International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

# Extracting Meaningful Insights from User Research Videos

Simran Kaur Ghatoray & Yongmin Li

**To cite this article:** Simran Kaur Ghatoray & Yongmin Li (06 Feb 2026): Extracting Meaningful Insights from User Research Videos, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2026.2619613

**To link to this article:** https://doi.org/10.1080/10447318.2026.2619613

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2619613



# Extracting Meaningful Insights from User Research Videos

Simran Kaur Ghatoray and Yongmin Li

Department of Computer Science, Brunel University London, Uxbridge, UK

**ABSTRACT**

Recognising and tracking user emotions in research videos is vital to understanding user needs and expectations. Limited research exists on automating emotion extraction from multimodal videos in user experience (UX). This study proposes a conceptual framework for automated extraction of actionable insights using facial, speech-to-text, and text-based emotion recognition to capture nuanced emotional data. The multimodal approach integrates visible and spoken cues through temporal alignment and fusion techniques, enabling robust behavioural pattern detection. An interactive AI analyst tool is used to query the integrated data in natural language, reduce manual workload, and improve the efficiency and scalability of UX evaluation. A case study of the implementation of the proposed framework is also provided with details of individual components, such as facial emotion recognition, speech-to-text, text-based emotion recognition, temporal alignment and fusion, and insight extraction via interactive AI.

**KEYWORDS**

User experience; insight extraction; facial emotion recognition; text-based emotion recognition; interactive AI

## 1. Introduction

In today's digital world, user emotions play a vital role in developing user-centric designs. The emotional state of a person can affect concentration, decision-making skills, and task-solving skills (Kołakowska et al., 2014). The growing importance of recognizing emotions can be seen in different fields such as healthcare, medical, customer service, human-computer interactions, education, gaming, etc. Among its vast applications, in recent years, focus has been made toward User experience and its evaluation methods.

User experience (UX) is a growing field that includes disciplines such as User Interface Design. Better user design leads to greater user satisfaction following good revenue and retention (Souza et al., 2022). UX evaluation is carried out to ensure better user satisfaction and meet user requirements through various methods and techniques. Usability and UX are intertwined terms, but serve different purposes. UX methods focus on improving user satisfaction by achieving pragmatic and hedonic goals, while usability methods aim to improve human performance (Souza et al., 2022). With the advancement in the field of Artificial Intelligence, there always remains a question of how AI methods and techniques can improve UX evaluation?

Different techniques, methods and modalities, now, make it possible to capture the whole picture of user emotions, allowing to change the traditional ways of UX evaluation. User emotions being a crucial part for the betterment of systems, can be detected and recognized with the help of machine learning and deep learning methods. Various modalities such as facial emotion recognition, speech emotion recognition, text-based emotion recognition allow one to automate the process of analyzing user emotions be it in an image or video, revolutionizing the ways of user feedback analysis.

Researchers have made efforts to implement AI-driven solutions to make UX evaluation easy and efficient (Aviz et al., 2019; Souza et al., 2022). The literature on improving individual elements such as

CONTACT Yongmin Li yongmin.li@brunel.ac.uk Department of Computer Science, Brunel University London, Uxbridge, UK

facial recognition in UX by Afriansyah et al. (2021), emotion detection using EEG signals (Gannouni et al., 2023), eye tracking and tracking of mouse clicks, etc. represented by Novák et al. (2023), Qu et al. (2017), and Souza et al. (2022) could be found. Some work in building a framework for UX evaluation by Cordeiro et al. (2024) and Drungilas et al. (2024) can be seen.

Emotions are complex, and it might not be possible to grasp the intentions and feelings of the user through a single modality of recognizing emotions. Integration of data gathered through the combination of different modalities such as face, speech, etc. provide deeper and meaningful insights. It also helps to capture any anomalies detected in the video. As demonstrated by Cordeiro et al. (2024), the integration of facial and speech emotion recognition methods can be a powerful tool to improve UX.

Among all the applications of emotion recognition technology, little research and solutions have been provided focusing on user emotions in an end-to-end workflow. The practice of manually analyzing the change and evolution of user emotions while reviewing a service/product in a recorded session still remains. UX designers re-watch the whole session number of times to keep track of emotion nuances, limiting the insights available to them.

To address the above mentioned problem, this study aims to automatically extract meaningful and actionable insights from user research videos, focusing on the use of facial, speech-to-text, and text-based emotion recognition modalities for data gathering and integration to capture nuances in user emotions expressed in the video, and to further extract and generate insights through a conversational analyst tool. The main contributions of the work include:

1. A novel conceptual framework is introduced for automating insight extraction from user testing videos, reducing the need for manual review.
2. It performs multimodal analysis using facial, speech, and text-based emotion recognition to capture rich emotional data.
3. Data integration through temporal alignment and fusion combines multimodal inputs for cohesive understanding and detection of behavioral patterns.
4. Interactive insight extraction is enabled via interactive AI, allowing designers to query emotional data naturally and efficiently, improving the scalability and effectiveness of UX evaluation.

## 2. Literature review

Emotion Recognition (ER) has been studied for more than a decade. Various methods have been introduced to recognize emotions. Some of which include recognizing emotions from face, speech, tone, text, physiological signals, body language, etc. In Saxena et al. (2020), extensive research has been conducted on various techniques for recognizing emotions through artificial intelligence (AI) in the past decade. The paper covers four modalities for ER, namely face, text, audio and physiological signals, where the experiments have shown that majority of the work has been done in facial, followed by textual and audio emotion recognition. Knowing these technologies and their better use could be an efficient way to handle complex problems on healthcare, media, customer service, and especially Human-Computer Interaction (HCI).

Among all the applications, its importance in User experience (UX) research and usability testing needs to be emphasized. UX evaluation (User Testing) is a process to gain insight into user satisfaction with using / reviewing a product or service. It analyzes how well the product has met the expectations of the user. Their valuable response helps in identifying strengths, weakness, and areas of improvement. According to Novák et al. (2023), the user experience combines the physical and technical aspects of the product with the cognitive processes of the user, focusing on the emotional impacts and satisfaction during the interaction with the product, while usability tests usually focus on the performance of tasks such as the execution time of a task and the number of clicks (Vermeeren et al., 2010). Various methods have been discovered to evaluate the UX process. In Novák et al. (2023), it is stated that UX is usually measured quantitatively on objective data at its core or qualitatively, where usability testing is considered at its core. Nearly 96 methods are reported in Vermeeren et al. (2010) for UX evaluation through comprehensive research, emphasizing the difference between usability testing and UX evaluation. According to the authors, the relationship between UX and usability testing is intertwined, but

objective usability testing is not a sufficient measure for subjective UX evaluation, as it focuses on how the user feels about a system/service.

Many applications of AI technology in the field of UX have focused on usability testing where they keep track of eye, keyboard input, number of mouse clicks, etc. Work has been done in objective evaluation of UX with very little focus on subjective part. In Souza et al. (2022), a framework is presented for UX evaluation focusing on various tracking techniques such as eye and mouse tracking, keyboard inputs, self-assessment questionnaire to categorize users in terms of performance profile (Qu et al., 2017). proposed an eye tracking technology to objectively evaluate UX for smartphone APPs.

### 2.1. Emotion models and theories

Understanding emotion models and theories helps contextualize implementation. For decades, psychologists and researchers have proposed different theories and models related to human emotion. In Ong et al. (2015), model of a lay theory of emotions explains how the observer, called an agent by the lay theory, infers about the target of reasoning. The presence of emotional stimuli and the interaction of these stimuli with other mental states such as goals, generate emotions within the agent. External manifestations of these emotions include speech, body language, facial expressions, and future actions. In Lopatovska and Arapakis (2011), the emotion theories are categorized into two different views as manifestation and structure.

The author states that emotional reactions can arise from either cognitive judgment or bodily responses when focusing on the manifestation point of view, whereas the structural point of view follows discrete and continuous approaches. The discrete approach as a consideration of universally recognized basic emotions (such as fear, anger, disgust, happiness, sadness, and surprise) is described in Ekman (1992). The continuous approach considers two or more dimensions that describe different emotions, as can be seen in the circumplex model of affect proposed by Russell (1980). The model distributes the emotions in a two-dimensional circular space containing arousal and valence. The vertical/horizontal axes represent the arousal/valence, where the center of the circle represents the medium level of arousal and the neutral level of valence. In addition to this, six basic emotions on the border of Russell's proposed model are added in Fernández-Caballero et al. (2016).

### 2.2. Emotion recognition modalities

A facial recognition (FR) system automatically detects and identifies human faces from a digital image or video frames from a video source (Li et al., 2001, 2003; Sharma & Gupta, 2013). Researchers approach the problem of facial recognition systems in different ways. In Kortli et al. (2020), the researcher presents state of the art of existing facial recognition techniques through three different approaches such as local – uses features with partially defined face such as Local Binary pattern (LBP), holistic – uses features which describes complete face as a model such as Principal Component Analysis (PCA), Eigenfaces, Support Vector Machine (SVM), Convolutional Neural Network (CNN) and hybrid – combination of local and holistic (Li et al., 2000, 2001, 2003), whereas in Setiowati et al. (2017), the researcher divides the solution to the FR problem into two categories of non-deep learning methods– Eigenface, Fisherface, SVM, LBP and deep learning (DL) methods – Multi-layer Perceptron (MLP) and CNN.

In line with author's approach in Setiowati et al. (2017), researchers in Canal et al. (2022), Ko (2018), and Moolchandani et al. (2021) also categorize facial emotion recognition (FER) methods into the conventional and DL approach. The general steps involved in developing FER model include data pre-processing, feature extraction, classification and finally results and validation. Three types of data pre-processing steps, that is, gray scale conversion, face detection, and dimensionality reduction, are presented by Canal et al. (2022). Whether a conventional approach or DL is being used, it is usually decided by the methods used for feature extraction and classification. Researchers have different opinions about which approach works best. According to Kortli et al. (2020), local feature techniques are better in terms of complexity, rotation and accuracy, whereas in Setiowati et al. (2017) experimentation

concludes that DL methods are more promising for facial recognition reporting an accuracy of 94.67% of low-high complexity for DL methods and 90.6% of low complexity for non-DL methods.

Moving forward to other emotion recognition technologies for speech and text, researchers have made an impressive contribution. A Speech Emotion Recognition (SER) system extracts and classifies the existing emotions of the target speaker from a pre-processed speech signal. A comprehensive survey on various SER methods is demonstrated in Wani et al. (2021). SER systems also follow the same workflow of data pre-processing, feature extraction, classification, and evaluation. Different ways of speech processing such as framing, windowing, normalization, noise reduction, etc. Different classifiers such as Artificial Neural Networks (ANN), KNN, SVM, deep neural networks (DNN), recurrent neural networks (RNN), CNN, etc. are broadly discussed in Wani et al. (2021). Another technology that could be used is speech-to-text (STT), which recognizes speech from an audio or video and converts it into text.

Different methods for STT and text-to-speech (TTS) have been reviewed in Nagdewani and Jain (2020). The basic process discussed in the paper for STT involves feature extraction, word matching using acoustic word models, sentence matching using syntax and semantics and finally language modeling to text.

Recognizing emotions from text is often classed as Sentiment Analysis (SA), however, if the textual classification is done on more than just classifying it into positive, negative and neutral, could provide more depth and context to the text. In Acheampong et al. (2020), three ways to approach text-based emotion detection as rule construction, ML and hybrid where rule-based approach uses the grammatical and logical rules to find emotions from a document, the ML approach uses ML algorithms to classify text into emotions and hybrid approach is the combination of both are presented.

In recent years, with advances in the DL methods, most of the research has been done using convolutional neural networks in a varied field of applications. A fine-tuned VGGNet architecture, in Khaireddin and Chen (2021), based on CNN is used to achieve the state-of-the-art single network precision of 73. 28% on the FER–2013 dataset. Teja Chavali et al. (2023)'s researcher uses SVM, CNN and pre-trained VGG–16 models to recognize emotions from facial expression along with age and gender prediction. The author of Pomazan et al. (2023) explores the potential of CNN-based emotion recognition in marketing research for advertising and gains insight into consumer behavior. In Zbaida et al. (2023), CNN-based technology is incorporated into e-learning platforms by comparing five different algorithms (VGG16, VGG19, RESNET50V2, EfficientNETB0, and EfficientB7) on their accuracy to find the best-suited algorithm. However, among all these CNN applications, a recent trend of Vision Transformers (ViT) has emerged for computer vision tasks.

The potential of ViT-based models is explored in Bobojanov et al. (2023) for FER on three datasets; RAF-DB, FER2013 and a clean, augmented and balanced dataset using images from the FER dataset. They conducted a comprehensive evaluation of 13 ViT models on these datasets and concluded a promising success of these models for FER tasks. A ViT method introduced in Huang et al. (2021) recognizes driver expressions by performing data augmentation based on a parallel imaging framework using the StarGAN network with CK + and KMU-FED data sets, achieving a higher recognition rate than CNN and ResNet18.

Following the trend and being inspired by Huang et al. (2021), this paper explores a different ViT based model in the application of user testing to enhance the user experience. Text-based emotion detection would provide more contextual understanding by providing context to emotions. A combination of STT and Text-based emotion detection would provide deeper and nuanced emotional insights than directly using the SER which only focuses on the tone of the speaker to recognize emotions.

### 2.3. Emotion recognition in user testing

When reviewing different applications of emotion recognition, it was observed that little has been done in the field of user testing. Emotion recognition would not only provide deeper meaningful insights by capturing verbal and non-verbal emotions but also help in improving design decisions leading to better product performance. Limited literature was found in which the problem of manually keeping track of user emotions, be it facial or speech, has been addressed. The author in (de Souza Veriscimo et al., 2021) also agrees with the fact that little has been done in the evaluation of UX through automated

emotion recognition. The author presents a systematic review to implement FER in UX evaluation, thus concluding the need to address the lack of standardization and modernization of tools, procedures, and evaluation criteria for emotion recognition in UX.

Several efforts have been made to narrow this gap. In Razzaq et al. (2023), a hybrid multimodal emotion recognition framework was proposed for UX evaluation, achieving an average accuracy of 98.19% in detecting four emotions: happiness, neutral, sadness, and anger. An UX evaluation model, called UXAPP, that uses an automated tool to calculate the user experience rating of a digital product or service based on positive, negative, and neutral points was implemented and validated in Cordeiro et al. (2024). The author used state-of-the-art emotion recognition modalities to carry out implementation and conducted experiments to generate a report through nine individuals who carried out the designated task and finally compared the results with the generated report of UXAPP.

Even after implementing different modalities to automate the process of manually keeping track of user emotions, it would not make sense for UX designers to handle gathered data manually. In order to automate the data analysis/insight generation process and being inspired by Cordeiro et al. (2024), rather than generating a positive, negative and neutral point system, a way to make the user ask questions directly from the gathered data rather than manually analyzing was explored in this project.

The concept of a whole process from evaluating the video to gathering insights and analyzing the gathered data through automation has been demonstrated using this project.

## 3. Conceptual framework and case study

The proposed conceptual framework to address the stated problem is illustrated in Figure 1, including three stages of user research input video, data preparation, and user interaction.

First, the UX designer or researcher provides a prerecorded video session of usability testing to the framework. In the data preparation phase, the input video is separately fed into a facial emotion recognition (FER) model and a speech-to-text (STT) model. The FER model detects the face in the video frame-by-frame and recognizes emotions expressed by the person in the video. The emotions detected are recorded and presented as probability in each frame. The data gathered is saved in a CSV file. The STT model converts the spoken language into written text i.e., audio transcription while recording the start and end time of each statement of the speaker. The output text is then fed to the Text-based emotion detection model which classifies the text into different emotions described in the model used for classification with the confidence score of the detected emotion. The final output is then saved in a separate CSV file. The data from both files are then integrated into a final data file. This is further fed to the AI chatbot through the conversational analyst tool for user interaction with the data. Finally, the user can interact with the chatbot for any queries related to the video provided, reducing the manual labor of re-watching the video for emotion tracking. The integration of facial and speech expression data would also enhance the chances of finding discrepancies and anomalies in the video to get in-depth insights.

### 3.1. Facial emotion recognition

Facial emotion recognition (FER) is used to detect and recognize faces in an image or video and classify them into basic emotions, providing deeper insights into a person's feelings through facial expressions. MTCNN network, which outperforms state-of-the-art CNN networks (Zhang et al., 2016), was used to detect faces and facial emotions were detected using ViT, noted for its efficiency and strong performance on image classification tasks (Dosovitskiy et al., 2020). These choices align with evidence highlighting the superior accuracy of DL techniques (Setiowati et al., 2017).

In this study, a pre-trained ViT model was used for facial emotion recognition as shown in Figure 2. In order to choose a suitable model from 10 pre-trained working ViT models available on website called HuggingFace, the models were evaluated on the three benchmark datasets i.e., FER-2013, AffectNet and CK + 48.

The FER − 2013 dataset contains 28,709 training and 5404 testing $48 \times 48$ pixels grayscale images of 7 emotions namely angry, disgust, fear, happy, sad, surprise and neutral. The AffectNet dataset contains

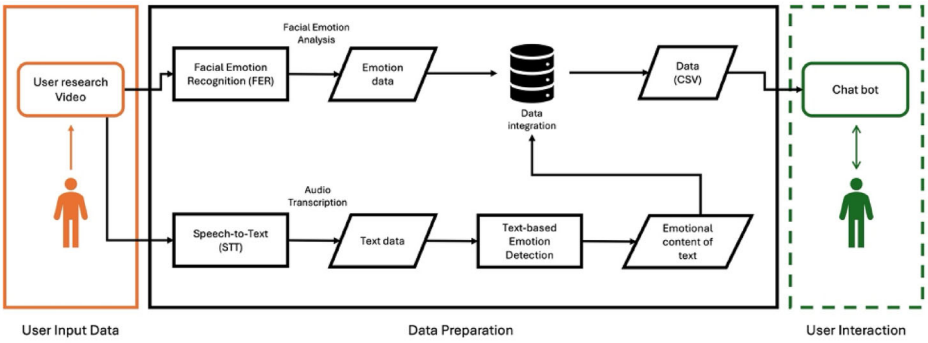


**Figure 1.** Project architecture.

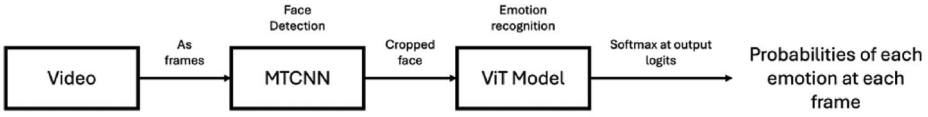


**Figure 2.** Facial emotion recognition workflow.

high dimensional colored 37,553 training, 3200 testing and 800 validation images of 8 emotions namely angry, disgust, fear, sadness, happy, surprise, contempt and neutral. The CK + 48 dataset contains 981 ($48 \times 48$ pixels grayscale) images of 7 emotions namely angry, disgust, happy, sadness, fear, surprise and contempt.

The model selection was done by evaluating the available models on the three datasets on the basis of accuracy, F1-score, precision and recall. The performance of models on different datasets can be seen in Table 1 below:

The Performance variations across datasets can be attributed to domain differences and pre-processing factors. CK+ contains controlled grayscale images, whereas AffectNet includes diverse, real-world color images, leading to potential domain-shift effects. Furthermore, variations in face alignment and color handling may have influenced model performance. Observing the performances on different datasets, Model 1 (trpakov/vit-face-expression) was chosen for Facial Emotion Recognition process proving to have more generalization ability and therefore can be used in different scenarios. The specifications of the model can be found at Trapkov (2024).

It is important to note that different class sets are used across datasets (e.g., presence of "contempt"), and no explicit label mapping or harmonization was applied before computing these metrics. Therefore, the reported results should be taken only as a reference for model selection rather than for cross-dataset evaluation.

As the focus of this work is on the conceptual framework of UX insight extraction, we have only included the point estimates in this paper. Additional analyses such as confusion matrices, domain shift, variance analysis, confidence intervals, and significance tests are indeed important for comprehensive performance evaluation. Details of these can be found in Bobojanov et al. (2023), Canal et al. (2022), de Souza Veriscimo et al. (2021), and Khaireddin and Chen (2021).

Commercial products such as FaceReader have been reported to outperform other facial expression recognition systems in accuracy and robustness for standardized expressions (Noldus Information Technology, 2024). However, they are not open-sourced and require licensing. Therefore, our study focuses on the open-source alternatives discussed above.

The user testing videos for the demonstration of the concept were chosen from YouTube. The two chosen videos were labeled as Testing_video1 and Testing_video2 for anonymity. Usability testing videos are usually large and, due to hardware limitations, the frame-rate of video file was decreased from 30 fps to 15 fps for Testing_video1.mp4 along with some video quality compressions. The Testing_video1 was also cut short to 9 mins and 32 s, removing the introduction and 1 task performed by the participant, from 14 mins and 39 s. It is important to note that reducing frame-rate and compressing

**Table 1.** Performance evaluation of models on different datasets.

| | Performance evaluation on datasets | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FER–2013 | | | | CK + 48 | | | | AffectNet | | | |
| Model No | Accuracy | Precision | Recall | F1-Score | Accuracy | Precision | Recall | F1-Score | Accuracy | Precision | Recall | F1-Score |
| 1 | 0.7115 | 0.7076 | 0.6935 | 0.6997 | 0.3649 | 0.2648 | 0.3060 | 0.2802 | 0.2675 | 0.3177 | 0.2675 | 0.2233 |
| 2 | 0.3536 | 0.3433 | 0.3094 | 0.2751 | 0.2120 | 0.3116 | 0.2460 | 0.1991 | 0.2128 | 0.1594 | 0.2128 | 0.1604 |
| 3 | 0.0906 | 0.6176 | 0.1417 | 0.0619 | 0.0479 | 0.5827 | 0.1164 | 0.0199 | 0.1019 | 0.0418 | 0.1019 | 0.0542 |
| 4 | 0.1106 | 0.1308 | 0.2222 | 0.0945 | 0.3191 | 0.3331 | 0.4253 | 0.2347 | 0.1297 | 0.0980 | 0.1297 | 0.1055 |
| 5 | 0.1209 | 0.5090 | 0.1346 | 0.0977 | 0.1458 | 0.5146 | 0.1876 | 0.1162 | 0.1316 | 0.0620 | 0.1316 | 0.0795 |
| 6 | 0.1123 | 0.1124 | 0.2244 | 0.0772 | 0.4139 | 0.4155 | 0.4239 | 0.2776 | 0.1263 | 0.0839 | 0.1263 | 0.0910 |
| 7 | 0.0741 | 0.1318 | 0.1961 | 0.0650 | 0.3848 | 0.4488 | 0.4274 | 0.2629 | 0.1097 | 0.0872 | 0.1097 | 0.0954 |
| 8 | 0.0911 | 0.1203 | 0.2002 | 0.0723 | 0.1804 | 0.1682 | 0.2523 | 0.1300 | 0.1013 | 0.1027 | 0.1013 | 0.0918 |
| 9 | 0.1180 | 0.1074 | 0.2203 | 0.0876 | 0.3170 | 0.3058 | 0.3959 | 0.2549 | 0.1141 | 0.0995 | 0.1141 | 0.1037 |
| 10 | 0.0984 | 0.1182 | 0.0751 | 0.0733 | 0.4098 | 0.3601 | 0.4354 | 0.2859 | 0.1062 | 0.0955 | 0.1062 | 0.0983 |

Model names are as follows: 1. trpakov/vit-face-expression, 2. jayanta/vit-base-patch16-224-in21k-emotion-detection, 3. Hector001/emotion-vit-model-hector, 4. hilmiatha/ViT-Emotion-Classifier, 5. StoneSeller/emotion-classifier-vit, 6. andikamandalaa/vit-base-patch16-224-in21k-emotion-classification, 7. gabrielganan/vit-emotion_classification, 8. andikamandalaa/vit-base-patch16-224-in21k-emotion-classification3, 9. evanrsl/vit_facial_emotion, 10. asyafalni/vit-emotion-classifier.

video quality can alter facial micro-expression dynamics, so such modifications may not be suitable for real-world applications when high-performance hardware is available.

The input video is fed as video frames to the MTCNN architecture. MTCNN group model detects the face and if a face is found, coordinates of the bounding box drawn on the face are recorded, and the face is cropped from the video frame (each frame can be seen as an image). This cropped face is preprocessed using an auto-feature extractor and fed to the pre-trained ViT model. The model recognizes the faces in each frame and classifies them as basic emotions as defined in the model configuration. To get the probabilities of each emotion in each frame, a softmax loss function is applied to the output logits. "Id2label" attributes of the pre-trained ViT model are retrieved from the configuration and class labels (emotions defined in the model) are mapped to their respective probabilities. The generated output data is saved in a CSV file for later use.

The total duration of the generated video was 9 min and 16 sec. All data generated by facial emotion recognition was stored in a CSV file called "emotion_probabilities.csv."

Figure 3 shows that the participant predominantly experienced "sad" emotion for approximately 100 s (400–500 s in the graph). Although some instances of "happy" and "angry" emotions were detected, the participant generally maintained a "neutral" expression (those with low peaks in Figure 3) while performing the assigned tasks. These subtle emotion variations provide important information for UX designers, enabling them to identify specific moments when the participant felt positive or negative while completing tasks.

### 3.2. Speech-to-text and text-based emotion detection

In this project, relevant NLP techniques were applied to perform Speech-to-Text, or Automatic Speech Recognition (ASR), that converts spoken language to text, with applications across domains from media to healthcare.

Text-based emotion detection identifies and analyses the underlying emotions in a text, often regarded as a subset of sentiment analysis. With its various applications in the fields of healthcare, customer support, marketing, and HCI, both technologies are used to unravel the underlying emotions expressed by the participant's speech during user testing. The overall framework is illustrated in Figure 4.

For ASR, the Whisper model was implemented, transcribing the participants' speech. According to Vásquez-Correa and Álvarez Muniain (2023), Whisper outperforms Wav2vec2.0 in accuracy, particularly under uncontrolled acoustic conditions. Based on its robustness and adaptability, Whisper was selected for ASR in this project. The detailed description of architecture and flow of the Whisper model has been discussed in Radford et al. (2022).

For the small dataset used in the case study, no errors were noticed in the transcriptions from Whisper and they were directly used for further text-emotion estimation. With large scale of user research video data, it is potentially a significant issue as mis-transcriptions could propagate to text-emotion errors.

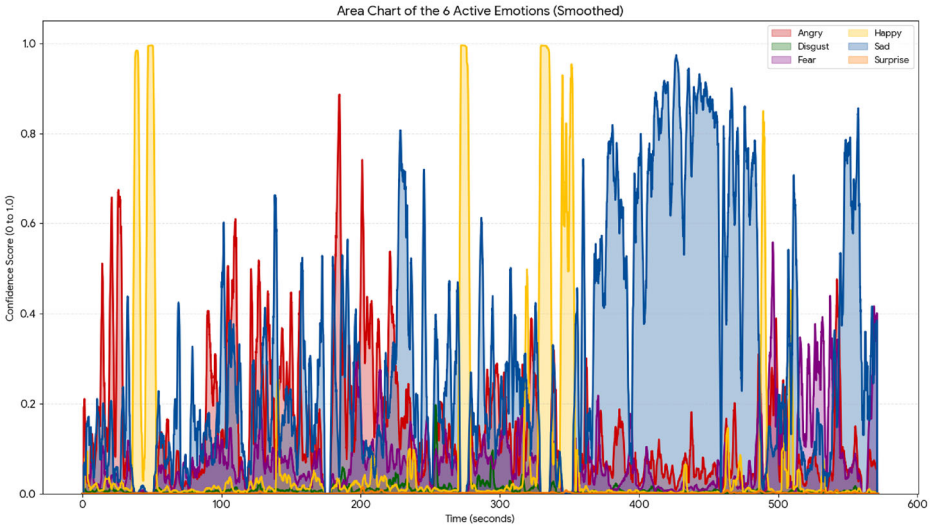


**Figure 3.** Active facial emotions (angry, disgust, fear, happy, sad and surprise) detected at each frame in Testing_video1. The emotion "neutral" is not included for visualization purpose.

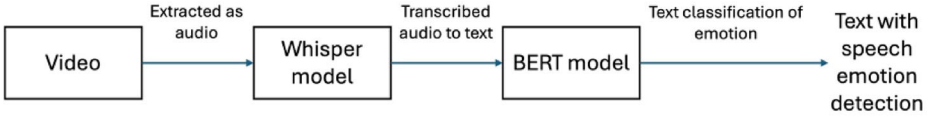


**Figure 4.** Speech-to-text and text-based emotion detection.

The BERT (Bidirectional Encoder Representations from Transformers) model introduced by Devlin et al. (2019), pre-trains deep bidirectional representations from unlabeled text using transformer-based architecture using only the Encoder component. In Devlin et al. (2019), the model is developed based on the original implementation of Vaswani et al. (2017). A detailed description of the model is provided in Devlin et al. (2019).

In this implementation phase, the video was loaded using the "VideoFileClip()" function of the "moviepy.editor" package. The audio was extracted and written in the ".wav" extension. After extracting the audio, the speech recognizer was initialized. The base version of the whisper model was loaded transcribing the recognized speech in the audio file. The balance between performance and computational efficiency is provided by the "base" version of the model. Since the resultant text was time-stamped word-by-word and segmented, it was concatenated to form the complete transcribed text and then tokenized into sentences. This step was necessary for later analysis in the Data Integration part.

A pre-trained classifier and tokenizer model called "bhadresh-savani/bert-base-go-emotion" from HuggingFace was used for text-based emotion classification. In (Savani, 2024), the authors evaluated the model, achieving an accuracy of 96.14% with evaluation and training losses of 0.116 and 0.12, respectively. The model was trained on 169,208 instances, using 3 epochs with 16 as batch size and 31,728 optimization steps. A function was created that took the segments and the classifier as input and classified each segment of text within the list of segments using the given classifier. The results included a list of dictionaries in which each dictionary contained the original text, start and end time of the sentence, emotion class of the text, and its confidence score. Table 2 illustrates the resultant classification of recognized text from audio into emotions. The resultant data was saved in a CSV file named "speech_probabilities.csv."

Figure 5 illustrates that the participant's statement had different emotions over the course of the video. Between 369 and 391 s, the emotion of "confusion" was quite dominant, indicating the participant's discomfort in completing the task during that period. Similar to the results from facial emotions in Figure 3, most of the textual emotions are in "neutral," although at places the user also had "approval," "sad" and "excitement" expressions, thus providing deeper insights.

**Table 2.** Text classification into emotions from recognized speech of Testing_video1.

| Text | Start | End | Emotion | Confidence |
|---|---|---|---|---|
| We'll be going to be posting that next part of the task. | 166.92 | 170.52 | Neutral | 0.3332 |
| And I'm just going to kind of get us back to the homepage so we can restart. | 171.8 | 176.72 | Neutral | 0.2554 |
| Okay. | 179.0 | 179.22 | Caring | 0.4461 |
| So next task. | 179.8 | 181.2 | Neutral | 0.9552 |
| You're going camping this weekend but you don't have a tent. | 182.34 | 185.62 | Neutral | 0.2950 |
| You want to find and rent a two-person tent to use. | 186.02 | 189.0 | Neutral | 0.7209 |
| Use Surfboard Board to accomplish this. | 189.52 | 191.42 | Neutral | 0.8308 |
| Okay. | 191.42 | 192.68 | Caring | 0.4461 |
| So I'm going to get here this time. | 193.62 | 196.42 | Neutral | 0.3660 |
| And I'm looking for a tent. | 197.32 | 199.64 | Neutral | 0.7322 |
| I want a two-person tent. | 201.74 | 205.16 | Desire | 0.4894 |
| I'm not really sure if I should click the menu looking thing or take an icon. | 207.0 | 214.26 | Confusion | 0.4445 |

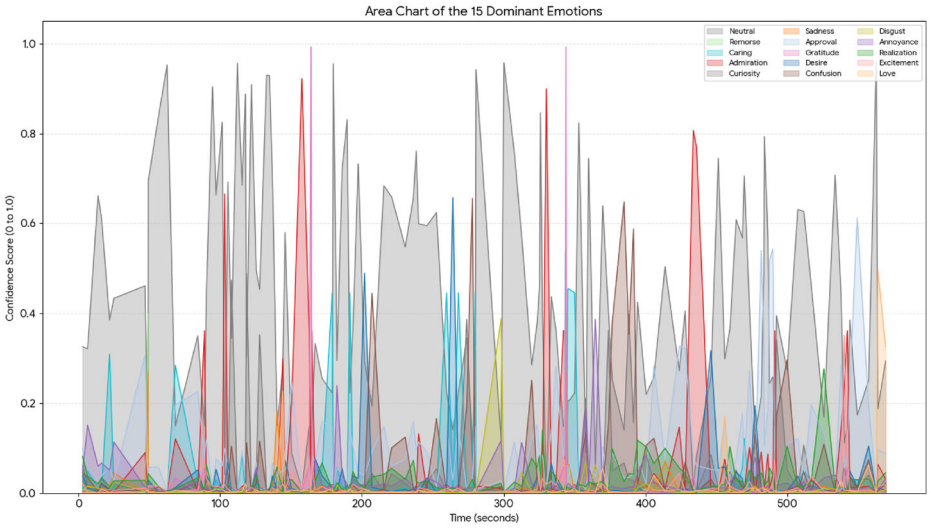


**Figure 5.** Emotions detected from participant's speech over time from Testing_video1.

### *3.3. Data integration*

Data integration is a crucial step in this project. Analyzing integrated data from different modalities such as facial expressions and speech emotion detection would have a comprehensive and insightful analysis of user research videos. A single model may not be sufficient to grasp the emotions of a person. As discussed in emotion theories and models, emotions can be expressed in different ways such as voice, body language, face, etc., and a combination of different modalities would help to capture the whole picture of emotions expressed in the video. Other benefit of joining the data is that it would help find any anomalies or patterns such as instances where the facial expressions is happy, but the speech suggests sarcasm or say otherwise, or instances where they align perfectly. This approach would provide a more accurate interpretation and reliability of the collected data.

Sometimes, data generated in one modality is prone to errors, missing or misleading data, thus the integration process would mitigate these errors by providing additional context and validation. For example, while detecting emotions from one of the user research videos (which was later not used in experiments), the FER model was not able to detect any face and reported 80% of the data as empty, that is, No face detected, while speech emotion detection went smoothly and classified the data into emotions. In such scenarios, even though some facial data were available, the speech data retained the emotions expressed by the participant, which was better than nothing.

In addition, data integration would help analyze emotions in context, leading to more meaningful insights. For example, a combination of sad expression with a negative speech tone would suggest a deeper level of dissatisfaction from the participant.

In terms of ethics and responsibility, the combination of different data sources would respect the complexity of human emotions by reducing misinterpretations, and from a practical perspective, data integration would reduce the complexity of interpreting and handling multiple streams of data

separately, allowing more streamlined data processing. In this project, data integration was implemented keeping in mind the benefits it offers.

For data integration, two concepts of Temporal alignment and Temporal fusion are used. Temporal alignment refers to synchronizing data from different data sources based on their time-related attributes, whereas Temporal Fusion refers to combining different time-aligned data from different sources into a single representation to obtain a more comprehensive view of the data.

With the help of AI, it is now possible to interact with the data directly without the need to learn SQL queries or hard coding, making the data analysis process automated. In this project, one of the possible solutions, a Python library called PandasAI, was used, which allows users to interact with the data in a natural language way, also known as a conversational data analyst tool. The link can be found at Sinaptik-AI (2024).

The tool allows to use different Large Language Models (LLMs) such as default BambooLLM, OpenAI models, Google PaLM, Google Vertexai, Azure OpenAI, HuggingFace via Text Generation, Amazon Bedrock models, IBM watsonx.ai models, and local models such as Ollama and LM Studio. Incorporating such technologies in the application of User Testing would enable designers to evaluate user research videos efficiently and in less time.

**Data collection**: In this project, two types of data have been collected: facial emotion data and speech emotion data. The former were collected from emotions detected by facial expression of the participant. The speech emotion Data were gathered using the combination of STT and Text-based emotion detection, which contain five columns, namely "text" – containing the participant's statements, "start" – the start time (in seconds) of statement made by the participant, "end" – the end time (in seconds) of statement made by the participant, "emotion" – the detected emotion of the text and "confidence" – the confidence score of the detected emotion.

**Data cleaning and preparation**: There were some instances where the model could not detect any face and returned NULL values. To handle these, they were converted to numeric 0 and not deleted considering their significance for later use. To obtain the dominant facial emotion in each frame, the emotion with the highest probability score was used. For data without a record, the dominant emotion was converted to the value "No face detected." The data file contains three columns, that is, Timestamp– time (in seconds) at which the emotion was recorded, "Highest Score" – the confidence score of the detected emotion and "Facial Emotion" – the dominant emotion at the time of detection.

**Temporal alignment** was implemented in the implementation stages of collecting facial emotion data and speech emotion data. The "Timestamp" variable was added to the facial emotion data and the "start" and "end" variable in the speech emotion data. To align the data according to time, the variables were converted to the H-M-S (hours-minutes-seconds) format using the datetime() function from the Pandas library prior to the integration process. This step was taken to accurately compare and combine the two different data sources.

**Temporal Fusion** was done by aggregating or fuzing the facial emotion data based on time intervals of the speech emotion data. In brief, for a participant's statement beginning at 2 s of recorded time and ending at 4 s, the number of frames present in that time interval were aggregated to represent the most frequently detected emotion within that same interval of time.

In addition, the fusion was done considering two scenarios. For the number of frames in the specified time interval, if the emotions detected in all the frames are same, then the average confidence score for facial emotion data at that time would be "mean" of all the data points representing the same emotion. Secondly, if the emotions detected in the facial emotion data are not same within the specified time interval, the most frequent emotion is extracted and the average of confidence score for only those particular frames would be considered as the final score for the detected emotion. For example, if within time interval of 2 s, 30 frames were captured, and for all 30 frames the emotion detected was "neutral," then the Facial Emotion for that time would be "neutral" with the confidence score as the average of Highest Score of all 30 frames. And if among 30 frames, 20 are detected as "happy" and 10 as "neutral," the Facial Emotion for that time would be "happy" as it is the most frequent one with confidence score as the average of Highest Score of only those 20 frames.

Furthermore, if the facial emotion data were not in alignment with the speech emotion data, logic would return the "None" value.

**Table 3.** Samples of the resultant integrated data.

| Text | Start | End | Emotion | Confidence | Avg FER Score | Dominant FER emotion |
|---|---|---|---|---|---|---|
| So I'm going to print it out. | 2.92 | 5.96 | Neutral | 0.326197386 | 0.815838384 | Neutral |
| So I'm going to click List Your Gear. | 6.44 | 9.24 | Neutral | 0.321020454 | 0.834265302 | Neutral |
| And that Own Attent. | 13.82 | 15.86 | Neutral | 0.661488712 | 0.736815128 | Angry |
| So I'm going to click Tent. | 16.3 | 17.94 | Neutral | 0.61031723 | 0.745623811 | Neutral |
| Give your listing a descriptive title. | 22 | 25 | Neutral | 0.385479093 | 0.807047625 | Neutral |
| And I'm going to click the check box listed. | 25 | 32.22 | Neutral | 0.433583885 | 0.77720179 | Neutral |
| Yeah, this is the Tent. | 47 | 48.3 | Neutral | 0.461419493 | 0.994635472 | Happy |
| Sorry. | 49 | 49.2 | Remorse | 0.399663627 | 0.995449468 | Happy |
| Now I'm going to write a description about my Tent. | 49.2 | 55.12 | Neutral | 0.69373101 | 0.955230098 | Happy |
| So this here. | 56.5 | 59.78 | Neutral | 0.839199185 | 0.809589095 | Neutral |
| And this is about three people. | 62.52 | 68.36 | Neutral | 0.952671409 | 0.870423921 | Neutral |
| Two, and if you move a lot in your sleep, it's pretty light. | 68.36 | 82.34 | Caring | 0.28432548 | 0.86390207 | Neutral |
| Okay, I'm going to type that correctly. | 84.12 | 87.24 | Neutral | 0.349999487 | 0.783816478 | Neutral |
| Cool. | 88.86 | 89.98 | Admiration | 0.361270905 | 0.577821005 | Neutral |
| And I'm going to set my price. | 90.32 | 92.08 | Neutral | 0.450031757 | 0.580378833 | Angry |
| Let's see. | 94.56 | 95.42 | Neutral | 0.903746367 | 0.521529078 | Neutral |
| Probably just put it up for like $5 per day. | 97.06 | 101.04 | Neutral | 0.663080513 | 0.725880786 | Neutral |
| It's like $100 Tent. | 101.42 | 102.7 | Neutral | 0.824869871 | 0.676134035 | Sad |
| That would be pretty cool. | 103.1 | 104 | Admiration | 0.665003479 | 0.732171007 | Neutral |
| I'll pick up Address. | 105.36 | 106.62 | Neutral | 0.692145646 | 0.714218024 | Sad |
| Where will people pick it up from you? | 107.9 | 109.84 | Curiosity | 0.474241048 | 0.640520512 | Angry |
| And you see a C. | 112.06 | 114.12 | Neutral | 0.956321597 | 0.821540532 | Neutral |
| And take a look at this. | 115.4 | 117.64 | Neutral | 0.684900999 | 0.881912405 | Neutral |
| Like 10. | 117.78 | 118.28 | Neutral | 0.887729764 | 0.955703162 | Neutral |
| What's this? | 118.54 | 119 | Curiosity | 0.488924116 | 0.960007565 | Neutral |
| And then we've got my product details. | 121.98 | 124.76 | Neutral | 0.90899092 | 0.666946242 | Neutral |
| I'm going to click here. | 125.34 | 126.4 | Neutral | 0.49662587 | 0.53009551 | Angry |

The implementation was done in Python version 3.12.3 in VS Code. Temporal alignment was performed using the datetime() function. A "for" loop was initiated by iterating through each row of the speech emotion data. After finding the facial emotion data that fall within the current time interval of the speech emotion data, temporal fusion was implemented.The fusion intervals were defined according to the start and end times of participants' speech segments in the recorded sessions. This approach ensured that facial emotion recognition results were temporally aligned with meaningful verbal expressions rather than arbitrary time windows. As speech naturally reflects discrete user responses or comments, segment-based fusion provides a more semantically coherent representation of multimodal affective states. A snippet of the final integrated data file is shown in Table 3.

### *3.4. Insight generation*

Among all the available options, OpenAI's GPT- 4o was used to evaluate the data collected by querying from the resultant integrated data. To get an overview of the emotions expressed by the participant/user, there is always a question about the dominant emotion detected. When asked the same question, it responded as shown in Table 4 Q1. The "neutral" emotion detected as the dominant facial emotion also aligns with previous findings that show a positive result of using the LLM as a conversational data analyst tool.

Knowing when the user felt happy helps to capture the strengths of the product. In Testing_video1, there were instances detected where the user was happy. Due to data integration, it is now possible to extract insights about the timing and part of the video when this happened. In Q2 in Table 4, it can be seen that the user was happy at 330.08 s in the video with the statement and the confidence score of the text.

Confusion or frustration are emotions that depicts when a person is not able to understand the task or felt annoying. Insights into these emotions help to find areas of improvement for UX designers. The response in Table 4 Q3 illustrates instances in which the user was unable to perform tasks conveniently.

Statements made by the user such as "I'm not really sure how to find," "I clicked it and then I wasn't really sure exactly how to filter three results." are meaningful insights for a UX designer that would allow them to track down the time when the task was being performed and which task was being performed that confused the user. From the above figure, it can also be observed that the facial emotion is sad and the speech emotion is confusion, align with each other, making a strong impression of the statement and providing an example of benefits of the data integration step performed earlier.

**Table 4.** The Q/a session through a conversational-AI tool.

| Number | Question & answers | | | | | | |
|---|---|---|---|---|---|---|---|
| **1** | **What was the most dominant facial emotion and how many times was it detected?** | | | | | | |
| | "type": "string," "value": "The most dominant facial emotion is 'neutral' and it was detected 63 times." | | | | | | |
| **2** | **At what time did the user feel the most happy? Can you provide the statement and the confidence score?** | | | | | | |
| | "type": "string," "value": "The user showed the most happy face at 330.08 s with the statement: 'Returning to the one I posted, it sounds pretty awesome' and confidence score: 0.899553418." | | | | | | |
| **3** | **How many times has the user felt confusion/frustration in the video, and can you list the corresponding data?** | | | | | | |
| | **index** | **text** | **start** | **end** | **emotion** | **confidence** | **avg_fer_score** | **dominant_fer_emotion** |
| | 50 | I'm not really sure if I should click the menu looking thing or take a icon. | 207 | 214.26 | confusion | 0.444 | 0.582 | neutral |
| | 63 | I'm not really sure how to find. | 277.82 | 279.86 | confusion | 0.655 | 0.497 | neutral |
| | 92 | I clicked it and then I wasn't really sure exactly how to filter three results. | 376.54 | 382.46 | confusion | 0.381 | 0.785 | sad |
| | 93 | I'm not sure what that was. | 384.84 | 387.62 | confusion | 0.648 | 0.633 | sad |
| | 95 | I'm not really sure these three I can't do. | 391.46 | 393.54 | confusion | 0.587 | 0.725 | neutral |
| | 121 | "So I'm never used Airbnb or anything, but I guess the most similar product I have used" | 499.8 | 507.4 | confusion | 0.296 | 0.727 | neutral |
| **4** | **Was there a time where no face was detected in the video?** | | | | | | |
| | **index** | **text** | **start** | **end** | **emotion** | **confidence** | **avg_fer_score** | **dominant_fer_emotion** |
| | 40 | And I'm just going to kind of get us back to the homepage so we can restart. | 171.8 | 176.72 | neutral | 0.255 | 0 | No face detected |
| | 84 | No worries. | 346.1 | 346.72 | caring | 0.454 | 0 | No face detected |
| | 132 | "I mean, I guess I like to see some plus video of like here either going one path or" | 557.34 | 562.74 | neutral | 0.251 | 0 | No face detected |
| | 133 | another. | 562.74 | 563.44 | neutral | 0.959 | 0 | No face detected |
| **5** | **Can you find anomalies in speech and facial emotions?** | | | | | | |
| | **index** | **text** | | | **speech_emotion** | **facial_emotion** | **avg_fer_score** | **confidence** |
| | 2 | And that Own Attent. | | | neutral | angry | 0.736 | 0.661 |
| | 6 | "Yeah, this is the Tent." | | | neutral | happy | 0.994 | 0.461 |
| | 7 | Sorry. | | | remorse | happy | 0.995 | 0.399 |
| | 8 | Now I'm going to write a description about my Tent. | | | neutral | happy | 0.955 | 0.693 |
| | 11 | "Two, and if you move a lot in your sleep, it's pretty light." | | | caring | neutral | 0.863 | 0.284 |
| | 13 | Cool. | | | admiration | neutral | 0.577 | 0.361 |
| | 14 | And I'm going to set my price. | | | neutral | angry | 0.58 | 0.45 |
| | 17 | It's like $100 Tent. | | | neutral | sad | 0.676 | 0.824 |
| | 18 | That would be pretty cool. | | | admiration | neutral | 0.732 | 0.665 |
| | 19 | I'll pick up Address. | | | neutral | sad | 0.714 | 0.692 |
| | 20 | Where will people pick it up from you? | | | curiosity | angry | 0.64 | 0.474 |
| | 24 | What's this? | | | curiosity | neutral | 0.96 | 0.488 |
| | 26 | I'm going to click here. | | | neutral | angry | 0.53 | 0.496 |
| | 28 | I put three. | | | neutral | angry | 0.603 | 0.93 |
| | 30 | But I'll put three again. | | | neutral | sad | 0.772 | 0.7 |
| | 31 | And I'm going to post my listing. | | | sadness | neutral | 0.717 | 0.18 |
| | 32 | Sweet. | | | admiration | neutral | 0.575 | 0.298 |
| | 34 | So now I'm just going to double check that my post looks good. | | | approval | neutral | 0.646 | 0.248 |
| | 35 | Awesome. | | | admiration | sad | 0.843 | 0.922 |
| | 36 | Cool. | | | admiration | neutral | 0.665 | 0.361 |
| | 37 | Thanks. | | | gratitude | neutral | 0.798 | 0.993 |
| | 38 | Cool. | | | admiration | neutral | 0.844 | 0.361 |
| | 40 | And I'm just going to kind of get us back to the homepage so we can restart. | | | neutral | No face detected | 0 | 0.255 |

(*continued*)

**Table 4.** Continued.

| Number | Question & answers | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 6 | **If I were to map the speech emotions into facial emotions, were there any anomalies where speech emotions did not match facial emotions?** | | | | | | | |
| | **index** | **text** | | | **emotion** | **mapped_fer_emotion** | **dominant_fer_emotion** | |
| | 0 | Sorry. | | | remorse | sad | happy | |
| | 1 | Where will people pick it up from you? | | | curiosity | surprise | angry | |
| | 2 | What's this? | | | curiosity | surprise | neutral | |
| | 3 | So now I'm just going to double check that my post looks good | | | approval | happy | neutral | |
| | 4 | Is there any foot salad? | | | curiosity | surprise | happy | |
| | 5 | Alright. | | | approval | happy | neutral | |
| | 6 | Sorry. | | | remorse | sad | happy | |
| | 7 | Was there anything that was hard or frustrating? | | | curiosity | surprise | sad | |
| | 8 | Because it kind of looks like a different thing, I guess. | | | approval | happy | sad | |
| | 9 | The price, yeah, one the one-sword, it's got a dark picture. | | | approval | happy | sad | |
| | 10 | Yeah, that's a stretch goal. | | | approval | happy | sad | |
| | 11 | Okay, so the feature, I guess, I really valued. | | | approval | happy | sad | |
| 7 | **Can you find where the user recommended any improvement or would like to have something different from what was presented?** | | | | | | | |
| | **index** | **text** | **start** | **end** | **emotion** | **confidence** | **avg_fer_score** | **dominant_fer_emotion** | **anomaly** |
| | 16 | Probably just put it up for like $5 per day. | 97.06 | 101.04 | neutral | 0.663 | 0.725 | neutral | False |
| | 17 | It's like $100 Tent. | 101.42 | 102.7 | neutral | 0.824 | 0.676 | sad | True |
| | 23 | Like 10. | 117.78 | 118.28 | neutral | 0.887 | 0.955 | neutral | False |
| | 56 | It looks like this is a two person tent. | 239.9 | 244.72 | neutral | 0.598 | 0.834 | neutral | False |
| | 68 | It has like results overlapped with this. | 307.42 | 312.08 | neutral | 0.758 | 0.76 | neutral | False |
| | 69 | But it went away like halfway and clicked something else. | 312.58 | 315.66 | neutral | 0.573 | 0.742 | neutral | False |
| | 98 | "Because it kind of looks like a different thing, I guess." | 405.88 | 409.46 | approval | 0.282 | 0.761 | sad | True |
| | 101 | "Yeah, but you can maybe make that like an icon in itself or something." | 427.9 | 431.66 | neutral | 0.405 | 0.86 | sad | True |
| | 106 | "Look at the next tent, because I look at this and I'd be like, " | 455.88 | 459.4 | neutral | 0.298 | 0.784 | sad | True |

(*continued*)

**Table 4.** Continued.

| Number | Question & answers | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 111 | "Other than that, yeah, it looks like this is a person." | 473.28 | 475.88 | neutral | 0.416 | 0.801 | sad | True |
| 115 | It's like a rating system. | 483.84 | 485.1 | neutral | 0.792 | 0.647 | sad | True |
| 122 | would be like a Lister Uber. | 507.4 | 509.26 | neutral | 0.63 | 0.678 | neutral | False |
| 127 | "You know, like Drew Persons." | 538 | 539.2 | neutral | 0.438 | 0.903 | neutral | False |
| **8** | **This data is based on user testing, as part of user experience study, can you tell me the key places where the experience can be improved?** | | | | | | | |
| | **index** | **text** | **start** | **end** | **emotion** | **confidence** | **avg_fer_score** | **dominant_fer_emotion** |
| | 50 | I'm not really sure if I should click the menu looking thing or take a icon. | 207 | 214.26 | confusion | 0.444 | 0.582 | neutral |
| | 63 | I'm not really sure how to find. | 277.82 | 279.86 | confusion | 0.655 | 0.497 | neutral |
| | 92 | I clicked it and then I wasn't really sure exactly how to filter three results. | 376.54 | 382.46 | confusion | 0.381 | 0.785 | sad |
| | 93 | I'm not sure what that was. | 384.84 | 387.62 | confusion | 0.648 | 0.633 | sad |
| | 95 | I'm not really sure these three I can't do. | 391.46 | 393.54 | confusion | 0.587 | 0.725 | neutral |
| | 121 | "So I'm never used Airbnb or anything, but I guess the most similar product I have used" | 499.8 | 507.4 | confusion | 0.296 | 0.727 | neutral |

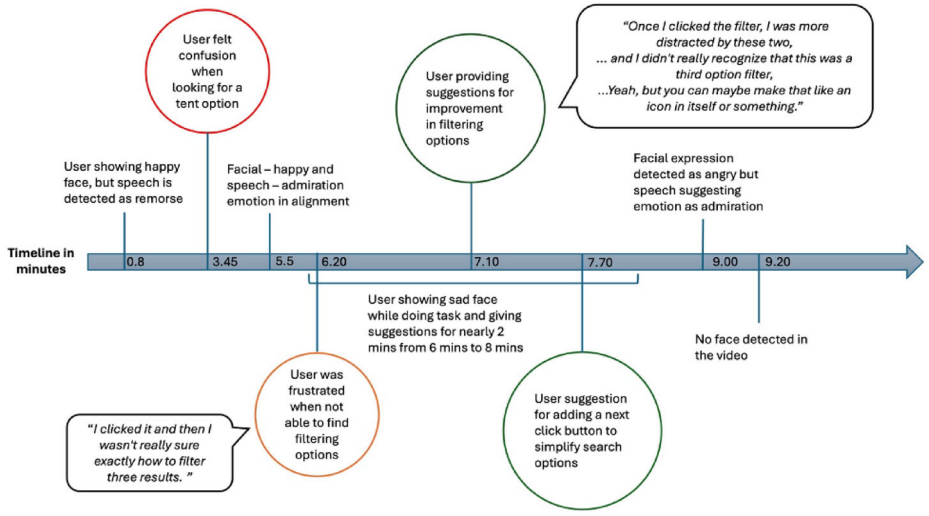


**Figure 6.** Timeline visualization of Testing_video1.

Sometimes, it is possible that data might be lost or not recorded, the simple question asked about whether there were instances when no facial expressions were recorded was made to investigate the same problem. In Table 4 Q4, it can be seen that there are some instances where the face was not detected in the video, but since there is speech emotion recorded, it might not affect the insight gathering process to a greater extent, thus again demonstrating integration of data from different modalities as a wise method for insight gathering.

One of the objectives of the project was to find anomalies through the implementation of different modalities for deeper insight extractions. LLM was able to find instances where the facial and speech emotions did not align as shown in Table 4 Q5 response. A total of 100 anomalies were found.

To further refine the response, a question was asked to which the LLM provided more detailed answer by displaying a generated column called "mapped_fer_emotion" as can be seen in Table 4 Q6. The LLM was able to manipulate the data itself and generate the desired response.

Detecting user speech has a very important benefit for the UX designer. Any suggestions, feedback and advice help to improve the service. Designers had to ask for feedback or conduct surveys after user testing. Keeping this in mind, the next question about instances in which user recommendations were requested and the response can be seen in Table 4 Q7. However, the suggestion made by the user as seen in index 101 is displayed, but with other irrelevant information, shows limitations in handling complex questions.

Through the final question in Table 4 Q8, it was observed that the current LLM is capable of generating responses to questions in which any data in the file are not explicitly mentioned. According to the findings, the designer should focus on areas or duration when the user was confused or annoyed.

A timeline visualization is also presented in Figure 6, illustrating the key highlights and insights found through the implementation of automating user emotion detection from Testing_video1. We make the following observations from the above example:

1. While most results are relevant, specific and accurate, at places they show limitations of the approach in handling complex questions. This is a challenge common to current LLMs.
2. By combining both facial and text information for user emotion recognition and integrating temporal alignment and fusion, we have notably reduced the hallucination effects of LLMs and improved the specificity of insight extraction.
3. The insights generated in Table 4 and timeline visulization in Figure 6, as the final output of the process, provide structured and interpretable summaries for UX analysis.

## 4. Conclusion

User testing or UX evaluation plays an important role in gathering insights for fulfilling the user's needs and expectations. A UX designer needs feedback and suggestions from its customers to improve

the service and meet expectations. One way of doing this is through user testing in which a task is assigned to a participant/user, and the user talks about the strengths and weaknesses he/she experienced while reviewing the product in a recorded session. The role of UX designer is to gather insights from the user testing video and keep track of user's changing emotions, key areas where the experience need to be improved, user's recommendations made while reviewing, etc. Traditionally, to meet all these requirements, designers had to re-watch the whole video again and again to keep track and note everything. And after performing this time-consuming and tedious task, they had to perform data analysis for their key findings.

To address the above problem, we have presented in this paper a conceptual framework for automated insight extraction from user testing videos, where insights are gathered from different emotion recognition modalities such as facial emotion recognition, speech-to-text and text-based emotion recognition. For FER process, a pre-trained ViT based model was implemented which was chosen on the basis of quantitative as well as qualitative analysis. Among 10 available pre-trained ViT models, the selected model performed well on three facial recognition benchmark datasets i.e., FER-2013, CK + 48 and AffectNet, displaying better generalization ability compared to other models. For speech-to-text conversion, OpenAI's Whisper model was chosen and for text-based emotion recognition process a pre-trained BERT based model trained on GoEmotions dataset from Google was implemented.

The FER model recognizes emotions into 7 basic emotions of sad, happy, neutral, angry, surprise, disgust and fear whereas the text-based ED model recognized 27 different emotions. For better understanding, the data integration step was implemented. Not only is this necessary for gathering richer data for analysis, but it was also equally important for UX evaluation. Data integration made it possible to extract deeper, meaningful insights by combining data from different modalities. This step also helps in finding any anomalies or patterns depicted in user behavior.

Data integration was performed using temporal alignment and temporal fusion methods in which the data from different modalities are aligned based on timestamps, and these temporally aligned data was then fused based on time. Finally, the integrated data are fed to a conversational data analyst tool i.e., PandasAI, which allows incorporating different LLMs, through which the designer could interact with the data in a natural language way without the need of learning SQL queries. This would reduce the burden of manually processing the data and make the UX evaluation process more time efficient and convenient. Among available options, Open AI's GPT- 4o was used for data analysis process in which the designer could directly ask questions about the user testing video.

We believe that the multimodal integration of facial and speech data offers a comprehensive approach to understanding user sentiment. Further incorporating a conversational data analyst tool, the proposed framework showcases a novel application of AI in UX research and evaluation, contributing to methodological advancements in the field. The gained insights on user emotions directly inform and improve UX design. Automating emotion detection and data processing reduces the need for manual review, making user testing more scalable and efficient for UX designers.

## Acknowledgment

We thank the Editor and anonymous Reviewers for their valuable and constructive comments. In many places, their input has directly contributed to the revised version of the manuscript.

## Author contributions

CRediT: **Simran Kaur Ghatoray**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; **Yongmin Li**: Conceptualization, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## ORCID

Yongmin Li http://orcid.org/0000-0003-1668-2440

## References

Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, *2*(7). https://doi.org/10.1002/eng2.12189

Afriansyah, Y., Nugrahaeni, R. A., & Prasasti, A. L., 7 (2021). Facial expression classification for user experience testing using k-nearest neighbor. In *Proceedings—2021 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2021* (pp. 63–68). Institute of Electrical and Electronics Engineers Inc.

Aviz, I. L., Souza, K. E., Ribeiro, E., de Mello Junior, H., & Seruffo, M. C. D. R. (2019). Comparative study of user experience evaluation techniques based on mouse and gaze tracking. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web* (pp. 53–56). ACM.

Bobojanov, S., Kim, B. M., Arabboev, M., & Begmatov, S. (2023). Comparative analysis of vision transformer models for facial emotion recognition using augmented balanced datasets. *Applied Sciences (Switzerland)*, *13*(22), 12271. https://doi.org/10.3390/app132212271

Canal, F. Z., Müller, T. R., Matias, J. C., Scotton, G. G., de Sa Junior, A. R., Pozzebon, E., & Sobieranski, A. C. (2022). A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, *582*, 593–617. https://doi.org/10.1016/j.ins.2021.10.005

Cordeiro, R., Sant', G., & Van Erven, A. (2024). Uxapp: Evaluation of the user experience of digital products through emotion recognition.

de Souza Veriscimo, E., Bernardes Júnior, J. L., & Digiampietri, L. A. (2021). Facial emotion recognition in UX evaluation: A systematic review. In M. Kurosu (Ed.), *Human-Computer Interaction. Theory, Methods and Tools (HCI 2021), Lecture Notes in Computer Science* (Vol. 12762, pp. 521–534). Springer.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929. https://arxiv.org/abs/2010.11929

Drungilas, D., Ramašauskas, I., & Kurmis, M. (2024). Emotion recognition in usability testing: A framework for improving web application UI design. *Applied Sciences*, *14*(11), 4773–4775. https://doi.org/10.3390/app14114773

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3-4), 169–200., 5 https://doi.org/10.1080/02699939208411068

Fernández-Caballero, A., Martínez-Rodrigo, A., Pastor, J. M., Castillo, J. C., Lozano-Monasor, E., López, M. T., Zangróniz, R., Latorre, J. M., & Fernández-Sotos, A. (2016). Smart environment architecture for emotion detection and regulation. *Journal of Biomedical Informatics*, *64*, 55–73. https://doi.org/10.1016/j.jbi.2016.09.015

Gannouni, S., Belwafi, K., Aledaily, A., Aboalsamh, H., & Belghith, A. (2023). Software usability testing using EEG-based emotion detection and deep learning. *Sensors*, *23*(11), 5147. https://doi.org/10.3390/s23115147

Huang, Z., Yu, Y., & Gou, C. (2021). *Driver facial expression recognition based on vit and stargan* [Paper presentation]. In 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI) (pp. 254–257). IEEE. https://doi.org/10.1109/DTPI52967.2021.9540071

Khaireddin, Y., & Chen, Z. (2021). Facial emotion recognition: State of the art performance on FER2013.

Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. *Sensors (Switzerland)*, *18*(2), 401. https://doi.org/10.3390/s18020401

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., & Wróbel, M. R. (2014). Emotion recognition and its application in software engineering and games. In R. Wrembel, & M. Szwoch (Eds.), *Human-Computer Systems Interaction: Backgrounds and Applications 3* (pp. 51–62). Springer.

Kortli, Y., Jridi, M., Al Falou, A., & Atri, M. (2020). Face recognition systems: A survey. *Sensors*, *20*(2), 342. https://doi.org/10.3390/s20020342

Li, Y., Gong, S., & Liddell, H. (2001). *Constructing facial identity surfaces in a nonlinear discriminating space* [Paper presentation]. In IEEE Conference on Computer Vision and Pattern Recognition (Vol. 2, pp. 258–263). IEEE.

Li, Y., Gong, S., & Liddell, H. (2001). *Video-based online face recognition using identity surfaces* [Paper presentation]. In The Second International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (pp. 40–46). IEEE.

Li, Y., Gong, S., & Liddell, H. (2003). Constructing facial identity surfaces for recognition. *International Journal of Computer Vision*, *53*(1), 71–92. https://doi.org/10.1023/A:1023083725143

Li, Y., Gong, S., & Liddell, H. (2003). Recognising trajectories of facial identities using kernel discriminant analysis. *Image and Vision Computing*, *21*(13-14), 1077–1086. https://doi.org/10.1016/j.imavis.2003.08.010

Li, Y., Gong, S., Sherrah, J., & Liddell, H. (2000). Multi-view face detection using support vector machines and eigenspace modelling. In *KES'2000. Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Proceedings (Cat. No. 00TH8516)* (Vol. 1, pp. 241–244). IEEE.

Lopatovska, I., & Arapakis, I. (2011). Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. *Information Processing & Management*, *47*(4), 575–592. https://doi.org/10.1016/j.ipm.2010.09.001

Moolchandani, M., Dwivedi, S., Nigam, S., & Gupta, K. (2021). *A survey on: Facial emotion recognition and classification* [Paper presentation]. In Proceedings—5th International Conference on Computing Methodologies and Communication, ICCMC 2021 (pp. 1677–1686). Institute of Electrical and Electronics Engineers Inc.

Nagdewani, S., & Jain, A. (2020). A review on methods for speech-to-text and text-to-speech conversion. *International Research Journal of Engineering and Technology*, *7*(5), 3616–3620.

Noldus Information Technology. (2024). *Facereader: Facial expression analysis software*. Retrieved October 26, 2025, from https://noldus.com/facereader

Novák, J. Š., Masner, J., Benda, P., Šimek, P., & Merunka, V. (2023). Eye tracking, usability, and user experience: A systematic review. *International Journal of Human-Computer Interaction*, *40*(2), 336–354. https://doi.org/10.1080/10447318.2023.2221600

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162. https://doi.org/10.1016/j.cognition.2015.06.010

Pomazan, V., Tvoroshenko, I., & Gorokhovatskyi, V. (2023). Development of an application for recognizing emotions using convolutional neural networks. *International Journal of Academic Information Systems Research*, *7*(7), 25–36.

Qu, Q.-X., Zhang, L., Chao, W.-Y., & Duffy, V. (2017). User experience design based on eye-tracking technology: A case study on smartphone APPs. *International Journal of Industrial Ergonomics*, *61*, 35–43. https://doi.org/10.1016/j.ergon.2017.05.011

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.

Razzaq, M. A., Hussain, J., Bang, J., Hua, C. H., Satti, F. A., Rehman, U. U., Bilal, H. S. M., Kim, S. T., & Lee, S. (2023). A hybrid multimodal emotion recognition framework for UX evaluation using generalized mixture functions. *Sensors*, *23*(9), 4373. https://doi.org/10.3390/s23094373

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Savani, B. (2024). bert-base-go-emotion.huggingface.

Saxena, A., Khanna, A., & Gupta, D. (2020). Emotion recognition and detection methods: A comprehensive survey. *Journal of Artificial Intelligence and Systems*, *2*, 53–79. https://doi.org/10.33969/ais.2020.21005

Setiowati, S., Zulfanahri, Franita, E. L., & Ardiyanto, I. (2017). *A review of optimization method in face recognition: Comparison deep learning and non-deep learning methods* [Paper presentation]. In 2017 9th International Conference on Information Technology and Electrical Engineering, ICITEE 2017 (Vol. 2018-January, pp. 1–6). Institute of Electrical and Electronics Engineers Inc.

Sharma, A., & Gupta, A. (2013). A review paper on facial recognition. *International Journal on Recent and Innovation Trends in Computing and Communication*, *1*(4), 224–228.

Sinaptik-AI. (2024). Github - sinaptik-ai/pandas-ai: Chat with your database (sql, csv, pandas, polars, mongodb, nosql, etc). pandasai makes data analysis conversational using llms (gpt 3.5 / 4, anthropic, vertexai) and rag.

Souza, K. E. S. d., de Aviz, I. L., de Mello, H. D., Figueiredo, K., Vellasco, M. M. B. R., Costa, F. A. R., & Seruffo, M. C. d. R. (2022). An evaluation framework for user experience using eye tracking, mouse tracking, keyboard input, and artificial intelligence: A case study. *International Journal of Human–Computer Interaction*, *38*(7), 646–660. https://doi.org/10.1080/10447318.2021.1960092

Teja Chavali, S., Tej Kandavalli, C., Sugash, T. M., & Subramani, R. (2023). Smart facial emotion recognition with gender and age factor estimation. *Procedia Computer Science*, *218*, 113–123. https://doi.org/10.1016/j.procs.2022.12.407

Trapkov. (2024). trpakov/vit-face-expression.huggingface.

Vásquez-Correa, J. C., & Álvarez Muniain, A. (2023). Novel speech recognition systems applied to forensics within child exploitation: Wav2vec2.0 vs. Whisper. *Sensors*, *23*(4), 1843. https://doi.org/10.3390/s23041843

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. arXiv:1706.03762.

Vermeeren, A. P. O. S., Lai-Chong Law, E., Roto, V., Obrist, M., Hoonhout, J., & Väänänen-Vainio-Mattila, K. (2010). *User experience evaluation methods: Current state and development needs* [Paper presentation]. In Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, NordiCHI '10, New York, NY (pp. 521–530). Association for Computing Machinery.

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE Access*, *9*, 47795–47814. https://doi.org/10.1109/ACCESS.2021.3068045

Zbaida, A., Kodad, M., Mohamed, Y., Benmoussa, N., & Achraf, Z. (2023). *Deep learning and emotion recognition for an e-learning platform*. ResearchGate. https://www.researchgate.net/publication/372230451_Deep_Learning_and_emotion_recognition_for_an_E-Learning_platform

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342

## About the authors

**Simran Kaur Ghatoray** is a research student at the Department of Computer Science at Brunel University London. Her focus of research lies on human computer interaction, artificial intelligence and UI/UX design.

**Yongmin Li** is a Professor at the Department of Computer Science, Brunel University London, UK. He received his PhD degree from Queen Mary, University of London, MEng and BEng from Tsinghua University, China. His research interest covers the areas of data science, artificial intelligence and human AI interaction.