International Journal of Future Generation Communication and Networking
Vol. 13, No. 3s, (2020), pp. 1721–1727

Fuzzy LDA for Topic Modeling: An Overview

Harshali Patil and Prof. Sushila Palwe

School of computer engineering & technology, MIT World Peace University, Pune, Maharashtra, India

Abstract

Topic modeling is underway from text-mining as well as data mining techniques for determining the suppressed semantic assembly in a collection of various dataset. In the conception of text mining every document is engendered from gathering of topics. Subject modeling is constructed on probabilistic modeling, it has a enormous, variety of solicitations such as morphological sympathetic, image detection, involuntary music creativeness identification etc. Topic modeling is implemented in many grounds such as software engineering development, civil engineering, bio medical environment etc. We describe a topic modeling using fuzzy LDA (Latent Dirichelt Allocation).Basically fuzzy logic algorithm generates the probability for LDA which cultivates the classification accuracy of structured as well as semi structured data. The system illustrates topic modeling on synthetic as well as real time data and evaluates the extensive performance analysis with various parameter tuning methods.

Keywords *: Topic Modeling, Text, Corpus, LDA, Fuzzy LDA, Topic modeling, Twitter sentiment analysis, Fuzzy Logic.*

I. INTRODUCTION

Data mining and data processing is most important data processing framework basically used in big data environment. According to In [3] The Natural Language Processing (NLP) has used to initial data preprocessing and feature extraction purpose like word tokenization, Lemmatization, and TF-IDF. Topic modeling approach has already developed in various authors with good accuracy [1], and brief description has mentioned in [2]. The extracted features has evaluated with exact classification algorithm. Digital document stored in much large quantities and increasing all the time requires relate degree automatic technique that enables users to seek out data. The effective method is bunching the documents by its class. To work out the class of a document [5], it has to extract the distinctive options contained within the document .The results of feature extraction are processed to work out the class of document so that similar documents are often collected during a cluster. Therefore, the tactic of extracting options plays a very important role in generating cluster of documents.

II. RELATED WORK

In an extensive perspective, Topic modeling approached according to LDA has been practical to NLP, Data Mining,

Social media analysis, Data Extraction Process. Topic models are conspicuous for representative of multi attribute data; also, provides a creative method to find concealed structures (semantics) in enormous data.

Text mining can be stated as the technique of statistics and machine learning with the aim of recognizing sample and uncovering hide information from text documents. They analysis medical and health application of fuzzy clustering and topic modeling with some key concepts.

ISSN: 2233-7857 IJFGCN

Topic models have so many appeal in NLP concepts. Number of reports and articles has been announced locate on fuzzy clustering and topic modeling techniques in various field such as linguistic science, social network and software engineering.

A. *Fuzzy Approach Topic Discovery*

- The aim of classification is to assign a label data to documents and train a corpus with predefined labels. In [1] Clustering set a cluster to all text document in a corpus based on dissimilarity and similarity between clusters. In wide range medical and health text data actually has been stored and created. Five available medical and health datasets in this research are Springer conducted two Corpus, Unlabeled corpus of 2,434 nursing notes, Ohsumed Collection , Health news from Twitter, Aggregation of text files to trace the bad eff ect. They are using "Fuzzy C-means clustering technique (FCM)" which is reduce overall interval from a cluster prototype to each data.
- Fuzzy clustering method has been applied for survey diabetic neuropathy , distinguishing stroke subtypes , assuming the answer to treatment with citalopram in alcohol dependence , finding early diabetic retinopathy. The main disadvantage in [1] is time complexity issues on large data.
- According to [6] They evaluated fuzzy co reference cluster graph algorithm using DUC technique for doing summarization. They include mainly two multi-document reports and focused summaries of sets between interval of 10 and 50 documents in size. They display ordinary and different topics of text document sets that can be detected based on which is informative and explanatory than simple keyword. Compare to [1] They make a highly strung data structure for topic analysis and summarization. They used to produce how fuzzy set theory can be applied to NLP, which improve robustness.

B. *Latent Dirichlet Allocation*

- The mixture of unigrams model allows for capturing different documents which are derived from different topics, but fails to capture the probability that a document may occurs in multiple topics. So, LDA capture probability, and does so with an increase parameter count of one parameter.

- According to [3] LDA have been proposed with main three parameters which are nothing but Gibbs sampling, Variational method, expectation propagation.
- In[4] they are using the topic models to characterized bad objects from non-failed bank-year objects.
- They have been produced some selected topics which is generated by PCA, NMF, KATE and latent dirichlet allocation for the 800 and 1000 data of 578 bank. In LDA, they simply using substitute of topic-word matrix as the word representation matrix.

ISSN: 2233-7857 IJFGCN

Table: 1 Bank-year datasets:

Dataset Names	**8000**	**10000**
Training Set Validation Set Test Set	2338 200 1866	1787 200 2164
Training Set (failed) Validation Set (failed) Test Set (failed)	39 3 23	40 4 29
Training Set (Not failed) Validation Set (Not failed) Test Set (Not failed)	2299 197 1843	1787 196 2135

The last task of classifying banks by topic representation of bank-year showed that KATE was very competitive than other methods .

- According to [3] NLP is a stimulating investigation part in computer science as well as information technology and permitting computers to get importance from social linguistic dispensation in text-data. Topic modeling approaches are very influential smart practices that extensively realistic in natural language processing to topic detection from imbalance documents or unlabeled dataset.
- In [7] Term frequency has used for classification on given 2000 documents into the separate document. The term-document situation is a matrix prominent the terms found among all the 2000 documents into the individual documents. Each row of the term of each document matrix resembles to a term, and every column parallels to a text. The standards in the matrix cells existing the incidence of each term in each object of document.

C. *Literature review*

- In [Karami A et.al.] Fuzzy Approach Topic Discovery in medical and health corpora , Device explain new fuzzy latent semantic analysis method in health care dataset. In [1] Time complexity issues on large data as well as it works only structured dataset.
- In [Subhasree Basu et.al.] Fuzzy Clustering of Lecture Videos Based on Topic Modeling, Fuzzy LDA and LDA has been described for classification of multimedia records for video search engine method. It always done with semantic classification method where large feature vector cause with high space difficulty.
- In [Hamed JElodar et.al.] A Survey of Topic modeling in Text Mining , In this paper many topic modeling approaches have been defined for text mining. We give mainly four type of

ISSN: 2233-7857 IJFGCN

technique such as Correlated topic model , Latent semantic analysis , Latent dirichelt allocation , Probabilistic latent semantic analysis.

- In [Chen Y et.al.] Comparative Text Analytics via Topic Modeling in banking, Comapre and generate numerous topic modeling methods for banks. We are using the mainly four topic modeling approaches to a corpus of 8000 and 10000 text documents . We are using the data from 2005 to 2016 of 578 banks.
- In [Filipe Rodrigues et al .] Learning Supervised Topic models for Classification and Regression from Crowds , It is used for develop the stochastic version inference technique that is going to balance big datasets. In this paper they experientially display, using real and simulated annotators from amazon mechanical turk. The submitted model is able to get state of the art method in many real world issues, such as new stories , new images, assuming the total number of stars of restaurant , assuming the total rating of movies based on feedback, classify posts. There is two supervised techniques are used to produce BK which calculate the classification ambiguity. The calculated classification ambiguity helps to improve positive and negative ratio.
- In [Write R et al] Fuzzy Clustering for Topic Analysis and Summarization of Document Collections, presented a fuzzy set theory that need to applied on processing natural language, which will enumerate vigorous, simplify utilization, and adjustable to the approach we are using, in the mean period shifting of objective and algorithm from the field of information management to tallying computation; aggregation from large-scale evaluations on data from the international Document Understanding Conference (DUC) contest prove how near we are with current narration systems. Exploration of document could impact on various margin.
- In [Jennifer Sleemanet al.] Discovering Scientific Influence using Cross – domain Dynamic Topic modeling, the priority on change in climate and use of technical records of the Intergovernmental Panel on Climate Change . For the IPCC use case, the transcript was based on 410 pages document and a vocabulary of 5911 terms while the citation prototype was of 200K paper of research and at least of 25K terms in vocabulary. Dynamic Topic Modeling (DTM) uses abstractions state space models and maintain the natural parameter for the subject matter for distribution and document circulation. Sometimes It generate the issue of data leakage when there is some misclassified data encounters.
- In [Anamata Sajid et al.] Automatic Topic Modeling for Single Document Short Text, In this presentation the process towards the automating of process for extraction of matters and head tiltle from a single-document short text. Three recommended algorithm for subject Extraction are Frequent Words Only Approach, Nouns Only Approach, PVAN (proper noun, Verb, Adjective and Noun). Calculation shows that the Noun only approach effective than the other three as the F measure is in between 0.5-0.

III. SYSTEM ARCHITECTURE

Figure 1 describes the projected framework of feature demonstrations for learning the taxonomy prototypes. In our research approach, we estimated among two various feature demonstrations applying the feature selection as well as extraction approach and finally executes the classification algorithm to gain the final accuracy of system.

According to given a web page data collection, the describes an of text processing is applied to extracted terms, sometime such terms contains such invalid data. The set of term is invented by smearing the topic model based on proposed LDA algorithm. The outcome is the topic possibility

ISSN: 2233-7857 IJFGCN

demonstration for each documents object. We apply various supervised learning base classification algorithm in train as well as test respectively to entire classification model. The various models we can use to estimation the performance of class extrapolation.

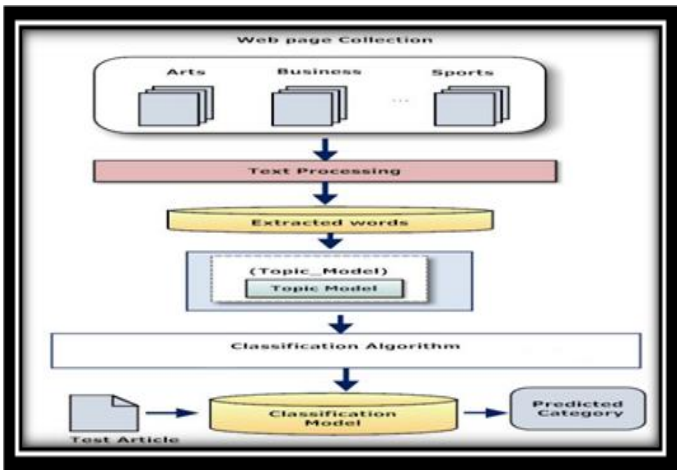

Fig 1 : Execution model

IV. ALGORITHM

Feature Selection is used for dimensionality reduction of original feature set to get the more relevant feature space for classification.

Input: Threshold , Test Dataset holds TestDBList[],Train Dataset holds TrainDBList[].

Output: HashMap with Class Label and Weights.

Step 1**:** Read all test instances using given formula.

Testfeature(m) = ∑ (featureset[A[i]…A[n]] ← TestDBlist)

Step 2: Fetch all features as an input from Testfeature(m) using given formula.

Extracted_featureset_x [t...n] = ∑ (t) ← Testfeature (m)

Step 3 : Read all train instances using given formula.

Trainfeature(m) = ∑ (featureset[A[i]…A[n]] ← Trainfeature(m)

Step 4 : Fetch all features as a input from Trainfeature(m) using given formula.

ISSN: 2233-7857 IJFGCN

Extracted_featureset_y [t...n] = ∑ (t) ← Trainstfeature (m)

Step 5 : Map all inputs instances of test feature set with train feature set.

Weight = calcSim (Extracted_featureset_x || ∑ Extracted_featureset_y)

Step 6 : Return Label with weight.

V. DATASET USED

Basically three dataset has used for experiment analysis of entire system, and split data using 10 fold cross validation. Some additional dictionaries has used for NLP process like stopword removal dictionary, wordnet , vocabul.

Table 2. well known datasets

Id	Name	Records/Samples
1	New York Times News Dataset	150
2	Twitter	10000
3	IEEE papers PDF dataset	300

VI. DISCUSSION

In the initial stage, Fuzzy LDA algorithm begins with the perusal of a set of words from the result of re-filtering. The algorithm produced the number of occurrences each word in each document. The algorithm also construct fuzzy output curve based on the number of topics were set during the inference. There are still much to be improved using LDA method for topic modeling for text datasets. Also thinking to improve efficiency of the Topic modelling applying LDA model on Speech datasets Using Part of speech tag filters. We use the audio files of developers speech from their conference videos or YouTube videos, which is never used before for analyzing topic modeling.

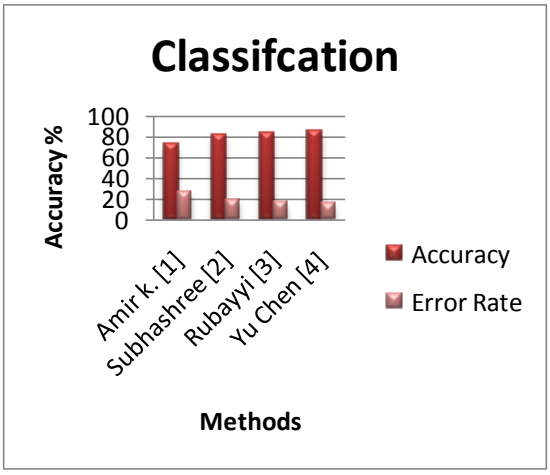

Fig 2 . Classification accuracy of various systems

The below figure 2 shows some existing classification accuracies of topic modeling with various dataset using different classification algorithms.

ISSN: 2233-7857 IJFGCN

CONCLUSION

To recover the presentation of text classification founded on the bag of words feature demonstration. This investigation , The LDA procedure is used to gathering the term structures into a set of dormant topics. We find , The identical subject will has terms that are semantically connected. We associated amongst the dispensation methods of NLP and the topic model. In the research , fuzzy LDA base topic modeling approach for classification. In further , we are going develop high spirited and hierarchize topic models with fuzzy output. In assembling , fuzzy latent semantic analysis develop social media text information to trace public emotions and going to use for message service recognition and online analysis.

REFERENCES

[1] Karami A, Gangopadhyay A, Zhou B, Kharrazi H. Fuzzy approach topic discovery in health and medical corpora. International Journal of Fuzzy Systems. 2018 Apr 1;20(4):1334-45.

[2] Subhasree Basu, Yi Yu, Roger Zimmermann. Fuzzy Clustering of Lecture Videos Based on TopicModeling. Tokyo 101-8430, Japan.

[3] Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. HamedJelodar,Yongli Wang,Chi Yuan ,Xia Feng , Xiahui Jiang , Yanchao Li, Liang Zhao.

[4] Chen Y, Rabbani RM, Gupta A, Zaki MJ. Comparative text analytics via topic modeling in banking. In 2017 IEEE Symposium Series on Computational Intelligence (SSCI) 2017 Nov (pp. 1-8). IEEE

[5] Filipe Rodrigues, Mariana Lourenço. Learning Supervised Topic Models for Classification and Regression from Crowd. IEEE 05 January 2017.

[6] Witte R, Bergler S. Fuzzy clustering for topic analysis and summarization of document collections. InConference of the Canadian Society for Computational Studies of Intelligence 2007 May 28 (pp. 476-488). Springer

[7] Jennifer Sleeman, Milton Halem, Tim Finin,Mark Cane. Discovering Scientific Influence using Cross-Domain Dynamic Topic Modeling. 15 January 2018 , IEEE.

[8] Automatic Topic Modeling for Single Document Short Texts. Anamta Sajid , Sadaqat Jan, Ibrar A. Shah. 2017 International Conference on Frontiers of Information Technology (FIT).

ISSN: 2233-7857 IJFGCN