Design and Development of Topic Modeling for Probabilistic Recurrent Neural Network

P. Lakshmi Prasanna, D. Rajeswara Rao

Abstract*: A topic model is a probabilistic model that discovers the main themes in a collection of documents. The basic idea is to treat the documents as mixtures of topics in the topic model, and each topic is viewed as a probability distribution of the words. In this paper we proposed LDA Algorithm and Probabilistic recurrent neural network algorithm (PRORNN) to classify the text documents. Topic modeling refers to the task of Discovering Latent Topics in the text corpus set, where the output is commonly represented as top terms appearing in each topic. Out algorithm is implemented by taking 20 news group data set and all the results related to LDA algorithm and probabilistic recurrent neural network (PRORNN) are tabulated. We compared our model with the state of art of algorithms of text classification.

Index Terms*:: Text, Topics, neural network, drichlent allocation, corpus.

I. INTRODUCTION

Topic modeling is a framework for analyzing large datasets where words are collected into groups. Although topic modeling has been focused on social science, biology, and computer vision, it has been mostly used in text datasets where documents are categorized as groups of words [1]. In a topic model, the words of each document are assumed to be exchangeable; their probability is appeared based on occurrences of the words. This simplification has proved useful for deriving efficient inference techniques and quickly analyzing very large corpora [8]. Thus, at a high level, one can think of a topic model as a black box with two outputs: the assignment of words to topics and the assignment of topics to documents. The first output, the topics, is distributions over words, and the second output, the documents, is distribution over topics

ARCHITECTURE

In the Architecture Contains Three Parts ,The First Part contains preprocessing and second Part contains Statistical Method and the Third part contains Machine learning Model. In the First part take the data from the corpus and apply the preprocessing techniques like transformation, filteration, steming, and lemmatization.After Preprocessing Filtered words to apply the LDA and PRORNN Algorithms.

Revised Manuscript Received on June 19, 2019

P.Lakshmi Prasanna,Research Scholar,Computer science and Engineering, KL University, Vijayawada, India.

Dr.D.Rajeswara Rao, Professor,Computer science and Engineering, KL University, Vijayawada, India.

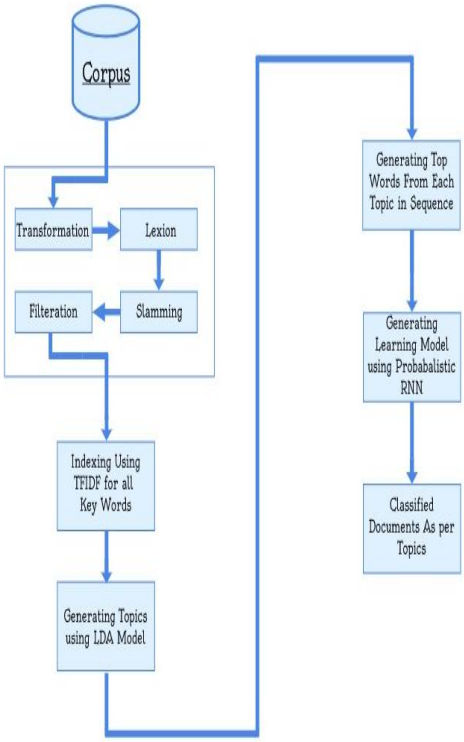

Figure 1 Architecture of Topic modelling using Probabilistic Recurrent neural network

LDA Algorithm

Latent drichlent allocation is mainly used in analyzing text documents .it assumes that there N topics according to documents are generated and each topic is represented by multinomial distribution over y words in the vocabulary [6, 7]. A document wd={wdt}dtt=1 is generated by sampling a mixture and these topics and sampling of words from the mixture.

A formal process of LDA is as follows

1 .the each topic n=1,2,3....N

Draw a word proportion ~~O~~n~drichlet (β)

2. For each document d=1,2...D

Draw topic proportions θd ~drichlet (α)

3. For each word t=1,2.....dt

Draw a topic assignment pdt~catogorical (θd)

Draw a word wdt~catogorical (~~O~~ Pdn) In this algorithm the first step it shows that number of topics and the second step represents every word temporary allocates to topics and this process done randomly and sometimes same words may be applied to different topics[8,9]. The third step shows that

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

update the topic assignment based on there probability based on the two criteria's:

1. The first criteria is how prevalent is that word across the topics it can be termed as P(w/t).
2. The second criteria is how prevalent are topics in the document P(t/d).

Symbol	Description
N	No of Topics
Y	No of Unique words in the Vocabulary
D	No of Documents
DT	No of words in the document DT
θ_d	Proportion of Topics Specific To Documents
~~Θ~~n	Proportion of Words Specific To Topic N
P_{dn}	Identify the topics of nth word in document d
W_{dt}	Identify tth word in document d
α β	Parameters of drichlent distribution

Table 1 Symbolic notations of LDA

II. Architecture Flow of Recurrent Neural Network and Feed Forward Network

In a Feed forward Network have 3 layers that are input layer, hidden layer, output layer. In this network don't have the capability of remember the previous outputs or previous inputs why because it don't have the memory capability but in a recurrent neural network have a memory capability to remember previous inputs and previous outputs and the sequence of words it can remember[10]. In this recurrent neural network also have same three layers input layer, hidden layer, output layer but some differences are there in a RNN have feedback connection to hidden layer and looping to the hidden layer[2] .By using this looping connection every time takes the inputs for entire data and it gives the output. This recurrent neural network it is very useful for Text Data. In Figure 2 it shows the architecture Flow of Feed forward Networks and Recurrent Neural Networks. In this Model we are Focusing on the probabilities of words so this model is probabilistic Recurrent Neural Network[5].

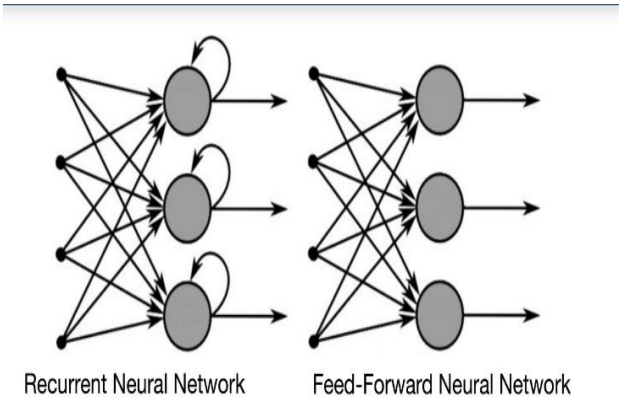

Figure 2 Architecture Flow of Recurrent Neural Network and Feed Forward Network

IV ALGORITHM FOR PROBABILITSTIC RECURRENT NEURAL NETWORK(PRORNN)

1. Represent dtm as vector form

   dtm~w,T_t

2. Represent output variable as Y and to specify the range 0 to n documents.
3. Represent the input f_{dtm} where dimensions as n to p
4. Set Labeled as document 1 to q as t_0

   $_{1,q=}$~ t_0

5 set labeled as document q+1 to n as t_1

   q+1,n=~ t_1

For Training the network

1. To set as ή as x and hidden states are h_t
2. Set the no of epochs are E.

For Testing

1. for testing to specifying Range of s documents to r.
2. To classify the documents for 1 to s it is t_0 and s+1 to r it is t_1

Symbol	Description
dtm	Document Term Matrix
T_t	Top terms of the Documents
X,Y	Input and Output variable
n	No of Documents
p,q	Range Specified as Input
tt_1	Topic Assignments
s,r	Set of Range of Documents
ή	Learning Rate
E	No of Epochs

Table 2 Symbolic Notations of PRORNN

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V .LITERATURE SURVEY

s.no	Name of the author	Statistical techniques	Classification techniques	Clustering techniques	Neural networks	Other techniques	Name of Journal/year
1	Amir Karami Aryya Gangopadhyay Bin Zhou Hadi Kharrazi	fuzzy latent semantic anal-ysis (FLSA)	Random forest	---	---	----	2013
2	Rene Witte1 and Sabine Bergler2	----	----	fuzzy clustering	----	----	2013
3	Subhasree Basu, Yi Yu† , Roger Zimmermann	LDA	----	Fuzzy Clustering, Fuzzy C-Means, k-Means	----	----	2016
4	Rubayyi Alghamdi,Khalid Alfalqi	Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Correlated Topic Model(CTM).	----	-----	-----	-----	2015
5	Yu Chen, Rhaad M. Rabbani, Aparna Gupta†, and Mohammed J. Zaki	NMF, PCA, LDA and KATE	----	----	----	----	2015
6	Ken Gorro1, Jeffrey Rosario Ancheta2, Kris Capao1, Nathaniel Oco2, Rachel Edita Roxas2, Mary Jane Sabellano1, Brandie Nonnecke3, Shrestha Mohanty3, Camille Crittenden3, and Ken Goldberg3	cosine similarity	----	----	----	----	2017
7	Zhenxing Niu,Gang Hua,Le Wang,Xinbo Gao.	LDA	----	----	----	----	2018
8	Anamta Sajid, Sadaqat Jan and Ibrar A. Shah	----	----	----	----	Automatic Extraction	2017
9	Jennifer Sleeman, Milton Halem, Tim Finin, Mark Cane	----	----	Cluster analysis	----	----	2017
10	Xiaoping Sun	TFIDF model	----	clustering accuracy	----	----	2017
11	Ichsani Mursidah, Hendri Murfi	singular value decomposition (SVD).	----	fuzzy c-means	----	----	2017
12	Tomoharu Iwata, Tsutomu Hirao, and Naonori Ueda	Gibbs sampling.	----	Unsupervised clustering analysis	----	----	2017

13	Yueting Zhuang, Hanqi Wang, Zhongfei Zhang	discriminatively objective-subjective LDA (dosLDA)&BOW	----	----	----	----	2017
14	Hanqi Wang, Fei Wu, Weiming Lu, Yi Yang, Xi Li, Xuelong Li.	identified objective–subjective latent Dirichlet allocation (LDA) (iosLDA)	----	----	----	----	2018
15	Filipe Rodrigues , Mariana Lourenc¸o, Bernardete Ribeiro,Francisco C. Pereira.	regression	supervised topic models.	----	----	----	2017
16	Yuan Wang, Jie Liu, Yalou Huang, and Xia Feng.	Hashtag graphs.	----	----	----	hashtag) clustering and hashtag classification problems.	2016
17	Jia Zeng, Zhi-Qiang Liu, and Xiao-Qin Cao	LDA's likelihood function.	----	----	----	expectation-maximization (EM) algorithm.	2016
18	Su-Jin Shin and Il-Chul Moon	Dirichlet Forest priors,	----	hierarchical clustering	----	----	2017
19	Hongzhi Tao, Jianfeng Li and Tao Luo,Cong Wang	----	----	K-means algorithm	----	----	2017
20	Ma. Shiela C. Sapul, Than Htike Aung and Rachsuda Jiamthapthaksin	Latent Dirichlet Allocation(LDA)	----	k-means, CLOPE clustering	----	----	2017
21	A. S. M. Ashique Mahmood	(LDA	----	----	----	----	2017
22	1MINO GEORGE, 2P. BEAULAH SOUNDARABAI, 3KARTHIK KRISHNAMURTHI	Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA).	----	----	----	----	2017
23	Hamed Jelodar1, Yongli Wang1, Chi Yuan1, Xia Feng2	Latent Dirichlet Allocation (LDA), Gibbs Sampling	----	----	----	----	2017
24	PengtaoXie,Eric P.Xing_	Integrating Document Clustering and Topic Modeling	----	----	----	----	2016

25	Wongkot Sriurai	IMPROVING TEXT CATEGORIZATION BY USING A TOPIC MODEL	SVM 79% Accuracy,comparision SVM,DT Naïve Bayes (NB), Decision tree (Dtree) and Support Vector Machines (SVM), are	----	----	----	2011
26	Pengtao Xie, Eric P.Xing	Integrating Document Clustering and Topic Modeling	----	LDA+Kmeans, LDA+Naive, CTM and MGCTM	----	----	2012
27	S. Sendhilkumar, Nachiyar S Nandhini, G.S. Mahalakshmi	NOVELTY DETECTION VIA TOPICMODELING IN RESEARCH ARTICLES	----	----	----	hierarchical Pachinko Allocation Model	2012
28	Pema Gurung1 and Rupali Wagh	----	----	----	----	----	2017
29	Clint P. George, Daisy Zhe Wang,Joseph N. Wilson, Liana M. Epstein, Philip Garland, and Annabell Suh	A Machine Learning based Topic Exploration and Categorization on Surveys	LDA and HDP. Conventional topic modeling	----	----	----	2012
30	Soon Jye Kho, Hima Bindu Yalamanchili, Michael L. Raymer, Amit P. Sheth	A Novel Approach for Classifying Gene Expression Data usingTopic Modeling	----	----	----	LDA,LPA,PLSA Medical lungs datatset	2017
31	Aakanksha Sharaff1, Anshul Verma2, and Hari Shrawgi2	Generic Document Classification Using Clustering, Centrality and Voting	feature words classification				2017

32	Paolo Missier, Alexander Romanovsky, Tudor Miu, Atinder Pal, Michael	Tracking Dengue Epidemics using Twitter ContentClassification and Topic Modelling	---	----	----	oisy, perhaps because LDA is not very effective on text.content that is pre-filtered for a specific set of keywords.	2016
33	Zhengyin Hu Shu Fang Tian Liang	Empirical study of constructing a knowledgeorganization system of patent documents using topic modeling	---	----	----	personalized KOS made up of topics can represent the technology information in a more	2014

VI RESULTS

Data Pre processing: In this paper we are taking 20 news group data .in this data set contains 20,000 documents. In this paper we are focusing on 2000 documents this 2000 documents belongs to 2 topics. for that 2000 documents we are Appling data pre-processing techniques we can get top terms of each document. in a pre-processing we apply some tasks that are 1.convert uppercase to lowercase ,2.removing special characters and dividing tokens3.remove stop words4.stemming 5.construct the document term matrix and at the end we are displaying top terms in the word cloud. In the figure it shows the word cloud of the top terms of all documents [3].

The data preparation process had the following steps:

- All White Spaces are removed in the Documents
- All Special Charecters to be removed
- All words divided into tokens by using tokenization
- All these words to apply stemming process
- All the words to apply lemmatization concept
- By apply all these things we can get words and these words will store it document term matrix.
- It is not possible to take all the words for optimization purpose to take only top words[4].

Top Terms of the Documents

```
     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"
[11,] "re:"                      "apr"
[12,] "just"                     "article"
[13,] "from:"                    "sender:"
[14,] "date:"                    "get"
```

Figure 3 Top Terms of the Documents

All these terms it can be represented in word cloud. In the figure 1.3 It shows the word cloud of the top 300 terms.

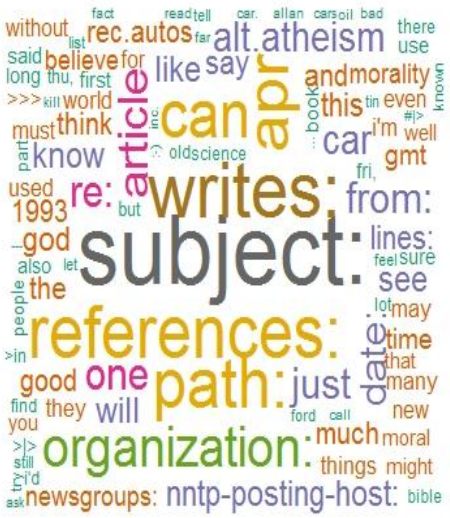

Figure 4 word cloud of the Top Terms

My Model is probabilistic recurrent neural network we are taking the all the probabilities values we are taking as a inputs and based on these probability values we want to find each document assigned to topics. in the 20 news group dataset we are taking 2000 documents in that 1000 are related to automobiles and 1000 are related to athisem .first 1 to 1000 document we are labeling in topic 0 and 1001 to 2000 documents labeled to topic 1.in the figure 2.2 shows the topic assessment in the few document and we are taking learning rate as 0.6 and number of hidden layers are 3 and figure 3 shows the histogram of test output.

```
 [2,]    0          [1001,] 1
 [3,]    0          [1002,] 1
 [4,]    0          [1003,] 1
 [5,]    0          [1004,] 1
 [6,]    0          [1005,] 1
 [7,]    0          [1006,] 1
 [8,]    0          [1007,] 1
 [9,]    0          [1008,] 1
[10,]    0          [1009,] 1
[11,]    0          [1010,] 1
[12,]    0          [1011,] 1
[13,]    0          [1012,] 1
```

Figure 5 topic 0 & Topic 1 Assignment

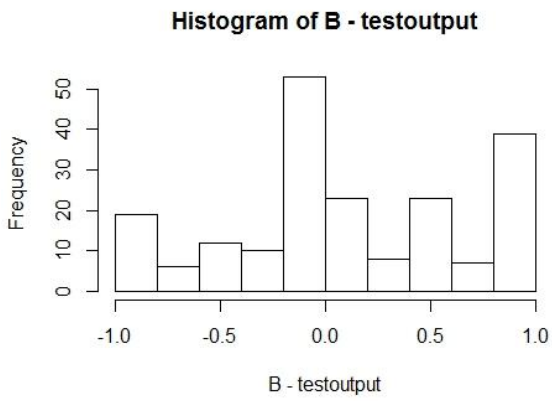

Figure 6: Histogram Plot on Test output Comparison of Different Machine Learning Algorithms

Machine Learning Algorithms	Accuracy
Naïve Bays	71.4
Support Vector Machines	72.6
Probabilistic Recurrent Network	76.3

Table 3: Comparison of Different Machine Learning Algorithms Accuracy

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

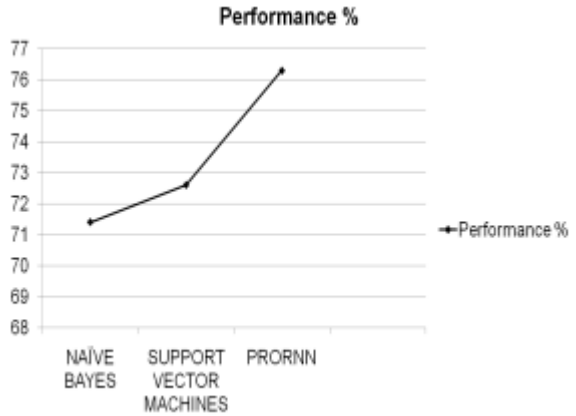

Figure7: Plot on Machine Learning Algorithms Accuracy

VII. CONCLUSION

In this paper we proposed a novel method for topic modeling using Probabilistic recurrent neural network and we show the topic assessment of Documents based on probability values.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 of model 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. Our experimental results showed that probabilistic recurrent neural network model improved the accuracy and performance of text classification compared to the other machine learning algorithms

REFERENCES

1. A Topic model on passion Decomposition haixinjiang 2019
2. Using Recurrent Network to predict Customer Behavor from interaction data by Daniel sanchez santoly July 2017
3 Mixture of topic models for analyzing short text documents with user information by yusaku imai on march 2016
4. Inference and applications for topic models by Christophe depuy on july 2018.
5. language modeling with recurrent neural networks by anna lena pokers on march 2018
6. incorporating domain knowledge in latent topic models by david Michael andrzejewski on 2010
7. structured topic models for language by hanna M.Wallach on 2008
8. Linguistic Extensions of topic models by Jordan boyd-graber on September 2010
9. topic models by Brandon Malone on February 2014
10. A Recurrent neural network with long range semantic dependency by adjiB.Dieng on February 2017.