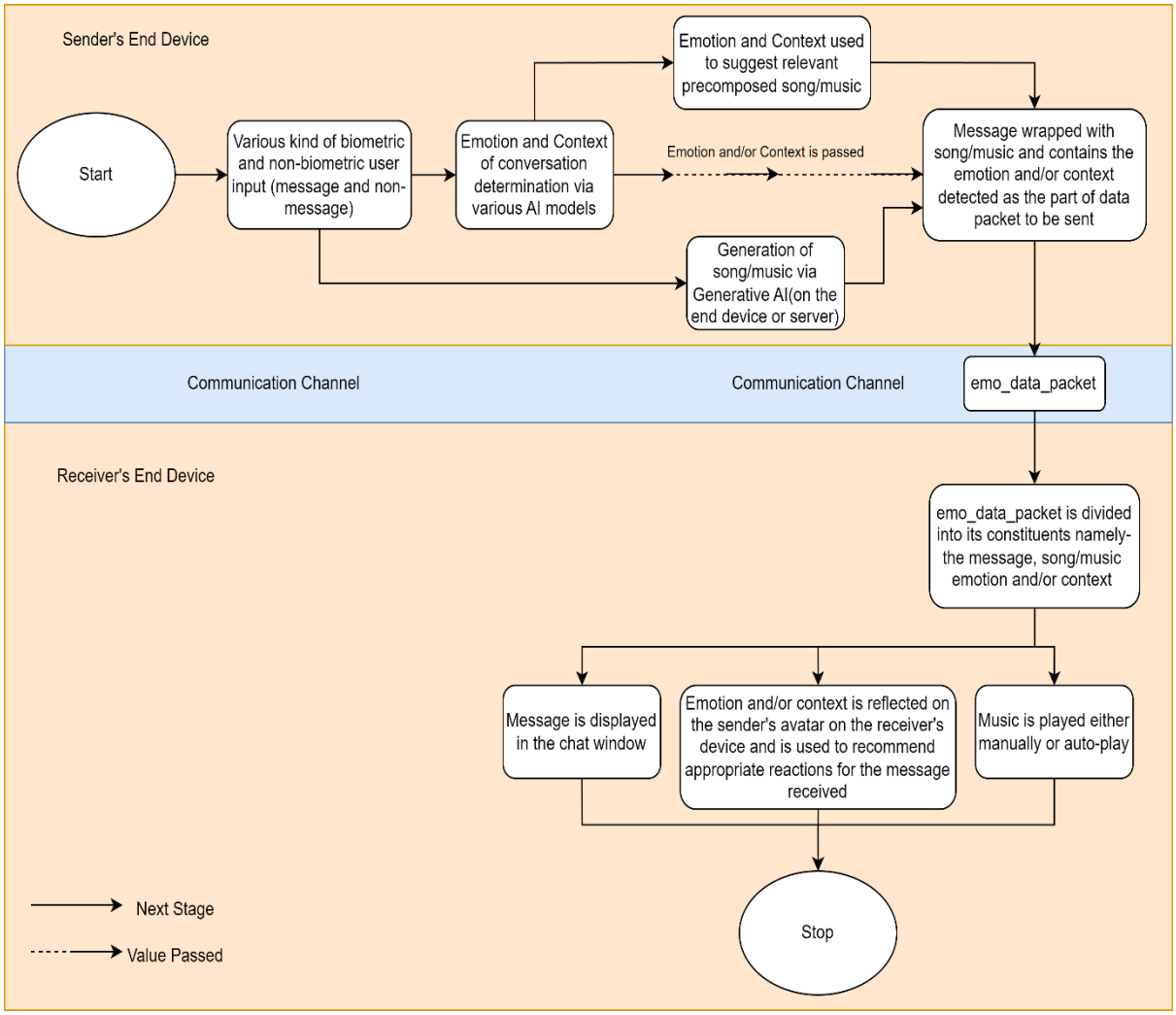

Sender's End Device
Start
Various kind of biometric and non-biometric user input (message and non-message)
Emotion and Context of conversation determination via various AI models
Emotion and Context used to suggest relevant precomposed song/music
Emotion and/or Context is passed
Message wrapped with song/music and contains the emotion and/or context detected as the part of data packet to be sent
Generation of song/music via Generative AI(on the end device or server)
Communication Channel
Communication Channel
emo_data_packet
Receiver's End Device
emo_data_packet is divided into its constituents namely- the message, song/music emotion and/or context
Message is displayed in the chat window
Emotion and/or context is reflected on the sender's avatar on the receiver's device and is used to recommend appropriate reactions for the message received
Music is played either manually or auto-play
Stop
Next Stage
Value Passed

Figure 1

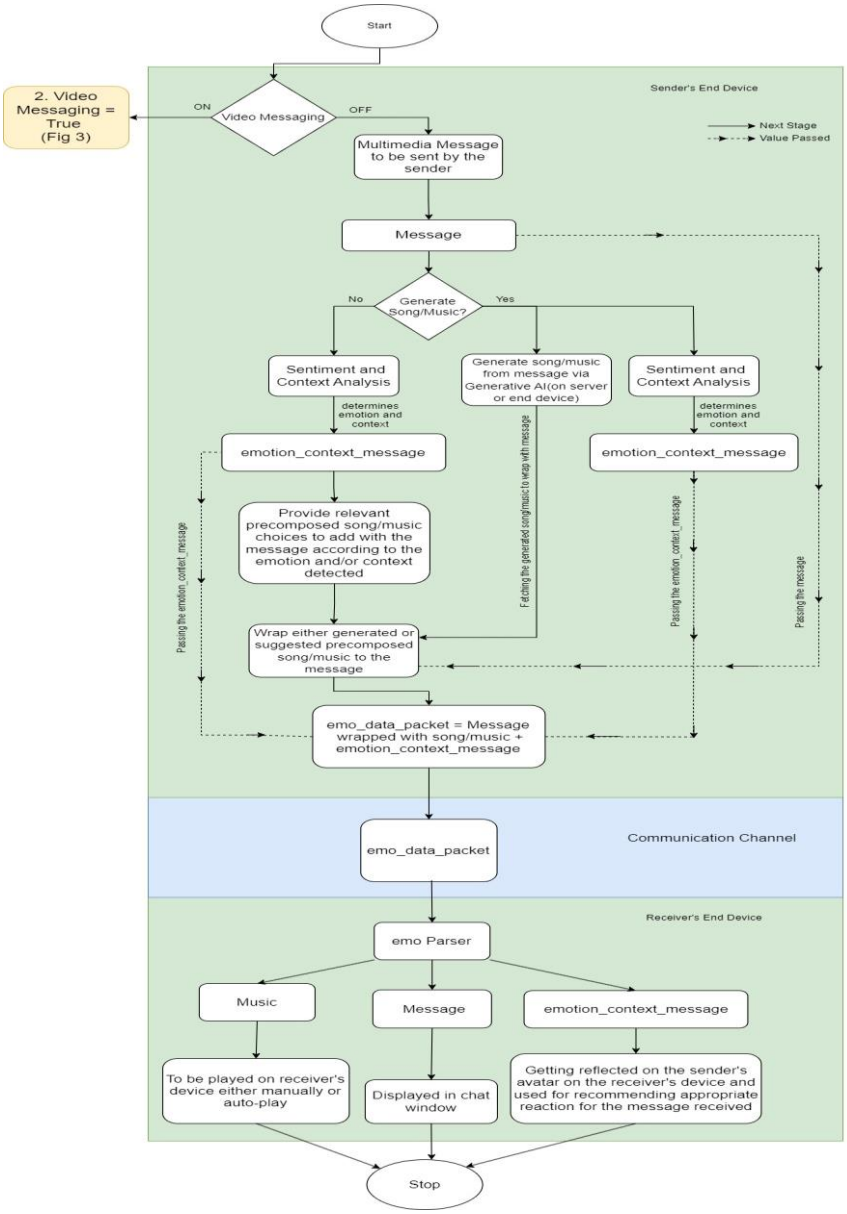

Start
2. Video Messaging = True (Fig 3)
ON
Video Messaging
OFF
Sender's End Device
Next Stage
Value Passed
Multimedia Message to be sent by the sender
Message
No
Generate Song/Music?
Yes
Sentiment and Context Analysis
Generate song/music from message via Generative AI(on server or end device)
Sentiment and Context Analysis
determines emotion and context
determines emotion and context
emotion_context_message
emotion_context_message
Provide relevant precomposed song/music choices to add with the message according to the emotion and/or context detected
Passing the emotion_context_message
Fetching the generated song/music to wrap with message
Passing the emotion_context_message
Passing the message
Wrap either generated or suggested precomposed song/music to the message
emo_data_packet = Message wrapped with song/music + emotion_context_message
emo_data_packet
Communication Channel
Receiver's End Device
emo Parser
Music
Message
emotion_context_message
To be played on receiver's device either manually or auto-play
Displayed in chat window
Getting reflected on the sender's avatar on the receiver's device and used for recommending appropriate reaction for the message received
Stop

Figure 2

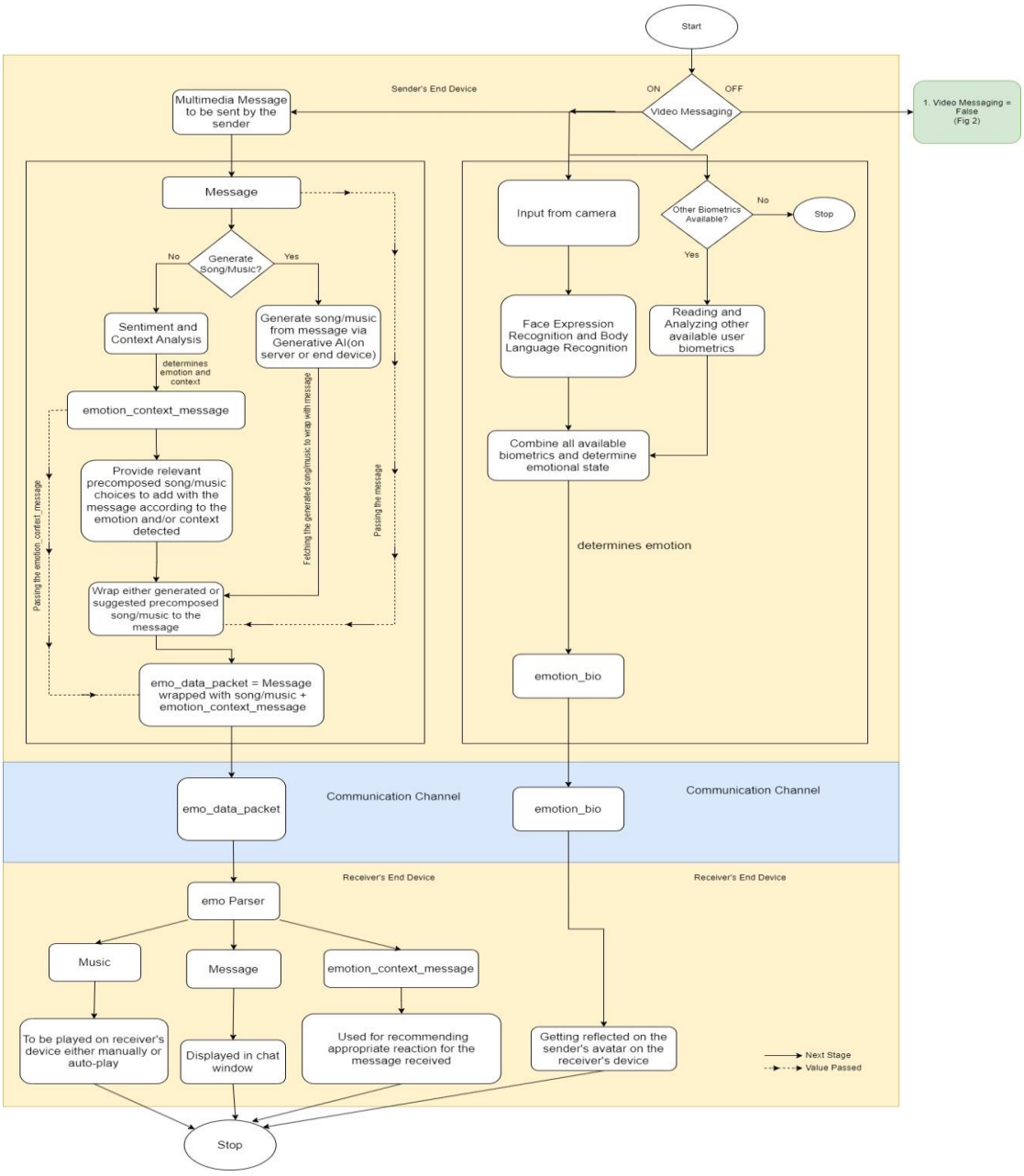

Start
Sender's End Device
Multimedia Message to be sent by the sender
ON
OFF
Video Messaging
1. Video Messaging = False (Fig 2)
Message
No
Generate Song/Music?
Yes
Sentiment and Context Analysis
Generate song/music from message via Generative AI(on server or end device)
determines emotion and context
emotion_context_message
Provide relevant precomposed song/music choices to add with the message according to the emotion and/or context detected
Passing the emotion_context_message
Fetching the generated song/music to wrap with message
Passing the message
Wrap either generated or suggested precomposed song/music to the message
emo_data_packet = Message wrapped with song/music + emotion_context_message
Input from camera
Other Biometrics Available?
No
Stop
Yes
Face Expression Recognition and Body Language Recognition
Reading and Analyzing other available user biometrics
Combine all available biometrics and determine emotional state
determines emotion
emotion_bio
emo_data_packet
Communication Channel
emotion_bio
Communication Channel
Receiver's End Device
Receiver's End Device
emo Parser
Music
Message
emotion_context_message
To be played on receiver's device either manually or auto-play
Displayed in chat window
Used for recommending appropriate reaction for the message received
Getting reflected on the sender's avatar on the receiver's device
Next Stage
Value Passed
Stop

Figure 3

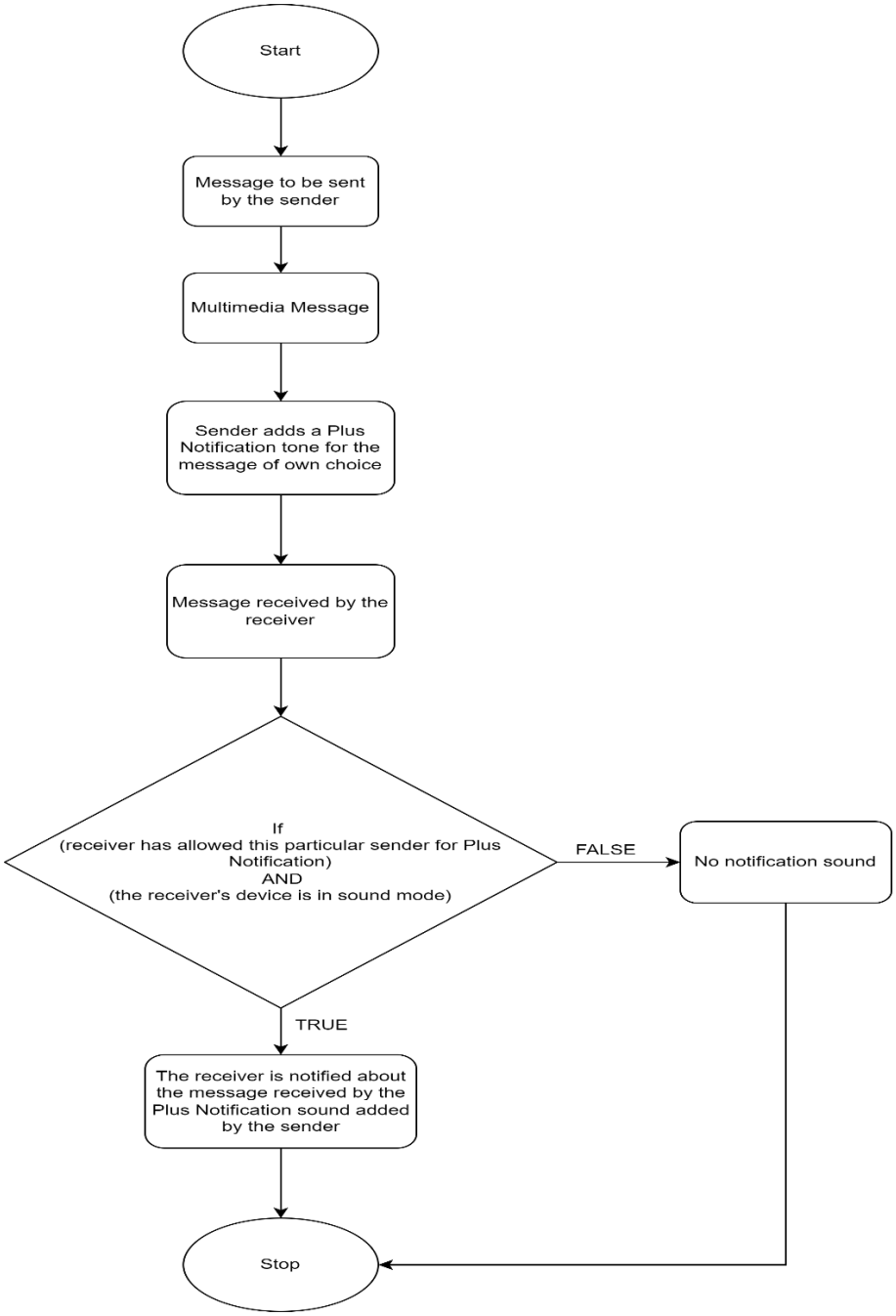

Start
Message to be sent by the sender
Multimedia Message
Sender adds a Plus Notification tone for the message of own choice
Message received by the receiver
If
(receiver has allowed this particular sender for Plus Notification)
AND
(the receiver's device is in sound mode)
FALSE
No notification sound
TRUE
The receiver is notified about the message received by the Plus Notification sound added by the sender
Stop

Figure 4

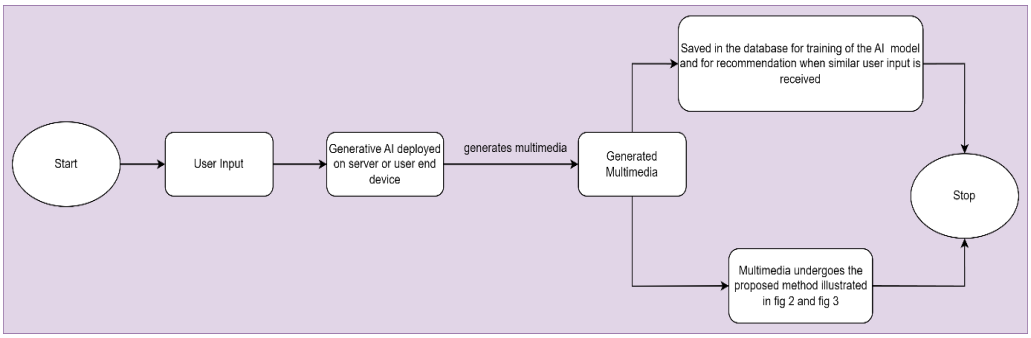

Start
User Input
Generative AI deployed on server or user end device
generates multimedia
Generated Multimedia
Saved in the database for training of the AI model and for recommendation when similar user input is received
Multimedia undergoes the proposed method illustrated in fig 2 and fig 3
Stop

Figure 5

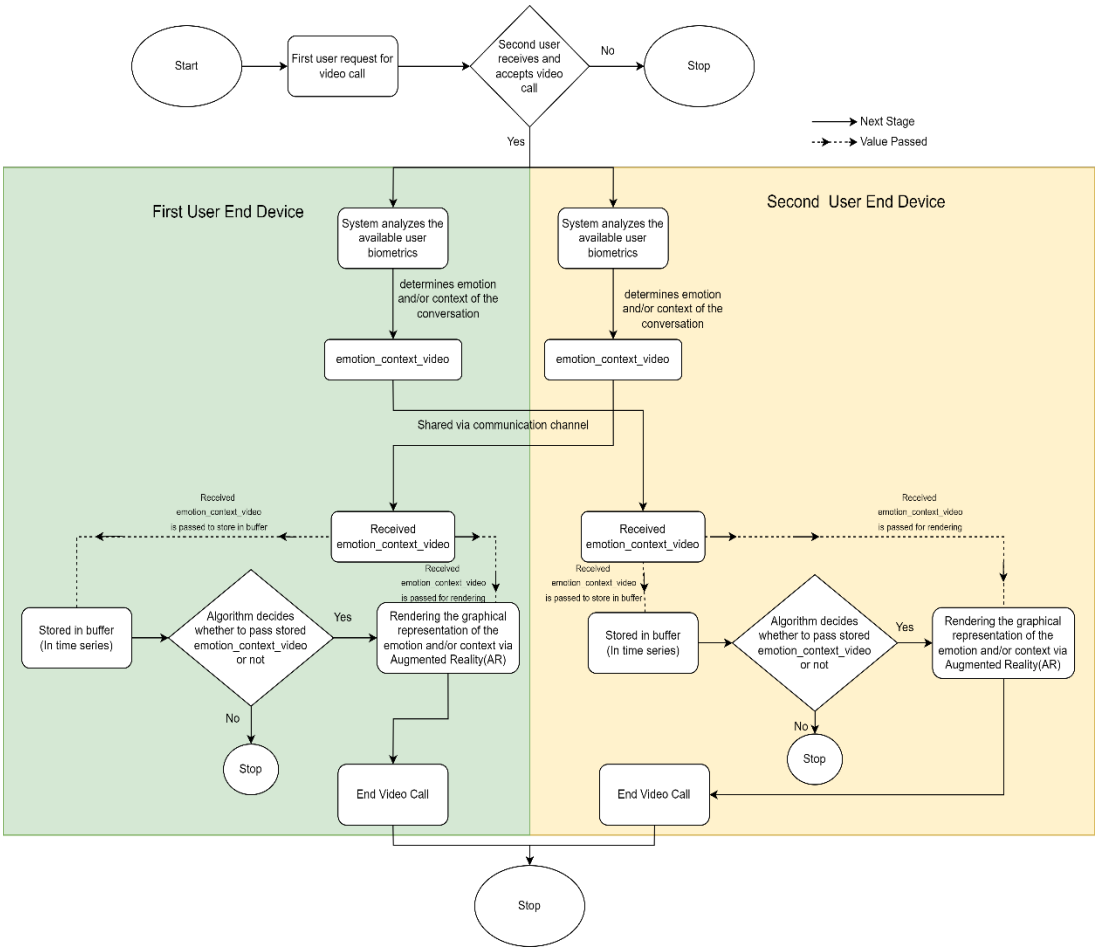

Start
First user request for video call
Second user receives and accepts video call
No
Stop
Yes
Next Stage
Value Passed
First User End Device
Second User End Device
System analyzes the available user biometrics
System analyzes the available user biometrics
determines emotion and/or context of the conversation
determines emotion and/or context of the conversation
emotion_context_video
emotion_context_video
Shared via communication channel
Received emotion_context_video is passed to store in buffer
Received emotion_context_video
Received emotion_context_video is passed for rendering
Received emotion_context_video
Received emotion_context_video is passed to store in buffer
Received emotion_context_video is passed for rendering
Stored in buffer (In time series)
Algorithm decides whether to pass stored emotion_context_video or not
Yes
Rendering the graphical representation of the emotion and/or context via Augmented Reality(AR)
Stored in buffer (In time series)
Algorithm decides whether to pass stored emotion_context_video or not
Yes
Rendering the graphical representation of the emotion and/or context via Augmented Reality(AR)
No
Stop
No
Stop
End Video Call
End Video Call
Stop

Figure 6

FORM 2

THE PATENTS ACT, 1970

(39 of 1970)

&

THE PATENT RULES, 2003

COMPLETE SPECIFICATION

(See Section 10 and Rule 13)

Title of invention:

AN NLP AND COMPUTER VISION-BASED COMMUNICATION SYSTEM FOR INSTANT MESSAGING APPLICATIONS

Applicant:

Lovely Professional University

An Indian University having address as:

Jalandhar-Delhi G.T. Road, Phagwara 144411,

Punjab, India

The following specification particularly describes the invention and the manner in which it is to be performed.

CROSS-REFERENCE TO RELATED APPLICATIONS AND PRIORITY

[001] The present application does not claim priority from any patent application.

TECHNICAL FIELD

[002] The present subject matter described herein, in general, relates to sentiment analysis based on natural language processing (NLP), sentiment analysis, machine learning, deep learning, computer vision, and multimedia integration. It combines techniques from these areas to enable emotion-based messaging. The invention leverages pre-trained models, such as CNN and BiLSTM, and transformers for sentiment analysis of messages.

BACKGROUND

[003] In recent years, the world has witnessed an exponential growth in instant messaging technology, transforming the way individuals communicate and interact with each other. With the help of instant messaging, people connect across the globe, enhancing productivity, fostering social relationships, and revolutionizing customer service, instant messaging has become an essential tool in today's interconnected world. Users can engage in text messaging, allowing them to send and receive written messages in real time. This feature alone has revolutionized communication by eliminating the delays associated with traditional methods such as postal mail or email. Users can have conversations back and forth effortlessly, making communication more dynamic and engaging. Over the past decade, numerous messaging platforms have emerged, revolutionizing the way people communicate. The likes of WhatsApp, Facebook Messenger, WeChat, and Telegram have amassed billions of users worldwide, enabling instant and seamless communication. These platforms provide features such as text messaging, voice calls, video calls, file sharing, and group chats. The accessibility, convenience, and user-friendly interfaces of these platforms have contributed to their widespread adoption and dominance.

[004] Furthermore, Instant messaging technology has connected people across the globe like never before. Geographical boundaries and time zones are no longer barriers to communication. Individuals can easily connect with friends, family, and colleagues, fostering a sense of community and belonging. Additionally, instant messaging has bridged the gap between cultures, allowing people to share experiences, exchange ideas, and broaden their perspectives. The integration of instant messaging in the workplace has significantly improved productivity and efficiency. Communication between team members has become seamless,

enabling quick decision-making, collaboration, and problem-solving. Employees can share files, documents, and updates instantly, reducing the need for time-consuming meetings and emails.

[005] As above discussed, instant messaging applications have gained popularity, but one of the critical issues they face is the lack of emotional expression and connection in text-based communication. Traditional messaging apps often offer limited options for users to convey emotions, leading to potential misunderstandings and difficulties in accurately gauging the emotional state of the person they are communicating with.

[006] The lack of emotional expressions causes a major lack of communication among the users. As Emotional expression is a fundamental component of human communication, facilitating understanding, empathy, and connection. Non-verbal cues such as facial expressions, tone of voice, and body language play a significant role in conveying emotions and intentions. In face-to-face interactions, these cues allow individuals to interpret messages accurately and establish a deeper emotional connection. In face-to-face interactions, non-verbal cues provide valuable context and nuance to the spoken word. Facial expressions, such as a smile, a furrowed brow, or raised eyebrows, can convey happiness, concern, surprise, or skepticism. The tone of voice, including pitch, volume, and intonation, adds layers of meaning to words and helps convey emotions like excitement, anger, or sadness. Body language, encompassing gestures, postures, and movements, contributes to the overall message being communicated. All such human expressions lack instant text messages which cause great discomfort for the users.

[007] As These non-verbal cues work in harmony with verbal communication to provide a more complete understanding of the underlying emotions and intentions behind a message. They allow individuals to interpret messages accurately, helping to avoid miscommunication, misunderstandings, and potential conflicts. Non-verbal cues add depth and subtlety to the communication process, making it easier to discern sarcasm, irony, or sincerity. They enable individuals to gauge the emotional state of the person they are interacting with, fostering empathy and connection.

[008] To address this issue while instant messaging applications have introduced features such as emojis, stickers, voice messages, and multimedia content to enhance emotional

expression, they are still limited compared to face-to-face interactions. These features aim to bridge the gap by providing visual and auditory cues that add emotional context to messages. However, they cannot fully replace the depth and complexity of non-verbal cues in in-person communication. Many messaging apps now offer voice and video messaging capabilities, enabling users to infuse their messages with tone of voice and facial expressions. This feature provides a closer approximation to face-to-face communication, allowing for more nuanced emotional expression and connection.

[009] The limited emotional expression in instant messaging applications causes several negative consequences. Firstly, misunderstandings and misinterpretations can occur due to the ambiguity of text messages, leading to conflicts and strained relationships. Without the ability to accurately gauge the emotional state of the other person, individuals may struggle to respond appropriately or empathetically, further exacerbating communication breakdowns. Moreover, the lack of emotional connection in text-based communication can hinder the formation of meaningful relationships. Emotional expression helps establish trust, intimacy, and a sense of shared understanding. Without this emotional connection, interactions through messaging apps may feel shallow or impersonal, diminishing the overall quality of relationships and hindering the development of empathy and rapport.

[0010] Therefore, it is very evident to provide the solution by advancing the technology that helps to solve the issues above-discussed problems which shall provide the receiver with a real-time facial expression of the sender and act as a bridge for the persons, as the above-mentioned alternatives such as emoji addition in instant messaging or voice or video call addition may create an inference for the emotion but it can be faked easily, and also voice and video don't solve the problem properly, hence a system shall be developed which is easy to use and can address the issue at a small level such as library level.

[0011] However, the above-mentioned problems have been addressed and tried to technically resolve at various levels but different approaches apart from the above-mentioned alternatives, but the complete solution is not provided yet by any of the measures with the improved and dedicated pointing.

[0012] US20220343576A1 illustrates Systems and methods for presenting an avatar that speaks sign language based on the sentiment of a speaker disclosed herein. A translation

application running on a device receives a content item comprising a video and audio, wherein the audio comprises a first plurality of spoken words in a first language. The video comprises a character speaking the first plurality of spoken words in the first language. However, the present invention doesn't perform the sentiment analysis of the user input(text, voice, and facial expressions) and doesn't extract the emotion from the input and reflects the emotion on the sender's avatar via animation at the receiver's end device, and is more related for the assistance of deaf person which may experience by providing graphical representations of a real-time live avatar that speaks sign language and exhibits an emotion of a speaker on a display screen of the computing device.

[0013] WO2020054945A1 discloses a robot and a method for operating the same, according to one aspect of the present invention, may acquire data related to a user, recognize emotional information on the basis of the acquired data related to the user, and provide an emotion-based service; and may generate an avatar character by mapping the recognized emotion information of the user to face information of the user, thereby automatically generating a character representing the emotion of the user. However, the present invention is to provide a robot capable of recognizing a user's emotion (emotion) and providing an emotion-based service and a method of operating the same in which the robot is also capable of expressing an emotion using the generated emotion characters and a method of operating the same but doesn't detect real-time facial expressions of the user.

[0014] US9898849B2 discloses a video call system and method for transmitting the emotion of the other party and determining the emotional state of the other party through a specific gesture and voice recognition taken by the other party during the video call, and to determine the state or operation of the avatar according to the identified emotional state It is for controlling and delivering to the user. According to the present invention, when the user terminal initiates a video call with the counterpart terminal, the user terminal outputs both the avatar corresponding to the counterpart and the video call data received from the counterpart terminal. Analyzing the video call data received by the user terminal to determine the emotional state of the other party through a specific gesture and voice recognition taken by the other party and adjusts and outputs at least one of the states and the operation of the avatar according to the determined emotional state. However, the present invention takes video, voice, or text as input processes the input generates animation messages and renders the animation on avatars and generates a video and has Text Based Facial Expression Engine which generates animation messages having facial expression parameters based on the result of the analysis of the text input. The present invention generates responses from the inputted video and doesn't work for

the instant messaging application for addressing the lack of facial expressions in real-time messaging.

[0015] KR101179465B1 related to Apparatuses, methods, and storage medium associated with creating an avatar video are disclosed herein. In embodiments, the apparatus may have one or more facial expression engines, an animation-rendering engine, and a video generator. One or more facial expression engines may be configured to receive video, voice, and/or text inputs, and, in response, generate a plurality of animation messages having facial expression parameters that depict facial expressions for a plurality of avatars based at least in part on the video, voice and/or text inputs received. However, the present invention provides a video call system in which the emotional state of the other party is determined through specific gestures and voice recognition taken by the other party during the video call and controls the state or operation of the avatar according to the identified emotional state to convey the emotional state of the other party but doesn't provide any solution for the instant message communication.

[0016] The present invention addresses the above shortcomings of the prior art. However, the present invention is entirely different from the prior art in terms of novelty and technological advancement.

OBJECT

[0017] The present invention has the object of providing a novel system that provides real-time responses based on analysis of various biometric, non-biometric, and multimedia input and represents in the form of an animation of the Avatar of the user. The present invention allows users to express their emotions and intentions more effectively, providing a more authentic and personal experience when communicating online with the help of Interactive Avatar Based chat, Wrap Music, Plus Notifications, Interactive video call, modes of chat, Avatar Based group chat, and Video Messaging all of which enhance the overall user experience and increase emotional connection while messaging on the instant message application.

SUMMARY

[0018] Before the present systems and methods, and embodiments are described, it is to be understood that this application is not limited to the particular systems, and methodologies described, as there can be multiple possible embodiments that are not expressly illustrated in the present disclosures. It is also to be understood that the terminology used in the description is for the purpose of describing the particular versions or embodiments only and is not intended to limit the scope of the present application. Further, it is also to be understood that the

terminology, if not defined by the present disclosure, should be construed as the general or known meaning/definition of the terminology, known to the person skilled in the art.

[0019] The disclosure of the present invention discloses a novel and inventive communication system that incorporates sentiment analysis, music, and avatars to enhance messaging experiences. When a user creates a text, voice message, emoticons, image, video, gif, sticker, or any other form of multimedia message, it undergoes analysis to determine the underlying emotion and context. Based on the emotion and/or context detected, relevant song/music is suggested to accompany the message, creating an angle to experience the underlying emotion and context. The message, emotion and/or context, and song/music are packaged into a data packet and sent to the recipient.

[0020] In another implementation, after receiving the packet, the recipient's device displays the message in a chat window, reflects the emotion and/or context on the user's avatar through animation, and allows the recipient to play the received song/music either manually or autoplay. This system introduces features like video messaging, where the camera analyzes the user's facial expressions, body language and various other biometric input reflecting the detected emotions and/or context of the conversation on both avatars. This enables users to communicate through video-like experiences while messaging.

[0021] In another implementation, the process involves pre-processing text and converting it into numeric vectors using word embedding techniques. For voice messages, speech signals are fragmented and converted into text through phoneme matching. For facial expressions and body language, the input from the camera undergoes pre-processing, and a deep learning model determines the emotions from the face and body language. For all other forms of messages like images, videos, emoticons, songs/music gif, and stickers, the input undergoes a deep learning analysis module that determines the emotions or context of the message. The system also incorporates song/music selection based on emotions and/or context of conversation and animations for avatars corresponding to different emotions and/or context of communication. Plus, notifications are implemented by allowing senders to choose specific notification sounds for their messages, which are played on the recipient's device if certain conditions are met.

[0022] Various objects, features, aspects, and advantages of the present invention will become more apparent from the following detailed description of the embodiments of the invention, along with the accompanying drawings in which numerals represent components.

BRIEF DESCRIPTION

[0023] The foregoing detailed description of embodiments is better understood when read in conjunction with the appended drawings. For the purpose of illustrating the present subject matter, an example of the construction of the present subject matter is provided as figures; however, the invention is not limited to the specific method disclosed in the document and the figures.

[0024] FIG 1: Illustrates the overview of the invention.

[0025] FIG 2: Illustrates a general structure of the system while the video messaging is off.

[0026] FIG 3: Illustrates a general structure of the system while the video messaging is on.

[0027] FIG 4: Illustrates the working of the plus notification in the system.

[0028] FIG 6: Illustrates the working of Interactive video call in the system.

DETAILED DESCRIPTION

[0029] Some of the embodiments of this disclosure, illustrating all its features, will now be discussed in detail. The words "comprising," "having," "containing," and "including," and other forms thereof, are intended to be equivalent in meaning and be open-ended in that an item or items following any one of these words is not meant to be an exhaustive listing of such item or items or meant to be limited to only the listed item or items. It must also be noted that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural references unless the context clearly dictates otherwise.

[0030] The following description includes the preferred best mode of one embodiment of the present invention. It shall be clear from this description of the invention that the invention is not limited to these illustrated embodiments but that the invention also includes a variety of modifications and embodiments thereto. Therefore, the present description should be seen as illustrative and not limiting.

[0031] The present invention provides a solution by modern-day technology that is based on different machine learning and deep learning models for emotion and/or context detection: Multimedia Sentiment and Context Analysis which includes Text Sentiment and context Analysis, Voice Sentiment and Context Analysis, Face Expression and body language Analysis and other biometric and non-biometric analysis models that lie in the domain. These models analyze user inputs to determine the underlying emotions and/or context accurately. The machine learning and deep learning implemented herein analyze the sentiments and/or context of the users based on the user input and converts it into the avatar response.

[0032] In one embodiment, the present invention relates to a system and method for emotion-based messaging, including multimedia messaging, video messaging. It also includes the implementation of plus notifications for messages. The description also involves emotion and context detection modules for analyzing text, voice, emoticons, images, videos, gifs, stickers, facial expressions, body language, and all kinds of biometric and non-biometric input that lie in the domain and the user provides to determine the underlying emotions and/or context of communication accurately. The system incorporates song/music integration, allowing users to select and attach relevant song/music to their messages.

[0033] In another embodiment, the analysis module analyzes the content of the message and identifies the associated emotion and/or context. Based on this determined emotion and/or context, relevant song/music selections are suggested to complement the message. The suggested song/music is then integrated with the message, forming a packet that is transmitted to the recipient, and at the receiver's end, upon receiving the packet, the message is displayed in the chat window, allowing the recipient to read and comprehend the content. Additionally, the sender's avatar on the receiver's end device is animated to reflect the emotion and/or context received, providing a visual representation of the sender's emotional state. Moreover, the recipient has the option to listen to the integrated music either manually or auto-play.

[0034] In another embodiment, novel features have been implemented such as video messaging and plus notification. The video messaging-based messaging system incorporates distinguishing features such as video messaging, which enables users to communicate through visual clues while concurrently exchanging the multimedia messages. When the user activates video messaging, the system activates the camera of the user's device to capture their facial expressions and body language. These facial expressions are analyzed to determine the underlying emotion, which is then displayed on the user's avatar on the receiver's device. Here other available user biometrics is also considered like heartbeat rate, pupil dilation for the determination of user's emotional state. This process is reciprocated by the other party as well. Users can engage in video messaging when they are unable to attend a video call due to specific circumstances.

[0035] In another embodiment, video messaging allows users to communicate via video-like interactions. The emotions inferred from the facial expressions are reflected on the avatars of the users, creating a similar experience to face-to-face conversations. Users can simultaneously communicate through multimedia messages, providing a means of communication without disrupting their respective environments.

[0036] In another embodiment, the system enables the user through plus notification in which Communication through messaging applications occurs in various contexts, ranging from casual conversations to work-related matters or urgent and critical issues that require immediate attention. To cater to these diverse scenarios, the messaging system incorporates a feature called plus notification. Senders have the ability to attach a notification sound to their messages, which serves to provide additional context and urgency. Upon receiving a message with Plus notification enabled, the receiver's device plays the specified notification sound.

[0037] In another embodiment, the text message module in the system enables the user when he/she creates a text message, the system incorporates a sentiment and context analysis module to analyze the textual content and determine the underlying emotion and/or context associated with it. The sentiment analysis module plays a crucial role in processing text messages to extract meaningful emotional and contextual information. Initially, the module pre-processes the text by eliminating unwanted elements such as punctuation marks and stop words, which are commonly occurring words with little semantic significance. After the pre-processing step, the sentiment and context analysis module leverages a technique called Word Embedding. This technique involves converting the pre-processed text into numeric vectors. Word Embedding assigns vectors of proximity to words with similar meanings, allowing the model to capture semantic relationships between words. These numeric vectors are then fed into a pre-trained deep learning model comprising Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) layers or transformers.

[0038] In another embodiment, the working of the sentiment analysis module involves utilizing a pre-trained deep learning model that comprises Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) layers or transformers. This model plays a crucial role in analyzing the numeric vectors derived from the pre-processed text and determining the emotion and/or context category associated with the text message. The CNN layer is a fundamental component of deep learning models designed for analyzing visual data or sequential data such as text which consists of multiple filters that slide over the input data, extracting relevant features at different spatial locations. For the sentiment and context analysis module, the input to the CNN layer is the numeric vectors obtained from the pre-processed text in which each filter within the CNN layer applies a series of mathematical operations, such as convolutions and pooling, to capture important patterns and relationships within the numeric vectors. By applying these filters across the entire input, the CNN layer learns to extract significant features that are indicative of various emotions and context.

[0039] In another embodiment, BiLSTM layers process the data in both forward and backward directions simultaneously. The bidirectional processing enables the model to capture dependencies and contextual information from past and future timesteps, thereby enhancing the understanding of the input sequence. For the present invention, the BiLSTM layer operates on the output of the CNN layer, which has already extracted relevant features from the numeric vectors. By considering the sequential nature of the input data and utilizing both past and future context, the BiLSTM layer can effectively model the complex relationships between words and their emotional and the contextual implications in the text message.

[0040] In another embodiment, the CNN and BiLSTM layers are combined to leverage their complementary strengths in feature extraction and sequential modeling. The output of the CNN layer, which represents the extracted features from the numeric vectors, serves as the input to the BiLSTM layer. The BiLSTM layer processes these inputs, incorporating both forward and backward information to capture the temporal dependencies and contextual nuances within the text. By integrating the CNN and BiLSTM layers, the machine learning model gains the ability to capture both local and global patterns in the text, facilitating a comprehensive understanding of the emotional and contextual content.

[0041] In another embodiment, the deep learning model processes the numeric vectors and generates an output corresponding to the emotion and/or context category associated with the text message. The output signifies the emotion_context_message that underlies the message. It is important to note that the deep learning model used for sentiment and context analysis is trained on a comprehensive dataset obtained from diverse sources such as review sites and social media posts. The training dataset enables the model to learn patterns and associations between textual content and emotions and context, thereby facilitating accurate emotion and context detection. After the emotion and/or context is determined the messaging system suggests relevant song/music choices to the sender that align with the detected emotion and/or context. Additionally, the sender can also generate the song/music based on the input. The sender can then select an appropriate song/music to wrap with the message. After the song/music selection, an emotional data packet (emo_data_packet) is generated. This packet encapsulates the original text message, now wrapped with the chosen song/music, along with the associated emotion_context_message.

[0042] In another embodiment, the Transformer model is incorporated into the system to ascertain the emotional and contextual expression conveyed in user input which comprises of two components encoder and the decoder. The encoder operates on the input sequence, while the decoder operates on the target output sequence. By employing sequence-to-sequence

learning, Transformers process a series of tokens and predict the subsequent token in the output sequence through iterative processing within the encoding layers.

[0043] In another embodiment, the encoder generates encodings that signify the relevant relationships among different parts of the input sequence. These encodings are subsequently passed to the subsequent encoder layer. The decoder incorporates all of these encodings and utilizes the derived context to generate the output sequence. Transformers employ an attention mechanism, which provides contextual information regarding items within the input sequence.

[0044] In another embodiment, transformers can process multiple sequences in parallel, thereby reducing the training time required for the model. This parallelization facilitates more efficient computation and optimization, leading to enhanced performance and faster processing of data.

[0045] In another embodiment, to ensure the security and privacy of the message during transmission, the emo_data_packet undergoes encryption before being transmitted to the receiver over the internet. Encryption involves converting the contents of the packet into an encoded form that can only be deciphered with the appropriate decryption key. This encryption process safeguards the emotional message from unauthorized access or interception. Upon reaching the receiver's device, the encrypted emo_data_packet is decrypted using the corresponding decryption key. Decryption reverses the encryption process, allowing the receiver to access the original contents of the packet. The decrypted emo_data_packet is then processed to extract its individual components.

[0046] In another embodiment, each above-disclosed module is incorporated for each other purposes such as the original text message, wrapped with the chosen song/music, is displayed in the chat window of the receiver's device. This enables the receiver to read and comprehend the textual content of the message. The receiver has the capability to listen to the song/music that accompanies the message via autoplay or manually by clicking the play button associated with the message. This allows the receiver to experience the emotional impact of the combined text and song/music and the emotion_context_message associated with the message is reflected on the sender's avatar, which is displayed on the receiver's device. The emotion_context_message is also used for recommending appropriate reactions for the message received. This visual representation of the sender's emotional state enhances the communication experience by providing an additional emotional angle to the message.

[0047] In another embodiment, the text message module is incorporated with the wrap song/music in a similar manner the system extends its emotion and context-based capabilities to voice messages and other multimedia messages. When a user records a voice message, the

system performs sentiment analysis on the recorded audio to determine the underlying emotion and/or context (emotion_context_message) conveyed by the sender. The voice message undergoes a series of processing steps to extract emotional and contextual information. Initially, the speech signal is fragmented into smaller fragments, each spanning 10 milliseconds. These fragments facilitate the analysis of specific sections of the voice message, aiding in the extraction of emotional and contextual cues. For each fragment, the power spectrum is computed, and the resulting spectral representation is mapped to cepstral coefficients. These coefficients help in converting the speech signal into a textual representation.

[0048] In another embodiment, the textual representation represented of the voice message is then inputted into the Text and Context Sentiment Analysis module described earlier. This module processes the textual representation and employs the same sentiment analysis and context techniques used for text messages to determine the emotion and/or context category associated with the voice message. Based on the detected emotion and/or context, the system suggests relevant song/music choices that are suitable to be integrated with the voice message. The sender can select the desired song/music from the suggestions provided by the system. The sender has the option to generate the song/music based on the message by the help of Generative AI and wrap the generated song/music with the message. Subsequently, the emo_data_packet is generated, which consists of the voice message, now combined with the selected song/music and the corresponding emotion_context_message. The emo_data_packet is encrypted for secure transmission as in the text messaging process then it is transmitted to the receiver, decrypted at the receiver's end, and the individual components are extracted.

[0049] In another embodiment, after a voice message is sent from the sender’s end at the receiver’s end the extracted components serve different purposes such as the voice message, integrated with the chosen song/music, is played in the chat window of the receiver's device. This allows the receiver to listen to the voice message and grasp the emotional content conveyed by the sender. The receiver has the ability to listen to the song/music associated with the voice message via autoplay or manually by activating the play button associated with the message as complementary background music for the message. This provides an enriched audio experience and enhances the emotional impact of the message. The emotion_context_message associated with the voice message is reflected on the sender's avatar, displayed on the receiver's device. The visual representation of the sender's emotional state further augments the communication experience and provides an additional emotional angle to

the message. The received emotion_context_message is also used to recommend appropriate reactions for the message received.

[0050] In another embodiment, at the receiver's end of the messages the user holds control over the plus notification feature. They can manage and regulate the behaviour of this feature according to their preferences. Specifically, the receiver can choose to enable or disable plus notifications for specific senders. This functionality empowers the receiver to customize their messaging experience based on their individual preferences and requirements. To ensure a convenient user experience, the notification sound associated with plus notifications is only played if the receiver's device is in sound mode. This consideration prevents unnecessary interruptions in situations where the receiver prefers to receive messages silently or when their device is set to a non-audible mode.

[0051] In another embodiment, the Text Sentiment and Context Module in the emotion and context detection module which in analyzing text messages and determining the emotions and/or context conveyed within them. To begin the analysis, the module pre-processes the text message by removing unwanted elements such as punctuation marks and stop words. This pre-processing step ensures that the subsequent analysis focuses on the relevant content of the message. After the pre-processing step, the text is transformed into numeric vectors using a technique called Word Embedding. Word Embedding assigns numeric representations (vectors) to words based on their semantic meaning and context. These vectors capture the underlying relationships between words, allowing the machine learning model to understand the meaning and sentiment expressed in the text.

[0052] In another embodiment, the resulting numeric vectors representing the pre-processed text are then fed into a pre-trained machine-learning model. This model consists of Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) layers or transformers. The CNN layer is responsible for extracting important features from the numeric vectors, while the BiLSTM layer analyzes the sequential nature of the text, and considers both past and future context. By processing the numeric vectors through these layers, the machine learning model predicts the emotion and/or context category associated with the text message.

[0053] In another embodiment, the voice sentiment and context module focuses on analyzing voice messages and discerning the underlying emotions and/or context. When a voice message is received, the speech signal is fragmented into smaller segments, each lasting for 10 milliseconds. These fragments are essential for capturing the nuances and variations in the voice, allowing for a more accurate analysis of the emotional content. To convert the voice

message into a format amenable to sentiment analysis, the power spectrum of each fragment is mapped to cepstral coefficients. These coefficients provide a representation of the speech signal that facilitates the identification of phonemes, the fundamental units of speech sounds. By matching these phonemes, the voice message is converted into textual representation. The resulting text, derived from the voice message, is then processed by the Text Sentiment and context Analysis module described earlier. This module utilizes the same machine learning model, comprising CNN and BiLSTM layers, to determine the emotion and/or context category associated with the voice message. By leveraging the power of the pre-trained model, the voice sentiment module accurately recognizes and classifies the emotions and/or context conveyed through the voice message.

[0054] In another embodiment, a face expression and body language analysis module to detect emotions based on the user's facial expressions and body language. This module utilizes the camera of the user's device to capture their facial expressions and body language in real time. For analyzing the facial expression module, the input of the face and body is to be taken of the user from the camera which undergoes a series of pre-processing steps to ensure accurate analysis. These steps involve detecting and isolating the user's face from the captured image and detecting body landmarks. By focusing specifically on the face and body landmarks, the system can effectively analyze facial expressions, and body language and extract the relevant features for emotion detection.

[0055] In another embodiment, the pre-processed data, consisting of the isolated face and body landmarks, is then fed into a pre-trained deep learning model based on a Convolutional Neural Network (CNN) architecture. This deep learning model has been trained on a labeled dataset containing images of human faces, body language, and their corresponding emotions. By comparing the user's facial expressions and body language with the patterns learned during training, the model accurately determines the emotion expressed by the user.

[0056] In another embodiment, the song/music as per the detected emotion is also integrated into the communication process. The system organizes song/music files into separate directories based on different emotion and/or context categories. When the sender selects the music button while composing a message, the system suggests a list of song/music options relevant to the detected emotion and/or context. Once, the song/music is selected, the chosen music's name and a link to access the song/music file are added to the message. At the receiver's end, the song/music name is displayed alongside the message. By clicking the play icon associated with the song/music, the receiver can listen to the selected music using the provided link or the music is to be played automatically after getting received by the receiver’s device if

the autoplay is ON. The user has the option to generate song/music as well based on the input provided, by the help of Generative AI and wrap the generated song/music with the message. The integration of song/music adds a multi-sensory dimension to the messaging experience, enriching the emotional expression and connection between the sender and receiver. A color notation is also integrated into the application which notifies the sender whether the receiver has listened to the song/music or not, if the receiver has not played the song/music, the song/music audio to be shown in black in the chat for both the user and in green if the music is played by the receiver.

[0057] In another embodiment, Avatar animation is to be incorporated into the mobile application which enhances the visual representation of emotions and context of the conversation. The avatar implementation reflects the user's facial expression and body language in visual representation. Predefined animations are corresponding to different emotions and contexts are stored as .json files. These animations are designed to reflect the specific characteristics and expressions associated with each emotion and context category.

[0058] In another embodiment, once the emotion and/or context is expressed in the message, the system searches for the appropriate animation file based on the received emotion and/or context. The system then applies the corresponding animation to the sender's avatar, visually representing their emotional state and context of the conversation. This implementation leverages the Lottie library, which is specifically designed for Android applications, to facilitate smooth and dynamic animation rendering. The Scenform SDK is to be used for 3d avatars rendering and animations. Additionally, game engines like Unity, Unreal and animation softwares like Blender is also incorporated for building avatars and avatar responses. The proposed invention encompasses 2d as well as 3d avatars and avatar responses. User has the option to choose an emoticon-based avatar, humanoid avatar or any other form of avatar. Emoticon-based avatar is the 2d or 3d animated emoticon representation of user's emotional state and/or context of the conversation, whereas humanoid avatar has human-like features. Alternative technologies is employed for devices running on different operating systems, ensuring compatibility across various platforms.

[0059] In another embodiment, the plus notification feature provides an additional layer of communication by allowing the sender to select a notification sound to accompany their message. After creating the message, the sender can specify a particular notification sound that they desire the receiver to hear upon receiving the message. At the receiver's end, the notification sound is played if the sender has been granted permission for plus notification by the receiver. Additionally, the receiver's device must be in sound mode to ensure that the

notification sound can be heard. By employing these conditions, the system ensures that the receiver is notified with the specified sound, creating a distinctive and engaging messaging experience.

[0060] In another embodiment, to maintain the confidentiality of users' private data, all processing of user input, including text, voice, song/music, images, video, gif, stickers, camera feed and all other kind of non-biometric and biometric user input, is conducted exclusively on the user's end device. Machine learning and deep learning models are deployed on the user's end device, utilizing the private user data to generate the desired output. These models are trained using the user's private data on their own device and periodically share the training weights with other devices either through a central server or a peer-to-peer network. These shared weights are then used by the end device-deployed models for their training to improve their accuracy with time.

[0061] In another embodiment, of this, the system ensures that the user's private data remains secure and protected. By keeping the processing and training within the user's control, the system mitigates the risk of any potential leakage of private data in the event of server compromise. Users have full control over their data and can maintain confidence in the security of their personal information, and decentralized approach to processing and training not only enhances data privacy but also provides better performance and efficiency. By utilizing the computational power of multiple devices, the system can leverage distributed computing to expedite processing tasks and improve overall system performance.

[0062] In another embodiment, to ensure the security and privacy of the message content, the emo_data_packet, which includes the text, voice, images, videos, gif, stickers and all other kind of biometric and non-biometric messages, the associated emotion and/or context, and any selected song/music, undergoes end to end encryption before being transmitted over the internet. Encryption algorithms are applied to the packet to convert it into an unreadable format, protecting it from unauthorized access during transmission. Once encrypted, the emo_data_packet is transmitted from the sender's device to the receiver's device through the Internet. This transmission can occur via various communication channels, such as messaging applications or email services. The encrypted packet is securely delivered to the receiver's device, ready for further processing and extraction.

[0063] In another embodiment, the decrypted emo_data_packet is then processed at the receiver's end to extract the individual components: the text voice message, images, videos, gif, stickers, and all other kinds of biometric and non-biometric messages, the associated emotion and/or context, and any selected song/music. Each component is extracted and

prepared for display and playback, providing a comprehensive messaging experience for the receiver.

[0064] In another embodiment, the components are to be extracted based on their utility based on their nature such as the text message displayed in the chat window of the receiver's device, allowing the receiver to read the content and understand the sender's message. If song/music was added to the message by the sender, the receiver can listen to the associated song/music via autoplay or manually by clicking the play button associated with the message. This enables the receiver to engage with the message on an auditory level and further experience the intended emotions conveyed by the sender, and the emotion and/or context associated with the message is reflected on the sender's avatar displayed on the receiver's device. Through animation and visual cues, the receiver can perceive and interpret the emotional state of the sender, creating a more immersive and expressive messaging experience.

[0065] In another embodiment, the song/music generation feature of the application allows users to create original song/music compositions that reflect their unique input. Generative AI is leveraged to enable users to generate music, images, videos, gif, sticker from various input sources. Users can share the generated media with other users and have the option to wrap the generated song/music with messages before sending them. The application provides users with photo editing and creating tools, allowing them to edit their photos and videos within the messaging platform.

[0066] In another embodiment, Users are provided with the option to wrap the generated song/music with messages before sending them to recipients which enhances communication and expression by allowing users to convey their messages with a personalized musical backdrop. Upon receiving a message, the song/music is played automatically using an autoplay function or can be manually triggered by pressing a dedicated play button, ensuring that the recipient experiences the intended emotional impact.

[0067] In another embodiment, the messaging and media generation application offers a comprehensive set of photo editing and creating tools. Users have the liberty to import their photos into the application or capture new ones using their device's camera. The application provides a wide array of editing features, including filters, cropping, adjustments, special effects, and more, enabling users to enhance and modify their photos according to their artistic vision. The creating tools within the application empower users to compose and edit videos directly within the messaging platform. Users can import video clips, add transitions, apply filters, and incorporate music to create visually stunning and compelling video content. These

tools enable users to unleash their creativity and produce captivating videos without the need for external editing software.

[0068] In another embodiment, the application ensures seamless sharing of the generated music, images, and videos among users, promoting a collaborative and engaging environment. Users can share their creations directly within the messaging platform, allowing for immediate and interactive feedback from recipients. This fosters a sense of community and encourages creativity among users.

[0069] In another embodiment, the user is provided with the flexibility to indicate the relationship of the second person in their contact list, such as work, family, or friend. The user interface (UI) of the chat window is tailored specifically to the relationship between the first and second users. For instance, if the second user is a family member or the user's significant other, the first user has the option to specify this relationship, resulting in a chat window UI that reflects this connection. The avatars in the chat window are to be depicted in informal casual clothing, and the overall style, colors, and design elements of the chat window are to be adjusted to match the context and emotional tone of the conversation, emphasizing informal communication. If the second person is a colleague from work, the chat window UI is to be adjusted accordingly. The avatars to be dressed in formal attire, and the visual elements of the chat window, including colors, style, and layout, aligns with a more formal and professional setting. It is important to note various visual elements and features are adjusted to suit the respective relationship and mode of communication.

[0070] In another embodiment, the proposed invention provides the feature of an avatar-based group chat. The avatar of all the users is to be represented in an elliptical fashion. The avatar of the user who has posted the last chat in the group appears in the center and shows a large size while the avatar of the other users is to be sidewise and small. All the other features as mentioned provided in one-to-one chat are available in group chat as well.

[0071] In another embodiment, based on the emotion and/or context detected from the conversation based on machine learning and deep learning the theme of the chat conversation gets changed in the window. The interactive avatar-based chat functionality is achieved through the integration of advanced technologies such as augmented reality (AR) and real-time data transmission. The application employs sophisticated algorithms and models to generate realistic and expressive avatars that accurately replicate the users' facial expressions, gestures, body language and emotions. When a user initiates an interactive avatar-based chat session, the application utilizes the device's camera and sensors to track and capture the user's facial movements and gestures and other available biometrics that lie in the domain in real time. This

data is then processed and mapped onto the avatar, allowing it to mirror the user's actions and emotions. The application also employs voice recognition technology to capture the user's speech and synchronize it with the avatar's lip movements and body language for a more immersive experience.

[0072] In another embodiment, to ensure smooth communication between users, the application employs efficient data transmission protocols and technologies. The avatar animations, emotions, and other data are encoded and transmitted in real time to the recipient's device. The receiving device then decodes and renders this data, allowing the second user to observe the interactive avatar's movements, expressions, and emotions in synchronization with the conversation. The application provides a user-friendly interface that allows individuals to easily enable or disable the interactive avatar-based chat feature according to their preferences. Users can access the settings menu or toggle switch to activate or deactivate the interactive avatar-based chat functionality.

[0073] In another embodiment, the utilization of advanced technologies such as Federated Learning and End-to-End Encryption (E2EE). The application comprises an analysis module with AI models deployed on the user's end device, ensuring that all processing of user input occurs locally on the device itself. To train the AI models, user data is utilized on the end device, ensuring the privacy of the user's personal information. The models are trained using the user's private data, and the resulting trained weights are securely shared with other computing devices. This sharing of trained weights occurs either through a peer-to-peer network or a centralized server, enabling the models to continuously improve over time. By keeping the training process decentralized and utilizing the user's private data locally, the system ensures that the privacy of the user's personal data is safeguarded.

[0074] In another embodiment, to maintain user privacy, all private messages exchanged and video call sessions within the messaging application are protected using End-to-End Encryption (E2EE). This means that private messages remain encrypted throughout the entire communication process and are only decrypted on the recipient's end device. But the generative AI models responsible for generating music, images, and videos an other multimedia are deployed on the server. When users request the generation of media content, the generative AI models process the data on the server. Once the generation process is complete, the resulting data is securely transmitted to the user's end device. This centralized approach allows for efficient computation and generation of media content while minimizing the computational load on the user's device. However, if the application size and the

computation power of the user's end device is not a concern then the generative AI models are to be deployed on the user's end device as well.

[0075] In another embodiment, the system incorporates an analysis module that processes user input, including text, emoticons voice messages, images, video, song/music, gif, stickers and various other multimedia to determine the emotion and context of the conversation. This information is visually represented on the avatar of the first user (sender) on the second user (receiver) end device. The appearance of the avatar dynamically changes according to the detected emotion and context. The avatars also synchronize their lip movements and body language with voice messages, enhancing the realism and immersion of the communication experience. The system extends the analysis of context and emotion to all types of user messages, including text, voice, photos, videos, GIFs, stickers, and more. The system can determine the emotional state of the user through various biometrics such as heart rate, pupil dilation, body language, and facial expressions.

[0076] In an exemplary embodiment, if the conversation indicates that the first user is going to the gym, the avatar of the first user on the second user's device depicted in gym clothes, running, and jogging. Additionally, the style of the chat window adapts to reflect the context and emotion of the conversation.

[0077] In another embodiment, the system analyzes user input during a video call, including speech, body language, hand gestures, and other available biometrics, to understand the context of communication. Based on this analysis, the system graphically represents the context and emotions on the screen. The computing device, which may vary from screens to holographic devices, augmented reality (AR) devices, virtual reality (VR) devices, and future inventions, accommodates a wide range of output devices.

[0078] In an exemplary embodiment, if the first user becomes angry during the video call, the system interprets the speech input, facial expressions, and body language and other available user biometrics to determine the anger and visually represents it on the second user's device.

[0079] In another embodiment, by effectively utilizing the extracted components, the receiver can fully engage with the message, understand the textual content, experience the intended emotions through song/music, and visually perceive the sender's emotional state via avatar animated responses.

[0080] In another embodiment, the system incorporates a recommendation feature based on the user input and associated emotions and/or context. Depending on the content and emotion and/or context of the input, the system suggests appropriate emoticons, GIFs, stickers, images, videos, song/music or other means for either generating the multimedia message or reacting to

the message received. This enhances the communication experience by providing users with relevant and expressive options for generating messages and responding to messages.

[0081] In another embodiment, the proposed invention encompasses various types of output devices, including hologram devices, augmented reality (AR) devices, virtual reality (VR) devices, and any other devices that may emerge in the future. This ensures the versatility and adaptability of the system to accommodate different technological advancements and user preferences.

[0082] In another embodiment, in addition to providing music/song suggestions based on the determined emotion and context from user messages and generating music/song from user input, the proposed system also offers users music/song editing tools. These tools allow users to perform various modifications on the music/songs, such as slicing, adjusting bass and treble, modifying tone, and editing other properties of the music/song. The music/song editing capabilities encompass various editable properties within the domain of music editing.

[0083] In an exemplary embodiment, of the proposed invention, the system possesses the capability to determine both the emotion and context of the user input. However, the decision regarding what should be reflected on the user avatar through animation can be made either by the algorithm or by the user's preference. This embodiment aims to provide a personalized and dynamic representation of the user's emotional state and contextual activities.

[0084] In another exemplary embodiment, if the analysis module determines from the conversation that the user is about to go to the gym, and various biometric and non-biometric user inputs indicate a happy emotion, the avatar represents that the user is to be animated by wearing gym clothes, engaged in running and jogging activities, with a cheerful expression and positive body language. On the other hand, if the analysis module determines a sad emotion from the user's inputs, the avatar still be animated in gym clothes, running, and jogging, but with a facial expression and body language reflecting sadness. To determine the emotion consider various sources of information, including biometric data such as heart rate, pupil dilation, body language, and facial expressions, along with non-biometric data such as text and voice messages. The system's analysis module thoroughly processes these inputs to determine the user's emotional state and context. By incorporating both biometric and non-biometric data, the system ensures a comprehensive understanding of the user's emotions and activities.

[0085] It is important to note that while the determination of emotion and context is automated, the system allows for user input or preferences to influence the animation displayed on the user

avatar. This flexibility ensures that users have control over the representation of their emotions and activities, making the overall experience more personalized and engaging.

[0086] The exemplary embodiment demonstrates the ability of the proposed invention to dynamically adapt the avatar's appearance, expression, and body language based on the user's emotional state and contextual information. By providing a visually rich and interactive representation, the system enhances the communication experience and facilitates a deeper understanding of the user's emotions and activities during conversations.

[0087] In another exemplary embodiment, if from the conversation, it is determined that the user wants to read books and the emotion determined via various biometric or non-biometric means is that the user is in a neutral emotional state then the avatar renders the animation of just reading books.

[0088] In another embodiment, the present disclosure consists of different machine learning and deep learning models that determine the underlying emotion and/or context from various multimedia and biometric inputs such as text analysis model, Voice Analysis, and Face Expression and body language Analysis. The Text Sentiment and Context Analysis model focuses on analyzing text inputs to determine the underlying emotions and/or context. When a user creates a text message, the message undergoes preprocessing to remove unnecessary elements like punctuation marks and stop words. The preprocessed text is then transformed into numeric vectors using Word Embedding, which assigns vectors of close proximity to words with similar meanings. These vectors capture the semantic meaning of the text. The numeric vectors are fed into a pre-trained machine learning model comprising Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) layers. The model analyzes the vectors and outputs the emotion and/or context category associated with the text message. The machine learning model is trained on a dataset collected from various sources such as review sites and social media posts, enabling it to learn patterns and correlations between text and emotions and/or context of the conversation.

[0089] In another embodiment, the Voice Sentiment and context Analysis model focuses on analyzing voice inputs to determine the underlying emotions and/or context. For voice messages, the speech signal is fragmented into smaller segments, typically 10-millisecond fragments. Each fragment's power spectrum is mapped to cepstral coefficients, which provide a representation of the acoustic characteristics of the speech. These coefficients are then matched to phonemes, which help convert the speech into textual representation. The resulting text is processed by the Text Sentiment and Context Analysis module described above. By utilizing the same model as the Text Sentiment and Context Analysis, the Voice Sentiment and

Context Analysis model leverages the learned patterns from text inputs to determine the emotion and/or context category associated with the voice message. This approach allows for consistent and coherent emotion and/or context analysis across different modes of communication.

[0090] In another embodiment, the Face Expression and Body Language Analysis model focuses on analyzing facial expressions and Body language captured by the camera to determine the underlying emotions and/or context. The input from the camera undergoes pre-processing, including face detection and cropping, and body landmark detection, to ensure accurate analysis. The pre-processed data is then fed into a pre-trained deep learning model based on Convolutional Neural Network (CNN) architecture. This deep learning model is trained on a labeled dataset containing images of human faces, body language, and their corresponding emotions. By analyzing facial features, such as eyebrow position, eye openness, mouth shape, and body language, the model determines the emotion expressed by the user. The Face Expression and Body langauge Analysis model provides a visual dimension to emotion analysis, enabling the system to capture emotions conveyed through facial expressions and body language.

[0091] In an exemplary embodiment, a comprehensive range of biometric and non-biometric user inputs is considered. Biometric inputs encompass facial expressions, body language, voice, heartbeat rate, and other relevant inputs within the domain. Non-biometric user input encompasses text, emoticons, images, videos, music/songs, gifs, stickers, and all other forms of user input within the domain. Message input consists of various multimedia elements explicitly sent by the sender to convey a message, while non-message input includes indirect user inputs during communication sessions, such as facial expressions, body language, and heart rate. It illustrates the capability of the proposed invention to leverage diverse user inputs, employ AI models for emotion and context analysis, suggest and generate song/music, and seamlessly integrate these elements into a comprehensive communication experience. By considering various forms of user input and providing dynamic recommendations and reactions, the system enhances user engagement and fosters deeper emotional connections in digital conversations.

[0092] In another exemplary embodiment, the collected input is then processed through various AI models, including machine learning and deep learning, to determine the user's emotional state and contextual information underlying the conversation. Based on the emotion and/or context, the system suggests relevant precomposed songs/music that can be wrapped with the multimedia message. The system employs generative AI to generate music based on

the user input, which can also be wrapped with the message. The generated music is stored in a database for future suggestions of relevant precomposed songs/music. Users have the option to choose between wrapping a pre-composed song/music or generating their own based on the input.

[0093] In another embodiment, the federated learning model gets trained in a local environment and gets personalized according to the specific user which means the model is able to understand the underlying language patterns specific to the user hence, is able to understand the tone, patterns used by the specific user and able to predict the emotions and/or context more accurately.

[0094] In another embodiment, Recurrent Neural Network (RNN) is used to determine the consideration of the emotion and or context lying in the time series. In RNN the output from the previous step is fed as input to the current step. The hidden state/ Memory State in RNN remembers information about the sequence. The RNN has multiple activation function units for each time stamp. Each unit has an internal state which is called the hidden state of the unit. This hidden state signifies the past knowledge that the network currently holds at a given time step. This hidden state is updated at every time step to signify the change in the knowledge of the network about the past. Backpropagation is applied in the training process to make the model accurate as the training advances. The RNN is trained on the dataset which consists of similar emotion and/or context-related data in textual format that appear in a video call session. The Biometric analysis module analyses the available biometrics of the user during the video call session. The emotion and/or context determined from the biometrics is then saved into a buffer. From the buffer the emotion and/or context is fed to the RNN which then determines what emotion and/or context has to be rendered graphically through AR.

[0095] In another exemplary embodiment, If the user wants to send a multimedia message when Video Messaging is OFF :

a. After the sender has created the multimedia message, Sentiment and Context Analysis of the message is done and the underlying emotion and/or context of the message is determined as emotion_context_message.
b. Based on the emotion and/or context detected, the sender is provided with the relevant precomposed song/music choices that is to be wrapped with the message
c. The user has the option to generate the song/music based on the input via Generative AI which can be further wrapped with the message. The generation of music is done either on a server or on the user's end device.

d. After a relevant song/music is selected by the sender, emo_data_packet is generated which contains the 'Message wrapped with music' and emotion_context_message.
e. End to end Encryption is provided to emo_data_packet and sent to the receiver via the communication channel.
f. emo_data_packet is received by the receiver's end device and is decrypted.
g. emo Extractor extracts the constituents of the emo_data_packet.
    - Message – Displayed in the chat window
    - Music – To be played at the receiver's device either manually or auto-play.
    - Emotion and/or Context – Reflected on the sender's avatar on the receiver's device and is used to recommend appropriate reaction for the message received.

[0096] In another exemplary embodiment, If the user wants to send a multimedia message when Video Messaging is ON:

a. Input from the camera is taken and Face Expression Recognition and Body Language Recognition is done of the user. Other user biometrics if available is considered. Other user biometrics depends on the availability. For example if the user is wearing a fitness band then the heartbeat rate is fetched from the wearable. Pupil dilation is also considered subjected to lighting conditions and camera module.
b. Combining all the available user biometrics, user's emotional state is determined as emotion_bio.
c. emotion_bio is end to end encrypted and is sent to the receiver via the communication channel.
d. emotion_bio is received at receiver's end device and is reflected on the sender's avatar.
e. Concurrently, the sender sends multimedia message following the method discussed previously except the corresponding emotion_context_message is only used for recommending the appropriate reaction for the message received.

emotion_bio is reflected on the sender's avatar on the receiver's device.

[0097] In another exemplary embodiment, the process of initiating and conducting a video call between two users is described. The first user initiates a video call request to the second user, and upon receiving and accepting the request, the video call session is established. During the

video call session, the system deployed on the end device of the first user analyzes the biometrics of the first user, including voice input, facial expressions, body language, and hand gestures, as well as other available user biometrics within the domain, to determine the emotion and/or context, denoted as "emotion_context_video." Similarly, the system deployed on the end device of the second user analyzes the biometrics of the second user to determine their emotion and/or context, also denoted as "emotion_context_video." The emotion_context_video of each user is encrypted and exchanged between them. Upon receipt, the emotion_context_video is stored in a buffer and used to render graphical representations using Augmented Reality (AR) technology.

[0098] In another embodiment, an AI model is sent to the device at the time of installation and that model is trained on that device's data. This trained model is then shared to a central server which aggregates the models received from various edge devices and pushes a new global trained model to all the edge devices, and in Decentralized Federated Learning, the devices/nodes share the model among themselves to obtain a global model. This setting is affected by the network topology in which the devices lie.

[0099] In yet another embodiment, a sensing mean may include not limited to a camera and may include any kind of biosensor such as remote photoplethysmography (rPPG) technology to detect heart rate that captures user's health data and other like, and visual sensors that denotes all sensors which, similar to the human eye, pick up at least a two-dimensional image of the object being measured such as the monochrome model, and the color model that applies image processing to images captured by a camera, the vision sensor calculates the characteristics of an object, such as its area, center of gravity, length, or position, and outputs the data or judgment results.

[00100] In another embodiment, the above disclosure is a description of the invention and is not intended to limit the scope of the invention. Other variations and modifications of the above-described embodiment shall be apparent to those skilled in the art and are intended to fall within the scope of the invention as defined in the following claims.

Dated this …….. Day of June 2023

Dr. Monica Gulati

Registrar

Lovely Professional University

We Claim:

1. An NLP and computer vision-based communication system for instant messaging applications, comprises:
    a. an avatar-based communication system for the communication between the sender and the receiver;
    b. a generative music module for wrapping the music as per the context of input;
    c. an expression analysis module through different biometrics captured by a camera, visual sensors or other sensing means for analyzing the expressions and emotions of the receiver and sender in messaging and in video call sessions;
    d. a multimedia module for the generation of multimedia through generative AI as per the context of the input;
    e. a security module for the secure transmission of the data;
    f. a mobile application with different modes of User interface for implementing the different sets of modules;
    g. a cloud server for storage purposes and to deploy the trained model.
2. An NLP and computer vision-based communication system for instant messaging applications, as claimed in claim 1, wherein the avatar-based communication system through lip-syncing of the customized avatar of the user provides sentiment analysis of the message, the context of the message, and sentiment of the facial expressions based on the biometrics, whereas the detection of the facial expressions is to be represented via multimedia such as images, videos, gif, and any other form.
3. An NLP and computer vision-based communication system for instant messaging applications, as claimed in claim 1, wherein the generative music module is further integrated with the Generative Artificial Intelligence that enables to generate the music/song/tune or any other based on the analysis of the context of the conversation either in multimedia messaging and video call sessions simultaneously for the context, and further stores it to the cloud-based server for future references, the generative music module enables to wrap the message by generated content.
4. An NLP and computer vision-based communication system for instant messaging applications, as claimed in claim 1, wherein the facial expressions analysis module identifies the facial features of the users engaged in the conversation either multimedia messaging or video call sessions, and either receiver or sender, the sentiment analysis further integrated to analyse the facial expressions of the users as recorded through any

of the cameras, visual sensor, or by other sensing mean, the system monitors the user activity and determines the context and/or emotions and represents graphically on screen with the help of AR.

5. An NLP and computer vision-based communication system for instant messaging applications, as claimed in claim 1, wherein the multimedia module further generates the multimedia based on the context by Generative Artificial Intelligence, and a multimedia editing tool is incorporated for the usage of the customized multimedia for the user, and the recommendation module that is integrated with the generative AI also recommends the appropriate multimedia as per the context of the conversation.
6. An NLP and computer vision-based communication system for instant messaging applications, as claimed in claim 1, wherein the security module for the secure transmission is integrated with the Federated learning that processes the user's private data and also shares the training weights for the future improvisation of AI models with the user of E2EE so that the confidentiality can be provided.
7. An NLP and computer vision-based communication system for instant messaging applications, as claimed in claim 1, wherein a mobile application with different kinds of User Interfaces that is based on the relationship of the sender and receiver for the implementation of all the modules, and to provide a real-time experience to the users of the conversations, the mobile application provides plus notification by which the sender can specify the notification sound for each message for the receiver.
8. An NLP and computer vision-based communication system for instant messaging applications, as claimed in claim 1, wherein the appearance of the avatar keeps changing based on the emotion/context of the conversation in any manner, and also based on the data received from the facial expression analysis module.
9. An NLP and computer vision-based communication system for instant messaging applications, as claimed in claim 1, wherein a remote cloud server is incorporated with the system for the storage, user authentication, real-time database, deploying Generative AI and other large-size AI models used in the system, and for supporting federated learning, the cloud-server also stores all the multimedia generated through Generative AI, and the conversations for backup purposes.
10. A method for communication in an NLP and computer vision-based communication system for instant messaging applications, comprising the steps of:
    a. establishing communication between a sender and a receiver through an avatar-based communication system.

b. wrapping music based on the context of text messages using a generative music module.
c. analyzing the expressions of the receiver in text messages and video calls through a facial expression analysis module using different biometrics captured by a camera.
d. generating multimedia content through generative AI based on the context of the text messages using a multimedia module.
e. ensuring secure transmission of data using a security module.
f. implementing different sets of modules through a mobile application with different modes of the user interface.
g. storing data and deploying trained models on a cloud server for storage and deployment purposes.

wherein the system performs sentiment analysis of the message, the context of the message, and sentiment of the facial expressions based on biometrics through lip-syncing of the customized avatar of the user, and also represents the detected facial expressions via multimedia, such as images, videos, gifs, or other forms of multimedia content.

Dated this …….. Day of June 2023

Dr. Monica Gulati
Registrar
Lovely Professional University

ABSTRACT

AN NLP AND COMPUTER VISION-BASED COMMUNICATION SYSTEM FOR INSTANT MESSAGING APPLICATIONS

The present invention is a communication system that incorporates advanced AI models to analyze and determine underlying emotions and context from text, voice, images, video, gif, sticker, facial expressions, body language, and all other kinds of biometric, non-biometric and multimedia user inputs. By analyzing the semantic meaning of the text, acoustic characteristics of speech, facial features, body language and other biometric, non-biometric and multimedia user inputs, the system can identify the emotions and context expressed by the users. The invention also integrates song/music suggestions based on detected emotions and context, allowing users to enhance their messages with suitable song/music. The system provides avatar animation reflecting the identified emotions and context and offers control over plus notifications. It promotes effective and emotionally engaging communication. Its application aligns with the SDGs of Good Health and Well-being, Quality Education, and Industry, Innovation by facilitating enhanced emotional expression, improved communication experiences.

Dated this …….. Day of June 2023

Dr. Monica Gulati

Registrar

Lovely Professional University

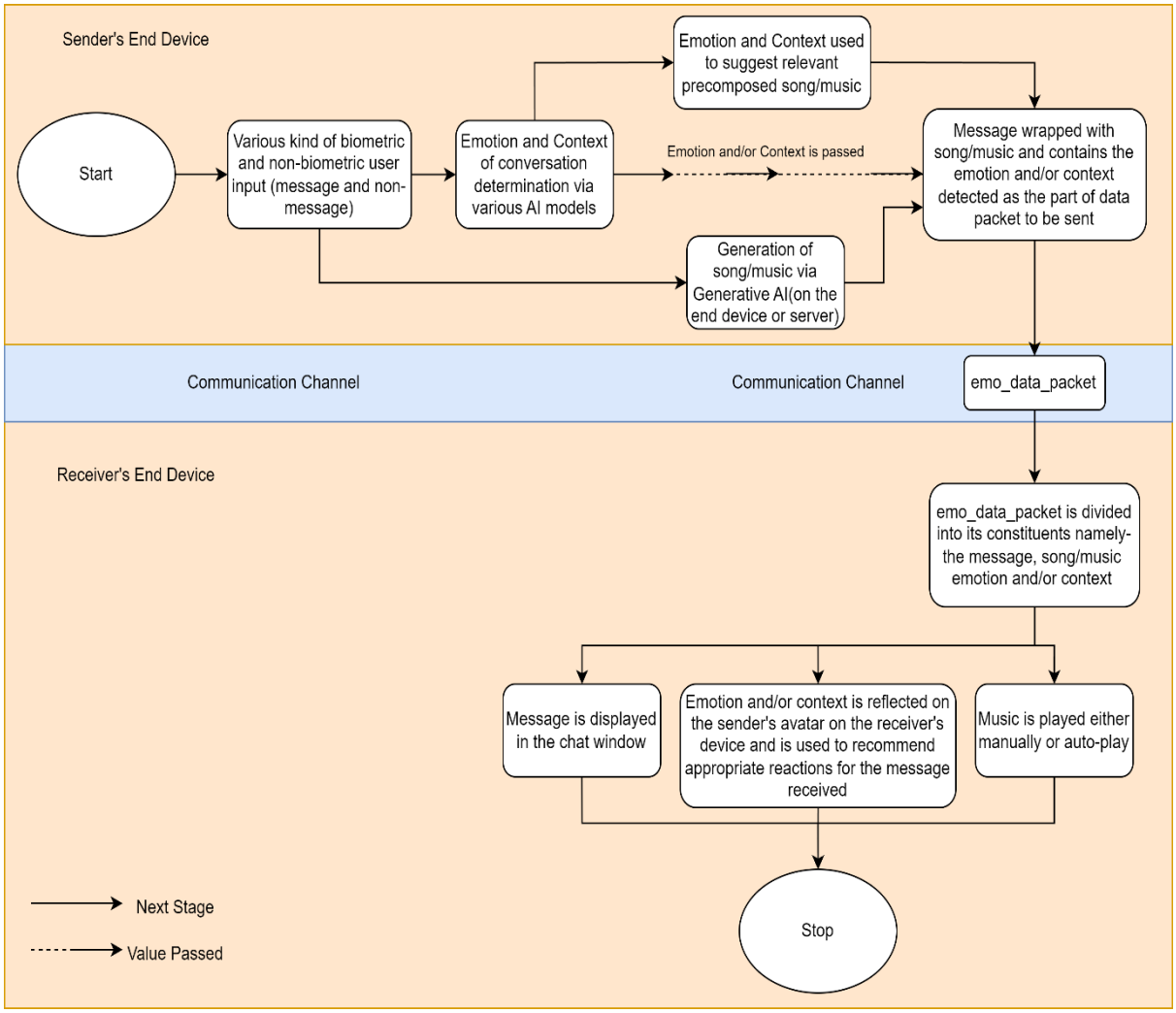


Figure 1

Dr. Monica Gulati

Registrar

Lovely Professional University

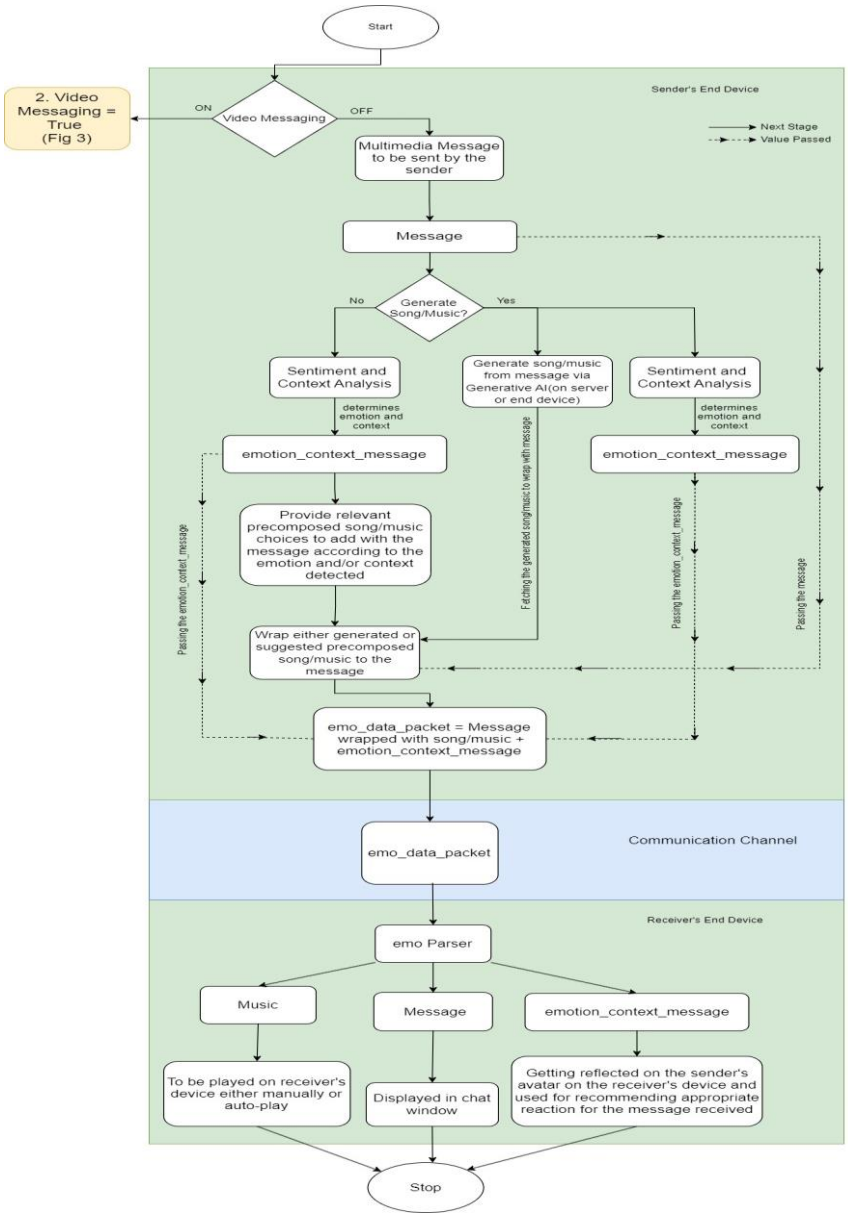


Figure 2

Dr. Monica Gulati

Registrar

Lovely Professional University

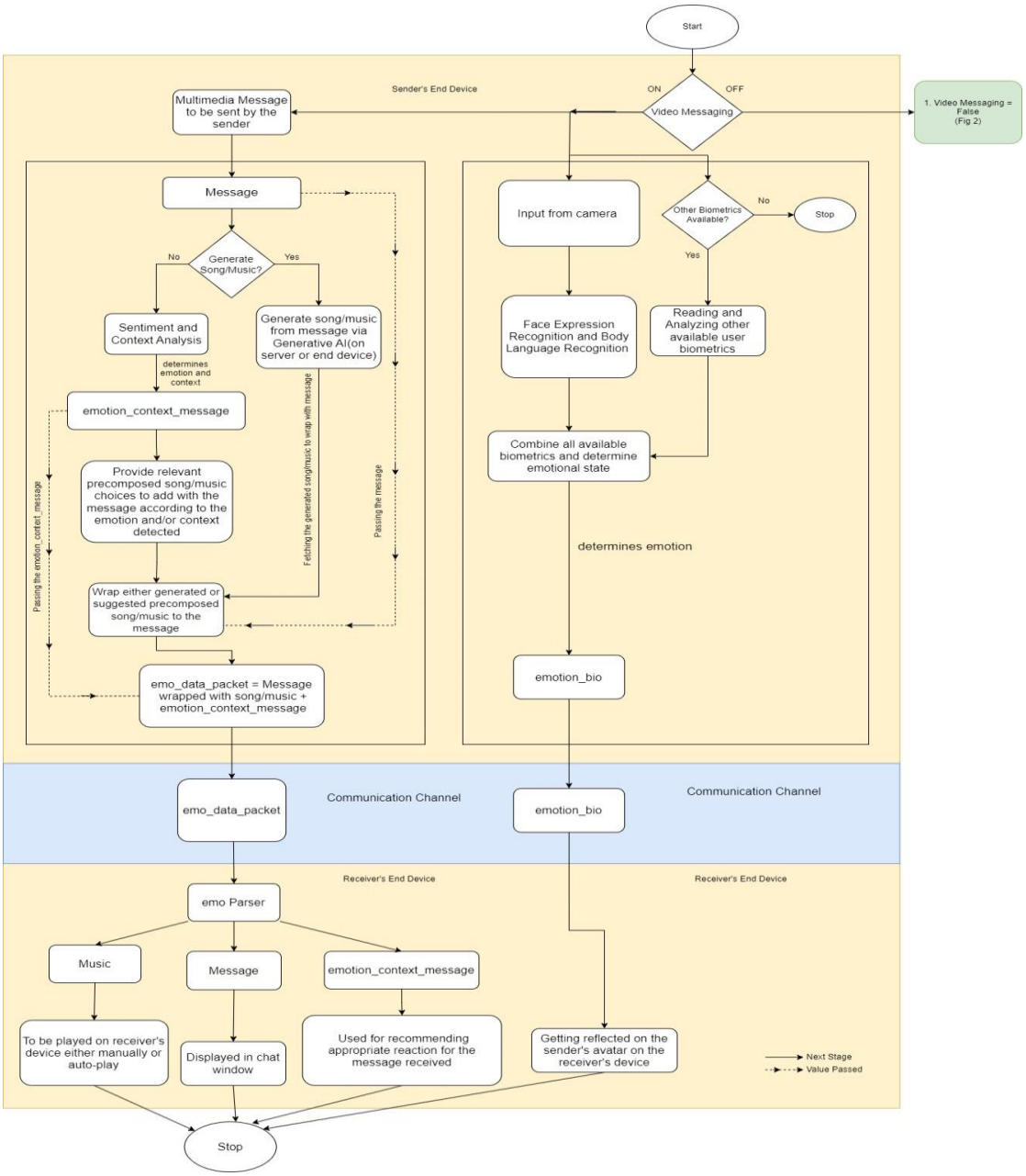


Figure 3

Dr. Monica Gulati

Registrar

Lovely Professional University

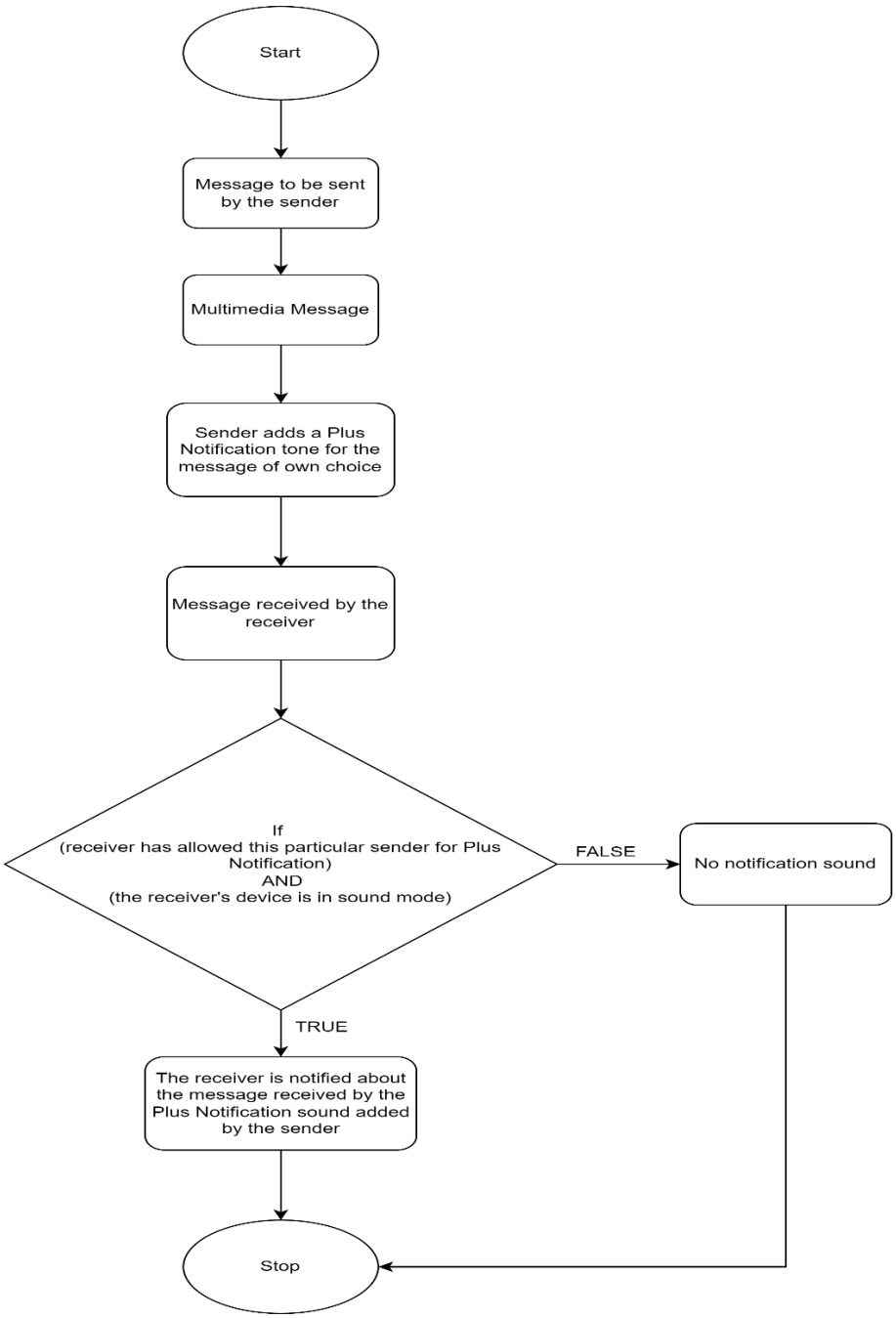


Figure 4

Dr. Monica Gulati

Registrar

Lovely Professional University

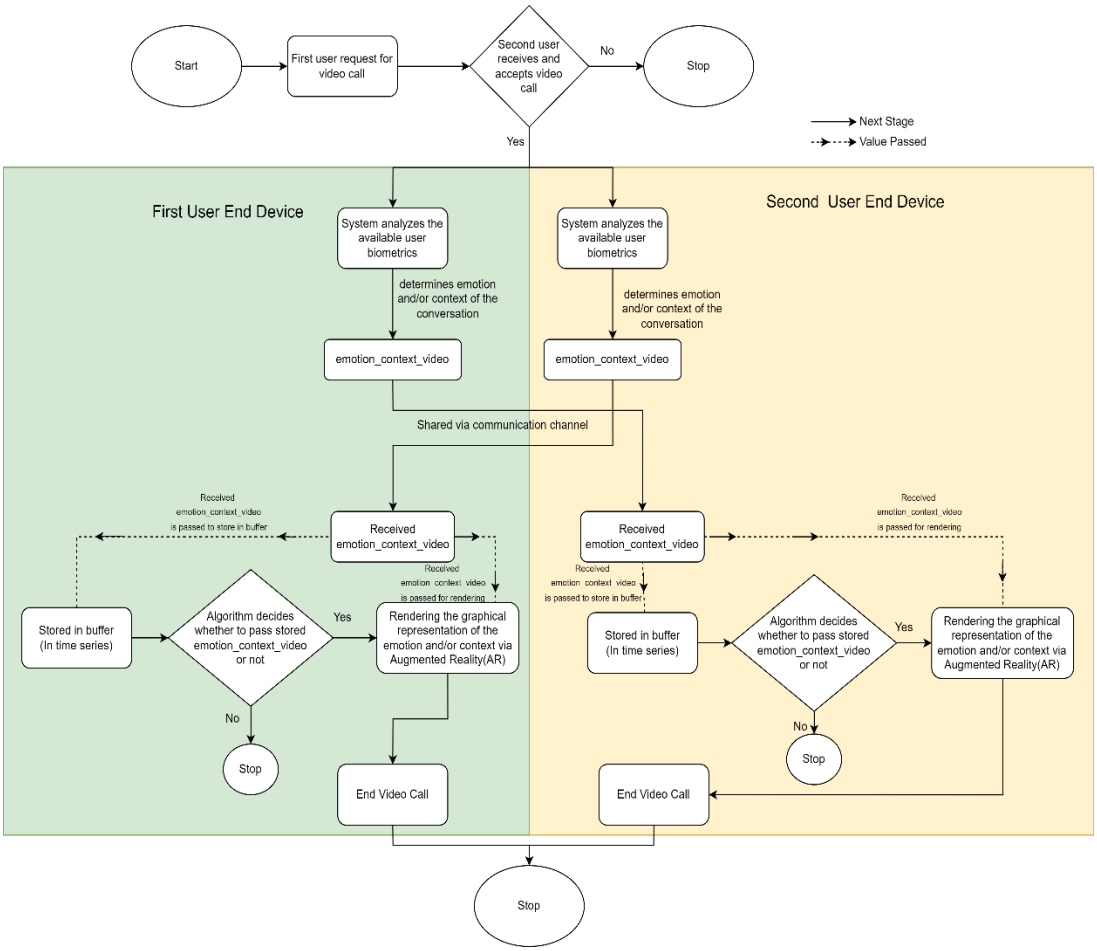


Figure 6

Dr. Monica Gulati

Registrar

Lovely Professional University