TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 8021 - 8028

Sentiment Analysis based Recommender System for Reforming Indian Education using Multi-Classifiers

Jabeen Sultana[1], M. Usha Rani[2] and M.A.H. Farquad[3]

[1]Department of Computer Science, Sri Padmavathi Maha VisvaVidyalayam, Tirupati
jabeens02@gmail.com

[2]Department of Computer Science, Sri Padmavathi Maha VisvaVidyalayam, Tirupati,
musha_rohan@yahoo.com

[3]INI Labs. Waterloo, Ontario, Canada

Article Info
Volume 83
Page Number:8021 – 8028
Publication Issue:
May-June 2020

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *18 May 2020*

***Abstract*:**
Social media sites have become major source of communication over the internet pertaining to discussion on various subjects. With increase in the growth of internet users, lotsof huge data is generated through social networking sites like Twitter and Facebook. Tremendous amount of data is being generated from educational sector too. Users share their learning experiences like difficulties faced, quality of content and teaching and this resulted in analyzing their valuable sentiments towards Education. This data needs to be mined and classified properly so that knowledge is drawn regarding sentiments. The obtained knowledge from public sentiments can be analyzed before making any decisions in educational reforms. The study aims to use the search key word titled Education in India for extracting tweets related to education from twitter database by means of twitter API. Tweets are pre-processed and further analyzed and classified into three types of sentiments i.e., positive, negative and neutral based on polarity scores. Machine Learning techniques like SVM, Naïve Bayes, Decision Tree, KNN and MLP are employed to predict and classify the tweets by extracting the hidden knowledge. In this work, our foremost objective is to discover the efficient classification technique in order to reform the educational society by considering valuable sentiments and views from the twitter data related to education. Results achieved are evaluated on parameters like Accuracy, Specificity, Sensitivity, Confidence Interval and Kappa Statistics. SVM outstands in predicting students‟ sentiments compared to other techniques.

I. INTRODUCTION

The advent of internet gave rise to immense bloom in social media. Networking sites of social media turned out to be widespread in modern years because of affordability of the data rates and devices. High demand among the users to interconnect with users all over the world, browsing the internet to solve day to day tasks in various walks of life gave birth to production of millions of devices. Tablets, laptops, home systems and smart phones are increasing in numbers leaded to the massive growth of social media. *Social networking* apps like Twitter, YouTube and Facebook makes use of Internet to get connected across various networks. Users share their thoughts, and express their sentiments freely over the social media, resulting in big data. To name a few on which users share their sentiments are related to various devices, products, current designs, education and organizations. The sentiments obtained from various social networking sites aided in reforming the businesses, governments and educational sectors etc. and also lead to decision making [1, 2].

Twitter is one of the popular social media which facilitates online interactions through which different age groups of people use to share their information. It can expressed be in the form of feelings, sentiments and opinions by tweeting. The data which is in the form of sentiments is available instantly as soon as users share sentiments. Formerly tweets comprise of 140 characters but by the year 2017-character size of the tweets was doubled to 280 characters. Tweets character size is still 140 for the Languages-Korean, Chinese and Japanese. Twitter averaged at 330 million online users per month in the worldwide from 1st quarter 2010 to 1st quarter as of 2019 [3]. Rapid increase in the number of users to use twitter resulted in vast extent of big data and that too in unstructured form. Mining the data comes into picture as lots of meta data is generated by billions of devices worldwide and is too difficult to analyze or understand the user‟s sentiments. According to an investigation carried out by International Data Corporation, 80% of data is unstructured occupying major part over the digital devices in comparison with structured data which occupies 20% only [4]. Therefore, data preprocessing is required to understand the data quickly.

Educational data mining and learning analytics has been evolving since recent past as huge data is generated from schools, colleges, and universities offline and online. Much emphasis is laid to focus, explore and analyze structured data to reform the Education by making good decisions as per the need [6].Coercive method is the one which is followed in the developing countries. The ruling government forces the academic staff to follow as per the rules without making any changes. No reforms in education can be made by the academic staff and they are forced to apply the reforms without further questioning. In addition to this, politicians involve with the government to meet their political agenda in planning education reforms. Other people who has to be involved and are part of the organization like Academic administrators, stakeholders, teaching staff, parents and students are left ignored [7].

Monopolization in the area of education with slight or no involvement of other participants, indicated us to thoughts of carrying this research by analyzing public sentiments on real time basis. To avoid limitations and difficulties in carrying out reforms in education, it is suggested to consider the opinions of public, students; academic staff and other stake holders so that public needs are satisfied.However, to my information, no one has researched exactly like mine and analyze public sentiments related to education in India based on tweets from Twitter. We use Machine learning concepts for our research work, and we opt supervised machine algorithms like SVM, decision tree, KNN, Naive Bayes and MLP. These are best according to the Machine learning experts and developers to perform sentiment analysis.

To meet our aim. The below interrogations are examined:

1) Whether sentiments and opinions of public guide are going to influence education reforms?

2) What factors must be considered from their sentiments, when making education reforms?

II. LITERATURE SURVEY

Opinions or feelings or thoughts or sentiments have always been an important and integral part of our society since our sentiments express what we are by impelling our actions. Considering public sentiments is given prime importance when making decisions for any product improvement or planning for new product or to make changes in the rules or reforms in educational organizations. Sentiment analysis states about organizing sentiments, thoughts, feedbacks about particular text and classifying them into classes like positive, negative and neutral [8]. The key role of sentiment analysis is to spontaneously decide the sensitive track of consumer criticisms so that new decisions can be made [9]. Some of the areas where sentiment analysis is performed are political debates, marketing, e-commerce, movie reviews etc. Buyers prefer to look at the public sentiments and analyze the opinions in advance so that decision can be made to buy the product or look for some other products. In similar fashion, marketers too perform sentiment analysis to study sentiments of public and towards their products and services and analyze buyer happiness level [10].

Mining methods were employed to gather and explore US higher educational twitter accounts. Machine learning methods are used to categorize sentiments extracted from tweets into 3 class labels like positive, negative or neutral [11]. Feedback of students was considered for accomplishing sentiment analysis. Classifiers like Naive Bayes, maximum-entropy SVM, and complement NB by taking unigrams as features [12].

There was drastic shift found on the Internet as individuals incline to swing from old- fashioned information exchange tools to modern tools and services. Users started posting their sentiments towards particular things or devices and facilities they use. Also, sentiments of religions, cultures, politics and weather are expressed, resulting in the source of publics" valuable opinions and sentiments [13]. Twitter data was used to accomplish semantic analysis resulting in effective classifier. Authors suggested the difficulties faced by researchers to analyze the data obtained from twitter, concentrated on analyzing the sentiments and special focus was laid on problems related to classification [14].

Presently, researchers are showing their research interest towards predicting sentiment analysis in learning tract of students by means of various data mining, deep learning and NLP techniques. Data mining techniques have been employed to natural language processing with some success [15].Sentiment classification of online learning posts and comments using a hybrid supervised technique was proposed and also found that the chi-statistics method dominates the other Feature selection methods. Sentiments of students towards learning environment and the difficulties faced by them were analyzed. Moreover, features like Worry, nervousness and irritation were figured out from other classes of sentiments [16]. Tweets containing casual discussions of Engineering students were pulled from the Twitter. Preprocessing was done and classified the data into 5 classes. Sentiments of the students towards the subjects, difficulties faced and heavy load were analyzed so that students perform well in the future [17]. Teacher Parent meetings feedback along with student"s opinions were considered and classified into as positive or negative. The engineering tool suggested for this was General Architecture for Text Engineering and its ANNIE application to classify feedbacks [18].

By analyzing the above studies, we propose efficient classification techniques to classify public sentiments of higher education in India and the sentiments were extracted from Twitter.

III. DATA SET AND PREPROCESSING

Twitter database is used in our research for data collection. R and Rstudio were installed on the Windows Operating System, possessing configuration of i7 core-4510 CPU with speed of 2.00 GHz and 8 GB of RAM. [19,20]. Once R is opened and R studio starts, we create a new R script and give the commands and provide access key and tokens to pull the tweets from twitter. We generated keys by using Twitter developer account. Real time tweets are being pulled from the Twitter. Performed pre-processing to analyze the data from the extracted tweets. Later, tweets are classified into classes based on polarity scores. The below steps describe the preprocessing methods to attain clean set, document term matrix, plot and wordcloud.

Pseudocode for Preprocessing

- Collecting the tweets using Twitter API.
- Pre-processing the tweets.
- Analysing the pre-processed data & classifying the tweets based on polarity scores.

Collecting the Tweets

First of all, Created Twitter user account and then applied for Twitter Developer Account. Once, permission was granted, a new app was created and automatically keys and access tokens were generated. After getting keys and tokens, we get access to twitter data and collected the tweets related to higher education in India. Programming was done in R using keys, access tokens and search key word, tweets were collected from Twitter.com considering the duration of 3 days using Twitter API [3]. Keywords used for extracting the tweets were “Education in India”, and saved as .CSV file. R is an open source tool for doing statistical computations and plotting graphs. R possesses powerful text mining tools and comes with lots of packages. It was developed by core team of R [20].

Pre-processing the Tweets

We need to pre-process the Twitter data as the data obtained from Twitter is in Unstructured Form. Corpus is used as preprocessing method. It consists of text documents, on which we apply text processing to extract meaningful information for further analysis. The whole text is converted into lower case using tm package in R. Noise is removed, url"s are eliminated. The tm package consists of functions to create corpus from files, vectors, removing punctuations, special characters deriving from stop words and other textual data. etc.

Stemming on preprocessed data is done after preprocessing using SnowballC package. For instance “sample" and "samples" area unit each stemmed to "sampl". Also, the data set is cleaned, white spaces are removed. A Term- document matrix is built after necessary stemming is done. Classification, association rule mining can be done on the Term matrix to identify the repeated or frequent words and make associations between them.

From the above step, maximum frequent words can be derived by calculating the number of times a word frequently occurring and ggplot2 package is used to find the frequent terms.

IV. ANALYSE THE PRE-PROCESS DATA & CLASSIFY THE TWEETS BASED ON POLARITY SCORES

Word cloud is formed by considering very frequent terms obtained from all preprocessing steps. It forms a cloud like structure with focus on more frequent terms in large size in comparison with less frequent occurring in very small size.

The below pie diagram shows the sentiment analysis of public sentiments and opinions, 45% positive sentiments of public shows that they are interested in giving suggestions, if taken their views. And 34% neutral sentiments were obtained and 21% negative sentiments were analyzed. From this, we understand that majority of public are interested to participate in making education reforms.

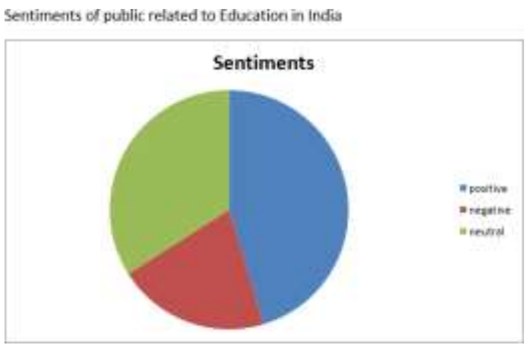

Fig 1. Pie Diagram showing sentiments of public

V. RESEARCH METHOD

Overview about the Proposed Methodology to Reform Education

In this research work, we propose classification of Indian educational tweets using twitter as the source of data. Public opinions regarding Education are considered to make necessary reforms in the education. We classify the tweets collected from twitter into 3 classes namely; positive, negative and neutral. Tweets are pulled online from twitter database and immediately preprocessing is done to analyze the data.

The tweets are further preprocessed by removing hash tags, url‟s, reducing the stems, cleaning the stop words and tokenization. Modulated words are reduced to their root stem in the stemming process. Some of the stop words are detached from text documents during the cleaning process of stop words, part of preprocessing. Tokenization involves breaking down stream of text into words, idioms or codes. The tf-idf values are evaluated. Data set is divided into two parts; training and testing. A model is obtained after training with multiclass. A confusion matrix is observed and the Sensitivity rate and specificity rates are calculated. Support Vector Machine, Naive Bayes, KNN, Decision tree and MLP classifiers are used and compared with one another on different parameters. Polarity of tweets is calculated i.e., positive, negative and neutral by the classifiers. Then unclassified set of data is fed to the model obtained after training the input so that data is categorized into proper class. Parameters like accuracy, true positive, false positive, F-score etc. are calculated.

Aim of this research:

1) Sentiment analysis on Twitter data is carried out to find observations and sentiments of the public on education in India.
2) Pre-processing the tweets and classifying into three classes based on the polarity score of the tweets.
3) Applying machine learning classifiers to find out the efficient classification technique based on accuracy measure and other evaluation parameters.

```
Algorithm: Sentiment analysis on Twitter data using Multi-Classifiers
Input: Twitter data set
Output: Classification of tweets whether tweet implies positive sentiment, negative sentiment or neutral.
Step 1: Tweets are loaded into a .csv file, data set is loaded
Step 2: Apply preprocessing steps-Corpus, stemming, term document matrix and form word cloud
Step 3: Partition the tweets into 3 classes based on the polarity score of the tweets
```

Step 4: Partition the data set as training and test set
Step 4: trained by SVM, Naïve Bayes, Decision trees, KNN and MLP
Step 5: The test data set for testing is given to the Model obtained after training
Step 6: Measure the accuracy of the Models based on some parameters

The below figure shows our proposed framework to analyze educational sentiments of twitter.

Proposed Frame Work

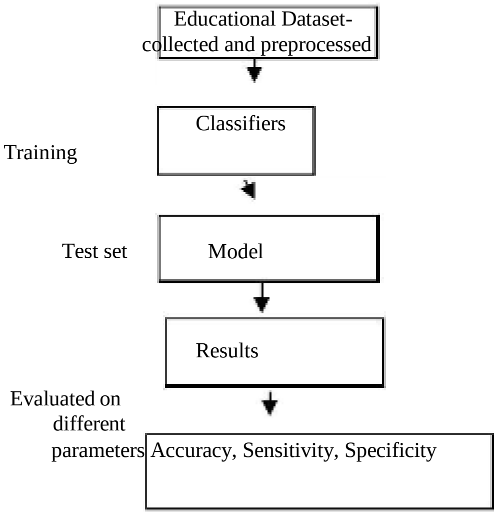

Fig 2. Evaluated on Different Parameters

Brief discussion on different classification techniques used

Dataset was collected from Twitter. Experimentations were done using open source tool R. R comes with many packages, we need to install and run the packages. It can proficiently work with more amounts of data and used in data mining, sentiment analysis and other tasks. Data is pre-processed by using corpus, cleansing and stemming is done. Term document matrix is calculated and frequent items are observed. Polarity score is obtained and Classification is done. Data set is taken in .csv format, Training and testing was done. The results were deeply analyzed on different parameters like Accuracy, Kappa-statistic, Confidence Interval, Specificity, Sensitivity, and ROC curve area. R is widely used software to perform classification, association, clustering and regression tasks.

Description of Techniques

MLP: MLP, kind of neural network which uses propagation algorithm to train multilayer perceptron‟s. Activation tasks or functions in MLP are carried out by Logistic and hyperbolic tangent sigmoid functions. A biased weighted sum of inputs is performed by each element and further sent to activate over a transfer function in order to produce result in the neural network [21].

KNN: K-Nearest Neighbors is very simple and easy to implement. It considers training data and build‟sa new model, based on this model obtained, test data is imparted on the model. KNN classifies test data based on distance measures like Euclidean distance, Manhattan distance [22].

SVM: is a very good machine learning technique compared to other machine learning techniques especially for solving classification and regression problems. Support vector machine used for classification and regression analysis. It constructs optimal hyperplane on the training data, further classifies new instances based on this hyperplane. SMO classifier is used [23].

Naïve-Bayes: It is used for Classification, classifies the data based on probabilities by applying Bayes theorem. It estimates classes by considering numeric precision values; Group of Features classified by this classifier is independent of each other [24].

Decision Trees: Decision trees are built starting from the root and continue until it reaches to its leaf nodes. J48 algorithm is employed to implement. [25].

Sentiments of public towards education in India are analyzed and presented in this research paper to reform education by applying machinelearning techniques. The obtained results are assessed using few parameters like Accuracy, Kappa- statistic, Sensitivity, Specificity and Confidence Interval.

VI. RESULTS AND ANALYSIS

Sentiments of public towards education are exposed by extracting sentiments from the twitter, a social media site. It is accessed by users for sharing their feedbacks and sentiments based on their subjects or hash tags. The code to extract tweets is done in R and query with search words is passed by providing access and token keys.

The tweets are loaded into .csv file and extracted to the system in some particular location i.e., desktop by using twitter.

Different classifiers were chosen in this research work and comparative analysis of their performance was done using R tool. R is a programming language used by data analysts, statisticians and miners for mining the data. R comprises bunch of functions, compiled data and sample data and are located in the R library. Dataset obtained from twitter was pre-processed and later fed to neural network-MLP, SVM, K- NN classifiers, Decision Tree and Naïve-Bayes. Training and testing was performed on each classifier; resulting in an accurate Model and data for testing was applied on this model. The obtained results from the models built were measured in different terms like Accuracy, Specificity, Sensitivity, Positive predictive value and negative predictive value values by drawing confusion matrix. It measures inaccurate and accurate estimation of the data into classes by the classifier. 95% Confidence Interval is calculated from the ROC curve.

The below tables gives detailed explanation of results on different parameters.

Table 1. Showing result analysis of twitter data using various classifiers

Classifier	Kappa Statistic	Positive predictive value	Negative predicti ve value
Decision tree	0.187	0.22	0.93
Naive Bayes	0	NaN	0.9
KNN	0	NaN	0.916
SVM	0.27	0.31	0.95
MLP	0.33	0.33	0.95

From the above results, we conclude that SVM performed well in analyzingpublic sentiments yielding maximum accuracies of 91%, followed by Decision trees and KNN with 90% Accuracies and MLP yielded low accuracy. TP, TN values of SVM are 0.70, 0.97 high compared to rest of the classifiers. Confidence Interval of SVM is very high ranging from 0.73-0.98 followed by Decision trees and KNN. Other Evaluation parameters likeKappa-Statistic is 0.27 for SVM, PPV is 0.31, NPV is 0.95 for SVM followed by decision tree and MLP.

VII. CONCLUSION

Vast volume of data is shared by the users on various subjects through different platforms like social media, forums, and blogs. Bulks of data is obtained from the most popular social media site i.e., twitter as it offers wide range of sentiments on various topics shared by the users. Users express their sentiments on a particular product or movie and this sentiment helps in analyzing the views and ultimately leads to decision making. In this research work, text mining and data mining techniques are used to preprocess and mine educational tweets, sentiments are analyzed for reforming education. Real time twitter tweets were extracted based on the subject Education in India using R. Further, tweets were Pre-processed and classified into 3 classes namely; positive, negative and neutral. Classifiers like SVM, Naïve Bayes, Decision Tree, KNN and MLP were employed. Training on the dataset was done and a model was obtained. Testing was done and the obtained results shows that SVM outstands in the performance by yielding 91% accuracy followed by Decision Tree and KNN-with 90% accuracy, and NaïveBayes is 68%. We conclude that SVM achieves best result in comparison with other classifiers.

Deep understanding of the extracted tweets exposes that the community needs teaching to be practical oriented with more examples to solve. Frequent breaks are needed in the lectures for every 45 minutes. Employment opportunities must be raised so that more number of students put their best efforts to make out their destined jobs. They must be awared of ethics and professional practices by making ethics as compulsory subject at graduate level. Considering these sentiments from public through twitter helps in effective decision making while revising education reforms.In future, we aim to work on online education ands also different kinds of data can be considered other than the textual data.

REFERENCES

[1]. B. Liu, "Sentiment Analysis and Opinion Mining,"*Morgan & claypool publsihers*, 2012.

[2]. O. Mwana, *et al*, "Mining tweets for education reforms," pp. 416-419, 2017.

[3]. https://twitter.com (https://dev.twitter.com/streaming/overview).

[4]. G. Chakraborty and M.K Pagolu., "Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining,"*SAS Global Forum, Washington D.C.*, March 2014.

[5]. J. Singh, *et al*, "A Review of Sentiment Analysis Techniques for Opinionated Web Text," *Csi Transactions in ICT*, vol. 4, no. 2-4, pp. 241-247, 2016.

[6]. C. Romero and S. Ventura, "Educational Data Mining: A Review of the State-of-the-Art,"*IEEE Transactions on Systems, Man, and Cybernetics,* vol.40, no.6, pp. 601-618, 2010.

[7]. W.P. Muricho andJ.K. Chang‟ach, "Education reforms in Kenya for innovation,"*International Journal of Humanities and Social Science,* vol.3 no.9, pp.123-145, 2013.

[8]. A.V. Yeole, *et al,*"Opinion Mining for emotions determination," ICIIECS, *IEEE International Conference Innovations Information, Embedded Communication Systems*, 2015.

[9]. F. Luo, *et al*, "Affective-feature-based sentiment analysis using SVM classifier," *IEEE 20th International Conference Computer Support Cooperation*, pp. 276-281, 2016.

[10]. A. Go, *et al*, "Twitter Sentiment Classification Using Distant Supervision", pp.1-6, 2009.

[11]. R. Kimmons, *et al*, "Institutional uses of Twitter in US higher education,"*Innovative Higher Education*, Springer, New York, vol.42, issue-2, pp. 97-111, 2017.

[12]. M. Munezero, "Exploiting Sentiment Analysis to Track Emotions in Students‟ Learning Diaries,"*Proceedings of 13th Koli Calling International Conference Computing Education Research*, pp. 145–152, 2013.

[13]. A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining,"*Proceedings of the Seventh Conference on International Language Resources and Evaluation*, Valletta, Malta, pp.1320-1326, 2010.

[14]. A. Bifet and E. Frank, "Sentiment Knowledge Discovery in Twitter Streaming Data,"*University of Waikato,* Hamilton, New Zealand.

[15]. J. Sultana, *et al,* "An Extensive Survey on Some Deep learning Applications,"*Springer Nature, Singapore. Advs in Intelligent Syst., Computing,* vol.1054, 2019.

[16]. Z. Kechaou, *et al,*"Improving e-learning with sentiment analysis of users' opinions," *Global Engineering Education Conference*, (EDUCON), IEEE. pp. 1032-1038, 2011.

[17]. W. Xindong, *et al,*"Data Mining with Big Data", *IEEE Transactions on Knowledge and Data Engineering,* vol.26, no. 1, pp. 97-107, IEEE, 2014.

[18]. T. Patela, *et al,*"Sentiment Analysis of Parents Feedback for Educational Institutes," *International Journal* of Innovative and Emerging Research in Engineering, vol.2, issue-3, pp.75-78, 2015.

[19]. S. Kohli and H. Singal, "Data Analysis with R," IEEE 7th International Conference on Utility and Cloud Computing, 2014.

[20]. https://www.r-project.org/

[21]. H. Walter, *et al,*"Recent Developments in Multilayer Perceptron Neural Networks", *Proceedings of the 7th Annual Memphis Area Engineering and Science Conference*, MAESC vol.699, 2005.

[22]. Moldagulova and R. B. Sulaiman, "Using KNN algorithm for classification of textual documents," *8thInternational Conference on Information Technology,* pp. 665-671, 2017.

[23]. J. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization,"*In B. Schoelkopf and C. Burges and A. Smola, editors, Advances in Kernel Methods-Support Vector Learning,* 1998.

[24]. R. Kohavi, "Scaling Up the Accuracy of Naïve-Bayes Classifiers: a Decision Tree Hybrid,"*In Proceedings of KDD-96*, Portland, USA, pp.202-207, 1996.

[25]. R. Quinlan,"Induction of decision trees. *Machine Learning*," vol.1, 81-106, 1986.

BIOGRAPHIES OF AUTHORS

	Ms. Jabeen Sultana Research Scholar at Sri Padmavati Maha VisvaVidyalayam, Department of Computer Science Engineering, Tirupati, 517502, India Research Area: Data Mining, Machine Learning. Published Research papers in the International Journals and also participated and presented papers in the International Conferences.Ad Hoc Reviewer for conference and journals.
	M. Usha Rani Professor at Sri Padmavati Maha VisvaVidyalayam, Department of Computer Science, Tirupati, 517502, India Research Area: Data Mining, Information Retrieval Systems, Cloud Computing and Artificial Intelligence. Published Research papers in the International Journals and also participated and presented papers in the International Conferences. Ad Hoc Reviewer for many conference and journals. Completed Funded Research Project: Pattern discovery through data mining techniques on NREGS (National rural employment guarantee scheme) data of A.P. Guided M. Phil and PhD Students. Published Books:5
	M.A.H. Farquad INI Labs, Waterloo, Ontario, Canada Research Area: Data Mining, Information Retrieval Systems, Cloud Computing and Artificial Intelligence. Published Research papers in the International Journals and also participated and presented papers in the International Conferences. International Travel Grants by Government of India to travel United Kingdom for research work presentation Research fellowship at Institute for Development and Research in Banking Technology a Reserve Bank of India initiative for research in banking industry. Guiding 3 research scholars from Jawahar Lal Nehru Technological University, Hyderabad, India. Ad Hoc Reviewer for many conference and journals.