International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-2, July 2020

Exploration of Twitter Sentiments and Classification by using Deep CNN and Naive Bayes

Nijil Raj N, Abilash Babu Philipose, Dency Dominic, Indu S

Abstract: ***Sentiment evaluation of tweets help the enterprises to evaluate public emotion towards the activities or products associated with them. Most of the research targeted to obtain sentiment capabilities with the help of analyzing syntactic and lexical features which can be expressed through sentiment phrases, emoticons, exclamation marks etc. In the proposed paper we introduce a phrase embedding received by means of unsupervised learning(deep learning) on large twitter texts which uses contextual semantic relationships and co-occurrence statistical characteristics between words in tweets and also consider the emojis to categorise the emotions whether it is positive or negative by the use of Naive Bayes. In the preceding paper which used usnsupervised learning approach for classification, has an accuracy of 87% and supervised has an accuracy of 89%. According to our context, Naive Bayes has given an accuracy of 100% and CNN has given an accuracy of 100%. As compared to machine learning. It has a higher performance on the accuracy, precision and recall.***

Index Terms: Tweets, Sentiment analysis, Word Embedding, Convolution Neural Network, Naive Bayes

I. INTRODUCTION

Twitter, now become a platform for organizations and individuals to express their political, social and economic interest in maintaining and enhancing their clout and reputation. Sentiment analysis provides these organizations with the ability to surveying various social media sites in real time. Text Sentiment analysis is an automatic process to determining whether a text segment contains objective or opinionated content, and it can furthermore determine the text's sentiment polarity. The Twitter sentiment classification goal is to determine whether a sentiment of a tweet's polarity is negative or positive.

Manuscript received on May 25, 2020.
Revised Manuscript received on June 29, 2020.
Manuscript published on July 30, 2020.
***** Correspondence Author
Dr. Nijil Raj N, Professor and Head, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Kollam, Kerala, India. E-mail: csithod.nijil@ycet.ac.in
Abilash Babu Philipose, B. Tech Student, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Kollam, Kerala, India. E-mail: abilash.4ever09@gmail.com
Dency Dominic, B. Tech Student, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Kollam, Kerala, India. E-mail: dencydominic98@gmail.com
Indu S, B. Tech Student, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Kollam, Kerala, India. E-mail: indusobhana123@gmail.com

The existing method of analyzing of Twitter sentiment classification, like the one proposed by Pang et al.[1] and apply machine learning algorithms to build a classifier and evaluate the tweets with manually annotated sentiment polarity label. In recent years, there has been a hike in the use of deep learning techniques, which can increases the accuracy.

In this article, we apply convolution algorithm on Twitter sentiment analysis to train deep neural network and also machine learning technique, in order to improve the accuracy and analysis speed. First we shall know about global vectors for word representation which converts the word sentiment information as the words embeddings. Afterwards, we concatenate these word representation with the prior polarity score feature. These feature sets is combined and fed into an deep convolution neural networks to train and predict the sentiment classification labels of the tweet. A model is presented which classify the tweet into negative or positive sentiment categories. The experiment results conclude that our approach has performed better in classifying.

II. EXISTING SYSTEM

A. Sentiment Classification using Machine Learning Tech- niques

The issue of arranging records not by point, however by and large notion, e.g., deciding if an audit is positive or negative. Utilizing film audits as information, Bo Pang, Lillian Lee and Shivakumar Vaithyanathan[1] locate that norm AI proce- dures completely outflank human delivered baselines. In any case, the three AI techniques utilized are Naive Bayes(NB), Maximum Entropy classification(ME), and Support Vector machines(SVM) don't proceed also on feeling grouping as on conventional theme based order. A human would handily identify the genuine slant of the survey, however pack of high- lights classifiers would apparently discover these occurrences troublesome, since there are numerous words characteristic of the contrary assessment to that of the whole audit. On a very basic level, it appears that some type of talk examination is fundamental. Hence, they presume that an significant fol- lowing stage is the ID of highlights demonstrating regardless of whether sentences are on subject (which is a sort of coreference problem) we anticipate tending to this test in future work.

B. Sentiment Analysis of Short Informal Texts

Kiritchenko and Zhu[3] delineate a front line end examination system that recognizes (a) the finish of short easygoing abstract messages, for instance, tweets and SMS (messagelevel task) and (b) the speculation of a word or an articulation inside a message (term-level task). The system relies upon a managed verifiable substance game plan approach using a variety of surface form, semantic, and feeling features.

The notion features are chiefly gotten from novel high-incorporation tweet unequivocal sentiment word references. These vocabularies are thus created from tweets with assumption word hashtags and from tweets with emojis.

To sufficiently get the speculation of words in refuted settings, an alternate notion jargon is delivered for invalidated words. So made a controlled verifiable assessment structure that recognizes the sentiment of short easy going abstract mes- sages, for instance, tweets and SMS (message level endeavor) similarly as the inclination of a term (a word or an express) inside a message (term-level task). The structure situated first in a long time at the SemEval-2013 contention 'Inclination Investigation in Twitter'. Moreover, it displayed the condition of the workmanship execution on two extra data sets: the SemEval 2013 SMS test set and a corpus of film review separates.

C. Tweet Sentiment Analysis with Classifier Ensembles

In this paper, da Silva and Hruschka E.R[4], present a methodology that naturally orders the supposition of tweets by utilizing classifier groups and dictionaries. Tweets are named either positive or negative concerning an inquiry term. This methodology is valuable for buyers who can utilize notion investigation to look for items, for organizations that target observing the open estimation of their brands, and for some different applications. In reality, slant characterization in microblogging administrations, similar to Twitter through classifier ensembles and vocabularies, has not been very much investigated in the writing. Investigations on an assortment of open tweet sentiment datasets show that classifier troupes shaped by Multinomial Naive Bayes, SVM, Random Forest, and Logistic Regression can improve order accuracy. The utilization of classifier outfits for tweet assumption investiga- tion has been underexplored in the writing. Here showed that classifier ensembles framed by expanded segments uniquely if these originate from distinctive data sources, for example, literary information, emojis, what's more, dictionaries can give best in class results for this standard circular space. Additionally thought about promising systems for the por- trayal of tweets (i.e., bag-of-words and highlight hashing) also, indicated their points of interest and disadvantages.

D. Contextual Semantics for Sentiment Analysis of Twitter

In this paper Saif H and Fernandez et al[6] present SentiCircles, a word reference set up procedure for thought examination regarding Twitter. This methodology considers the revelation of thought at both substance level and tweetlevel. They survey their proposed approach on three Twitter datasets using three assorted presumption word references to gather word prior ideas. Results show that their approach in a general sense defeats the baselines in precision besides, Fmeasure for component level subjectivity (fair versus polar) moreover, furthest point (positive versus negative) acknowledgments. For tweet-level end area, this approach performs better than the bleeding edge SentiStrength by 4-5% in precision in two datasets. In this paper, they proposed a novel semantic suppo- sition portrayal of words, called SentiCircle, which can rele- gate setting explicit feeling direction to words. here depicted the utilization of SentiCircles for vocabulary based opinion distinguishing proof at both substance level and tweet-level utilizing various techniques. This proposed approach beat other vocabulary naming methods for both substance level and tweet-level opinion identification. For tweetlevel feeling discovery, gave a superior generally speaking outcome thanthe cutting edge dictionary based methodology SentiStrength on two out of three datasets.

E. Combining Semantic and Prior Polarity for Boosting Twit- ter Sentiment Analysis

In this article, Jianqiang Z and Xueliang [5] center around the more typical casual printed correspondence on the Web, for example, online conversations, tweets and informal community remarks and propose an instinctive, less area explicit, unadministered, vocabulary based methodology that gauges the degree of enthusiastic power contained in text so as to make a pre-word usage. This methodology can be applied to, and is tried in, two distinctive however reciprocal settings: subjectivity identification what's more, extremity characterization. Broad investigations were conveyed on three genuine world data sets, separated from online social Web destinations and commented on by human evaluators, against condition of the craftsmanship managed approaches. The outcomes exhibit that the proposed calculation, despite the fact that unaided, outflanks AI arrangements in most of cases, by and large introducing an extremely powerful and dependable answer for estimation investigation of casual correspondence on the Web. In this article, proposed an unaided, vocabulary based classifier that gauges the degree of enthusiastic valence in text so as to make a forecast, expressly intended to address the issue of supposition investigation in such conditions. They have included a broad rundown of phonetically deter- mined functionalities to the classifier, for example, refuta- tion/capitalization discovery, intensifier/diminisher identifica- tion and emoji/outcry discovery, all of which add to the last expectation. The proposed calculation is appropriate in two diverse however reciprocal settings: conclusion identification (i.e., distinguishing whether the content contains an outflow of sentiment or is goal) and extremity recognition.

F. Comparison Research on Text Preprocessing Methods on Twitter Sentiment Analysis

In this paper Jianqiang Z and Xiaolin[7] assessed the impacts of text pre-directing methodology on speculation depiction execution in two sorts unmistakably of development undertakings, and included up the gathering presentations of six pre-overseeing strategies utilizing two zone models and four classifiers on five Twitter datasets. The assessments show that the exactness moreover, F1-level of Twitter hy- pothesis gathering classifier are improved while utilizing the preoverseeing structures for extending contracted structures and abrogating invalidation, yet barely changes while dis- charging URLs, crippling numbers or stop words.

The Naive Bayes and Random Forest classifiers are more delicate than Logistic Regression and bolster vector machine classifiers while fluctuating pre-preparing procedures were applied. The cutoff focuses utilized here are Word N-grams and Prior farthest point score.

This paper inspects that six specific pre- overseeing perspectives impact end farthest point gathering in the Twitter. Therefore quick an improvement of assessments utilizing four classifiers to check the attainability of a couple pre-arranging perspectives on five Twitter datasets. Starter results show that the arrival of URLs, the outing of stopwords and the flight of numbers unnecessarily impact the presentation of classifiers; likewise, emptying nullification and widening shortenings can improve the system precision. In this way, discharging stop words, numbers, and URLs is real to lessen change regardless doesn't impact execution. Repealing refutation is unfathomable for doubt assessment. So picked fitting handling techniques and highlight models for various classifiers for the Twitter slant gathering task.

G. Deep Convolution Neural Networks for Twitter Sentiment Analysis

In this paper Jianqiang, Zhao, Gui Xiaolin[8] present a word embeddings strategy acquired by solo learning dependent on huge twitter data, this technique utilizing idle logical semantic connections and co-event factual qualities between words in tweets. These word embeddings are joined with n-grams high- lights and word opinion extremity score highlights to frame an assumption include set of tweets. The list of capabilities is coordinated into a profound convolution neural system for preparing and foreseeing the feeling grouping marks. It was concluded with CNN technique producing better results than the previous works.

III. MATERIALS AND METHODS

A. Material used for this project

In deep learning, a convolutional neural system (CNN, or ConvNet) is a class of deep neural systems most usually used for analyzing visible imagery. They are also called shift invariant basically dependent on their common load structure. They have utilized in picture and video acknowledgment, recommender frameworks, photograph arrangement, clinical photo examination, and natural language processing.

Algorithm: a. Convolution Neural Network(CNN), pow- erful set of techniques for learning in neural networks. It is a class of ML calculations which will utilizes a numerous layers for logically separate more significant level of highlights from the input. Example, picture processing, lower layers will discover edges and higher layers that become aware of the concepts relevant to a human such as letters or digits. These networks is applied to fields consisting of speech recognition, herbal language processing, laptop imaginative etc where In this undertaking we exhort a profound convolution neural sys- tem model to ordering tweet as negative or positive estimation.

b. Naive Bayes classifiers are a lot of type calculations dependent on Bayes Hypothesis. It isn't algorithm, anyway a group of calculations in which every one of them rate a common guideline, For example each pair of capacities being sorted is autonomous of each other. This set of rules is normally used in text category and with issues having more than one classes. Naive Bayes strategies are a fixed of supervised gaining knowledge of algorithms primarily based on utilizing Bayes' hypothesis with the "guileless" presumption of contingent independence between each pair of highlights given the cost of the class variable. It become at first conveyed for textual content classification obligations and still is utilized as a criterion. Unsupervised getting to know is a gadget technique, in which you would prefer not to manage the model. Directed becoming acquainted with permits you to gather records or produce a records yield from the past experience.

1) Datasets:

- Stanford Twitter Sentiment Gold(STSGd): This dataset contains 2034 tweets, which the three annotators agreed on their sentiment labels (positive/negative). It contains 3 fields:
  - Id: Id of the corresponding tweet.
  - Polarity: The sentiment polarity of the tweet (0: negative) (1: positive).
  - Tweet: The tweet text.
- Sentiment140 dataset: It contains tweets separated by the utilization of the twitter api. The tweets can be (0 = negative, 4 = positive) and they will be utilized to stumble on polarity. It conveys the accompanying 6 fields:
  - Target: The sentiment of the tweet (0 = negative, 4= positive).
  - Id: Tweet id
  - Date: The date of the tweet.
  - Flag: The query (lyx). If there is no query, hence this value is NO QUERY.
  - User:The user of the tweet
  - Text: Sentence.

2) Features:

- N-gram: An N-gram is various items n from a bit of textual content or speech. The objects may be letters, phonemes, syllables, words or base pairs in step with the usage

- Word Sentiment polarity Score: The word assessment extremity score is a vocabulary based conclusion high-light, and a few approaches usually use it as a conclusion include for tweet notion examination.
- GloVe Model: Global Vectors, is a version for distributed phrase representation. The model is an unsupervised getting to know set of rules for obtaining vector rep- resentations for words.

3) *Standard Measures:*

- Precision: In a classification task, the precision for a classification is the no of positives (i.e The assortment of things viably arranged as having a place with the great class) separated by methods for the general scope of components sorted as having a place with the enormous class (For example The total of true positives and false positives, are the things that are inaccurately classified having a place with the class).

Precision = $TP/TP + FP$ (1)

- Recall: Recall is called as the no of true positive separated by the absolute no of components that are effectively have a place with the positive class.

$$Recall = TP/TP + FN \quad (2)$$

- F1 score: F1 Score is a normal of Recall and precision. Both the false positives and false negatives considers. Naturally it isn't generally as simple to comprehend as precision, however F1 regularly gives more exactness which works pleasant if false positives and false negatives have comparable Precision and Recall ficial than preci- sion, predominantly if you have a not even proportion of realities that must be prepared, while still as it ought to be and totally describing the original data set.
- Word Vectorisation: The feature extraction begins by

$$F1Score = \frac{2 \times (Recall \times Precision)}{(Recall + Precision)}$$

IV. METHODOLOGY

A. Data Cleaning

Using changing the tweets to numerical value. One of the strategies is known as bag-of-words approach. The bag of phrases version overlooks sentence structure and request of expressions. When we have a corpus (text records) at that point initial, a list of vocabulary is made based at the

Tweets are loaded with deficient articulation, an assort- ment of commotion, unstructured sentences on account of the normal nearness of abbreviation, strange syntax, badly formed words and non-word reference terms. The Noise and unstructured Twitter measurements will influence the general execution of tweet conclusion classification. Before to work choice, a progression of preprocessing are done to tweets to reduce the commotion inside the miniaturized scale blog text. The preprocessing is:

- Expulsion all non-ASCII and non-English characters in the tweets.
- Expulsion all URL joins. The URLs don't contain the supposition data of tweet, so there will be erased from tweets.
- Expelling of numerical worth. The numbers by and large don't contain supposition data, so it are pointless when estimating conclusion and are erased from tweets so as to refine the tweet content.
- Evacuating of negative references. Tweets contain differ- ent thoughts of negation. Generally, refutation assumes an imperative job in making sense of the estimation extremity of the tweet. Here, the strategy for nullification is changing "won't", "can't", and "n't" into "will not", "cannot", and "not", separately.
- Explusion of stopwords. Stop words for the most part talk over with the most typical words in a language, for example, "the", "an", and "than". The great strategy de- pends absolutely on pushing off the stopwords acquired from precomplied records.

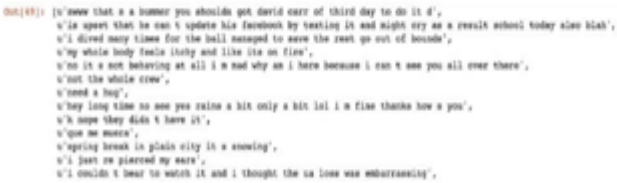

Fig. 1. After Data Cleaning Process

B. Feature Extraction

After the statistics cleansing method, characteristic extrac- tion takes place. If we want to apply textual content in system getting to know algorithms, we must convert them to a numerical representation. Feature extraction is a method of dimensionality reduction with help of which an underlying arrangement of raw data is diminished to more pieces for preparing. An element of these huge data sets is countless factors that require a great deal of figuring sources to way. Feature extraction is the name for techniques that select and join factors into features, productively diminishing the amount entire corpus. At that point each report or records section is spoken to as numerical vectors basically based at the vocabulary worked from the corpora. It is a technique for regular handling language preparing that removes the words (highlights) utilized in a sentence, record, site, and so on.

C. Convolution Neural Network

After information cleaning and highlight extraction, It is then passed to the convolution arrange layer. In the main convolution layer, convolution computation is performed uti- lizing utilize different channels with variable window size, and produce nearby estimation include vector for every conceiv- able word window size. Every convolution activity produces another setting neighborhood highlight vector in a word win- dow. The convolution channel creates a neighborhood include planning vector for every conceivable word window in the tweet, which is trailed by the finishing of the convolution activity to produce another vector.

After convolution activity K-Max pooling activity is performed. A while later the convolution activity, a k-max pooling activity is utilized on the new element vector created by the convolution layer. K-Max pooling planned the vector to a fixed length vector.

The length of the vector is a hyper boundary to be dictated by the client and compares to the quantity of shrouded units in the convolution layer.

In this work, k-max-pooling is used. For notion arrangement, the most unequivocal word or expression is frequently just a couple, yet not consistently dissipated all through the content. The k-max pooling is only the absolute most discriminative language pieces. The k-maxpooling select the top 'k' number of highlights comparing to different concealed layers, with the goal that the most significant conclusion includes data that can be held.

The yield layer of the architecture is a softmax layer that creates a likelihood estimation of positive or negative supposition. The yield layer utilizes a completely associated softmax layer to alter the supposition attributes of the info layer and gives a likelihood circulation of the assumption characterization names whether it is certain or negative

D. Naive Bayes Classification

a. Another part of this project is the classification of live tweets. This part is similar to process as that of CNN. But there is moderate difference.

Here the process begins by accepting a user id which is unique for every twitter user.

After entering the id, all the tweets posted by the user gets evaluated. Here emojis also taken in the tweet and not discarded. All other process in data cleaning and data pre-processing is the same as used n CNN model.

b. The Naive Bayesian calculation is utilized here to arrange the tweets. The model is trained separately with tweets and emojis which are usually used in the tweets.

c. Using Naive Bayes Emoji is classified into positive or negative. The emoji is taken as UTF-8 code and is trained. This is used for classifying the live tweets into negative or positive.

d. Both the methods, CNN and Naive Bayes algorithm used have executed successfully and predicted the inputs given to them. While CNN didn't use emoji to classify tweets, Naive Bayes method have done that properly to classify the tweets.

Fig. 2. Dataflow Diagram

V. RESULT AND DISCUSSION

- Twitter sentiment analysis frameworks permits to sort enormous arrangements of tweets and distinguish the sentiment of every tweets correctly Furthermore, the best part, it's quick and basic.
- In the proposed method CNN and Naive Bayes have been used. The CNN which uses dataset with pure texts has an accuracy of 100% and Naive Bayes which has text and emojis has an accuracy of 100%. The result obtained is given in Table I

Table I Result Table

Algorithm	Accuracy	Precision	Recall	F1Score
CNN	100	100	100	100
Naive Bayes	100	100	100	100

From the comparison of results in Table II and graph from Figure 3, it is observed that in the existing method, the accuracy of Unsupervised method was 87% and that of Supervised method was 89%. By the proposed method, the

TABLE II ACCURACY TABLE FOR ML

Existing System				
Algorithm	Accuracy			
Unsupervised	87			
Supervised	89			
Proposed System				
Algorithm	Accuracy	Precision	Recall	F1Score
CNN	100	100	100	100
Naive Bayes	100	100	100	100

accuracy for both, Deep Learning and Naive Bayes are 100%. By this a conclusion can be made that the proposed system is much better in classifying sentiments.

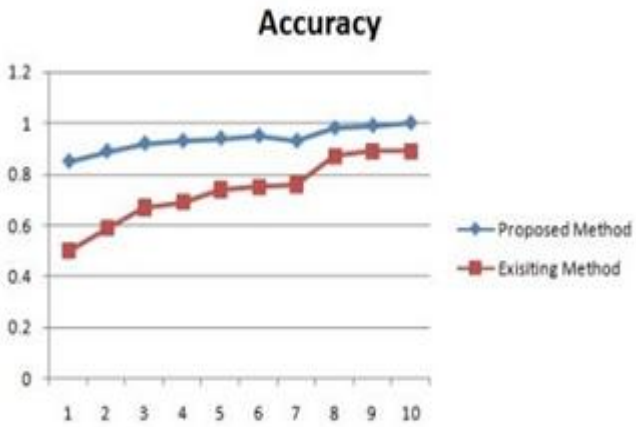

Fig. 3. Comparison of Accuracy

VI. CONCLUSION

Our system proposes a machine learning technique using Naive Bayes as well as deep learning method using Convolution neural network to classify the sentiments of tweets. In the most of the previous studies done, emojis were discarded while classifying. In our Naive Bayesian classification, we consider the emoji to obtain better accuracy.The CNN has also produced better results than the existing result.

ACKNOWLEDGEMENT

We acknowledge this open way to offer our certifiable gratitude to all those without whom this project would not have been a success. First of all, we owe our thanks to the almighty for providing us the strength and courage to complete the project. We express our deep and sincere gratitude to our guide Dr Nijil Raj N, Head of the Computer Science and Engineering Department, Younus College of Engineering and Technology for providing valuable advice and timely instructions, without which we could never have been able to complete the work in time.

REFERENCES

1. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sen- timent classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.
2. Georgios Paltoglou and Mike Thelwall. Twitter, myspace, digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):1–19, 2012.
3. Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.
4. Nadia FF da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179, 2014.
5. Xueliang C Jianqiang Z. Combining semantic and prior polarity for boosting twitter sentiment analysis. In *2015 IEEE International Confer- ence on Smart City/SocialCom/SustainCom (SmartCity)*, pages 832–837. IEEE, 2015.
6. Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. Contextual semantics for sentiment analysis of twitter. *Information Processing &Management*, 52(1):5–19, 2016.
7. Zhao Jianqiang and Gui Xiaolin. Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, 5:2870– 2879, 2017.
8. Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun. Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, 6:23253–23260, 2018.

AUTHORS PROFILE

Dr. Nijil Raj N is currently working as a Professor and Head of Computer Science and Engineering department, Younus College of Engineering and Technology, Kollam, Kerala. He is having 16 years of teaching experience in UG and PG level. He has completed the Doctoral Degree from M. S. University, Tamilnadu, in Computer Science and Information Technology. His research area is based on Computational Biology, Machine Learning and AI, and Image Processing. He has published more than 15 papers in national and international journals. He was post graduated from M. S. University, Tamilnadu in M.Tech and MCA also MBA from MG University Kottayam, Kerala.

Abilash Babu Philipose is currently pursuing BTech degree in Computer Science and Engineering from Younus College of Engineering and Technol- ogy, Kollam, Kerala under A P J Abdul Kalam Technological University, Trivandrum, Kerala.

Dency Dominic is currently pursuing BTech degree in Computer Science and Engineering from Younus College of Engineering and Technology, Kollam, Kerala under A P J Abdul Kalam Technological University, Trivandrum, Kerala.

Indu S is currently pursuing BTech degree in Computer Science and Engineering from Younus College of Engineering and Technology, Kollam, Kerala under A P J Abdul Kalam Technological University, Trivandrum, Kerala.