International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Emotion-Driven Facial Animation for Chat-Bots

Pulkit Juneja, P. Gayathri, Anshupriya Srivastava, Hemant Jain

Abstract: ***Facial animation is quickly becoming an important feature of virtual assistants and is gaining traction as the preferred communication technique between man and machine. In addition to lip movement, facial expressions such as those of the eyes and cheeks help in conveying the sentiment and context of what is being spoken. This paper aims to present a new methodology to create an emotionally expressive virtual AI that is capable of understanding the sentiment of the conversation and displaying emotions during conversations. To achieve this the system uses a generative chatbot and combines it with a 3D talking head that is animated parametrically. This work could be beneficial to virtual assistants and help facilitate more lifelike interactions, holding significance in environments that require the user to feel more comfortable with their interactions. The complexities lie in developing a domain specific chat-bot that will not only provide valuable replies but also recognize and display appropriate facial expressions while communicating.***

Keywords: ***Animation, Chat-bots, Computer Graphics, Machine Learning.***

I. INTRODUCTION

Virtual worlds and environments are becoming increasingly popular and the interactions between computers and humans has moved from simple text-based exchanges to complex audio-visual exchanges. Various domains require such audio-visual exchanges and hence the need for advanced graphics and machine learning techniques emerged. This work aims to aid towards the transition to a realistic chat bot that is capable of displaying appropriate gestures and non-verbal ques that will make communication with the computer feel less distant. By creating a 3D model of a human face and animating it based on speech, we propose to develop a system capable of displaying emotions while speaking. The movement of the lips and other facial joints are modelled closely to those of humans to deliver a more personal conversation environment. With the ability to model and recognize personalities of users, conversational bots have grown rapidly in terms of understanding human preferences and reacting accordingly to their predicted personality needs.

The system proposed in this paper presents a simple and efficient method of combining facial emotion animation with a conversational bot. The prototype is capable of understanding human speech and translating that into text, which a generative chat-bot uses to generate responses. The core complexity lies in the animation of the 3D model to display human emotion via the mouth and eye movement. Instead of building a chat-bot that is hard-coded with a set of responses, the proposed model is capable of generating responses that it has never seen before. Instead of just identifying the context, intents and entities and using these to structure a response manually, it is capable of using sequential neural networks and deep learning to identify context and respond accordingly to the text input to it.

II. BACKGROUND

Facial animation is becoming an important communication technique between man and machine. In addition, it is pivotal in the development of synthetic actors/agents that can act like humans. Over the last three decades many techniques have been used to create convincing speech-synchronized facial animation. It has proven to be a difficult task due to the complexity of the system and the low tolerance for inconsistencies in the animation from a human audience.

One such technique was proposed by Cao, Y., Tien, W. C., Faloutsos, P., & Pighin in their paper ` Expressive speech-driven facial animation ` [2]. They propose a method to approach the problem of speech synchronized animation by using a machine-learning algorithm that utilizes a database which contains accurate facial motions mapped to speech components. Their aim was to derive a generative model that can appropriately show facial motion, control its emotions and also maintain correct synchronization between speech and lip movement. The end user provides an input to the system, which is the data that the system requires for analyses. This data can also be derived from an audio signal employing a Support Vector Classifier, which provides an effective solution to the problem of speech synchronicity.

In another Paper, King, Scott A., Richard E. Parent, and Barbara Olsafsky [3] proposed a B-spline surface-based approach to achieve speech synchronized lip motion. In their approach they used B-spline surfaces to model both the inner and outer portions of the lip structure. These surfaces can then be controlled by specified muscle-based parameterization to allow for specification of a wide range of lip motion. The face is rendered using a procedural texturing approach allowing for increased realism.

Manuscript published on November 30, 2019.
* Correspondence Author
Pulkit Juneja*, Digital analyst at Mckinsey and company Bangalore.
Hemant Jain , System software developer in the TensorIT server team at NVIDIA.
Anshupriya Srivastava, pursuing a master's degree in data science from Duke university, North Carolina.
Gayathri Prakasam, Assistant Professor (Senior) in the School of Computing Science and Engineering at VIT University, Vellore, Tamil Nadu, India.

In yet another paper, Deng, Z., Neumann, U., Lewis, J. P., Kim, T. Y., Bulut, M., & Narayanan, S [4] presented an expressive facial animation synthesis system enabled by automated learning from facial motion capture data. The system used motion capture to record the movements of a subject as he recited a predesigned corpus with specific spoken and visual expression. They presented a novel motion capture and mining technique that they use to learn speech coarticulation models for diphones and triphones from the recorded data. Their approach uses the motion signal processing and principal component analysis dimensionality reduction technique to generate a reduced Phoneme-Independent Expression Eigenspace (PIEES) that encloses the dynamic expression signals created using a texture-synthesis based approach.

Besides that, there exists a lot of published literature exploring the applicability of virtual avatars in daily life. Sukhanya Kethuneni, Stephanie E. and August James Ian Vales proposed creating a virtual health care avatar, which interacts with a database containing data related to health care [5]. Adrian Horzyk, Stanis law Magierski, and Grzegorz Miklaszewski [6] performed an experiment to showcase the use of an adaptable chatbot in the context of a shop assistant. A sample Internet shop was created with a chatbot playing the role of a shop-assistant similar to a real shop-assistant in a traditional shop. Their chatbot would predict a customer's personality and based on that, adjust its model of actions.

III. BASIC KNOWLEDGE

A. Chat Bots

A Chat bot (also known as a conversational agent, Bot, Artificial Conversational Entity) is a computer program that conducts a conversation via auditory or textual methods. Such programs are often designed to convincingly simulate how a human would behave as a conversational partner, thereby passing the Turing test.

There are two types of chat bots:

- **Retrieval-based models [17]**: These models are easy to understand and develop. To train them, a predefined set of queries and responses are fed into the system. This works well for a closed domain and the chat bot eventually becomes effective at context and content understanding. These models involve a good extent of Machine Learning, but they are incapable of generating any new text or content. In essence, they select a definite text from a predefined set of texts and use the set that best serves the *intent* of the question. These models serve will for chat bots that perform a specific task. Simply put, it is a matter of selecting an answer from a large set (according to the requirements of the question).
- **Generative models [17]**: These models are hard to understand and build. They have no defined classes and self-generate entire answers of their own based on their understanding of a giving sentence. Since they generate entirely new responses from scratch, they are relatively complex. For a larger or more open domain, this model performs better. It appears more realistic and replies more similar to that of a human.

B. Phonemes

An abstract unit of the phonetic system of a language that corresponds to a set of similar speech sounds. More simply, phonemes are the individual sounds that make up speech. A naive facial animation system may attempt to create a separate facial position for each phoneme. The English language has 11 different phonemes.

C. Parametric animation

Parametric Animation is an animation technique that blends together two or more animations in real time to create a new animation.

IV. METHODOLOGY

A. System Overview

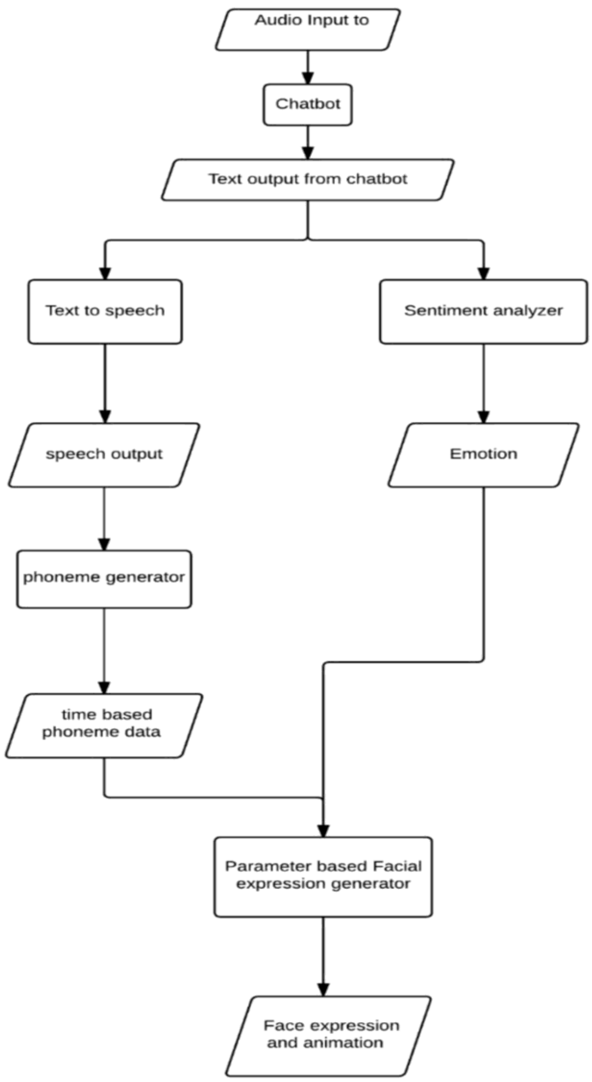

Fig - 1: Proposed architecture for the system

I. For the proposed system, as shown in Fig. 1, audio is taken as an input into the system and the voice recognition engine (speech to text) converts the audio into words/text that the conversational agent can understand.

II. The resulting text is given as an input to the generative chatbot which processes the text and generates a reply based on the user's query. The response from the chatbot is fed into two parallel workflows.

a. The response text is used to generate the audio data, using the text to speech converter, which is used by a phoneme recognizer to generate a temporal sequence of phonemes [14] of the audio.

b. In parallel, A sentiment analyzer is run on the response text and an estimate of the sentiment of the response is extracted

III. The extracted sentiment is then used to alter the facial expressions of the 3D avatar while it tries to simultaneously articulate the phoneme data derived from the response text's audio. Phoneme articulation is achieved by animating the lips of the model. To portray emotion, a number of facial features (such as the eyes, eyebrows as well as the mouth) are taken into account and animated.

B. Detailed Methodology

For our work, we would be using a generative model for speech analysis and sentiment derivation. At the base of the generative model, is a seq2seq model that is capable of taking a sentence as an input and generating another sentence as a response. The Seq2Seq model [10], as shown in Fig. 1, has two primary layers: the encoder and the decoder. The encoder is comprised of several layers of left-stack LSTMs. Similarly, the decoder side has several layers of right-stacked LSTMs. The purpose of the encoder is to encode the input sequence i.e. a sequence of words into an internal mathematical representation called 'context vector'. This context vector in turn is used as an input by the decoder to generate an output sequence. It is important to note that the lengths of input and output sequences can be different since there is no explicit one-on-one relation between them. This allows us to utilize this architecture in a variety of places such as machine translation and question answering. This makes this machine learning architecture suitable for conversational bots – that take one sentence as an input and generates another as output.

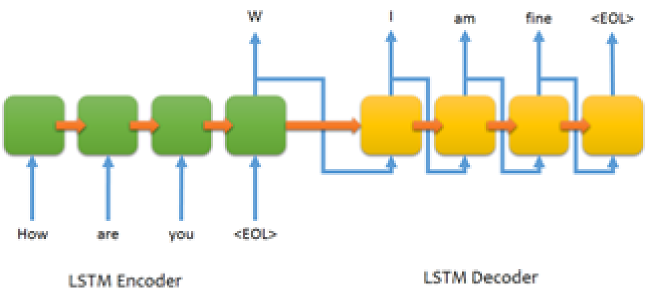

Figure 2. Architecture of the Seq2Seq model used in the chat-bot [10]

For Phoneme Recognition the open source pocket sphinx library is used. The library is provided the audio file for the text to be spoken and it outputs a list of CMUBets [11] and their respective time frames.

For animating the face, Blender [12] is used. However, the algorithm can be easily ported to any other 3d modelling program or game engine. The Blender instance runs a modal operator which opens a socket server and continuously listens for incoming phoneme data.

The Meshes for both the face and the jaw contain predefined blend shapes, which are used to animate the lips and create expressions on the face Each of these blend shapes either represents a phoneme or an emotion expression. The algorithm begins by iterating over the list of Phonemes and their time durations, assuming that the phoneme data received is in sequential order. This time duration is multiplied by the desire FPS and is rounded to the nearest integer. This gives the number of frames needed for that particular animation (nf_i). For each phoneme shape two frames are inserted into the scene timeline, the starting frame with a blend shape value of 0 and a final frame with a blend shape value of 1. The animation then creates the intermediate frames by interpolating between these. The algorithm begins by inserting the first frame of the first phoneme animation at frame 0. The final frame is inserted after nf_0 which is the number of frames required for the first animation. Further frames are inserted sequentially with the starting frame being inserted at the same position as that of the final frame of the previous shape.

Emotions are similarly mapped to different blend shapes which define all vertex positions starting from the basic face expression to create the overall expression. The face expression animation is performed exponentially at first, made stable and then returns to the rest state. Once the animation is created, Blender plays the animation by interpolating the vertex positions between the stored blend shapes in the key frame.

V. RESULTS AND ANALYSIS

A. Chat Bot Performance

Performance of the chat-bot [15] is evaluated based on the following criteria:

- **Confusion Triggers and Perplexity**: The types of input sentences for which the chat-bot says a version of 'I don't know' or fails to understand what an appropriate response should be.

    - **Perplexity** – Perplexity has been set as a standard method of evaluate any language model. Perplexity is calculated by calculating the probability distribution over an entire sentence or text. For example, if a test sample's sentence comprises of approximately 1,000 words and could be coded using a total of 8.2 bits per word, a model perplexity of 28.2 = 294 per word can be calculated. It can be said that the model encounters confusion when it has to choose amongst 294 possibilities for each word [13]

The loss and perplexity values are shown in Table 1. As we can see, the perplexity decreases as the data set is trained. The accuracy of the chat-bot increases gradually i.e. Loss and Perplexity values decrease as it is trained better as shown in Fig 3 and 4.

Table 1: Perplexity and Loss of the training dataset.

Step	Loss	Perplexity
1000	4.63	102.56
2000	4.36	78.41
3000	4.08	59.04
4000	3.7	40.51
5000	3.4	29.88
6000	3.57	35.53
7000	3.36	28.89
8000	3.4	29.95
9000	3.2	24.61
10000	3.22	24.94
11000	3.24	25.58
12000	3.04	20.97
13000	3.09	22.02
14000	2.96	19.36
15000	2.85	17.34
16000	2.83	17.02
17000	2.79	16.21
18000	2.8	16.37

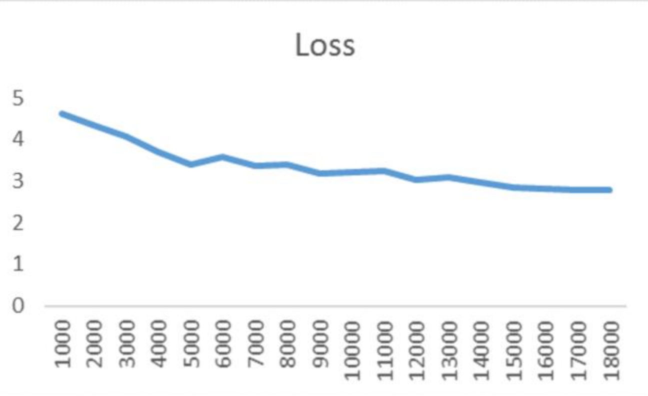

Fig – 3: Loss vs No. of Iterations while training the dataset

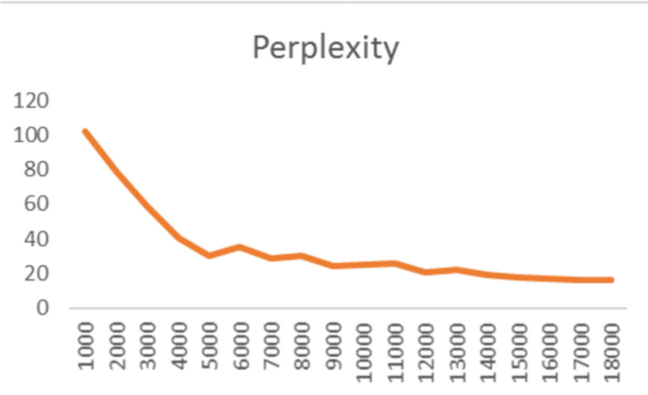

Fig – 4: Perplexity vs No. of Iterations while training the dataset

Figure 5 and 6 show the loss vs no. of iterations plot during the training of the model. Fig 5 shows it until 1800 iterations and Fig 6 shows it until 3800 iterations. The emotions displayed by the face are very subjective and therefore does not have any specific metrics.

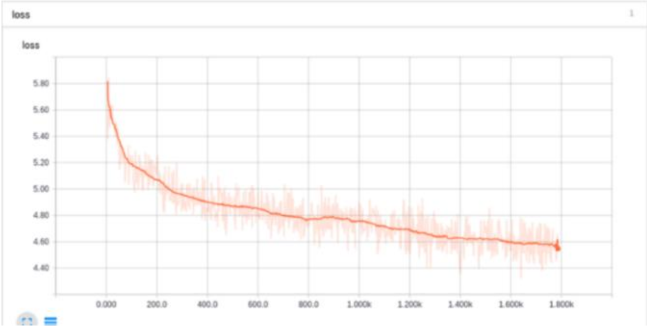

Fig – 5: The loss of the model during training until 1800 iterations

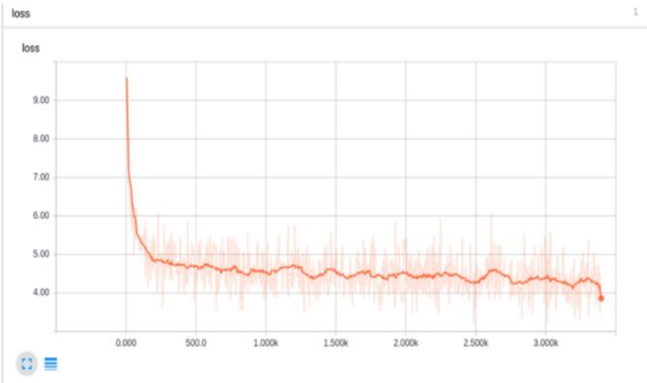

Fig – 6: The loss of the model during training until 3800 iterations

B. Animation Accuracy

We have obtained a chat-bot that provides with appropriate responses based on user queries. These responses are accompanied with expressions one can expect from a real human face. The face of the chat-bot is shown below in Fig. 5 and Fig. 6.

The sample animation for the same is available as the supplementary material [1] and it can be seen that the lip and facial animations are consistent with the words spoken and the emotion being expressed. For example, the face forms a sort of happy expression, with elevated eyebrows and a smile on the face, when the sentiment of the response is "Happy".

C. Conversation Sample

Below is a sample conversation between the chat-bot and the user:

EVA: Hello. How are you feeling today?
User: Hello Eva
EVA: What can I call you?
0.0192945215551
sad

User: You can call me Wally
EVA: Wally, nice to meet you! Are you a man or a woman?
0.978123099664
happy

User: I am a man
EVA: Thanks for telling me your gender, Wally. Most of my clients are men.
0.00427144330384
sad

User: Where are you from EVA?
EVA: I am originally from Silicon Valley Now I live in your computer. Where are you?
0.436275869649
neutral

User: I live in Bangalore
EVA: What is living in Bangalore like?
0.0546987658167
sad

User: It is nice. The weather is good here
EVA: It seems nice to me too.
0.439358303154
neutral

User: How are you?
EVA: I am functioning within normal parameters.
0.540199628776
neutral

User: What is it like to be a bot?
EVA: Much the same as to be a bot would be for you.
0.0259246464743
sad

User: You give ambiguous answers
EVA: It is always better to give than to receive.
0.950856559233
happy

In the above examples, the answer is shown, as well as the sentiment score and the overall sentiment of the response.

The sentiment and replies are accurately calculated, and the overall conversational agent appears more human like.

VI. CONCLUSION AND FUTURE SCOPE

Using the proposed system, we created a virtual chat bot agent that combines deep learning with parameterized animation. It is able to synchronize speech with lip movement and facial expression. Using the described natural language techniques, it understands the sentiment of the end user and maintains an engaging dialogue with them.

The ability to train this dialogue system for specific tasks with specific data allows it to be used in various domains [9] including psychiatric therapy [16], healthcare, education, games [7] and kiosks, for a more standardized and comfortable user experience. While chat bot technologies are not new, giving the agent a realistic face adds to its personality and increases user engagement. The aforementioned process makes the face feel more human-like and realistic, thereby making the entire process feel more natural.

Despite the system demonstrating a high performance, there is scope for improvement in terms of more complex emotional cues and reactions. Currently the face is able to express only the 6 basic emotions - happiness, sadness, anger, surprise, disgust and fear. To make the chatbot look more human-like, it is necessary to overcome this limitation both during sentiment understanding and facial animation. While we independently animate only the mouth and the eyebrows, extending the same to the nose, eyes and forehead could allow for a greater degree of expression utilizing a greater variety of visual cues [8].

ACKNOWLEDGEMENTS

We extend our heartfelt gratitude to Dr. P. Gayathri, Associate Professor, VIT for her guidance and insight. Her suggestions and feedback were invaluable and greatly assisted us in the writing of this paper.

REFERENCES

1. A working demo of the agent (http://bit.ly/2Kt3MS4)
2. Cao, Y., Tien, W. C., Faloutsos, P., & Pighin, F. (2005). Expressive speech-driven facial animation. ACM Transactions on Graphics (TOG), 24(4), 1283-1302. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
3. King, S. A., Parent, R. E., & Olsafsky, B. (2000, November). An anatomically based 3D parametric lip model to support facial animation and synchronized speech. In Proc. Deform 2000 (pp. 7-9).
4. Deng, Z., Neumann, U., Lewis, J. P., Kim, T. Y., Bulut, M., & Narayanan, S. (2006). Expressive facial animation synthesis by learning speech coarticulation and expression spaces. IEEE transactions on visualization and computer graphics, 12(6), 1523-1534.
5. James, S. K. S. E. A., & Vales, I. (2009). Personal healthcare assistant/companion in virtual world. In AAAI Fall Symposium Series.
6. Horzyk, A., Magierski, S., & Miklaszewski, G. (2009). An Intelligent Internet Shop-Assistant Recognizing a Customer Personality for Improving Man-Machine Interactions. Recent Advances in intelligent information systems, 13-26.
7. Slater, S., & Burden, D. (2009, March). Emotionally responsive robotic avatars as characters in virtual worlds. In Games and Virtual Worlds for Serious Applications, 2009. VS-GAMES'09. Conference in (pp. 12-19). IEEE.
8. Augello, A., Pilato, G., Gambino, O., Pirrone, R., Gaglio, S., & Cannella, V. (2011). An Emotional Talking Head for a Humoristic Chat-bot. INTECH Open Access Publisher.
9. https://www.quora.com/What-are-good-applications-for-chat-bots-other-than-customer-service
10. https://github.com/farizrahman4u/seq2seq
11. https://cmusphinx.github.io/wiki/phonemerecognition/
12. https://www.blender.org/
13. https://nlpers.blogspot.com/2014/05/perplexity-versus-error-rate-for.html
14. Gibert, G., Olsen, K. N., Leung, Y., & Stevens, C. J. (2015). Transforming an embodied conversational agent into an efficient talking head: from keyframe-based animation to multimodal concatenation synthesis. Computational Cognitive Science, 1(1), 7.
15. Foster, M. E. (2008, June). Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In Proceedings of the Fifth International Natural Language Generation Conference (pp. 95-103). Association for Computational Linguistics.
16. Pontier M., Siddiqui G.F. (2008) A Virtual Therapist That Responds Empathically to Your Answers. In: Prendinger H., Lester J., Ishizuka M. (eds) Intelligent Virtual Agents. IVA 2008. Lecture Notes in Computer Science, vol 5208. Springer, Berlin, Heidelberg
17. http://www.wildml.com/2016/04/deep-learning-for-chatbots-part-1-introduction/

AUTHORS PROFILE

Pulkit Juneja: Pulkit is working as a Digital analyst at Mckinsey and company Bangalore. He recently completed his Bachelor of Technology in Computer Science and Engineering at VIT university, Vellore in 2017 and has been working since. He specializes in application development, distributed systems, Augmented/virtual reality and his research interests include data science and computer graphics. While in college he was the technical advisor in the Computer soceity of India.

Hemant Jain: Hemant Jain is currently working as a System software developer in the TensorIT server team at NVIDIA. He previously completed his bachelor's degree in computer science and engineering from VIT University, Vellore in 2017 and later went ahead to pursue a master's degree in data science from the university of Washington where he graduated in 2019. His research interests include Data science and Machine learning

Anshupriya Srivastava: Anshupriya Srivastava is a graduate student currently pursuing a master's degree in data science from Duke university, North Carolina. She previously completed her Bachelor of Technology degree in computer science from VIT University, vellore in 2017. Her research interests include Data science and machine learning

Gayathri Prakasam: P. Gayathri is working as an Assistant Professor (Senior) in the School of Computing Science and Engineering at VIT University, Vellore, Tamil Nadu, India. She received her MTech in Computer Science and Engineering from VIT University. She has seven years of teaching experience. She is currently pursuing her PhD in VIT University, India. She is a life member of Computer Society of India. Her research interest includes data mining, information retrieval and soft computing.