A Hybrid Framework using Hierarchical Analysis for Classification of Mental Health Information

Infanta Jenifer R, Ramya R, Surya R
Assistant Professor, infanta.jenifer96@psnacet.edu.in
Assistant Professor,ramyarajamanickam2596@psnacet.edu.in
Research Scholar, suryarrrm@gmail.com

Abstract.

Social media has become one of the most important platforms to share the experiences of users. Reddit, A famous social networking site has facilities to express mental health problems. It can assist with medical decisions, public health policies, and improving health care. In a recent study, using Hierarchical approach mental health Reddit posts were evaluated and grouped into 11 disorder themes with Attention mechanisms in a recurrent neural network[24]. However, some posts were misclassified as existing approaches did not give importance to sequential characteristics of the text. In this paper, proposed a hybrid framework of hierarchical attention neural network to accurately classify the posts and to maintain the positional information in the sequence using attention mechanisms such as Multi head with global attention, Directional Self-attention. The results of the proposed system is evaluated using the PubMed medical abstracts and Reddit posts. The comparative analysis is done against other neural networks with Attention mechanism and each approach is evaluated using F- Measure.

Keywords. Hierarchical Attention network (HAN), Multi-head attention, Directional Self Attention, mentalhealth,hierarchicalapproach

1. INTRODUCTION

Text in the form of unstructured data is present everywhere like email, reviews, etc. But extracting the information from it is difficult due to its unstructured nature. The knowledge extracted from that large amount of data is useful for many applications and text classification helps to achieve that by classifying the problems into its predefined categories. It is one of the most significant jobs in Natural Language Processing (NLP), with applications in sentiment analysis, subject labelling, spam detection, and many other areas. Machine learning-based text classification learns to make classification decisions based on previous observations. The algorithm used the labelled dataset to train and based on that it will predict the label for particular unseen input.

Recent research in deep learning and machine learning provide solutions for a large dataset of health information. The approach not only used the data available from the medical organization such as Electronic Health Records (EHR) but also used the patient-generated text available on social media sites such as Twitter, Reddit, etc[17][18]. Mostly patient reports for health problem identification or drug suggestion and any user-generated reviews are used to classify in the healthcare field so it will be in document format. To classify the long sequence of text, existing approach lack accuracy and efficiency as it will not consider the overall characteristics of the sequence.

In this paper, how to categorize mental health-related reddit posts as well as medical abstracts is discussed. For a short sequence of text, a convolutional neural network and a recurrent neural network already exhibited good performance [16]. Moreover, to capture the long sequence of text Long Short Term Memory (LSTM) and Hierarchical approach used with attention mechanisms as attention helps to get the important words related to the sentence [9]. This approach works well for many sentences/document, but some misclassification occurs and computation speed is decreased. Also, the word order in the sentence is important to find the correct classification so positional information is added into this hierarchical network and parallel computation also utilized to improve the performance.

We apply a hierarchical attention network with two types of attention:

- Multi-head global attention to enhance computation performance by incorporating multiple heads.
- Directional Self Attention to capture positional information.

2. RELATED WORK

In this section, most of the traditional approach used for multiclass text classification in healthcare and other, as well as new research in hierarchical attention networks and their disadvantages, are described in this section.

2.1 The Traditional approach for multiclass text classification

The majority of previous text categorization research used a variety of machine learning classifiers, including logistic regression, Support Vector Machines (SVM) (Cortex and Vapnik, 1995), Nave Bayes, and Random Forest,Rule-based and many other approaches with different typesof features (eg: bag of words, TF-IDF, Topicmodeling (LDA) (Resnik et.al, 2015; Rumshinsky et.al).However, the accuracy in predicting the predefined categories isnot efficient. For the small dataset, thetraditional approach works well. As the size of the datasetincreases, then it will not be able to work well and alsoconsume more time while executing thealgorithm.

2.2 By applying Deep Learning Techniques

Convolutional neural network (CNN)generally used for computer vision (images), however, theyhave recently been applied to various NLP tasks (YoonKim, 2014)[21]in themedical domain. Recurrent neural network for shorttext developed to maintain the order of the describedevents for some time step. But the drawbacks in RNN is thatitcannot have the memory to storelongrange dependencies in the text as sequential characteristicsof text is important in predicting the classfor document/sentence classification and it also hasa vanishing gradient or exploding gradient problem.Long Short Time Memory (LSTM) networks [1] are a type of RNN extension that allows the RNN to retain its inputs for a longer length of time while using more memory. Gated Recurrent Units (GRU)[2] overcome the memory inefficiency problem in LSTM as well as the vanishing gradient problem to reach state-of-the-art performance in deep learning applications such as speech recognition, speech synthesis, natural language understanding, and so on. These above techniques can use the one hot encoder to get theinteger representation and feed into any of the above neural networks or any types of embeddingtechniques (Word2vec[3], Glove[4], fasttext[5], Universal sentenceencoder[6], Elmo[7] and ULMfit[8]) can be incorporated intoit.Furthermore, the attention mechanism [22] developed forthe above architecture to interpret the relevantinformation based on the context and this attention layer used on top of the neural network layers (CNN, LSTM, GRU,etc.)

2.3 By applying the Hierarchical approach

Initially, attention mechanism developed for machine translation then it is also used for image caption generation to focus only in the relevant region of the image (Xu et.al, 2015). In NLP, it is incorporated with RNN/CNN based architectures in text classification [22] to predict words which contribute more importance to the sentence. The approach done in a sequentialmanner for document classification consumes moretimeand accuracy in predicting the classes tendstodecrease significantly. Hierarchical approach captures the accurate information because of thehierarchical structure from word level to sentenceleveland then to document[9] shownin Figure 1. The attention mechanism is added on topof the word level and sentence level to focus only onthe relevant information. So thehierarchical classification methods not only increaseaccuracy but also increase greater understanding ofthe document from word and sentencelevel. Hierarchical network used for food recommendations [20] to find the similar users tend to eat and learning user preferences in various recipes. Multi head Attention mechanism with residual gated recurrent unit [19] to capture contextual information within long range and positional embedding is added to know the sequence of text. Bullying detection using hierarchical network [10] to find the representation of comments in instagram.

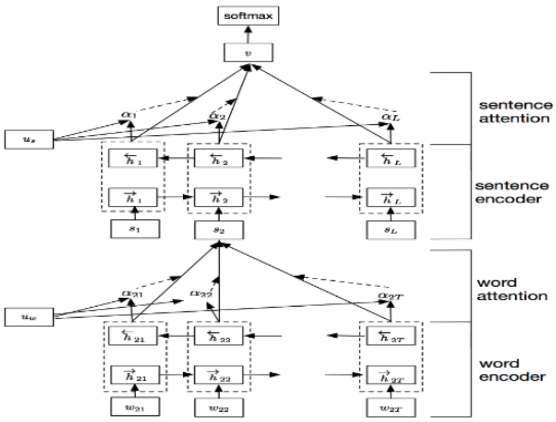

Figure 1. Hierarchical Attention Network [9]

3. PROPOSED METHODOLOGY

Apply the base diagram [1] the proposed changes are done asindicated in Figure 2. The dataset is preprocessed and convertedtointeger representation, then the hierarchicalnetwork approach is used along withattention mechanisms. The classification is done to getthe predefined categories. Each module inthe diagram is explained below.

3.1 Preprocessing

Dataset from different sources oruser-written text is not in standard format to process that text. To improve the data quality, the dataset needstobe preprocessed. So, the first thing to do is to remove the special characters and convert tolowercase letters. The algorithm should learn thestructure of the text as the machine cannot understand words. So, it is represented in numerical form. This can be done by using the kerastokenizer function. The function splitsthe sentence into words and keeps themost occurring words in the text corpus. Then the tokenizer also keeps an index of words whichcanbe accessed by the tokenizer. word_index to specify the maximum number of words. Themaxlen parameter should be specified so that eachsentence will be of the samelength.

3.2 Embedding

To represent each word with its similar context, embedding should be performed for the input sentence. In this approach, we have used Glove and fastText Pre-trained vectors. Glove embedding learns by creating a co-occurrence matrix (words X context) that counts the number of times a word appears in a context. It is trained on the Wikipedia corpus with a dimension of 300 and length of this dictionary is around billion, fastText[11] is an extension of word2vec model. It represents each word in input sentence as n-gram of characters. Pre-trained word vectors are generated by training with 2 million word vectors from common crawl. We match these two pre-trained word vectors with input sentence and extracted only the embedding of words that are in our word index and created an embedding matrix. We have compared the performance of two vectors in this approach.

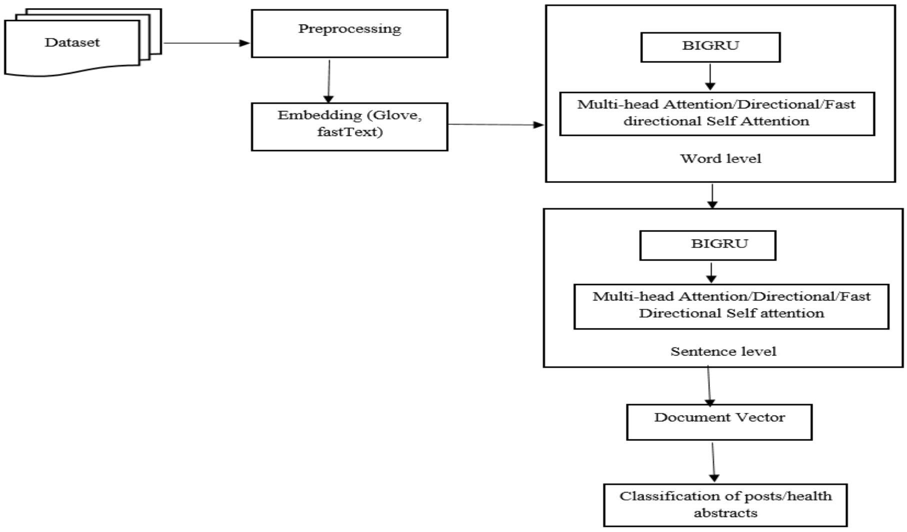

Figure 2. System design for the proposed method

3.3 Word Encoder

3.3.1 Bidirectional Gated Recurrent Unit (BI GRU)

In this step, the output obtained from the previous module is passedi.e) embedding word vectors in the embedding matrix.Then, to gather both forward and backward information, we employed bidirectional Gated Recurrent Units (BIGRU) and attention processes, as well as a gating mechanism to recall the sequences. Reset gate rt and update gate ztdetermine the data that should be transmitted to the output referred from [9].

3.3.2 Attention Mechanisms

The attention mechanism is important to focus only on the relevant information in the sentence. The information from the preceding layer (BIGRU) is transmitted to word-level attention mechanisms including multi-head and directional self-attention and combine the representations of those words to get a sentence vector. The traditional attention mechanism used cannot exploit the positional information of the input sentence. (Ex: I like dogs more than cats and I like cats more than dogs, here both looks similar but here the sentence compares two different entities). To solve this problem proposed attention mechanism used instead of traditional attention mechanisms.

3.3.2.1 Multi-head Attention

The multi-head attention [12] method repeats the standard attention process many times in parallel to increase the computation speed. It then separated into several heads, each of which executes parallel computations. Each head's attention outputs are simply added together and linearly translated into the required dimensions. It enables the overall context of the sentence to be derived from information from several representations at various points.

3.3.2.2 Directional Self Attention

Directional Self-attention[13], the input sentence is transformed into hidden state and then the multidimensional token2token [13] self-attention is used to compute the dependence between x_i and x_j for all elements in the input sentence to handle the diversity of contexts surrounding the same word.Then, positional masks is employed to attention distribution in both directions, i.e. forward mask and backward mask, to encode earlier structural knowledge, such as temporal order and dependency parsing. Thisdesign overcomes the disadvantage in traditional attention mechanism by modeling order information and takes full advantage ofparallel computing. Thearchitecture of Directional self- attention have a fewer parameters,less computation and easier parallelization.

3.4 Sentence Encoder

The sentence vector obtained from theword encoder is given as an input. We canget the document vector in a similar way by applying the attention that was applied in the wordencoder toobtaintherelevantinformationinsentence level. The sentence level attention in HAN is calculated by[9],

$$u_i = \tanh\left(W_s h_i + b_s\right) \quad (1)$$

$$\alpha_i = \frac{\exp\left(u_i^T u_s\right)}{\sum_i \exp\left(u_i^T u_s\right)} \quad (2)$$

$$v = \sum_i \alpha_i h_i \quad (3)$$

Where U_s is the context vector obtained by (1) and overall sentence level attention $\alpha_i$obtained by (2) and v the document vector, which encapsulates all of the data from sentence to document.

3.5 Classification

The output from the sentence encoder v, ie) the document vector with a representation of the document can be used as a feature to the softmax classifier to get theprobability and normalize the value to get the classification labelfor thedocument

4. Experiments

4.1. Dataset

We have used two datasets: 1) Socialmedia Reddit posts, 2) Pubmed medical abstracts. Theuser posted their problem or suggestion in reddit undera particular subreddit. i.e topic-specificcommunity within the platform. Total file of posts around24GB JSON data is publicly available. We havedownloaded the data and then extract only the health-relatedposts that are relevant to the 10 categories mentioned in Table 1. After retrieving,the dataset consists of 96147 posts. Then another datasetof PubMed medical abstracts is downloaded[2]. Itconsists of 210176 medical abstracts with attributes, labeland text.

Table 1. Number of rows in each subreddit

Total Categories:10	
Subreddit	**No ofrows**
Anxiety	11304
BPD	2970
Addiction	341
Autism	1573
Bipolar	9622
cripplingalcoholi	11758
Depression	25909
Opiates	28154
Schizophrenia	1919
Selfharm	1414

4.2 Training the model

The dataset is preprocessed and convertedinto integer representation and it is passed as an input intothe word encoder and then the sentence vector passed tothe sentence encoder and it outputs the document vectorand the classification is done using softmax classifier.We have used keras toolkit for the implementation.The hierarchical attention network architecture [9] is implemented. The dataset is split into two parts: 80 percent training and 20 percent testing. The input vocabulary was set to 30k, the maximum sentence length was set to 15, and the maximum number of words in a sentence was set to 100. The glove pre-trained vectors were used to construct the embedding matrix with a dimension of 100 and were trained using our data set (40000-word vectors with 100 dimensions). We train the model for 15 epochsand utilised early stopping to establish the halting condition for the iteration (epoch). Usually, loss tends todecrease after each epoch and accuracy get an increase. When val_loss tends to rise in some epochs and stays in the same condition for a long period of time, the training will come to an end. By using this we can find the correct number of epochs to train the model.We used Adam optimizer because it works best comparedtoother optimizers. We have used categorical cross entropy to output the probability for multipleclass.

5. RESULT AND ANALYSIS

We implemented the hierarchicalattention network with the attention mechanism such asmulti-head attention and directionalself-attention.Theevaluationisdoneforbothredditpostsand PubMed medicalabstractsandcompares theresultswiththe traditionalneuralnetworkapproach.Theresultsarepresentedin Table2 and Table 3.Theexperimental resultsforredditdataset with Glove word vectorsshows thatmulti-headwithGRU/LSTMlayerachieves 0.68Precision (PR)morethantheexistingHANapproach becauseofmultipleheadsincorporated into the hierarchicalstructurei.e)each headperforms computation inparallel.ButwithoutusingtheGRU/ LSTMlayer inthehierarchicalstructureyieldsless performance becausethefeaturesarenotcaptured effectively.Whileapplyingdirectionalself-attentionto thehierarchical structure,itachievescomparable performancewith existingapproachesbecauseof maintainingtheorder informationinthesentence..TheresultforPubMed datasetachieves0.83 precision inmulti-headattention andwhileapplying directionalself-attention tothisdatasetachieves0.82 precisionanditperformsfastercomputation thanthe multi-headattention.Misclassificationofthepostcan occurbecausesomepostmayfallundertwocategories, butinourapproach,itisreducedbetweenthecategories Another Word embedding used for the same attention mechanism i.e) fastText word vectors. The fastText embedding with simple BIGRU layer produced 0.60 precision for reddit dataset and 0.77 precision for pubmed medical abstracts. Multi-head with GRU/LSTM layer achieved 0.79 precision for pubmed abstracts higher than Reddit dataset. fastText word vectors with directional self- attention achieved 0.58 precision for both datasets. Comparing both word vectors Glove with different attention mechanism achieved higher performance for both datasets.

Approach	Precision	Recall	F1-Score	Accuracy
Convolutional Neural Network	0.82	0.82	0.82	0.82
Recurrent Neural Network	0.84	0.84	0.84	0.84
Hierarchical Attention Network with GRU	0.8	0.8	0.8	0.8
Hierarchical Attention Network with LSTM	0.81	0.82	0.82	0.81

Approach	Precision	Recall	F1-Score	Accuracy
Convolutional Neural Network	0.55	0.53	0.52	0.53
Recurrent Neural Network	0.61	0.6	0.58	0.6
Hierarchical Attention Network with GRU	0.61	0.61	0.59	0.58
Hierarchical Attention Network with LSTM	0.61	0.61	0.59	0.58
Multi-head Attention in HAN with GRU /LSTM	0.68	0.64	0.64	0.64
Multi-head Attention in HAN without GRU /LSTM	0.59	0.59`	0.58	0.58
Directional self-attention Network with RNN	0.6	0.6	0.6	0.6
Directional self-attention Network with HAN (GRU)	0.6	0.6	0.6	0.6
Directional self-attention Network with HAN (LSTM)	0.6	0,6	0.59	0.58

Table 2. Result of Reddit Dataset

.**Table 3.** Result of Pubmed Dataset

Multi-head Attention in HAN with GRU /LSTM	0.83	0.82	0.83	0.83
Multi-head Attention in HAN without GRU /LSTM	0.8	0.8	0.8	0.8
Directional self-attention Network with RNN	0.83	0.83	0.83	0.83
Directional self-attention Network in HAN with GRU/LSTM	0.82	0,82	0.82	0.82

6. CONCLUSION

Textcategorization isanimportanttaskinNatural LanguageProcessing (NLP)since it allows to classifytheproblems basedonthepredefinedcategories. The most difficult aspect of multiclass text classification is properly predicting the categories and having positional information in the sentence.Theattentionmechanism isincorporatedtofocusonlyontherelevantwords withinthewordleveltobepassedtothesentence level andpredict thecategories. Togethe positionalinformation andtoimprove the computationefficiencymulti-headwith global attention anddirectional/fast directionalself- attention isimplementedwiththehierarchical approach. Theproposedapproachachieves comparative performancewith theexisting approach. In future, directional attention mechanism performance needs to be improved for reddit dataset. Fastdirectionalself-attention and otherdifferenttypesofattentionmechanismcan beusedalongwiththehierarchicalstructure and compared with the existing approaches.To reduce dependability on the balance of training data, background knowledge extracted from external corpus can be used in hierarchical attention network

REFERENCES

[1] Hochreiter, Sand Schmidhuber, J, 'Long short-term memory', Neural Computation, 1997. https://doi.org/10.1162/neco.1997.9.8.1735.

[2] Cho, Kyunghyun; van Merrienboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi, 'Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation', EMNLP, 2014. https://doi.org/10.48550/arXiv.1406.1078

[3] Mikolov, Tomas; Sutskever, Ilya; Chen, Kai;Corrado, Greg S.; Dean, Jeff,' Distributed representations of words and phrases and their compositionality', Neural Conference on Neural Information Processing Systems, 2017.

[4] Jeffrey Pennington, Richard Socher, and Christopher D. Manning,' Glove: Global vectors for word representation', EMNLP,2014.Doi: 10.3115/v1/D14-1162

[5] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, 'Bag of Tricks for Efficient Text Classification', Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 2, January 2017. https://doi.org/10.48550/arXiv.1607.01759

[6] Cer, Daniel & Yang, Yinfei& Kong, Sheng-yi& Hua, Nan&Limtiaco, Nicole & St. John, Rhomni& Constant, Noah & Guajardo-Cespedes, Mario,et.al,' Universal Sentence Encoder', Arxiv,2018, https://doi.org/10.48550/arXiv.1803.11175

[7] Peters, Matthew E. and Neumann, Mark and Iyyer, Mohit and Gardner, Matt et.al, 'Deep contextualized word representations', Proc. of NAACL,2018. https://doi.org/10.48550/arXiv.1802.05365

[8] Howard, Jeremy and Ruder, Sebastian,' Universal Language Model Fine tuning for Text Classification', Association for Computational Linguistics, Vol.1, July, 2018. https://doi.org/10.48550/arXiv.1801.06146

[9] Diyi Yang, Chris Dyer,Xiaodong, Alex Smola, Eduard Hovy,'Hierarchical Attention Networks for Document Classification',Assoc.for Computer Linguistic, June, 2016. DOI:10.18653/v1/N16-1174

[10] Lu Cheng, RuochengGuo, et.al., 'Modeling Temporal Patterns of Cyberbullying Detection with Hierarchical Attention Networks', ACM/IMS Transactions on Data Science,2021..https://doi.org/10.1145/3441141

[11] Joulin, Armand and Grave, Edouard and Bojanowski, Piotr and Mikolov, Tomas, 'Bag of Tricks for Efficient Text Classification', 2016.

[12] Ashish Vaswani, Noam Shazeer,NikiParmar,JakobUszkoreit, et.al,'Attention Is All You Need', 31st Conference on Neural Information Processing Systems ,NIPS 2017.

[13] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, ShiruiPan,et.al,'Disan:Directional self-attention network for rnn/cnn-free language understanding', AAAI, 2018. https://doi.org/10.48550/arXiv.1709.04696

[14] Shen,Tao&Zhou, Tianyi& Long, Guodong, Jiang, Jing, Zhang, Chengqi,'Fast Directional Self-Attention Mechanisms',2018.

[15] Shang Gao Michael T Young John X Qiu Hong- Jun Yoon, James B et.al,'Hierarchical attention networks for information extraction from cancer pathology reports', Journal of the American Medical Informatics Association, Vol. 25, 2018,DOI: 10.1093/jamia/ocx131

[16] George Gkotsis, Anika Oellrich, SumithraVelupillai, Maria Liakata et.al,' Characterisation of mental health conditions in social media using Informed Deep Learning', Scientific reports, 2017. Https://doi.org/10.1038/srep45141

[17] Ive, Julia &Gkotsis, George & Dutta, Rina& Stewart, Robert&Velupillai, Sumithra,'Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health', Association for Computational Linguistics,69-77, 2018,. DOI:10.18653/v1/W18-0607

[18] George Gkotsis, AnikaOellrich, Tim J.P.Hubbard,RichardJ.B.Dobson, et.al,'The language of mental health problems in social media',CLPsych@HLT NAACL, 63-73,2016, Doi: 10.18653/v1/W16-0307

[19] R.Bharathi, T.Abirami," Energy efficient compressive sensing with predictive model for IoT based medical data transmission", Journal of Ambient Intelligence and Humanized Computing, November 2020, https://doi.org/10.1007/s12652-020-02670-z

[20] XiaoyanGao, Fuli Feng, et.al,’ Hierarchical Attention Network for Visually-Aware Food Recommendation’,IEEE Transactions on Multimedia;22(6): 1647 – 1659, 2019. 10.1109/TMM.2019.2945180
[21] Yoon Kim,’Convolutional Neural Networks for Sentence Classification’, EMNLP, Association for Computational Linguistic, 1746–1751, 2014.
https://doi.org/10.48550/arXiv.1408.5882
[22] Lyubinets, Volodymyr&Boiko, Taras& Nicholas, Deon,’Automated labeling of bugs and tickets using attention-based mechanisms in recurrent neural networks’, 2018 IEEE Second International Conference on Data Stream Mining & Processing, 2018DOI:10.1109/DSMP.2018.8478511

Biographies

Infanta JeniferR received the bachelor's degree in Computer Science and Engineering from MepcoSchlenk Engineering College, Sivakasi in 2017, and the master's degree in Computer Science and Engineering from PSG College of Technology in 2019. She is currently working as an Assistant Professor at the Department of Artificial Intelligence and Data Science, PSNA College of Engineering and Technology, Dindigul. Her research areas include Data Mining, Deep Learning, and Internet of Things.

Ramya R received the bachelor's degree in Computer Science and Engineering fromShanmuganathan Engineering College, Pudukottai in 2017, and the master's degree in Computer Science and Engineering from PSNA College of Engineering and Technology in 2019. She is currently working as an Assistant Professor at the Department of Artificial Intelligence and Data Science, PSNA College of Engineering and Technology, Dindigul. Her research areas include Data Mining, Cloud Computing, and Internet of Things.

Surya R received theM.Phil(Master of Philosophy) in mathematics from Department of Mathematics,Alagappa University, Karaikudi. Currently, she is pursuing her Ph.D (Doctor of Philosophy) in Department of Mathematics,Alagappa University, Karaikudi. Hercurrent research interests are Operations Research, Mathematical Modelling and Neutrosophic Inventory models.