International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

Sentiment Analysis using Bi-directional Recurrent Neural Network for Telugu Movies

Kumar R G, Shriram R

Abstract*: *Sentiment Analysis is the Natural Language Processing (NLP) is the active research area due to its vast application like stock market prediction, product re-views etc. The sentiment analysis in the regional languages are required for the film industries to increase their profit. Many existing methods has been applied on the sentiment analysis in the regional languages to increases the performance and still, it lags due in efficiency. In this research, the Bi-directional Recurrent Neural Network (BRNN) is applied to increase the performance of the sentiment analysis in the regional languages. The BRNN method has the advantages of rep-resenting the high and poor resources sentences in the common space and sentiment is analyzed based on the similarity measure. The proposed method is evaluated on the twitter data and compared this with the existing methods such as Random forest and Support Vector Machine (SVM). The proposed BRNN has the overall accuracy of 50.32%, while existing method of SVM has the overall accuracy of 38.73%.

Index Terms*: *Bi-directional Recurrent Neural Network, regional languages, Sentiment Analysis, Support Vector Machine and twitter data.

I. INTRODUCTION

Sentiment analysis is one of the most active research field in the Natural Language Processing (NLP) [1]. The twitter users provide their sentiment about the product or event in the twitter and with the more number of active users, more and more data are available in twitter like social media. The data mining techniques can be applied to these data to retrieve the user sentiment related to the product, news or event accurately [2]. Typically, sentiment analysis is applied to classify the emotions in three categories i.e., positive, negative and neutral [3]. The millions of active users are providing the information about their opinion in the social media that provides the effective feedback about the product or movies [4]. Sentiment analysis is process of NLP that extract the content in the text format and analyze the sentiment of the user [5]. Sentiment analysis has been used in the various applications like product, event or movie feedback [6].

Distributors in the motion picture industries are need to make appropriate decision to distribute the film to the theaters to earn more profits [7].

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Kumar R G*, Department of Computer Science & Engineering, Bharathiar University, Coimbatore, India.
Dr Shriram R, Department of Computer Science & Engineering, Bharathiar University, Coimbatore, India.

The film industries have been developed and provide the online streaming services include TV shows, movies etc.., [8]. The accurate forecasting of sentiment analysis in the Indian movie based on the user opinion in the microblog data helps these industries to earn more profit [9]. Many existing methods has been applied in the sentiment analysis and the standard methods like SVM, random forest provides the considerable performance [10]. In this research, the Bidirectional Recurrent Neural Network (BRNN) is applied in sentiment analysis of the Indian movie. The BRNN method is trained using all input information in the backward and forward data, simultaneously. So, this helps to increase the performance of the sentiment analysis in the Indian movie. The proposed BRNN method has the overall accuracy of 50.32%, while the existing method SVM method has the accuracy of 38.73%.

The paper is organized as, the section II presented Literature review of the existing method, the proposed BRNN in sentiment analysis is explained in the section III, experimental results is discussed in the section IV and conclusion is in the section V.

II. LITERATURE WORKS

This section is focused on the recent research in the regional language and movie sentiment analysis techniques and its efficiency. In this section we have analysed the limitation of the existing work in sentiment analysis in movie recommendation.

Shriya et al. [11] uses the supervised machine learning algorithm to classify the movie sentiment in two classes namely: positive and negative. Machine learning techniques such as the SVM, decision tree, Naïve Bayes and Maxent classifiers are applied to classify the Tamil movie reviews. The datasets are prepared from the different source of webpages and TamilSentiwordnet are used to select the features from the data. This study shows that the SVM has the higher performance in the classification of the Tamil reviews.

Kumar, et al. [12] uses the hybrid techniques of collaborative filtering and content-based filtering for recommend the movies based on the microblogging data. The data related to the movies are collected from the microblogging websites to understand the user reviews about the movie.

Mandal et al. [13] developed code-mixed data with standard Bengali-English language and provide the tag for the sentiment analysis.

The hybrid technique is used such as the combination of the rule based and supervised models for both sentiment and language tagging. The performance of sentiment analysis is need to be improved and applying the machine learning can improve the performance.

Li et al. [14] attempts to bridge the gap between the movie and TV series viewer's domain with the social media. They tried to solve the "cold-start" problem in TV series and movies recommendation system. The proposed method has the ability to handle the unstructured data in the social media dataset. The proposed method has the lower efficiency and running time is required to reduced.

Shalini et al. [15] uses the Convolution Neural Network (CNN) for the increase the performance in the sentiment analysis in the Indian movies. The classification has been made into three classes: (i) positive, (ii) negative and (iii) neutral. The data related to the Telugu movies are collected from the microblogs. The accuracy of the method is high for some data in the sentiment analysis. The reliability of the method is need to be increased for the sentiment analysis of the Indian movie.

Methods in the sentiment analysis are having the considerable performance in classifying the data. Most of the method applied to the Indian movie sentiment analysis is not efficient in microblogging data. Here, the BRNN technique is proposed to increase the performance of the sentiment analysis.

III. PROPOSED METHOD

Many existing methods in the sentiment analysis in regional languages have low efficiency and high computational time. In this research, BRNN techniques is proposed to increase the performance of the sentiment analysis in the Indian movie. The BRNN train the network in the forward and backward data. The tweets related to the Telugu movies are collected from the twitter and analyzed the sentiment based on BRNN. The microblogging data are informal data, which involves in the spelling errors, rare words and variation of the same words. The feature is extracted from the tweet and errors are need to be corrected. The character trigram is used to embedded the sentences instead of words. This technique handles the spelling errors and rare words in the informal data. To further handle the informal data, the morphology analyzer has been used that segment the words into its constituent morphemes. This method is the vector-based representation and provide as input to the BRNN. The result of the BRNN is classified as positive, negative and neutral in the sentiment analysis. The block diagram of the proposed BRNN in sentiment analysis is shown in the Fig. 1.

A. Bidirectional Recurrent Neural Network

Each sentence pair is mapped into $\left[l_i^1, l_i^2\right]$ such that $l_i^1 \in I¡^m$ and $l_i^2 \in ¡^n$, where m *and* n are total number of trigrams character in both languages respectively. BRNN model encodes the sentence twice, one is in the original (forward) order and another is in reverse (backward) order. BRNN measure the weight for both order independently. The algorithm works in the same way as the back propagation, except that the back-propagation is occurs in the hidden states of the unfolded timesteps. Element-wise Rectified Linear unit (ReLU) is applied to the output encoding of the BRNN. ReLU is expressed as: $f(x) = \max(0, x)$. The ReLU simplifies the back-propagation process, increases the learning rate and avoid saturation. The architecture of back-propagation last dense feed forward layer converts the ReLU layer into the fixed vector of $s \in ¡^d$. In this architecture, the value of d is set to 128. The overall model is formulated as in Eq. (1).

$$s = \max\left\{0, W\left[fw, bw\right] + b\right\} \tag{1}$$

Where W is a learned parameter matrix (weights), fw is the forward RNN encode the sentence, bw is the backward RNN and bias term is b.

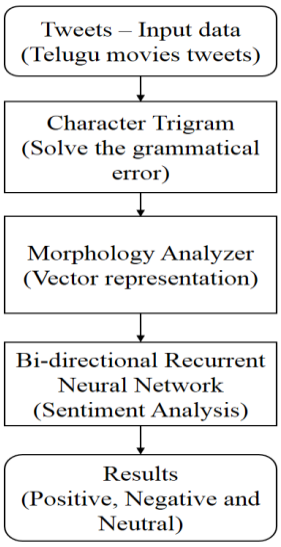

Fig. 1. The block diagram of BRNN in sentiment analysis

1) Training and testing Phase.

The pair of sentences from the twitter data are consider to train the model to find the similarity in their classes. The developed model is differing from the other deep learning model in the parameter sharing. The network is trained with the shared set of parameters, which not only reduce the computation and also helps to represent the data in the shared space to find the similarity. The shared parameter learns to minimize the distance between the sentence in the same class and increase the distance between the distance in the different class.

Consider an input, l_i^1, l_i^2 where l_i^1 and l_i^2 are sentence from the data and label $y_i \in \{-1, 1\}$, the loss function is expressed in the Eq. (2).

$$loss\left(l_i^1, l_i^2\right) = \begin{cases} 1 - \cos\left(l_i^1, l_i^2\right) & y = 1; \\ \max\left(0, \cos\left(l_i^1, l_i^2\right) - m\right) & y = -1; \end{cases} \tag{1}$$

Where m is the margin that measures the distance in the dissimilar pairs that moved from each other.

The value is generally varying from 0 to 1. The loss function is minimized such that the similar class has the value of 1 moves towards each other and the dissimilar class have the value of -1 moves away from each other in the problem space. The overall loss function in batch wise is used to train the network. The objective of minimize is given in the Eq. (3).

$$L(\Lambda) = \Sigma_{(l_i^1, l_i^2) \in CUC'} loss(l_i^1, l_i^2) \quad (3)$$

Where C contains the batch of the sentiment sentences in the same class and C' has the batch of sentiment sentences in the different classes. Back-Propagation Through Time (BPTT) updates the shared parameter using the BiLSTM sub-networks.

For testing, samples are randomly selected with a certain number of sentences for class R_{class} from the datasets consists of large amount of data. For each testing data, the training model is applied to find the similarity between the test and train the model to classify the respective data. Finally, the given test data R_{class} with the most similar data to classify the data.

B. Genre Classification

In addition to the sentiment analysis, this method provides the genre classification for the Telugu movies based on its discussion in the twitter. The dictionary is build and updated based on the Term Frequency-Inverse Document Frequent (TF-IDF) [17] and Latent Dirichlet Allocation (LDA) [17]. This method is used to classify the genre of the movies based on its discussion.

1) Term Frequency-Inverse Document Frequent.

The TF-IDF techniques is common way to represent the document and highly used in the document classification method. This is based on the bag of words, which is the collection of the words that are represented in the vector form. The equation of the TF-IDF is shown in the Eq. (4) [17].

$$TF - IDF_{ij} = tf_{ij} \times \log\left(\frac{N}{df_{i+1}}\right) \quad (4)$$

2) Latent Dirichlet Allocation.

The main aim of LDA is to find the latent themes or topics that are related to the specific data based on the frequent of words in the document and in this case, twitter data. The process is based on that the single document is generated by a pre-defined probabilistic process. Thus, this is considered as unsupervised generative topic model. The LDA consider that the genre of the movie is based on the distribution of words in the data [17].

$$p(\phi \mid \beta) = \Pi_{k=1}^{k} \frac{r(\beta_k)}{\Pi_{v=1}^{v} r(\beta_{k,v})} \Pi_{v=1}^{v} \phi\_k, v_{k,v}^{\beta-1} \quad for\ \phi_k : Dir(\beta) \quad (5)$$

The word probability distribution ϕ per topic are calculated using the given data considered that the movie genre is based on the word distribution, as shown in Eq. (5).

IV. EXPERIMENTAL SETUP AND RESULT ANALYSIS

Sentimental Analysis is the NLP technique in the microblog data and that helps in the stock market analysis, film industries etc. Sentiment analysis in the regional language based on the twitter data is difficult method. Various methods are provided for increases the performance of the sentiment analysis in the regional languages and that involves in low efficiency. In this research, the BRNN method is applied in the sentiment analysis in the regional languages. This section gives the description about the dataset, experimental setup, performance and comparison analysis.

A. Dataset

For evaluating the performance of the proposed BRNN method, the data is mined directly from the twitter. The tweets related to the Telugu movies are extracted from the twitter based on tool of twitter API. These data are used to measure the performance of the proposed BRNN method.

B. Experimental Setup

The proposed BRNN method in sentiment analysis is evaluated using the system having the configuration of Intel i7 processor, 8GB of RAM, and 500GB hard disk. The existing method is also evaluated in the same scenario as in the proposed method. The proposed method is developed and tested on the python 3.

C. Metrics

The metrics used in this method to evaluate the efficiency of the proposed BRNN method is accuracy, precision, recall and f-measure. The formula for measure the accuracy, precision, recall and f-measure are shown in the Eq. (6-9), respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (6)$$

$$\text{Precision} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{Re}call = \frac{TP}{TP + FN} \quad (8)$$

$$F - Measure = \frac{2 \times precision \times recall}{precision + recall} \quad (9)$$

Where TP represents the True Positive, TN represents the True Negative, FP represents the False Positive and FN represents the False Negative. The f-measure is the perfect combination of the precision and recall.

D. Performance Analysis

The BRNN is applied in the collected data in the Telugu language and classify the sentiment into three classes. The obtained results for the sentiment analysis in the Telugu movie by the proposed BRNN is shown in the Table 1.

The precision value is achieved as 77% for the positive tweets and 100% for the negative tweets. The f-measure of the proposed BRNN method in sentiment analysis has achieved as 75% for the neutral sentiment and 73% for the positive sentiment.

Table 1. The BRNN method in the Sentiment Analysis

BRNN Method	Neutral	Positive	Negative
Precision	0.66	0.77	1
Recall	0.87	0.68	0.2
F-Score	0.75	0.73	0.34

The BRNN method has the analyzed the sentiment from the twitter data. The sentiment analysis from the BRNN method is classified as positive, negative and neutral. The most of the tweets are neutral about the movies, positive tweets are more and negative tweets are less. In the genre classification, the "bahubali" movie is classified as the action movie based on the discussion in the twitter.

E. Comparative Analysis

The proposed method is compared with existing method in the sentiment analysis to analyze its performance. The existing method such as the random forest and SVM classifier has the higher efficiency in the sentiment analysis. Therefore, these two methods are considering for the comparison of the proposed BRNN method. The three classifiers are applied for the same twitter data in the same environment for the effective analysis of the proposed BRNN technique.

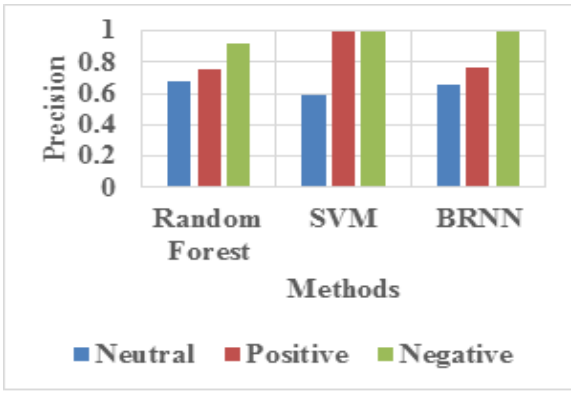

Fig. 2. The precision for the various methods

The precision value is measure for the proposed BRNN and the existing methods in the sentiment analysis of the Indian movies. The precision value is calculated for the three values such as positive, negative and neutral. The method has the higher performance in correctly classify the negative sentiment. The proposed BRNN has the higher performance than other two classifiers in the sentiment analysis, as shown in Fig (2). The SVM has the higher performance in classifying the positive value than the BRNN method. The overall performance is high for the proposed BRNN method than the SVM method in sentiment analysis. The proposed method has the higher performance due to representation of the resource rich and poor sentences in the common space based on the similarity between annotated tags. The proposed BRNN method has the higher precision value of the 100% in the negative sentiment and the existing method has the precision of 100% and 92% for SVM and random forest, respectively.

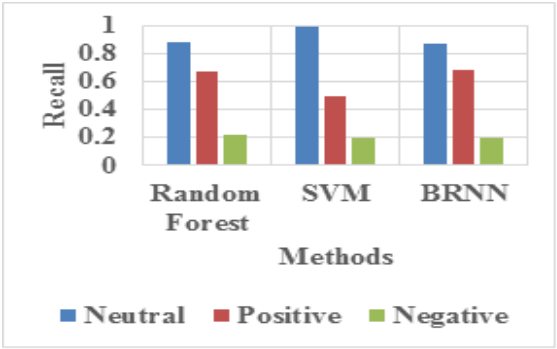

Fig. 3. The recall value for the various method in sentiment analysis

The recall value is measured for the different techniques in the sentiment analysis in Indian language. The metrics such as recall is measured for the three classes like neutral, positive and negative, as shown in the Fig. (3). The overall recall value is higher for the proposed BRNN method than the SVM method. The proposed BRNN method has the higher recall value for the sentiment analysis in Indian movie data. The recall value of the BRNN in sentiment analysis is achieved as 68% for the positive sentiment and the existing method recall value of the random forest and SVM method is 67% and 50%, respectively.

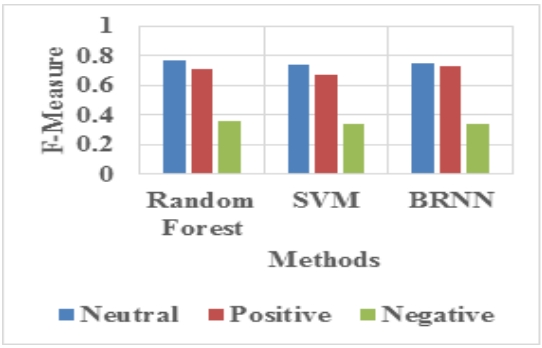

Fig. 4. The F-Measure for the various method in sentiment analysis

The F-measure is measured for the different methods in sentiment analysis and compared with the existing method such as random forest and SVM, as shown in Fig (4). The proposed BRNN method has overall higher performance than other existing methods. The proposed BRNN has the F-measure of 73% for the positive analysis while the existing method such as SVM and Random forest has the F-measure of 71% and 67%, respectively.

The overall accuracy of the various method in sentiment analysis of the Indian movies is measured and compared in the Fig. (5). This shows that the proposed BRNN method has the higher accuracy than the other existing methods in sentiment analysis. The proposed BRNN has the overall accuracy of 50.32% and the existing method such as Random forest and SVM has the overall accuracy of 41.21% and 38.73%, respectively.

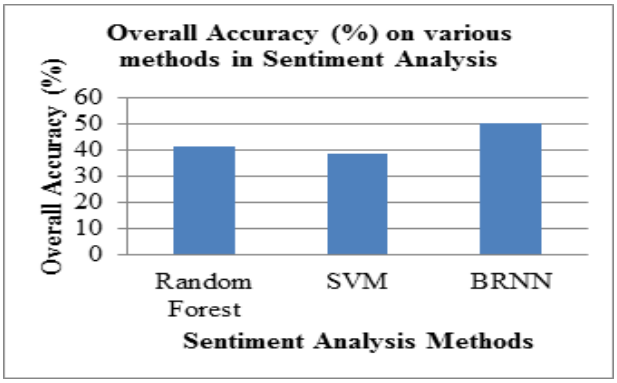

Fig. 5. The overall accuracy of the various method in sentiment analysis

Therefore, the proposed BRNN has the higher performance than other existing methods in the sentiment analysis. The proposed BRNN has the advantages of the representing the high and low resource sentences in the common space and classify the sentiment based on the similarity measure between the annotated tags.

V. CONCLUSION

Sentiment Analysis is used in many applications and process in the regional languages, useful for many applications. In this research, the BRNN is applied for the sentimental analysis in the regional languages to increase the performance. The twitter data related to the Telugu movies are collected and processed for the sentiment analysis. The proposed BRNN method has the advantages of the representing the low and high resources in the common space and classify them based on the similarity between the annotated tags. The different metrics are measured from the BRNN method for three classes. The proposed BRNN method has the f-measure of 73% in positive sentiment and existing method such as Random forest has the f-measure of 71%. The future work of the proposed method involves in using the effective representation of text and optimization technique to increase the performance of the sentiment analysis.

REFERENCES

1. C. Yang, H. Zhang, B. Jiang, and K. Li. (2019). Aspect-based sentiment analysis with alternating co attention networks. Information Processing & Management, 56(3), pp. 463-478.
2. M. Song, H. Park, and K. S. Shin, (2019). Attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean. Information Processing & Management, 56(3), pp. 637-653.
3. C. Wu, F. Wu, S. Wu, Z. Yuan, J. Liu, and Y. Huang. (2019). Semi-supervised dimensional sentiment analysis with variationalautoencoder. Knowledge-Based Systems, 165, pp. 30-39.
4. C. Diamantini, A. Mircoli, D. Potena, and E. Storti. (2019). Social information discovery enhanced by sentiment analysis techniques. Future Generation Computer Systems, 95, pp. 816-828.
5. O. Araque, G. Zhu, and C. A. Iglesias. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. Knowledge-Based Systems, 165, pp. 346-359.
6. S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi. (2019). Sentiment analysis based on improved pre-trained word embeddings. Expert Systems with Applications, 117, pp. 139-147.
7. M. Hur, P. Kang, and S. Cho. (2016). Box-office forecasting based on sentiments of movie reviews and Independent subspace method. Information Sciences, 372, pp. 608-624.
8. H. Li, J. Cui, B. Shen, and J. Ma. (2016). An intelligent movie recommendation system through group-level sentiment analysis in microblogs. Neurocomputing, 210, pp.164-173.
9. D. Zimbra, K. R. Sarangee, and R. P. Jindal. (2017). Movie aspects, tweet metrics, and movie revenues: The influence of iOS vs. Android. Decision Support Systems, 102, pp. 98-109.
10. A. C. Pandey, D. S. Rajpoot, and M. Saraswat. (2017). Twitter sentiment analysis using hybrid cuckoo search method. Information Processing & Management, 53(4), pp. 764-779.
11. S. Se, R. Vinayakumar, M. A. Kumar, and K. P. Soman. (2016). Predicting the sentimental reviews in tamil movie using machine learning algorithms. Indian Journal of Science and Technology, 9(45).
12. S. Kumar, S. S. Halder, K. De, and P. P. Roy. (2018). Movie Recommendation System using Sentiment Analysis from Microblogging Data. arXiv preprint arXiv:1811.10804.
13. S. Mandal, S. K. Mahata, and D. Das. (2018). Preparing bengali-english code-mixed corpus for sentiment analysis of indian languages. arXiv preprint arXiv:1803.04000.
14. H. Li, J. Cui, B. Shen, and J. Ma. (2016). An intelligent movie recommendation system through group-level sentiment analysis in microblogs. Neurocomputing, 210, pp. 164-173.
15. K. Shalini, A. Ravikurnar, R. C. Vineetha, R. D. Aravinda, K. M. Annd, and K. P. Soman. (2018). Sentiment Analysis of Indian Languages using Convolutional Neural Networks. In 2018 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-4.
16. N. Choudhary, R. Singh, and M. Shrivastava. (2018). Cross-Lingual Task-Specific Representation Learning for Text Classification in Resource Poor Languages. arXiv preprint arXiv:1806.03590.
17. D. Kim, D. Seo, S. Cho, and P. Kang. (2019). Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Information Sciences, 477, pp. 15-29.

AUTHORS PROFILE

Mr. Kumar R G has 7+ years of Academic experience and handles courses for CSE (B.Tech & M.Tech), MCA, he has 4+ years of research experience.

He has done his B.E. in CSE from Veltech Engineering College affiliated to Anna University and M.Tech in Software Engineering from B.S.Abdur Rahman University where he stood second top and received university Silver Medal for his PG course. Now he is pursuing his Ph.D. in Bharathiar University, Coimbatore. He has guided 10+ UG projects in his experience and 2 PG projects. Kumar acts as a Single Point of Contact (SpoC) for APSSDC (State level Skill Development Corporation) and Coordinator for ICT Academy for his working institution in Andhra Pradesh.

His area of interest are Mobile Computing, Sentiment Analysis and Data Mining related works. He also established Research Projects in the Department of CSE and also organized International, National Conference, FDP, seminar, Workshop, and symposium. He is a member of Indian Society for Technical Education (ISTE) since 2014, Internet of Society and has applied for IEEE membership.

Mr. Kumar R G is a Asst. Professor by choice and has a passion for teaching.

Dr. Shriram Raghunathan works on Natural Language Processing, Instructional Design and Gaming. He completed his PhD in Computer Science and Engineering in 2008 and did his Masters in Instructional Design (online) from Open University of Malaysia. His significant achievements include work on setting up India's first Centre of Excellence in Pervasive Computing in 2007 and completion of 4 funded projects (mobile Tamil interfaces, mobile keypad standardization and plagiarism checking) in Tamil Computing since 2004. Shriram has published widely in domains of cloud computing, natural language processing and instructional design and guided 4 Doctoral students for their PhD since 2014. Now he heads the Gaming Technology Division at VIT Bhopal.