iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.46025

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

CHATDOC(Medical Assistance)

Pallavi CV[1], K Prasanna Kumar[2], Akshath MG[3], Madhu Lokesh[4], Jeevan S V[5]
[1, 2, 3, 4, 5]*BNMIT*

Abstract: ***We are now living in a world where visiting hospitals for our regular health checkups are becoming difficult due to the afraid of getting infected by different Viruses(COVID 19). Also the number of patients consulting the doctor becoming more and more because of which the patients have to wait for a long time to get consultation from the doctor. Our idea to avoid these type of problems in the Health care field is to create a AI Medical Chatbot using Natural language processing which is used to understand the input which are usually symptoms of different dieseases provided by the user also with the help of Neural networks and different machine learning algorithms the system predicts the disease also it helps the user in providing the medication. The chatbot is designed to provide medication to mild diseases so that user can reduce to visit hospital for mild dieseases. Chatbots are programs that works on Machine learning, Natural language processing(NLP) which uses NLTK to understand the human language.***

Keywords: Medical/Healthcare chatbot,chatdoc, NLP, Machine learning, Neural networks, Artificial intelligence.

I. INTRODUCTION

As the world becoming more and more digitalized in many areas such as financial sector, Marketing sector etc..Now a days technology is also playing a big role in the Healthcare sector.Due to COVID-19 people are very scared to visit the hospitals for their routine check up, also the people visiting the Hospitals has seen a drastic raise in numbers due to which they have to wait for a long time to visit the doctor.To provide solutions for these problems in Healthcare field we have developed a web based application called CHATDOC.

With the help of different available technologies such as Machine Learning, Artificial intelligence,Neural networks, Natural language processing we thought of developing a user friendly chatbot dedicated to the Medical field.

Chatbots are computer programs that simulates and processes the human conversation either in the written or spoken format with the help of Natural language processing(NLP) which uses different techniques or algorithms to understand the human input provided to the chatbot. With regard to the Medical chatbot the input will be the symptoms of different dieseases. With the help of Neural networks that is LSTM which is a special kind of Reccurent neural networks, it will separate the words in the given input sentence and if it gets matched with the targeted tags, then it will select the best accurate response. Since we are talking to the chatbot there is no physical contact with the doctor thus the risk of getting infected by the virus is eliminated also it saves the time of the patient.

II. LITERATURE REVIEW

Research Paper's gives us the information about various technologies being used in building a user friendly medical chatbot. The different technologies include various Machine learning algorithms that includes Random forest, Naïve bayes, Decision tree etc .These algorithms mainly helps us to predict the most accurate disease based on the user symptoms given to the chatbot/machine which will be in the form of a sentence.With the help of various research paper's on can be clear of which algorithms he/she should use in building the chatbot so that they can get better accuracy.

The paper[1] A survey on Different Algorithms used in Chatbot gives us the information about various ML algorithms that are being used for different chatbots. It gives us the comparison of different Machine learning algorithms and the accuracy differences between the algorithms with the result of each ML algorithms. With this paper one can know that which Machine learning algorithm is best to use to get better accuracy.

The paper[2] A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Lives it mainly made use of four different supervised Machine learning algorithms, Decision tree, Naïve Bayes, Random forest and KNN Algorithms. This paper gives us the understanding and working of each supervised Machine learning algorithms with the accuracy of each algorithm.

The paper[3] Doctor Chatbot: Heart Disease Prediction System it mainly used to predict the presence of heart disease of the user. It mainly made use of Machine learning algorithms such as K-Nearest Neighbour(KNN), ANN, Support vector machine(SVM), Naïve bayes, Decision tree etc..This paper concludes that among all the different algorithms SVM on a Heart disease Dataset gives the best possible accuracy.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

The paper[4] Contextual Chatbot for Healthcare Purpose (using Deep Learning) it mainly made use of Neural networks, which is used to train the data to obtain the accurate results, It also made use of Natural language processing (NLP) with Deep learning for getting better results.They have combined the concepts of TensorFlow, TFlearn, NLTK and NumPy with the field of health care assistance.

The paper[5] A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks they have used different concepts such as, Neural networks which is to train the data, Deep Learning it is a type of Machine learning and artificial intelligence(AI) that imitates the way humans gain certain types of knowledge, Natural language processing(NLP) which is used to understand the input provided by the user. This paper concludes that with the help of deep neural networks one can achieve a better accuracy of the result.

III. EXISTING SYSTEM

Many of the applications available in the field of healthcare has direct live chat between the doctor and the patient and the chatbots can only book an appointment with the doctor and can only help the user to get comfortable with the application such as providing the doctor details, hospital details etc..

Also these chatbots focus on specific domain level diseases like cardiology (heart), gastroenterology, or others. For this reason these applications use NLP for pre-processing and machine learning algorithms like SVM, Naive Bayes, Random forest for disease prediction. Some chatbots were built on more robust methodology using neural networks but used traditional techniques like RNN and relied on dataset of conversations.

IV. PROPOSED SYSTEM

To overcome some of the drawbacks of existing models, we will be making use of Neural networks to train the dataset. In the training phase, there are three parts, which are Pre-processing of the input provided to the chatbot by the user, sentence-based representation which aims to encode the semantic information into a real-valued representation vector, which will be utilized in further sentence classification or matching tasks, and response selection where LSTM is used to get the accurate response by decoding.

In the first part that is Pre-processing, each input message turn into a dialogue which will be split into sentences and words and then the patterns are generated by replacing the words with the predefined tags. Finally each, message pattern is converted into the form of vector representation.

Then we use LSTM model to convert the content into sentence based representation vectors. Then finally the proper question pattern and the corresponding response is selected.

A. Sentence Pre-process

In order to convert the sentence into a comprehensible embedding vector, we first use the open source available segmentation tool which is used to separate words from sentences. Consider for example two sentences which has similar meaning, “Where to go after dinner” and “Where to go after lunch”.

Here we convert the sentences into patterns, thus reducing the complexity of the dataset. Then each word in the database is matched with the pre-defined tags, if the tag is found we can get the accurate output.

As the next process we use GloVe method to train word vector model. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

B. Sentence-Based Representation Model

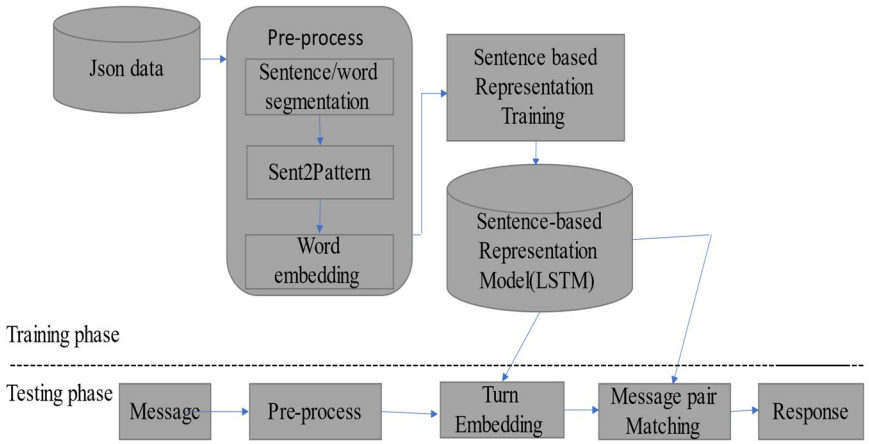
LSTM(long short-term memory) is a kind of Reccurent Neural network(RNN) architechture. Basically the architecture contains three gates they are, input gate, forget gate and the output gate. The output of each gate will go through sigmoid function, and gives output either 0 or 1. The forget gate is used to decide which information will be thrown away from the cell state, which is decided by the sigmoid layer.

The study uses two layers of LSTM model, the first layer accepts the word-based sequence to extract the semantic information between words in sentence. The second layer of LSTM accepts sentence-based sequence, which is used to extract the relation between the sentences.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

V. DATA FLOW DIAGRAM



VI. ALGORITHMS USED

A. Natural Language Processing

The field of NLP dates back decades and has been quite mature in the recent past. In the beginning these were limited to collecting data from these limited set of digital document, the emergence of the (WWW) World Wide Web has seen a burst of knowledge in different languages. Much work has been done in the field of (IR) informations retrieval, which is considered an application in the field of natural language processing. Prior to talking about IR technology, let's dive into the theoretical and practical aspects of NLP. The Traditional NLP approach followed the following discrete steps.

1) Text Pre-processing/Tokenization
2) Lexical Analysis
3) Syntactical Analysis
4) Semantic Analysis

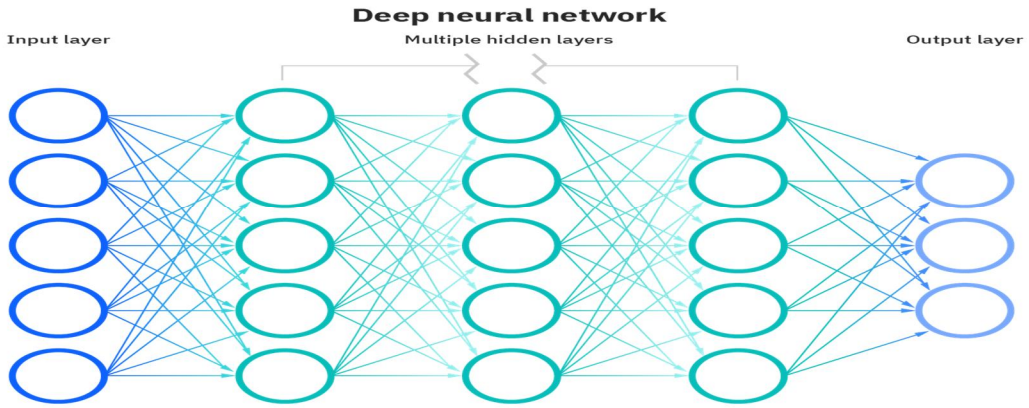
The first challenge of pre-processing/tokenization is to part of a given documents into sentence and words. The token of word, which was originally limited to programming language theory, is now same as splitting texts into word. Many languages uses spaces as separators, but it can be a bit tricky in some language. Although it seems simple, the challenges include decomposing words such as "I`m" into "I am" and deciding whether to decompose blocks such as "highimpact" into two words. Compounding the problem is the language of the document. The Unicode standard is very useful because each character is assigned a unique value, so the underlying language can be determined. Another concept frequently used by N experts is "regular RE [REDACTED])". RE is also derived from computer programming language theory, which specifies the format of the string to be checked. For example, a password string (token) that can only contain uppercase letters will be designated as [AZ], and a number-counting string will be designated as [09]. In addition to segmenting the text into tokens / words, the NLP field also places great importance on finding sentence boundaries. Although more language uses punctuation such as comma full stop to define sentence boundaries, remaining language are more complex in this regard.

B. Neural Networks

Neural networks, also known as simulated neural networks (SNNs) or artificial neural networks (ANNs) , are subdivisions of machine learning and are at the center of deep learning algorithms. Their structure and name are influenced by the human brain, enacting similar to how biological neurons signal to one another.

©IJRASET: All Rights are Reserved

Artificial neural networks are consists of an input layer, node layers, hidden layers, and an output layer. Every node, artificial neuron, links to a different and has a threshold and an associated weight.
If the output of someone node is higher than the desired threshold value, that node is activated, sends data to the following layer of the respective network. Or else, no data is transferred along to the subsequent layers of the network.

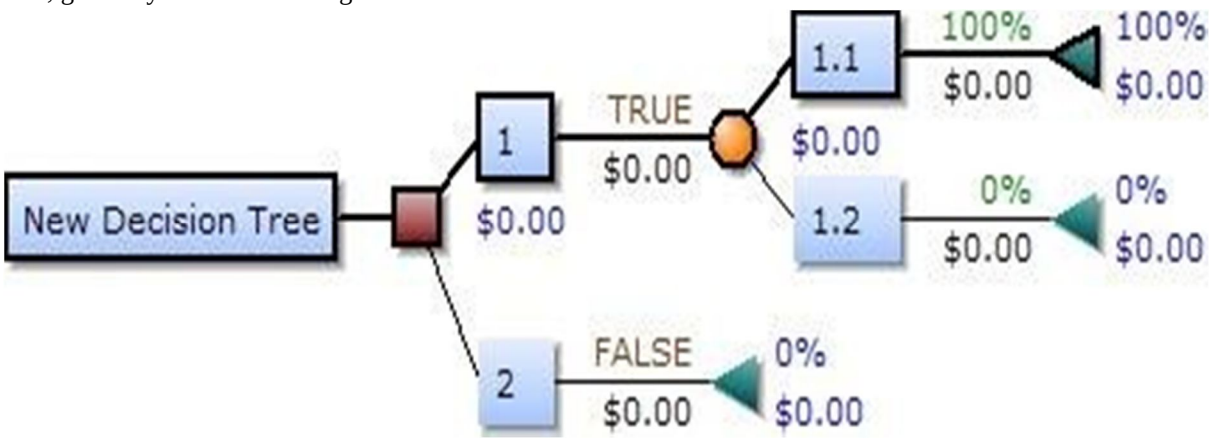


Neural networks depend upon training data to find out and increase their accuracy with time. However, some of these learning algorithms are adjusted for accuracy, these are powerful accessories in computer science and computing , which permits us to catogarize and group data which has high velocity. speech recognition or image recognition tasks may use minutes or hours when put next to the human identification by experts. one among the foremost popular neural networks is Google's search engine algorithm.

C. Decision Tree

A decision tree is a decision support tool that utilizes a tree-shaped model of decisions and some of their possible consequences , includes outcomes of chance events, utility and resource tools .It provides a way to pictures an algorithm that has only conditional control statements .
They are popularly utilized in operations research, to help in identification of a strategy most likely to touch a goal and in decision analysis, , but they are also a popular tools in machine learning humans. A decision tree structure is like a flowchart where internal node act as a test example head and tails of a coin , each segment represents the outcome of the each test, and each leaf node represents a class (final decision taken after all attributes) . The path from root node to leaf node are represented as classification rules.
In decision analysis, a decision tree and the similarly influenced diagram are utilized as a analytical and visual decision support tools, here the expected utility or values of other options are calculated. The decision tree is composed of 3 types of nodes:

1) Decision node, generally shown as a square
2) Chance node, generally shown as a circle
3) End node, generally shown as a triangle



D. Naïve bayes

Naive Bayes is a simple technique for building classifiers: a model that assigns class labels to problem instances, expressed as a vector of entity values, where class labels are extracted from a finite set. There is no single algorithm for training such classifiers, but a series of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given class variables. For example, if the fruit is red, round and about 10 cm in diameter, you can think of it as an apple. The Naive Bayes classifier believes that each of these features independently affects the probability that the fruit is an apple, regardless of the possible correlation between the color, roundness, and diameter features.

For certain types of probabilistic models, naive Bayes classifiers can be trained very effectively in a supervised learning environment. In many practical applications, the parameter estimation of the naive Bayes model adopts the maximum likelihood method; in other words, people can use the naive Bayes model without accepting Bayesian probability or using any Bayesian method.

Despite its naive design and obviously oversimplified assumptions, Bayes's naive classifier works well in many complex real-world situations. In 2004, analysis of the Bayesian classification problem showed that there are reasonable theoretical reasons for the seemingly implausible performance of the naive Bayes classifier. However, a comprehensive comparison with other classification algorithms in 2006 shows that Bayesian classification performance is better than other methods, such as powered trees or random forests.

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

The formula for Bayes' theorem

Where,

P (A | B) is the posterior probability: hypothesis A is the probability of observed event B.

P (B | A) is the probability of possibility: the probability of hypothesis is true, the probability of evidence.

P (A) is the prior probability: the probability of a hypothesis before viewing the evidence.

P (B) is the marginal probability: the probability of evidence.

E. Random forest

Random forest is one of the popular and most used algorithm in machine learning that belongs to the supervised technique. It is used instead of decision tree as it uses multiple trees among that it will choose the most accurate so that we can get the better accuracy than decision tree. Random forest can be used for both classification and regression problems in machine learning. The algorithm is mainly based on the concept of Ensemble learning, it is a process of combining multiple classifiers to solve more complex problem and to enhance the performance of the model.

Some of the advantages of Random forest compared to other Machine algorithms include, it has been observed that the training time taken by Random forest algorithm is less than compared to other algorithms, It predicts the output with high accuracy even for large dataset, and it maintains the accuracy even if the large proportion of data is missing.

Random forest basically works in two-stages,

1) First it is to create the random forest by combining N decision tree and

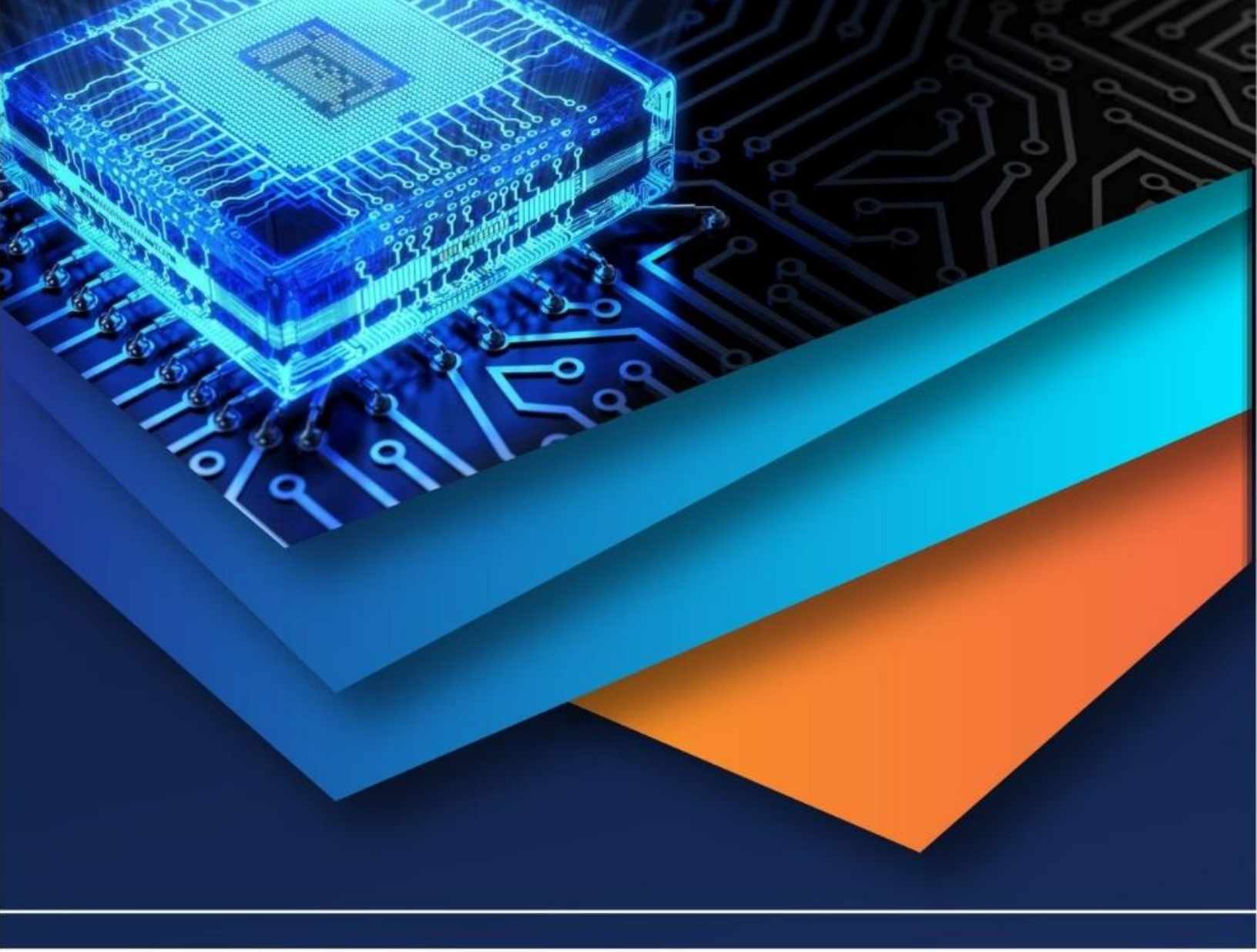
2) Second is to make predictions of each tree created in the first stage.

VII. EXPERIMENTAL RESULTS

System performance changes according to the number of training epoch.System cannot learn well if number of epoch is low. Training time might increase if epoch number is high. 500 epoch had been set and loss and accuracy was noted for each value . 150 epoch provided a gentle accuracy pushing epoch to 200 we did not record any loss hence set it to 200.

We calculated the turn embedding model with different hidden sizes of LSTM by loss function. Best effect was recorded when there was least loss value. The loss function is described below.

$$loss = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)