Wireless Personal Communications
https://doi.org/10.1007/s11277-021-08580-3

# Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM)

**Usha Devi Gandhi[1] · Priyan Malarvizhi Kumar[2] · Gokulnath Chandra Babu[3] · Gayathri Karthick[4]**

Accepted: 4 May 2021


**Abstract**
Twitter sentiment analysis is an automated process of analyzing the text data which determining the opinion or feeling of public tweets from the various fields. For example, in marketing field, political field huge number of tweets is posting with hash tags every moment via internet from one user to another user. This sentiment analysis is a challenging task for the researchers mainly to correct interpretation of context in which certain tweet words are difficult to evaluate what truly is negative and positive statement from the huge corpus of tweet data. This problem violates the integrity of the system and the user reliability can be significantly reduced. In this paper, we identify the each tweet word and we are assigning a meaning into it. The feature work is combined with tweet words, word2vec, stop words and integrated into the deep learning techniques of Convolution neural network model and Long short Term Memory, these algorithms can identify the pattern of stop word counts with its own strategy. Those two models are well trained and applied for IMDB dataset which contains 50,000 movie reviews. With huge amount of twitter data is processed for predicting the sentimental tweets for classification. With the proposed methodology, the samples are experimentally collected from the real-time environment can be discriminated well and the efficacy of the system is improved. The result of Deep Learning algorithms aims to rate the review tweets and also able to identify movie review with testing accuracy as 87.74% and 88.02%.

**Keywords** Sentiment analysis · Stop words · Word2vec · CNN · LSTM

✉ Priyan Malarvizhi Kumar
mkpriyan@khu.ac.kr

Usha Devi Gandhi
ushadevi.g@vit.ac.in

Gokulnath Chandra Babu
gokulkapoor@gmail.com

Gayathri Karthick
gk419@live.mdx.ac.uk

[1] School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India

[2] Department of Computer Science and Engineering, Kyung Hee University, Seoul, Korea

[3] Vellore Institute of Technology, Vellore, India

[4] School of Science and Technology, Middlesex University, London, England

Published online: 17 May 2021

## 1 Introduction

Natural Language Processing (NLP) refers to the artificial intelligence method of communicating with an intelligence system using the natural language. The most used applications of NLP is using "sentiment analysis" for Twitter or Facebook sentiment analysis is been used. Then next level of applications is chatbot, speech recognition, machine translation, spell checking, keyword searching, information extraction, and advertisement matching. With the help of this, it can develop the tweet application which will be able to understand the human readable languages. It is one of the part with data science which holds the automatic process of understanding, deriving and analyze the text data in a desired way. With the help of this, a person can manage the huge amount of text data, perform many typical tasks and also give the solution for particular range of problems for the above given next level applications of NLP.

Sentiment Analysis accumulates from the use of Natural Language Processing (NLP), tweet analyzing, and biological data to automatically establish and calculate the personal feelings of tweets data. The existing work of Sentiment Analysis is to control the opinion of tweets with positive, negative and neutral. Sentiment Analysis is widely available to the voice of public reviews and survey responses, and also including medical coding data that varies from marketing business to public service. It can be applicable to write the text data, images, videos, speech recognition, etc. Sentiment Analysis works for similar kind of tweet data, after getting this data first it will split the input into separate word or sentences. This is called as Tokenization.

A movie review is like a vocabulary which gives opinion about the movie and predicts the opinion from the public as positive or negative. From this we can able to get some idea to watch the particular movies or not. It will be a collection of director, stars, co-workers who are participators to work for movies. About movie success or failure is depending upon the public reviews posted in a particular movie websites. [1] Websites are the key features to the humans where they can showcase their feelings or opinion about the movie and posted their reviews in the internet. Therefore, public reviews are collected as a tweet data and then some special operations are performed for this data to identify the behavior of consumers who purchase the movies and also for predicting any new availability in the movie market field. This pattern is called as "opinion mining" or "sentiment analysis" [2].

Mostly, Sentiment Analysis focused on social media which contains the following sources such as, Facebook, Twitter, and IMDB [3] are increasing the demand of public opinions and collects it as a text format. Furthermore, predicting the tweet data with correct interpretation of text for movie review is a challenging task.

Deep learning is a subset of machine learning techniques based artificial neural networks which use multiple levels of layers for extracting the huge level features from the raw data. It has been used in the variety of fields like signal and image processing. Since, Deep Neural Networks holds the multiple neural networks, where the outcome of one network is an input to the next network and so on. Deep learning study about the text data features on its own way, and it also study about the multiple layers of features for predicting that data. The movie review dataset deep learning has been used in opinion mining or sentiment analysis.

This paper contains IMDB dataset of sentiment analysis for English movie reviews by using deep learning models and also classifies this reviews into positive and negative respectively.

Since the text data is being interchanged every second due to the massive amount of data, neural networks and machine learning techniques were analyzed for this text data is not sufficient to use for the huge corpora. In this paper, we propose the sentiment analysis on twitter data for movie reviews by using popular deep learning models. The main objectives of this paper is to understand each parameter of convolution neural network and long short term memory, understand how CNN and LSTM work with stop words and word2ec methods. We also preprocessing this NLP feature stop words, used to ignore the commonly occurred tweet words from the IMDB dataset. As, this words will occupy more space in our movie review data or it will not give meaningful sentence for those data. For this reason we remove those words by using this stop words. Word2vec is used to connect the tweet words into vector format in different dimensions with each separate tweet word, and assigned a corresponding vector space for that particular separated word. Gensim provides the Word2vec which works for huge amount of tweet data.

## 2 Related Work

Sanjeev Ahuja [3] used the SentiWordNet for obtaining the overall polarity of the movie review tweets which will get the tweet data information form WordNet. It includes sentiment of words and comprises into sentiment scores with positive, negative and also objective scores. Calculation of all these scores gives the relative strength into it. The State of Art approach has been included for increasing the performance and accuracy of the tweet data.

Rachana Bandana, [4] proposed the sentiment analysis with document level for the movie review classification applied by heterogeneous features, for supporting the supervised machine learning algorithms as Naive Bayes and Linear Support Vector Machine. With these the author learns and classified the text reviews into positive and also a negative list.

Purtata Bhoir, [5] used the subjectivity analysis with two methods as SentiWordNet and Naive Bayes. Among these Naive Bayes classifier is a probabilistic model theorem to calculate the probability with a class of individual sentences or text. Naive Bayes analyzed for extracting the feature and opinion from the public tweets and gives better solution compared to SentiWordNet.

M. Ali Fauzi, [6] shows the ensemble technique with Naive bayes for sentiment analysis of movies on Indonesian Twitter data. The ensemble features includes twitter specific features, textual features, part of speech features, and lexicon-based features, and Bag of Words. They came up with lexicon-based features which hold the percentage of positive and negative tweet words based on their part of speech.

V.K. Singh, R. Piryani, A. Uddin, [7] introduces the sentiment classifiers also called as IR formulations are used not only for reviewing the positive and negative tweet command, also used for highlighting the tweet products with a very short period of time used in different domains.

Vasilija Uzunova, [8] proposed sentiment analysis of film reviews in Macedonian using Naive Bayes. With the help of sentiment polarity they classified the film comment text and checked whether it is positive or negative based on opinion of the public reviews.

H. Ghorbel and D. Jacot, [9] introduced the sentiment analysis for French movie reviews with the help of shallow linguistic features. Three different categories are performed for this movie reviews. One is “lexical” approach determines from the word

unigrams and it find some similar tweets in positive and negative reviews. Another one as "morpho-syntactic" approach which reduces the tweet features into part of speech categories for improving the performance of data. Then the "semantic" approach uses the lexical feature of sentiwordnet to increase the polarity of tweet word and calculating the overall polarity score of the movie review for each part of speech tag.

Liang-Chu Chen et al., [10] proposed the sentiment analysis of social network for Military life board on largest online communities in Taiwan. The text mining occupied some systematic method for this online community along with the help of self-organized military sentiment vocabulary. A web crawler method used for this military life PPT board to extract the content of military posts and message blogs and also analyze this online community which calculates it from the different parameters.

Ishan Arora, [11] introduced the Naive Bayes classifier for calculating the conditional independence of tweet class as positive or in a negative form and the tweet words are conditionally independent with each other. With the help of this the accuracy of the tweet data classification will not get any issue and it gives a fastest performance for the classification problem.

Deepa Anand and Deepan Naorem, [12] works on filtering the sentiment analysis statements from the public reviews and grouping it with a corresponding aspects of categories. The review of sentence filtering is used to detect the accuracy of sentiment tweets and it is different from the subjectivity classification. The tweet sentence problem denoting a target of public other than the items being reviewed for the movie.

Barkha Bansal et al., [13] works on demonstrating the CBOW model compared with skip- gram model for customer reviews in mobile phone application. The semantic features using the cosine distance measure, for each input of mobile phone.

## 3 Proposed Methodolody

Deep Learning models are implemented for CNN and LSTM techniques and proposed work is showing in Figs. 1 and 2. Figure 1 explains the first step as loading the IMDB movie review dataset and then the next step is to extract the review data with natural language processing features such as stop words and word2vec. The result of this work would be further preceded by convolutional layer with 1 dimensional matrix, and global max pooling is applied after that. After performing fully connected layer the output for CNN test classification will produce the movie review output.

Figure 2 explain about the other model LSTM, it will review about each text data and it will convert the movie review data into tokenize form or in integers, then the embedding layer will converts this integers into some specific size of data. Then, LSTM layer will be defined by hidden state and depends on number of layers allocated. The Sigmoid layer performs the activation of all output values between 0 and 1. From this sigmoid output this model will produce the LSTM test classification result.

### 3.1 Preprocessing of Movie Review Tweets

From the above Figs. 1 and 2, the movie review is extracted using two main features which will be used for training the input data. Its main purpose is to remove the unwanted tags from the most highlighted texts, and removing the unwanted URLs, removing the unwanted

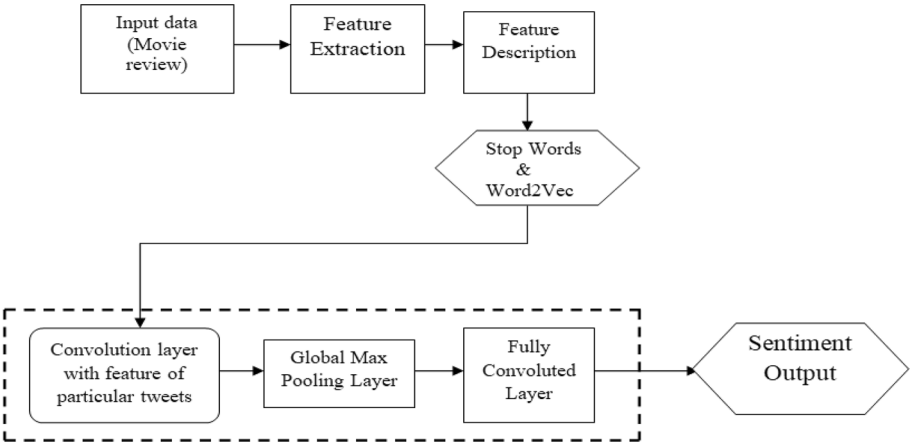


**Fig. 1** Proposed work for CNN Model

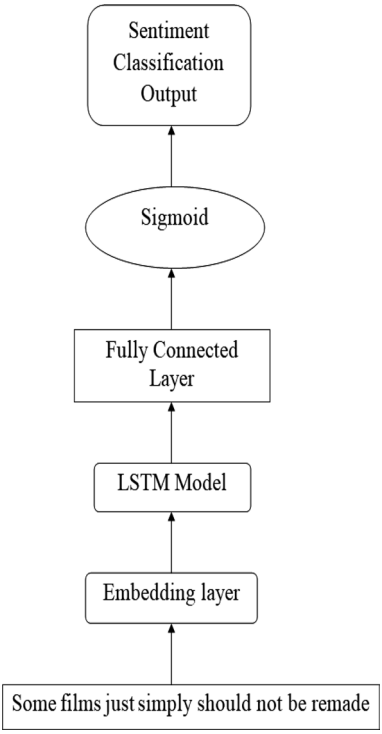


**Fig. 2** Proposed work for LSTM model

links, spaces, and brackets by using those features using nltk toolkit. After removing these all the movie review will show the reviews in lowercase alphabets. These features were used for training the further movie review network data. The following steps are involved for preprocessing:

1. The Tokenization process will split all the reviews in a text form into word form and it will be done for any character of the data, especially on space character.
2. Word2Vec is an embedding model which maps the review words into vector format, and then each and every single tweet word is converted into respective numerical vector form which will be taken from the movie reviews. This Vector format is used for training the CNN and LSTM models and it will join the dimensions into the output array of data.
3. Stop words is a collection of Natural Language processing which removes the unwanted sentences from the movie reviews or removing the repeated words which will be showing from the dataset. It will print the output of most occurring text words in an array format.
4. The tweet data which is passed to any kind of neural network in a binary vector format, the sequential model is an important process to rotate the numeric array representing data from the word2vec which embedding into a digitalized vector format. This format will be passed to build the further classification model of movie reviews.

### 3.2 Dataset for Movie Reviews

The dataset is collected from IMDB (Internet Movie Database) from kaggle.com which contains the information's about the films, director, and producer, cast and also for games and videos etc. It contains the movie review tweets with two attributes namely, review and sentiment.

## 4 Convolutional Neural Network (CNN)

Convolutional Neural Network is a collection of neural networks and gives sufficient information for the field of image processing, image classification and speech recognition areas. [14] The Fig. 1 displayed for CNN model is consist of following layers such as, Initially, the Sequential model is defined as a linear form of layers and passing the different list of layers into a constructor format. The input layer which takes the movie reviews encoded values as an input. The Sequence of movie review data contains different lengths and keras initially gives importance for the tweet inputs were it could be in vectorized form and all the inputs to be in same length. Later, we will pad all those movie review input and it contains the review length of 100. For this instance, we use pad sequences function. Then, we are defining our Embedding layer for training the neural networks. The Embedding layer has contains the vocabulary size of 92,768 and input length of 100 and will choose an embedding space of 300 dimensions. We can now see that our Embedding layer shows the output in a 100*300 matrix form.

Then, the Convolutional Neural Networks having the different level of dimensions as Conv1D, Conv2D, Conv3D. This paper explains about Conv1D to construct the one-dimensional convolutional neural networks for movie review time series of classification and training the review data is mathematically drastic and it consumes more time. It is very useful for calculating the text features from the in-build length of huge amount of data. From this 1D network flow, it holds the filter of 128 feature vectors and it fits the number of output filters used in the convolution network. With 1D convolution layer, a kernel size applied to kernel weight matrix of 5 which contains 5 feature vectors and gives the size of the convolutional window. Then, the next layer GlobalMaxpooling1D applied for 1D temporary data which takes the maximum of vector space compare to step by step dimensions

of neural network. It will collect the maximum vector value of movie review sentences in which the most used vector representing of a word will be 'movie' collecting from the IMDB dataset. The Dropout Layer is used to avoid the over fitting of data and the model showed the dropout rate of 0.2 from the movie reviews. [15] Finally, the activation model contains the name of activation function like relu and sigmoid. After performing convolution operation we have to apply these two activation function into it. Finally, this model compiled, trained and gives the testing accuracy of 87.72%.

## 5 Long Short Term Memory (LSTM)

LSTM is a deep learning based technique that designed for sentimental analysis, modeling the languages, predicting the text data and also for speech analysis. It is considered as most special artificial neural network and having capability to study about the long term dependencies of text data. But LSTM will overcome with this and works mainly to omitting the long term dependencies. [15]

From the Fig. 2, LSTM explains about the layers, as first the movie reviews will be tokenized which means the whole sentence will be separated into single words and convert the text words into tokens i.e. in integer form. Then Embedding layer will convert that tokenized word into some predefined size. Then, the movie tweets are passed into LSTM layer, having the filter size of 128 combined with hidden size of movie reviews. The final output comes from this each LSTM tweet word is combined with the respective movie reviews vector length. This vector length is connected to fully connected layer with sigmoid activation function and also it collect the tweet output from LSTM layer to produce the expected text output size. The sigmoid activation function will transfer all the input values into an output values of movie tweets between 0 and 1. The movie reviews classification output will collect from the sigmoid output, and considered as a final sentiment classification for this model. The output of test classification accuracy for LSTM model is 88.02%.

For summarizing above two models, LSTM gives the validated tweet accuracy of 88.02% whereas CNN gives testing accuracy of 87.72% based on the calculating the measurement of movie reviews (Fig. 3).

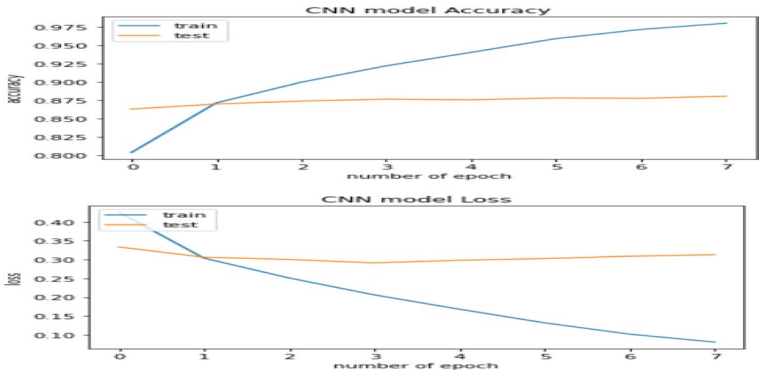


**Fig. 3** CNN accuracy and loss

## 6 Plotting Result for Movie Reviews-CNN and LSTM Accuracy and Loss

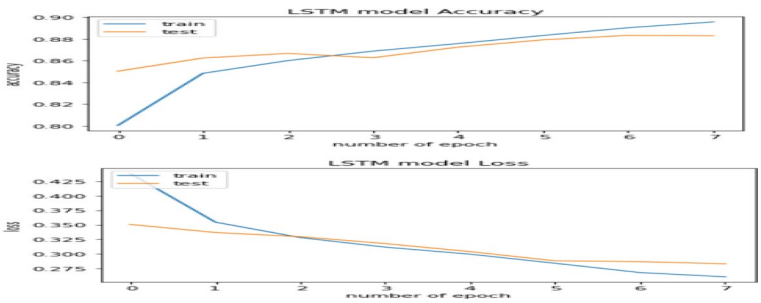


**Fig. 4** LSTM accuracy and loss

## 7 Conclusion and Feature Work

From this working model, we introduced new features that give the high impact on calculating the movie reviews and applied deep learning models with natural language processing methods. Then, we performed the classification for testing the review data and it is used for doing mathematical expression for CNN and LSTM models which are successfully achieved the better results for the feature work (Fig. 4).

For the future work, we will take consideration on CNN model with different techniques and try to work on this model with more complex features to improve the results of text data. Later, verifying this model with huge corpora of tweet data other than IMDB dataset which is able to differentiate our outcome with the implicit and explicit aspect level of deep learning methods.

**Data Availability** Based on the Request.

**Declarations**

**Conflict of Interest** The author declares that they no conflict of interest. The author of this research acknowledges that they are not involved in any financial interest.

**Ethical Approval** Author certifies that this material or similar material has not been and will not be submitted to or published in any other publication before. Furthermore, Author certifies that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript.

## References

1. Yasen, M., Tedmori, S. (2019). "Movies reviews sentiment analysis and classification", IEEE jordon international joint conference on electrical engineering and information technology (JEEIT).
2. Trivedi, S.K., Tripathi, A. (2016). "Sentiment analysis of indian movie review with various feature selection techniques", IEEE international conference on advances in computer applications (ICACA).

3. Lokesh, S., Kumar, P. M., Devi, M. R., Parthasarathy, P., & Gokulnath, C. (2019). An automatic tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. *Neural Computing and Applications, 31*(5), 1521–1531
4. Bandana, R. (2018). "Sentiment analysis of movie reviews using heterogeneous features", 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech) IEEE, (pp. 1-4).
5. Bhoir, P., Kolte, S., (2015). "Sentiment analysis of movie reviews using lexicon approach". IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), (pp. 1-6).
6. Manogaran, G., Shakeel, P. M., Hassanein, A. S., Kumar, P. M., & Babu, G. C. (2018). Machine learning approach-based gamma distribution for brain tumor detection and data sample imbalance analysis. *IEEE Access, 7*, 12–19
7. V.K. Singh, R. Piryani, A. Uddin, "Sentiment Analysis of Movie Reviews and Blog Posts- Evaluating SentiWordNet with different Linguistic Features and Scoring Schemes", IEEE, (2012)
8. Kanisha, B., Lokesh, S., Kumar, P. M., Parthasarathy, P., & Babu, G. C. (2018). Speech recognition with improved support vector machine using dual classifiers and cross fitness validation. *Personal and ubiquitous computing, 22*(5), 1083–1091
9. Ghorbel, H., & Jacot, D. (2011). *Sentiment Analysis of French Movie Reviews*. (pp. 97–108). Berlin Heidelberg: Springer.
10. Chen, L. C., Lee, C. M., & Chen, M. Y. (2019). *Exploration of social media for sentiment analysis using deep learning*. Springer.
11. Narayanan, V., Arora, I., & Bhatia, A. (2013). *Fast and accurate sentiment classification using an enhanced Naive Bayes model*. (pp. 194–201). Springer.
12. Anand, D., Naorem, D., (2016). "Semi-supervised aspect based sentiment analysis for movies using review filtering". 7th International conference on Intelligent Human Computer Interaction, IHCI.
13. Bansal, B., Srivasta, S., (2018). "Sentiment classification of online consumer reviews using word vector representations". International Conference on Computational Intelligence and Data Science (ICCIDS), (pp. 1147–1153).
14. Jianqiang, Z. H. A. O., & Xiaolin, G. U. I. (2018). AND Zhang XUEJUN, "deep convolution neural networks for twitter sentiment analysis." *IEEE Translations and content mining, 6*, 2169–3536
15. Heikal, M., Torki, M., EI-Makky, N., (2018). Sentiment analysis of arabic tweets using deep learning", 4th International Conference on Arabic Computational Linguistics, (pp. 114–122).




**Usha Devi Gandhi** is working as an Associate Professor in the School of Information Technology and Engineering, Vellore Institute of Technology University. He received his Bachelor of Engineering and Master of Engineering degree from the Anna University. His current research interests include big data analytics and wireless networks. He has published number of international journals and conferences

**Dr. Priyan Malarvizhi Priyan** is currently working in Department of Computer Science and Engineering, Kyung Hee University, South Korea. Before that he is Postdoctoral Research Fellow in Middlesex University, London, UK. He had completed his Ph.D. in the Vellore Institute of Technology University. He received his Bachelor of Engineering and Master of Engineering degree from Anna University and Vellore Institute of Technology University, respectively. His current research interests include Big Data Analytics, Internet of Things, Internet of Everything and Internet of Vehicles in Healthcare. He is the author/co-author of papers in international journals and conferences including SCI indexed papers. He has published 38 papers in which 2 in IEEE Access, 1 in IEEE Transactions, 1 in ACM Transactions, 6 in Elsevier Publication and 16 in Springer Publications. He is a reviewer for Elsevier, IEEE Access, IEEE Transactions, and Springer journal. He is a lifetime member in International Society for Infectious Disease, Computer Society of India and member in Vellore Institute of Technology Alumni Association. He has been given the Best Researcher Award for the year 2017 and 2018 at Vellore Institute of Technology University.

**Gokulnath Chandra Babu** is currently working as a Teaching cum Research Associate in VIT University. He is currently doing his Ph.D. in the Vellore Institute of Technology University. He received his Bachelor of Engineering and Master of Engineering degree from Anna University and Vellore Institute of Technology University, respectively. His current research interests include Big Data Analytics, Internet of Things, Internet of Everything and Internet of Vehicles in Healthcare. He is the author/co-author of papers in international journals and conferences including SCI indexed papers. He has published many papers in which 2 in IEEE Access, 1 in ACM Transactions, 3 in Elsevier Publication and 7 in Springer Publications. He is a reviewer for Elsevier ,IEEE Access, IEEE Transactions.,and Springer journal. He is a lifetime member in International Society for Infectious Disease, Computer Society of India and member in Vellore Institute of technology Alumni Association.

**Gayathri Karthick** is experienced Research fellow with a demonstrated history of working in the information technology and a teaching portfolio. Strong research professional with a Doctoral's degree focused in Computer Science from Middlesex University, London, United Kingdom. Extensively published in conferences, journals with relatively expertise in setting up Cloud computing practical Labs.