International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-5, January 2019

A Text Mining Research Based on Topic Modeling using Latent Dritchlent Allocation

P. Lakshmi Prasanna, D. Rajeswara Rao

Abstract*: *Topic modelling is started from text-mining technique for discovering the latent semantic structure in a collection of documents. In the concept of text mining each document is generated from collection of topics. Topic modelling is based on probabilistic modeling, it has a huge range of applications such as linguistic understanding, image detection, automatic music improvisation identification etc.. topic modeling is applied in various fields such as software engineering, political science, medical etc.In this paper we propose topic modelling using LDA (Latent Dirichelt Allocation).LDA is one kind of probabilistic model that work backwards to learn the topic representation in each document and the word distribution of each topic. this paper I will focusing on LDA algorithms and the results shown based on the 20 news group data set. I will also show how topic modelling works on news groups data set on R Tool. Topic Models to analysis news groups data set with tm and topic modelling package in R, to see what are those documents from different topics.

***Index Terms*: Topic Modeling,Text,Corpus,LDA,LSA,Gibbs sampling.**

I. INTRODUCTION

Natural language processing (NLP) is a challenging research area in computer science and information technology and enabling computers to obtain meaning from human language processing in text-documents. Topic modelling methods are very powerful smart techniques that widely applied in natural language processing to topic discovery from unordered documents or unlabeled documents[3]. In a wide perspective, Topic modeling methods based on LDA have been applied to natural language processing, text mining, and social media analysis, information retrieval. Topic models are prominent for demonstrating discrete data; also, give a productive approach to find hidden structures (semantics) in gigantic information. Topic models are applied in various fields including medical sciences, software engineering, geography , political science [2] etc. Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes. Topic modelling algorithms can be applied to massive collections of documents. Recent advances in this field allow us to analyze streaming collections, like you might find from a Web API. Topic modelling algorithms can be adapted to many kinds of data. Among other applications, they have been used to find patterns in genetic data, images, and social networks [4]. topic modelling started from the need to compress large data into more useful and manageable knowledge. There are a variety of different methods for topic modelling, using different sampling algorithms for word selection and topic creation. Examples of topic models include latent semantic analysis. This method is the most basic and looks at the frequency of words within a document and creates topics based on the frequencies of words occurring in each document. In topic modelling, a “topic” is viewed as a probability distribution over a fixed vocabulary. Topics are created based on the strength of correlations between words. Explicit semantic analysis adds words from a document to a matrix based on frequency and creates topics based on the frequency of co-occurrence between words. The amount of data available on the Internet is vast and will only increase over time. Topic modelling provides an easy way to process large amounts of information efficiently. It also allows for individual search topics to be discovered. Topic modelling methods are generally used for automatically organizing, understanding, searching, and summarizing large electronic archives. The main importance of topic modeling is to discover patterns of word-use and how to connect documents that share similar patterns. So, the idea of topic models is that term which can be working with documents and these documents are mixtures of topics, where a topic is a probability distribution over words. In other word, topic model is a generative model for documents. It specifies a simple probabilistic procedure by which documents can be generated.

Revised Manuscript Received on 30 January 2019.
***** Correspondence Author
P. Lakshmi Prasanna, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522502.
Dr. D. Rajeswara Rao, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522502.

S.No	Author	Title	Methodology	Result	Comparison	Datasets	Existing method	Year of pub
1	Amir Karami Aryya Gangopadhyay Bin Zhou Hadi Kharrazi	Fuzzy Approach Topic Discovery in Health and Medical Corpora	In this research we describe fuzzy latent semantic anal-ysis (FLSA), a novel approach in topic fuzzy perspective. FLSA can handle health & medical corpora redundancy issue and provides a new method to estimate the number of topics.	FLSA against LDA by document classi cation using Random Forest, document classi cation using Random Forest, document and execution FLSA against RedLDA by document modeling on redundanttime test. We also evaluate documents.	Document classification and FLSA	medical and Health	LDA Algorithm	2013
2	Subhasree Basu, Yi Yu† , Roger Zimmerman	Fuzzy Clustering of Lecture Videos Based on Topic Modeling	we propose to use LDA in fuzzy clustering to successfully cluster multimedia documents like lecture videos. One of theattempts to clustering lecture videos was to cluster them based on keywords useful to users.	where the keywords are matched with the query words used in video-search by the matched with the query words used in video-search by the users.	Fuzzy C-Means 0.453, PLSA,k-Means	We collected the videos from the YouTube channel forNational Programme on Technology Enhanced Learning	Clustering and classification techniques	2016

3	Rubayyi Alghamdi,Khalid Alfalqi	A Survey of Topic Modeling in Text Mining	Latent Semantic Analysis (LSA),Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA) and Correlated Topic Model(CTM).	Observed	survey on all 4 methods	Survey of methods	Introduction of LSA,PLSA,LDA,CTM	2015
4	Yu Chen∗, Rhaad M. Rabbani, Aparna Gupta†, and Mohammed J. Zaki	Comparative Text Analytics via Topic Modeling in Banking	Bank company Data Set applied to topic modelling	1 day 0.7 0.65 0.54 0.39	NMF, PCA, LDA and KATE for the 8-K and 10-K data, respectively	8-K and 10-K filings, from the years 2005–2016, of 578 bank holding companies	These methods include Principal Component Analysis, Non-negative Matrix Factorization, Latent DirichletAllocation and KATE	2017
5	Filipe Rodrigues , Mariana Lourenc¸o, Bernardete Ribeiro,Francisco C. Pereira.	Learning Supervised Topic Models for Classification and Regression from Crowds	two supervised topic models, one for classification and another for regression problems.	We develop an efficient stochastic variational inference algorithm that is able to scale to very large datasets.	Comparison of all supervised models	Crowdsoucing, multiple annotators.	Supervised topic models.	2017
6	Rene Witte and Sabine Bergler	Fuzzy Clustering for Topic Analysis and Summarization of Document Collections	fuzzy clustering algorithm for the analysis of document collections, as they could have resulted from a query posed to an Internet search engine or an intranet	Focused summary of ten documents based on a question, generated from a cluster graph	Fuzzy clustering algorithms	Survey on all fuzzy clustering algorithms	Fuzzy Clustering Algorithms	2017

			document server					
7	Zhenxing Niu,Gang Hua,Le Wang,Xinbo Gao.	Knowledge-Based Topic Model for Unsupervised	a novel knowledge-based topic model,	significantly improves	Unsupervised topic modelling based on e-based topic	Object discovery, object localization, latent Dirichlet allocation.	LDA with must-links	2018
8	Anamta Sajid, Sadaqat Jan and Ibrar A. Shah	Automatic Topic Modeling for Single Document Short Texts	novel approach to automate the process of extracting topic and main title from a single-document short text.	Nouns are more related, reliable, and suitable words for finding the topic of the text	compared to find the best approach for automatic extraction of a topic	relevance,novelty	topic modeling.	2017
9	Jennifer Sleeman, Milton Halem, Tim Finin, Mark Cane	Discovering Scientific Influence using Cross-Domain Dynamic Topic Modeling	Cross domain analytics to identify the correlations between the IPCC chapters and their cited documents.	predict the importance of its extracted topics on future IPCC assessments through the use of cross domain correlations, Jensen-Shannon divergences and cluster analytics.	Survey On Cross Domain Corrolation	cross-domain correlation; data integration; domain influence	assessment reports of the Intergovernmental Panel on Climate Change (IPCC)	2017
10	Xiaoping Sun	Textual Document Clustering using Topic Models.	Topic modeling	The simple method can achieve the comparable clustering accuracy and recall rate to those latest models and algorithms.	Compare them with the cluster-oriented topic model and other major clustering	Document,probabilistic	TFIDF model	2017

					methods.			
11	Ichsani Mursidah, Hendri Murfi	Analysis of Initialization Method on Fuzzy C-Means Algorithm Based on Singular Value Decomposition for Topic Detection	we examine a non-random initialization by using singular value decomposition (SVD).	SVD based initialization method solves the center of gravity problem in a certain degree of fuzziness and gives a better accuracy	Analysis on Fuzzy C-Means Algorithm	topic detection; fuzzy c-means; initialization; Singular Value Decomposition	TDT methods	2017
12	Zhen Hai, Gao Cong, Kuiyu Chang, Peng Cheng, and Chunyan Miao	Analyzing Sentiments in One Go: A Supervised Joint Topic Modeling Approach	efficient inference method for parameter estimation of SJASM based on collapsed Gibbs sampling.	the proposed model outperforms seven well-established baseline methods for sentiment analysis tasks.	Analysis on supervised joint Topic Models	aspect-based sentiment analysis, probabilistic	SJASM	2017
13	Tomoharu Iwata, Tsutomu Hirao, and Naonori Ueda	Topic Models for Unsupervised Cluster Matching	Efficient inference procedure for the proposed model based on collapsed Gibbs sampling.	The effectiveness of the proposed model is demonstrated	Anlaysis on Gibbs Sampling	Unsupervised object matching.	topic models for unsupervised cluster matching	2017
14	Yueting Zhuang, Hanqi Wang, Zhongfei Zhang	Bag-of-Discriminative-Words (BoDW) Representation via Topic Modeling	model named as discriminatively objective-subjective LDA (dosLDA) is proposed .	each document is appropriately represented as “bag-of-discriminative words” (BoDW).	Traditional approaches, discriminative power of each word in terms of its objective or subjective sense	discriminatively objective-subjective	latent Dirichlet allocation, objective and subjective classification, bag-of-discriminative-words representation	2017

15	Hanqi Wang, Fei Wu, Weiming Lu, Yi Yang, Xi Li, Xuelong Li.	Identifying Objective and Subjective Words via Topic Modeling	model named as identified objective–subjective latent Dirichlet allocation (LDA) (iosLDA) is proposed	observed that distinct words in a given document have either strong or weak ability in delivering facts	Objectives and subjectives of topic modeling	latent variable model.	Latent Dirichlet allocation	2018
16	Filipe Rodrigues , Mariana Lourenc¸o, Bernardete Ribeiro,Francisco C. Pereira.	Learning Supervised Topic Models for Classification and Regression from Crowds	two supervised topic models, one for classification and another for regression problems.	We develop an efficient stochastic variational inference algorithm that is able to scale to very large datasets.	Analysis on supervised and regression analysis on topic modelling	Crowd sourcing, multiple annotators.	supervised topic models.	2017
17	Yuan Wang, Jie Liu, Yalou Huang, and Xia Feng.	Using Hashtag Graph-Based Topic Model to Connect Semantically-Related Words Without Co-Occurrence in Microblogs	topic model to understand the chaotic microblogging environment by using hashtag graphs.	evaluate the effectiveness of HGTM on tweet (hashtag) clustering and hashtag classification problems.	Experiments on two real-world tweet data sets show that HGTM has strong capability to handle sparseness and noise problem in tweets.	Hashtag graph, sparseness of short text,	Hash Tag Based Words	2016
18	Jia Zeng, Zhi-Qiang Liu, and Xiao-Qin Cao	Fast Online EM for Big Topic Modeling	fast online EM (FOEM) algorithm that infers the topic distribution from the previously unseen	Within the stochastic approximation framework, we show that FOEM can converge to the local stationary point of the LDA's likelihood	Analysis on topic distribution of words	Latent Dirichlet allocation, big model,	expectation-maximization (EM) algorithm.	2016

			documents.	function.				
19	Junyu Xuan, Jie Lu, Guangquan Zhang, Richard Yi Da Xu, and Xiangfeng Luo.	Bayesian Nonparametric Relational Topic Model through Dependent Gamma Processes	nonparametric relational topic model using stochastic processes instead of fixed-dimensional probability distributions	capabilities of learning the hidden topics and, more importantly, the number of topics.	Analysis on non parametric and relational model	network analysis, Bayesian nonparametric	Bayesian Nonparametric Relational Topic Model	2017
20	Zheng Lin, Xiaolong Jin, Xueke Xu, Yuanzhuo Wang, Xueqi Cheng, Weiping Wang, and Dan Meng	An Unsupervised Cross-Lingual Topic Model Framework for Sentiment Classification	a novel cross-lingual topic model framework which can be easily combined with the state-of-theart aspect/sentiment models.	significantly improve the accuracy of sentiment classification in the target language.	With already existing models	Cross-language,	Sentiment Classification	2016

II. LATENT DIRICHLET ALLOCATION

Latent Dirichlet designation (LDA) are a kind of probabilistic generative model, which have been generally utilized for finding dormant points (topics) that swarm a given corpus. Each found theme commonly comprises of a set of related individual words, which are the real subjects examined or sentiments communicated in the corpus. In this way, theme models can be utilized in numerous imperative research regions[6]. Topic demonstrating offers a suite of valuable apparatuses that naturally take in the idle semantic structure of a gathering of reports or pictures, with inactive Dirichlet designation (LDA) [11] as a generally prevalent precedent. The vanilla LDA is an unsupervised model based on information substance of records or pictures. In numerous applications side data is frequently accessible aside from crude substance, e.g., client gave rating scores of an online survey or client created labels for a picture. Such side flag for the most part gives extra data to uncover the fundamental structures of the information in study. One of the reasons for its popularity is because it models each document as a multi-membership mixture of K corpus-wide topics, and each topic as a multi-membership mixture of the terms in the corpus vocabulary. This means that there is a set of topics that describe the entire corpus, each document can contain more than one of these topics, and each term in the entire repository can be contained in more than one of these topics. Hence, LDA is able to discover a set of ideas or themes that well describe the entire corpus.

III. DATA PREPROCESSING

The proposed text mining analysis starts with pre-processing the data obtained from 20 news groups data set Before presenting the data pre-processing tasks, the terminologies used throughout the paper are introduced. A document which is defined as a sequence of words and punctuation, refers to the news group data set. In our text mining analysis, there are 20,000 documents consistent with the number of records available in 20 news group data. I will take two classes consists of 2000 documents are identified in pre processing.A term is defined as a word, however term and word are interchangeably used throughout this paper The term-document matrix is a matrix projecting the terms found among all the 2000 documents into the individual document. The term-document matrix is a matrix projecting the terms found among all the 2000 documents into the individual documents. Each row of the term-document matrix corresponds to a term, and each column corresponds to a document. The values in the matrix cells presents the frequency of each term in each document [7].

Our data pre-processing includes five tasks namely, (1) convert to lowercase, (2) remove special characters and tokenize them into terms, (3) remove stop words, (4) stemming, and (5) construct term-document matrix. The details of each of these tasks are explained in the following. Format for all the documents. Step 2 removes the special characters including punctuation marks (e.g. !%$#&*?,/.;"\) and numbers, as they never contribute to our text mining analysis, and then tokenizes the document into terms. Step 3 is the major task in our data pre-processing. Stop words are generally a set of commonly used words in any language (here English) that should be removed from the document, in order to make the focus on the important words instead. In this analysis two types of stop words, namely, generic stop words (some common words such as "a", "an", "is", etc), and domain specific stop words, are defined.

There is no specific rule in identifying the domain specific words, and it requires a thorough consideration of the objectives of our text mining analysis as well as the domain knowledge. The aim of our analysis is to extract the latent topics from the news groups data , for finding topics I will give the topic number 0 and topic 1.the first step is to remove stop words.

The goal of stemming is to reduce the variation in the text data by converting words to their common base form/word stem. for example auto ,auto mobile, auto mobiles were converted into auto and another example universe, university ,universitization were converted into universe. The stemming step is very common in text mining analysis as it helps concentrate the analysis on the base form of the words, rather than differentiating between variations of the words that might cause confusion to the text mining algorithms. The last task in data pre processing is constructing the term-document matrix. This matrix presents the distribution/frequency of terms (rows) within documents (columns). Indeed, the matrix reduces documents (narratives) into vectors of unique words that are presented as the columns. The term-document matrix is mainly used as an input to most of the text mining algorithms, including LSA, and LDA[7] in my data set I can get the document term matrix it can displayed in figure 1.and figure 2 the top 10 terms of the documents in topic wise

And figure3 has shown that the probability values of each term in the topic wise.and Figure 4 shown the word cloud of these terms and the figure 5 shown that Word cloud

IV. RESULTS

```
> dtm
<<DocumentTermMatrix (documents: 2000, terms: 52948)>>
Non-/sparse entries: 276147/105619853
Sparsity           : 100%
Maximal term length: 311
Weighting          : term frequency (tf)
```

```
> dtm
<<DocumentTermMatrix (documents: 2000, terms: 4140)>>
Non-/sparse entries: 180532/8099468
Sparsity           : 98%
Maximal term length: 193
Weighting          : term frequency (tf)
```

Figure 1: document term matrix

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 2: top 10 terms of the documents

probabilities

subject:	message-id:	writes:
0.0172949546	0.0170884915	0.0135671864
references:	path:	apr
0.0122303777	0.0118613931	0.0114458514
can	organization:	article
0.0112066146	0.0088592080	0.0075160422
one	re:	just
0.0070988660	0.0067745911	0.0064550337
from:	date:	car
0.0063414342	0.0062106215	0.0061137902
alt.atheism	like	know
0.0060162284	0.0049348162	0.0048697613

Figure 3:Probabilities of each term in the document

Wordcloud :

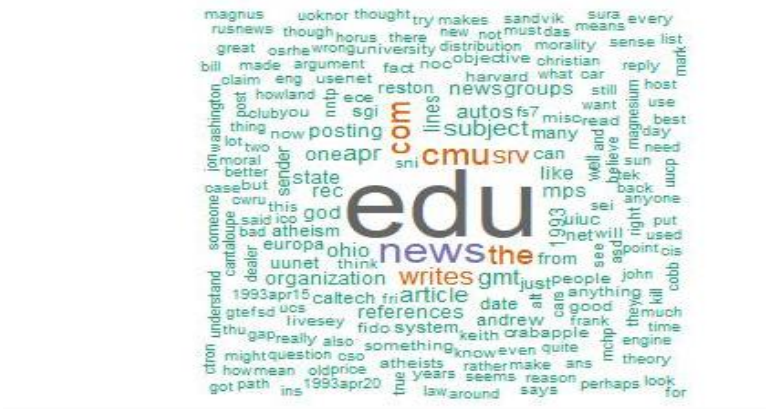

Figure 4: Word Cloud of the terms

V. .CONCLUSION

Topic modeling based on LDA was proposed in this paper to reveal useful information from 20 news group data. LDA generates semantically meaningful topics/clusters to summarize the terms into a mixture of topics that would not be possible by human annotations as reading large volumes of data and interpreting them is very time consuming and difficult. The topic popularity/time trend analysis of these topics can be useful in evaluating the effectiveness of the news group data . in this paper I am focusing on results Document Term Matrix, word cloud ,probability of terms, top terms of the20 news group data set using LDA .

REFERENCES

1. Latent Drichlent Allocation David M.Blei,Andrew Y 2003
2. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey Hamed Jelodar · Yongli Wang · Chi Yuan · Xia Feng · Xiahui Jiang · Yanchao Li · Liang Zhao
3. Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation. Journal of machine Learning research, 2003. 3(Jan): p. 993-1022.
4. Surveying a suite of algorithms that offer a solution to managing large document archives. by David M. Blei 2012
5. Fuzzy Approach Topic Discovery in Health and Medical Corpora, Amir Karami Aryya Gangopadhyay Bin Zhou Hadi Kharrazi 2013
6. A TEXT MINING RESEARCH BASED ON LDA TOPIC MODELLING Zhou Tong,Haiyi Zhang 2016
7. Latent Dirichlet Allocation (LDA) for Topic Modeling of the CFPB Consumer Complaints Kaveh Bastani1 ,*, Hamed Namavari1,2, Jeffry Shaffer 2016
8. Comparative Text Analytics via Topic Modeling in Banking, Rubayyi Alghamdi, Khalid Alfalqi
9. Yu chen, Rhaad M.Rabbani,aparna gupta,mohammad j zaki ,comparative text analytics via topic modeling in banking .
10. Fuzzy Clustering for Topic Analysis and Summarization of Document Collections.
11. Extraction of Unigram and Bigram Topic List by using Latent Dirichlet Markov Allocation and sentiment Classification. PreetChandan Kaur, TusharGhorpade, Vanita Mane Department of Computer Engineering RamraoAdik Institute of Technologhy.
12. Application of text mining to biomedical knowledge extraction: analyzing clinical narratives and medical literature,A Neustein, SS Imambi, M Rodrigues, A Teixeira, L Ferreira, DeGyter publication, 2014
13. Extraction Of Biomedical Information From Medline Documents –A Text Mining Approach", International Journal of Science, Environment and Technology, Vol. 2, No 2, pp 267 – 274, ISSN*: 2278-3687, 2013*