International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-4, February 2019

Probabilistic Recurrent Neural Network for Topic Modeling

P. Lakshmi Prasanna, D. Rajeswara Rao

Abstract*: Data storing, and retrieving is the most important task in the current situation. Storing can be done based on the topic that the document describes. To know the topics, we have to classify the documents, to classify we are using topic modeling. In this paper we proposed probabilistic recurrent neural network (PRORNN) gives the most prominent result in the classification. it's a Recurrent neural network (RNN)-based language model designed to directly capture the worldwide linguistics which means relating words during a document via latent topics. owing to their consecutive nature, RNNs square measure smart at capturing the native structure of a word sequence – each linguistics and syntactical – however would possibly face problem basic cognitive process long-range dependencies. As recurrent neural network fails to remember large dependencies, we are using topic modeling merged with probabilistic recurrent neural network which is called PRORNN. This PRORNN consists of all the merits of RNN and latent topic models. Thus, it gives most accurate classification as the result. The proposed PRORNN model integrates the merits of RNNs and latent topic models. In this paper we take the 20 news groups data set in that we take 2000 documents and we can labeled to two topics. to classify this 2000 documents and assigned 2 topics to for that documents and use the rnn package to execute recurrent neural network in R Tool.

***Index Terms*: PRORNN, Classification, Topic Modeling, local, RNN.**

I. INTRODUCTION

The system classifies the text using PRORNN which is more accurate compared to the neural network. There are several neural networks to classify the data, they are: 1) Recurrent Neural Network 2) Convolution Neural Network 3) Backpropogation Neural Network

II. RECURRENT NEURAL NETWORK (RNN):

Recurrent Neural Network is the first algorithm that remembers the input based on the internal memory and it is very useful in sequential data and in this algorithm is majorly used in deep learning in now days. This Network is powerful type of a neural network for these algorithms have an internal memory to store the input. In this network have 3 layers: input layer, hidden layer, output layer. In the input layer it take the input and Sending that input to hidden layer in that hidden layer it can process the cyclic and that data it can send to output layer. in Figure 1 it shows the Diagrammatical Representation of RNN.

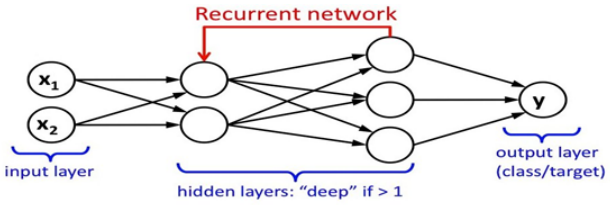


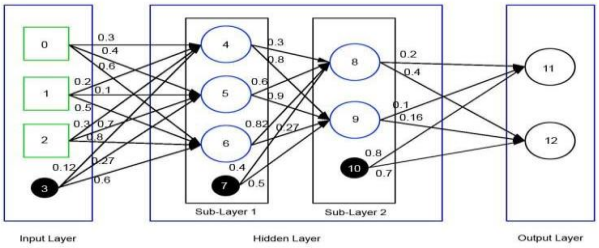
Fig.1.Diagram for recurrent neural network

There are different types of recurrent neural networks: a) Fully recurrent neural network b) Recursive neural network c)Neural history compressor d) Long short term memorye)Gated recurrent unit neural network f)Neural turing machines

a) Fully recurrent neural network:

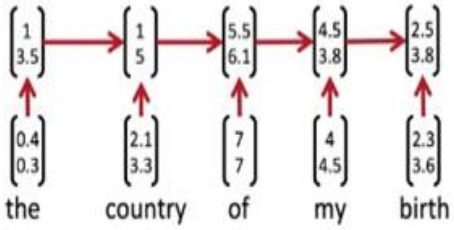
It is a multilayer perceptron, and every element having a weighted connection with every other element. It uses Back propagation neural network having a feedback connection to itself.

Diagram for fully recurrent neural network:



b) Recursive neural network: This recursive neural network is a type of deep neural network, in this we apply the same set of weights recursively over a structured input. Recursive neural network is successful in learning the sequences and tree structures in NLP.

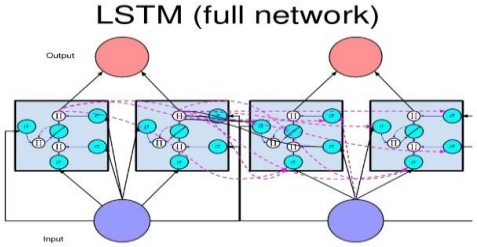
Diagram for recursive neural network:



c) Neural history compressor: In this neural history compressor the input layer can predict the nest input value from its previous input values.

Manuscript published on 28 February 2019.
***Correspondence Author(s)**
P. Lakshmi Prasanna, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522502
Dr.D.Rajeswara Rao, Koneru Lakshmaiah Education Foundation. Vaddeswaram, Guntur, Andhra Pradesh, India – 522502

d)Long short term memory It is called as building unit for layers of the recurrent neural network. LSTM consists of a cell, input gate, output gate, forget gate.

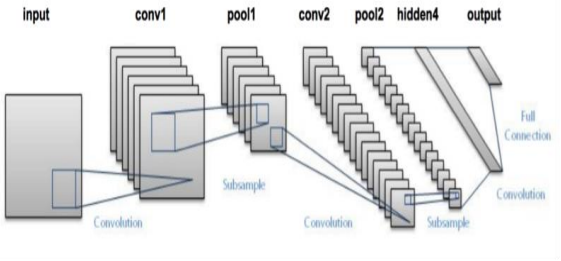


e) Gated recurrent unit neural network It works effectively in speech recognition, NLP, machine translation. It is similar to LSTM in the performance on polyphonic music and speech signal modeling.

f) Neural turing machine: It is a recurrent neural network which combines both fuzzy pattern matching capabilities and algorithmic power of programmable computers.

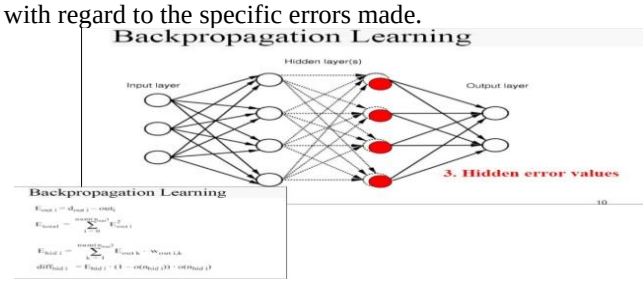
III. CONVOLUTIONAL NEURAL NETWORK

CNN has been successfully applied to analyze or classify the image. It is a feed forward artificial neural network. It uses a variated multilayer perceptron for minimal processing. the two main processes taken place in cnn are convolution and pooling. In cnn the main problem is overfitting.Generally to avoid overfitting pruning, cross validation are used.



IV. BACKPROPAGATION NEURAL NETWORK

In back propagation network the weights can be changed according to the output. If the generated output is not the required output, then the weights and connections can be changed by back propagating the network. It is a supervised learning algorithm that allows the network to be corrected with regard to the specific errors made.



Description

In this paper we are using PRORNN for classifying the text more accurately as PRORNN uses both latent topic and merits of RNN in classifying. Topic modeling is a frequently used text mining tool for discovering the semantic structures in the given text. It is a statistical model. Topic model can include context information such as timestamps, network and author information. In the below diagram every column represents the document and row represents the word. In this every cell stores the frequency of a word. Topic model group both documents which use similar words. The resulting patterns are called Topics.

V. WHY PREFERRING PRORNN BUT NOT RNN:

RECURRENT NEURAL SYSTEM (RNN)- BASED DIALECT DISPLAY INTENDED TO SPECIFICALLY CATCH THE WORLDWIDE SEMANTIC IMPORTANCE RELATING WORDS IN AN ARCHIVE THROUGH INACTIVE SUBJECTS. THE PRORNN COORDINATES THE BENEFITS OF RNNS AND IDLE THEME MODELS, IT CATCHES NEIGHBORHOOD (SYNTACTIC) CONDITIONS UTILIZING A RNN AND ALSO IT CAN STORE INPUT IN INTERNAL MEMORY.

VI. APPLICATIONS OF PRORNN

In many classification methods the main problem is to store the semantic structures of the words to overcome that Topic modeling is used as it can remember large frequencies, so that we can easily classify using their frequencies. A Topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering based on statistics in each document.

1 Personalized short text conversation
2 Speech recognition and machine translation
3 Language modeling
4 Context for dialogue modeling
5 Image captioning
6 Image recognition from text

VII. RESULTS

In this paper we are using 20 news group data set in that 20,000 documents are available and it is 20 set groups. each set consists of 1000 documents related to one particular content .in this paper we are using two set of documents in that one set documents are related to automobiles and another set related to education news group .each set consists of 1000 documents .so we are using 2000 documents .for the first set of documents labeled as topic 0 it is automobiles and second set of documents labeled as topic 1 and it is related to education group. we applied nlp techniques like removing white space, removing stop words, lemmatization, tokenization. These are all preprocessing techniques of the text mining. After the preprocessing 2000 documents 4140 terms are available We then applied the LDA Algorithm to reduce terms from 4140 to 300 terms.. The fig 1 shows the top 10 words from the topics. top5termsperTopic

	Topic 1	Topic 2
[1,]	"subject:"	"the"
[2,]	"message-id:"	"newsgroups:"
[3,]	"writes:"	"lines:"
[4,]	"references:"	"gmt"
[5,]	"path:"	"date:"
[6,]	"apr"	"from:"
[7,]	"can"	"1993"
[8,]	"organization:"	"re:"
[9,]	"article"	"organization:"
[10,]	"one"	"people"

Figure 1: top5 termsperTopic

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

After identifying top terms and we can apply the lda technique and then find the probability of each term based on the probability value we can find priority wise (top)terms. in figure 1 and figure 2 shown that some examples of top terms and the probabilities of the terms .

Probabilities

edu	news	cmu	com	the	srv
0.0607193003	0.0187586503	0.0139525799	0.0123143925	0.0111547800	0.0084064289
writes	subject	aprqmt	article	ohio	
0.0082522474	0.0074809080	0.0068539408	0.0068247050	0.0064686641	0.0050961229
autos	lines	organization	1993	mps	posting
0.0049854769	0.0046986507	0.0046422962	0.0045622272	0.0045329009	0.0044564308
state	rec	god	like	references	one
0.0044525331	0.0044292370	0.0043097829	0.0042613889	0.0041106451	0.0040938560

Figure 2: Probabilities of the terms

With these terms, we created document term matrix(dtm), where rows are topics and columns are top words from the both sets documents .To classify the data sets we applied the recurrent neural network from the package of rnn in R tool . Learning rate may be varied from 0.1 to 1. After the experiments we identified that the model performs well at 0.6 learning rate. The result are tabulated below in fig3 and represented graphically in fig 4.

```
Trained epoch: 1 - Learning rate: 0.6
Epoch error: 0.22413260830237
Trained epoch: 2 - Learning rate: 0.6
Epoch error: 0.0623705445027698
Trained epoch: 3 - Learning rate: 0.6
Epoch error: 0.0390665946306848
Trained epoch: 4 - Learning rate: 0.6
Epoch error: 0.0274639539940824
Trained epoch: 5 - Learning rate: 0.6
Epoch error: 0.0222470412120581
Trained epoch: 6 - Learning rate: 0.6
Epoch error: 0.0191627249097986
Trained epoch: 7 - Learning rate: 0.6
Epoch error: 0.0152465553434716
Trained epoch: 8 - Learning rate: 0.6
Epoch error: 0.0135154498748832
Trained epoch: 9 - Learning rate: 0.6
Epoch error: 0.0116343564033731
Trained epoch: 10 - Learning rate: 0.6
Epoch error: 0.0104846037631226
```

Figure 3: Trained Epoches

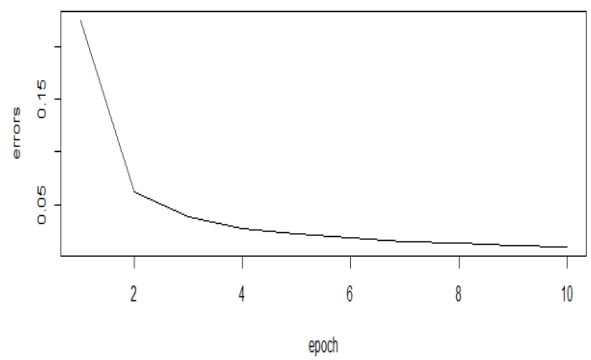


Figure 4: Plot of the Error Rate

VIII. CONCLUSION

we proposed a PRORNN method for improving the accuracy and performance of the neural networks in topic modeling and text classification .This algorithm is useful for classification of documents where semantic meaning of terms is to be considered. As we applied LDA for reducing the terms, the complexity of learning model decreased and the accuracy PRORNN algorithm is increased. The documents can be classified easily and there is no loss of information due to memory inconsistency as it can remember large words also.

REFERENCES

1. Topicrnn: A Recurrent Neural Network With Long-Range Semantic Dependency By Adji B. Dieng, Chong Wang, Jianfeng Gao, John Paisley.
2. Recurrent And Convolutional Neural Networks By Ji Young Lee, Franck Dernoncourt.
3. Neural Network Approach For Text Classification Usinf+G Relevance Factor As Term Weighted Method By Anuradha Patra And Divakar Singh.
4. Automatic Text Categerization Using Neural Networks By Mignel E.Ruiz.
5. Text Classification Using Artificial Neural Networks By Fraser Murray
6. Hierarchical Text Categorisation Based On Neural Networks And Dempster-Shafer Theory Of Evidence By Gertrud Jeschke And Mounia Lalmas
7. Generative And Discriminative Text Classification With Recurrent Neural Networks By Dani Yogatama, Chris Dyer, Wang Ling, And Phil Blunsom.
8. Fuzzy Approach Topic Discovery In Health And Medical
9. Corpora By Amir Karami _ Aryya Gangopadhyay _ Bin Zhou _ Hadi Kharrazi
10. Discovering Scientific Influence Using Cross-Domain Dynamic Topic Modeling By Jennifer Sleeman, Milton Halem, Tim Finin, Mark Cane
11. Textual Document Clustering Using Topic Models By Xiaoping Sun
12. Analysis Of Initialization Method On Fuzzy C-Means Algorithm Based On Singular Value Decomposition For Topic Detection By Ichsani Mursidah, Hendri Murfi
13. Analyzing Sentiments In One Go: A Supervised Joint Topic Modeling Approach By
14. Zhen Hai, Gao Cong, Kuiyu Chang, Peng Cheng, And Chunyan Miao
15. Topic Models For Unsupervised Cluster Matching By Tomoharu Iwata, Tsutomu Hirao, And Naonori Ueda.
16. Bag-Of-Discriminative-Words (Bodw) Representation Via Topic Modeling By Yueting Zhuang, Hanqi Wang, Jun Xiao, Fei Wu, Yi Yang,Weiming Lu, And Zhongfei Zhang.
17. Sequential Short-Text Classification With Recurrent And Convolutional Neural Networks By Ji Young Lee ,Franck Dernoncourt_
18. An Unsupervised Cross-Lingual Topic Model Framework For Sentiment Classification By Zheng Lin, Xiaolong Jin, Xueke Xu, Yuanzhuo Wang, Xueqi Cheng, Weiping Wang, And Dan Meng.
19. Trending Topic Discovery Of Twitter Tweets Using Clustering And Topic Modeling Algorithms By Ma. Shiela C. Sapul, Than Htike Aung And Rachsuda Jiamthapthaksin.
20. Impact Of Topic Modelling Methods And Text Classification Techniques In Text Mining: A Survey By Mino George, P. Beaulah Soundarabai, Karthik Krishnamurthi

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)