Process. Conf.*, 2009, pp. 1225–1229.

[13] B. Schuller, M. Lang, and G. Rigoll, "Robust acoustic speech emotion recognition by ensembles of classifiers," in *Tagungsband Fortschritte der Akustik-DAGA#*, vol. 5. Berlin, Germany: German Acoustical Society (DEGA), 2005.

[14] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Comput. Appl.*, vol. 9, no. 4, pp. 290–296, Dec. 2000.

[15] I. S. Engberg and A. V. Hansen, "Documentation of the Danish emotional speech database des," Internal AAU Rep., Center for Person Kommunikation, Denmark, Tech. Rep., 22, 1996.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1517–1520.

[17] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.

[18] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, and J. Shawe-Taylor, "TrueLearn: A family of Bayesian algorithms to match lifelong learners to open educational resources," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 565–573.

[19] J. Zhou and S. Bhat, "Modeling consistency using engagement patterns in online courses," in *Proc. 11th Int. Learn. Anal. Knowl. Conf.*, Apr. 2021, pp. 226–236.

[20] S. Chen, Y. Fang, G. Shi, J. Sabatini, D. Greenberg, J. Frijters, and A. C. Graesser, "Automated disengagement tracking within an intelligent tutoring system," *Frontiers Artif. Intell.*, vol. 3, Jan. 2021, Art. no. 595627.

[21] S. Bulathwela, M. Pérez-Ortiz, E. Yilmaz, and J. Shawe-Taylor, "Power to the learner: Towards human-intuitive and integrative recommendations with open educational resources," *Sustainability*, vol. 14, no. 18, p. 11682, Sep. 2022.

[22] S. Bulathwela, M. Perez-Ortiz, E. Yilmaz, and J. Shawe-Taylor, "VLEngagement: A dataset of scientific video lectures for evaluating population-based engagement," 2020, *arXiv:2011.02273*.

[23] M. P. Ortiz, S. Bulathwela, C. Dormann, M. Verma, S. Kreitmayer, R. Noss, J. Shawe-Taylor, Y. Rogers, and E. Yilmaz, "Watch less and uncover more: Could navigation tools help users search and explore videos?" in *Proc. ACM SIGIR Conf. Human Inf. Interact. Retr.*, Mar. 2022, pp. 90–101.

[24] M. R. Jimenez-Liso, M. Martinez-Chico, L. Avraamidou, and R. L. Lucio-Villegas, "Scientific practices in teacher education: The interplay of sense, sensors, and emotions," *Res. Sci. Technol. Educ.*, vol. 39, no. 1, pp. 44–67, Jan. 2021.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[26] J. Nias, "Thinking about feeling: The emotions in teaching," *Cambridge J. Educ.*, vol. 26, no. 3, pp. 293–306, Nov. 1996.

[27] M. Zembylas, "The power and politics of emotions in teaching," in *Emotion in Education*. Amsterdam, The Netherlands: Elsevier, 2007, pp. 293–309.

[28] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Automatic assessment of document quality in web collaborative digital libraries," *J. Data Inf. Quality*, vol. 2, no. 3, pp. 1–30, Dec. 2011.

[29] M. Warncke-Wang, D. Cosley, and J. Riedl, "Tell me more: An actionable quality model for Wikipedia," in *Proc. 9th Int. Symp. Open Collaboration*, Aug. 2013, pp. 1–10.

[30] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proc. 15th Int. Conf. World Wide Web*, May 2006, pp. 83–92.

[31] P. J. Guo, J. Kim, and R. Rubin, "How video production affects student engagement: An empirical study of MOOC videos," in *Proc. 1st ACM Conf. Learn. Scale Conf.*, Mar. 2014, pp. 41–50.

[32] K. R. Scherer and H. G. Wallbott, "Evidence for universality and cultural variation of differential emotion response patterning," *J. Personality Social Psychol.*, vol. 66, no. 2, pp. 310–328, 1994.

[33] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers Psychol.*, vol. 8, p. 1454, Aug. 2017.

[34] A. K. Anderson and E. A. Phelps, "Lesions of the human amygdala impair enhanced perception of emotionally salient events," *Nature*, vol. 411, no. 6835, pp. 305–309, May 2001.

[35] L. Anderson and A. P. Shimamura, "Influences of emotion on context memory while viewing film clips," *Amer. J. Psychol.*, vol. 118, no. 3, pp. 323–337, Oct. 2005.

[36] F. G. Ashby and A. M. Isen, "A neuropsychological theory of positive affect and its influence on cognition," *Psychol. Rev.*, vol. 106, no. 3, p. 529, 1999.

[37] E. Bartolic, "Effects of experimentally-induced emotional states on frontal lobe cognitive task performance," *Neuropsychologia*, vol. 37, no. 6, pp. 677–683, Jun. 1999.

[38] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," 2017, *arXiv:1710.03957*.

[39] R. E. Sutton and K. F. Wheatley, "Teachers emotions and teaching: A review of the literature and directions for future research," *Educ. Psychol. Rev.*, vol. 15, no. 4, pp. 327–358, 2003.

[40] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[41] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using Mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," 2010, *arXiv:1003.4083*.

[42] E. Sieber, "Teaching with objects and photographs: Supporting and enhancing your curriculum. A guide for teachers," Mathers Museum World Cultures, Indiana Univ., Bloomington, IN, USA, Tech. Rep., 2001.

[43] R. H. Kay and L. Knaack, "Exploring the impact of learning objects in middle school mathematics and science classrooms: A formative analysis," *Can. J. Learn. Technol.*, vol. 34, no. 1, pp. 1–19, Dec. 2008.
[44] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[45] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015, *arXiv:1506.02640*.
[46] P. J. Huber, "A robust version of the probability ratio test," *Ann. Math. Statist.*, vol. 36, no. 6, pp. 1753–1758, Dec. 1965.

SUJIT ROY received the Ph.D. degree in computer science from Ulster University in collaboration with IIT Kanpur. He worked as a Machine Learning Researcher at The University of Manchester in the domain of explainable AI. He is also the Co-Founder of Brainalive Research Pvt. Ltd. His research interests include explainable AI, computer vision, signal processing/synthesis, graph neural networks, and reinforcement learning.

VISHAL GAUR received the master's degree in information technology with a major in data engineering and machine learning from Tampere University. He is currently working as a Data Scientist at Brainalive Research Pvt. Ltd. He enjoys reading tech blogs and getting hands-on experience with the latest technologies. He also enjoys working in the domain of computer vision, time series modeling, and neural networks.

HAIDER RAZA (Senior Member, IEEE) received the bachelor's degree in computer science and engineering from Integral University, India, in 2008, the master's degree in computer engineering from Manav Rachna International University, India, in 2011, and the Ph.D. degree in computer science from Ulster University, U.K., in 2016. From December 2015 to June 2016, he worked as a Postdoctoral Research Assistant in brain–computer interface (BCI) for both magnetoencephalography (MEG) and electroencephalography (EEG) systems at Ulster University, Northern Ireland, U.K. From July 2016 to November 2017, he worked as a Research Officer (Data Science) with the Farr Institute of Health Informatics Research, Swansea University Medical School, U.K. He is currently a Lecturer with the School of Computer Science and Electronics Engineering, University of Essex, U.K. His Ph.D. Project was funded through the Vice-Chancellor Research Scholarship from Ulster University. During his Ph.D. degree, he won the Best-Literature Review Award sponsored by McGraw Hill.

SHOAIB JAMEEL received the Ph.D. degree from The Chinese University of Hong Kong, in 2014. He is currently a Lecturer at the School of Electronics and Computer Science, University of Southampton, U.K. He works with various technology startups in the U.K. spearheading their technical sphere, where his research outputs are directly applied to their production systems. His research interests include text mining, natural language processing, and computer vision. He is a fellow of the Higher Education Academy. He has served on the programme committees of various conferences, such as AAAI, IJCAI, and SIGIR. He is an Associate Editor of *AI Communications* journal.