Comparative Study Analysis on Research Papers Categorization using LDA and NMF Approaches

Bandi Rupendra Reddyy, Darukumalli Sai Tharun Reddy, Sandeep Preetham M C, Deepa Gupta

Department of Computer Science and Engineering, Amrita School of Engineering, Bengaluru,
Amrita Vishwa Vidyapeetham, India.

bl.en.u4aie19009@bl.students.amrita.edu, bl.en.u4aie190016@bl.students.amrita.edu, bl.en.u4aie19058@bl.students.amrita.edu,
g_deepa@blr.amrita.edu

***Abstract* - In the current digital world, there has been an exponential growth in the field of computers and information technologies all over the globe. Researchers are finding it relatively challenging to recognize and categorize their favorite research articles. Our main motive in this project is to come up with a comparison between to system model for research paper classification which further can group the research papers into their respective classes such that their publications are very likely to have subjects in order to solve these challenges. To extract sample keywords from each of the papers and themes and model unstructured data categorized into subjects, we use the topic modeling methodologies such as NMF and LDA in our project. LDA is knowingly used everywhere, In this paper we would like to compare its performance with another generative model (NMF) and see how they perform on our dataset. The dataset we are going to use consists of 1740 papers that were extracted from the NYC university website. We have further compared the two models by calculating the average coherence score for the LDA method which was 0.5282 with its optimal choice of topics being 22, which was comparatively higher than the coherence of the NMF model as it yields us a coherence score of 0.5012 with its optimal topics being 9.**

Keywords: Topic modeling, LDA, NMF, Coherence score

I. INTRODUCTION

A vast collection of online archives of the published scientific articles are available to researchers, however, it has become very challenging to bring connectivity and find relevant articles among all of them. Though this task can be achieved by putting in the human work, it will still end up being a very time-consuming task. It becomes increasingly difficult to efficiently cluster, organize and process the growing number of research publications as time goes on as we are aware of the relationships between the papers to be examined and classified being so complicated. As it is a complex task to grasp the subject of each research paper in a shorter time and it's comparatively complex to effectively classify the research papers with their comparable subjects with respect to the content correctly. To overcome this issue, a self-regulating processing approach to deal with such a large number of volatile research papers is required in order to classify them with lesser time complexity and to make sure it yields the results accurately. We hope to come up with an approach where these large number of research papers would be systematically classified into comparable categories, helping the users to find their relevant research papers quickly and easily. The term tagging, which also known as topic is modeling, is a method of assigning a unique identifier to research articles that aids in the recommendation and search process.

Topic models, have been used broadly in the field of statistics and NLP in order to extract the abstract "themes" which already exists in a collection of documents. Topic modeling is one of the finest text extraction approaches for analyzing the latent meaningful structures present in the document. These models also help in organizing and analyzing massive volumes of unstructured text bodies by providing insights and also finding the instructional patterns present in the data which could be very applicable in the field of genetic information, image data, and also in networking, originally designed as text-mining techniques. Given a text, which is about a particular topic, it is practically possible to expect certain words to occur more or less number of times (frequency) throughout the document in the corpus. These models are broadly used to categorize and classify the papers into specific categories. In the realm of text analysis, there are various topic modeling techniques that can effectively classify documents.

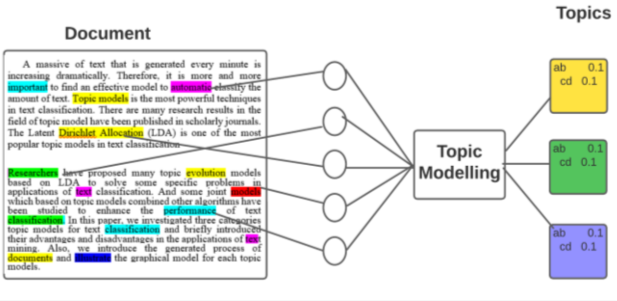


Fig. 1. Overview of Topic Modelling.

We propose a comparison between two of the most used generative topic modeling techniques for categorizing research papers into topics using both the LDA Latent Dirichlet allocation Model and NMF (Non-Negative Matrix Factorization) model. Both the models follow different approaches as LDA is a well-known probabilistic-generative-model whereas NMF (Linear algebraic model) is a matrix factorization technique. The research papers are chosen at random and based on a set of variables each document is assigned to a category. LDA is a statistical topic modeling approach that divides big clusters of text into fictional groups and categorizes them. Each text is made up of themes, each of which has a proportion and a weight assigned to it by each word. NMF divides a input matrix into two different matrices by considering only positive elements for both the original matrix and resultant two matrices. It makes it easier to see the matrices that have been obtained as there are only positive values and turn out to be a sparse matrix. It is widely in use in the fields of both linear algebra and multivariate analysis. In comparison with these two topic models respectively, the LDA model has been demonstrated to be effective, dependable, and simple to use when it comes to classifying textual data. However, NMF performs better when the input data is short and the documents have high coherency. The performance of LDA reduces as these two external conditions are taken under consideration.

II. LITERATURE SURVEY

LDA: It is a probabilistic generative model that is set for discrete data which can be widely used in documents and text corpora it belongs to. The main ideology behind this algorithm is to obtain a general overview of a large number of text(documents) in order to find an important statistical model with a good relationship that could further be used for classifying and summarizing the text and also to analyze text similarity and also for originality detection [3]. LDA can also be used to construct topics that summaries other research subjects where each term is derived from a fundamental concept [10]. They have also enhanced LDA in another investigation by using an author model to see which authors use which subjects [11].

Proposed [4] a document clustering approach which uses NMF to extract document features, also the K-means clustering algorithm which helps in grouping the documents. The proposed research's main focus is on reducing the dimension of the vectors which are created by the document term count, rather than a sophisticated classification which is based on a variety of subject words.

III. CONCEPTS

A. LDA

It is an unsupervised-learning technique that aims in categorizing the hidden observations from the input document into a large number of categories. These categories constitute a probability distribution over the traits in and of themselves. LDA is a generative probability model, which means it uses latent variables to try to produce a model for the distribution of outputs and inputs. LDA comes under the generative models whereas the other type (Discriminative models) mainly focus to learn how input data is being mapped to the outcomes.

The application of LDA is been vastly increasing from clustering buyers based on product purchases to automatic harmonic analysis in music, LDA can be used for a variety of tasks. However, it is most commonly associated with text corpus topic modeling. Documents are what we call observations. Vocabulary is the term for the feature set. A word in a respective document is a term that is used to describe a feature, likewise, Topics are the categories that are achieved from this process.

The LDA model consists of 3 hyperparameters, namely

α — It represents the prior probability estimated for a particular topic (the number of topics that can be used to categorize each document). If the alpha value is set to be high it indicates that each document consists of numerous topics, similarly, if the alpha value is low it represents that the document is made up of fewer topics.

β — It represents the topic-word density (The total number of words present in a topic)

K - It represents the total number of topics that can be used to classify all the documents present in the corpus. It is essential for us to choose the right number of topics to yield optimal results.

The purpose of training the model is to discover the hyperparameters that maximize the likelihood that the text corpus will be generated by the algorithm.

The working of LDA

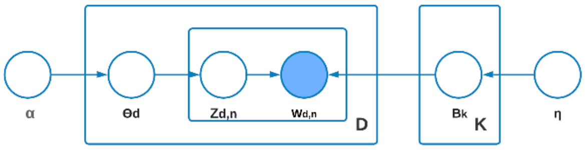


Fig. 2. LDA working model

Here,
K = No. topics
B_k = the distribution of topic over the vocabulary
D = total count of documents present in the corpus
Θ = per-document topic proportions,
N = Number of words present in the document.
$Z_{d,n}$ = Assigning each word to a separate topic.
$W_{d,n}$= observed word,
α, η = Dirichlet parameters

In the context of LDA, two types of approximation inference are used: Variational inference and collapsed Gibbs sampling.

B. NMF

It is one of the famous dimensionality reduction approaches for forming lower rank representations of matrices compared to the input higher dimensionality matrix that contain non-negative or positive elements which are most helpful in dimensionality reduction. These dimensionality reduction matrices has a vast range of application in the domains of Images, as we are aware of the fact that the values in the matrices are positive integer numbers indicating the pixel intensities. The Word-document matrices are also used to indicate the collection of document in information retrieval and also text extraction.

NMF decomposes the input matrix A with m x n representing the rows and columns of the matrix into 2 reduced matrices W and H (strictly positive elements) with their dimensions being m x k and k x n.

Matrix A is defined as follows:

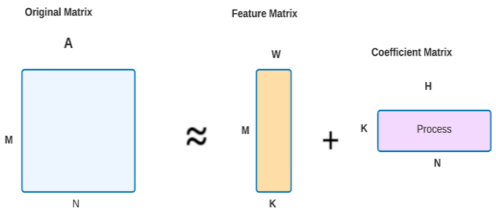


Fig: 3 NMF working

It is represented as $A_{m*n} = W_{m*k} + H_{k*n}$

here,

A = The given input matrix

W = The decomposed Feature Matrix

H = The decomposed Coefficient Matrix

k = Rank approximation for the given input matrix A such that (k less than or equal to min (m,n)).

The goal of NMF is to reduce dimensionality without losing the important data and extract the essential features for modeling. The main objective of the NMF is to find the two positive matrices such that W ∈ Rm*k and H ∈Rn* k, where k denotes the total number of topics which is a hyperparameter in this model.

As a result, we can generate factorized matrices with substantially fewer dimensions than the product matrix by employing non-negative-matrix factorization. The core assumption behind modelling NMF is that it considers the original input that consists of hidden features, wherein each respective column from the feature matrix (W)denotes the position of the respective data point in the matrix, and each column of the coefficient matrix(H) talks about the position of the data point' in the feature matrix. It also holds the weights associated with matrix W. As we are aware of its feature of only taking positive values, no other operations are allowed on the matrix. NMF is applicable in various fields which include image processing, text mining and many more. Due to its effective dimensionality reduction, NMF provides fast computational time, however, NMF is currently undergoing research to improve its efficiency and stability.

IV. IMPLEMENTATION

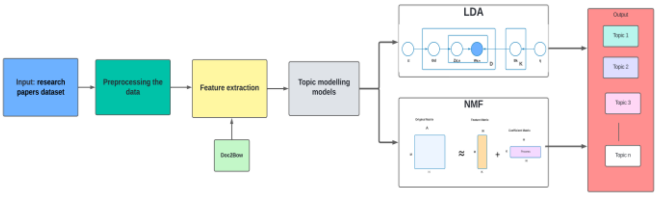


Fig 4: Flow Chart.

Dataset: The corpus consists of 1740 scientific research papers that were extracted from the NYC university website. The dataset consists of a title, authors, abstract, introduction, relevant work, methods, results, conclusion, and references, which are collected from various research papers and combined in one text file.

Data Preprocessing:

Firstly, we load our respective 1740 scientific research papers from our dataset. After that, we perform preprocessing steps for the data before diving into the topic modeling. The essential Natural language techniques have been performed on the dataset such as Tokenization, lemmatization of nouns, and removal of stop-words and single-character terms to keep things simple. We further move on to creating the bi-gram-based phrases and eliminated all the terms in the dataset that do not have weightage or value to the model and might end up disturbing the model's accuracy. This is being done by assigning a threshold to the words in the corpus such that all words that appeared less than 20 times or words that were in more than 50% across all documents are being eliminated. We further use the bag-of-words model to know the occurrence of each word present in the corpus. Now that our documents have been processed, we can begin the topic modeling stage with a better representation using the help of the bag of words model.

Models:

Further, we perform the two different topic modeling techniques on our respective research paper dataset that was extracted from the NYU website. Likewise, we also perform classification using both the topic modeling models LDA and NMF. Firstly, we performed LDA-based topic model methods on our research papers based corpus using the Gensim library and printed the topics for respective articles, and then calculated the coherence score to check the performance of our model. Further, we have also performed a sample test research paper to predict the most dominant topics. Likewise, for comparison between the LDA models, we have also performed LDA using an inbuilt mallet package which yields us a better coherence score.

Now we move on to building our second generative model that is widely used for topic modeling (NMF) to predict the respective dominant topics for each document after splitting the input A matrix into two matrices(document-topic and topic word matrices).

Evaluation:

Topic Coherence score: A statistical metric used that analysis the degree of similarity with respect to the semantics between the words with high score in an individual topic. Topic coherence can e classified into two different types which we use in our project to evaluate the models performance.

- Cv measure
- UMass

The C_v measure works on the principle of a moving window. With the help of the co-occurrence of the words in the document it creates vectors of these words and further evaluates the models score based on cosine-similarity and NPMI.

C_UMass: this approach comparatively performs better that C_v measure. Here we the score is being calculated based on the occurance of 2 words that appear consecutively in the document.

V. COMPARISION

In this section we are comparing the both models LDA and NMF for coherence score, time taken to compute the topics, number of topics for each topics, topic-word matrix and comparison of test sample for both models.

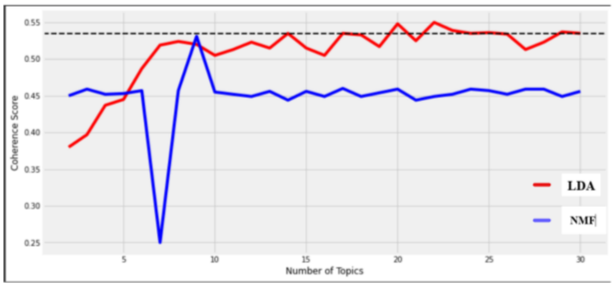


Fig. 5. Coherence score of research papers for LDA and MNF models.

Here we represented the coherence score for 2 to 30 topics for both models. In the graph red line represents the coherence score for LDA and blue line represent coherence score for NMF. By analyzing the coherence score from fig.5, we got the highest coherence score for LDA model for 22 topics. For NMF of topic 9 got the similar coherence score as we got in LDA.

Perplexity: It is a metric method for how confuse a topic model has been trained to predict the new test data. Perplexity is a diminishing function of the likelihood of new documents in LDA topic modeling of text documents. The less perplexity score indicates that it is a good topic model. We got -8.53533 perplexity of the LDA model using the LDA mallet package.

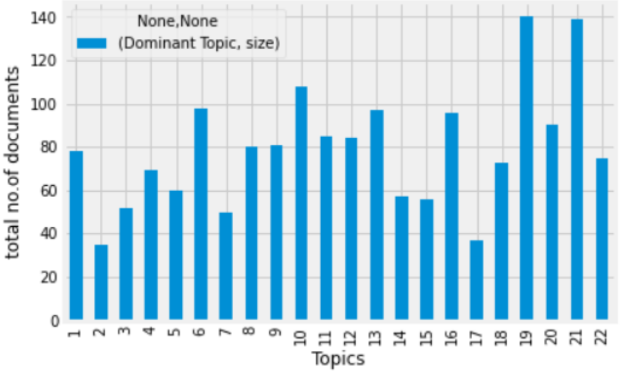


Fig.6. number of topics for each topic for LDA.

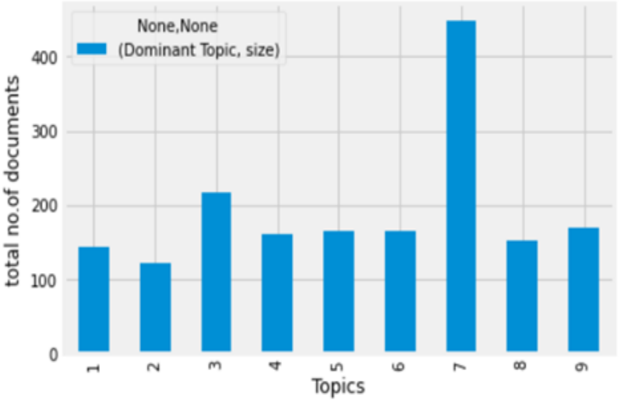
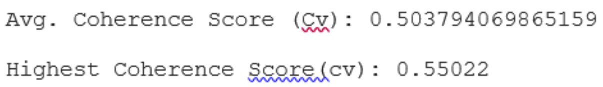


Fig. 7. Number of documents for each topic for NMF

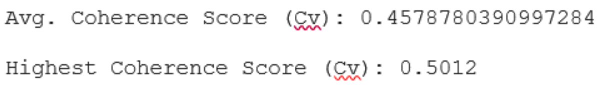
From fig 6, fig 7 by comparing the both models topics for 19 and 21 topics in LDA and 7th topic in NMF has the more topics which means that many topics are similar to one particular topic.



```
Avg. Coherence Score (Cv): 0.503794069865159

Highest Coherence Score(cv): 0.55022
```

Fig. 8. Coherence score for LDA.

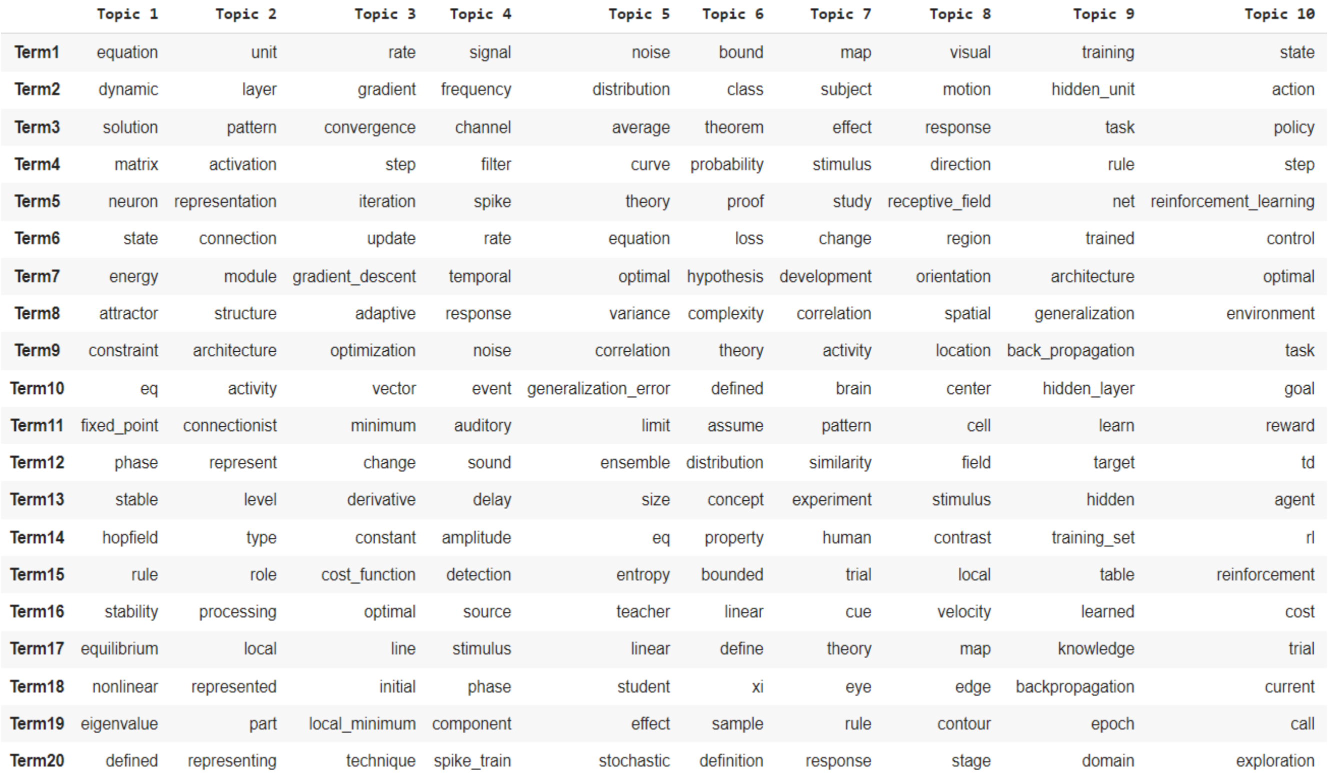


```
Avg. Coherence Score (Cv): 0.4578780390997284

Highest Coherence Score (Cv): 0.5012
```

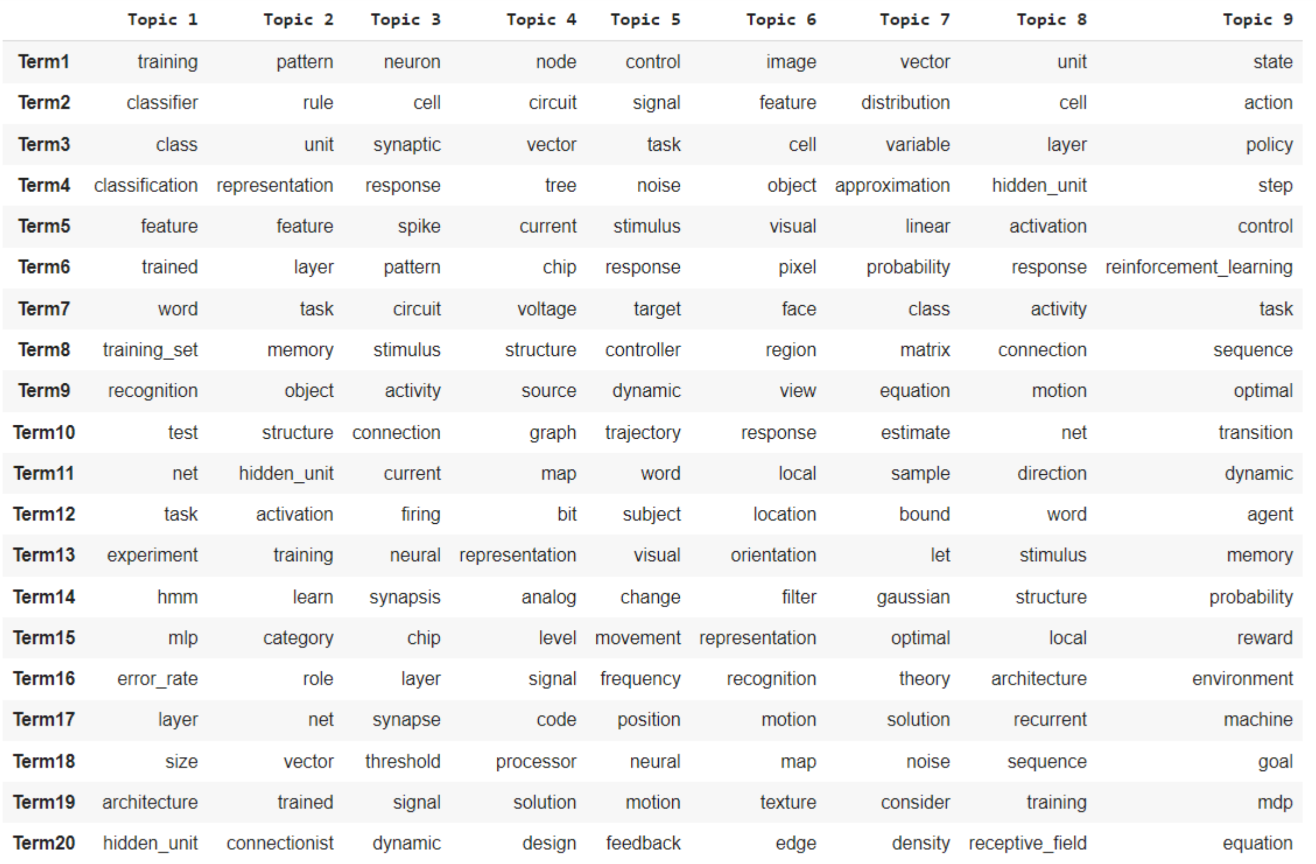
Fig. 9. Coherence score for NMF.

From fig 8 and 9 we can see that LDA has highest coherence and highest average coherence. From this we can conclude that LDA can perform well for any topics when compare to NMF. But if we take 9 topics for NMF it performs well and equal to the LDA model. Suppose if we have less data in each document in that situations NMF performs well when compares to LDA because it needs more data to predict correct topics.



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Term1	equation	unit	rate	signal	noise	bound	map	visual	training	state
Term2	dynamic	layer	gradient	frequency	distribution	class	subject	motion	hidden_unit	action
Term3	solution	pattern	convergence	channel	average	theorem	effect	response	task	policy
Term4	matrix	activation	step	filter	curve	probability	stimulus	direction	rule	step
Term5	neuron	representation	iteration	spike	theory	proof	study	receptive_field	net	reinforcement_learning
Term6	state	connection	update	rate	equation	loss	change	region	trained	control
Term7	energy	module	gradient_descent	temporal	optimal	hypothesis	development	orientation	architecture	optimal
Term8	attractor	structure	adaptive	response	variance	complexity	correlation	spatial	generalization	environment
Term9	constraint	architecture	optimization	noise	correlation	theory	activity	location	back_propagation	task
Term10	eq	activity	vector	event	generalization_error	defined	brain	center	hidden_layer	goal
Term11	fixed_point	connectionist	minimum	auditory	limit	assume	pattern	cell	learn	reward
Term12	phase	represent	change	sound	ensemble	distribution	similarity	field	target	td
Term13	stable	level	derivative	delay	size	concept	experiment	stimulus	hidden	agent
Term14	hopfield	type	constant	amplitude	eq	property	human	contrast	training_set	rl
Term15	rule	role	cost_function	detection	entropy	bounded	trial	local	table	reinforcement
Term16	stability	processing	optimal	source	teacher	linear	cue	velocity	learned	cost
Term17	equilibrium	local	line	stimulus	linear	define	theory	map	knowledge	trial
Term18	nonlinear	represented	initial	phase	student	xi	eye	edge	backpropagation	current
Term19	eigenvalue	part	local_minimum	component	effect	sample	rule	contour	epoch	call
Term20	defined	representing	technique	spike_train	stochastic	definition	response	stage	domain	exploration

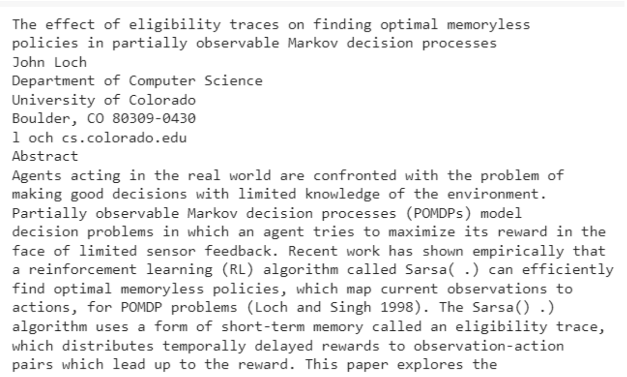
Fig. 10. Topic – word matrix for LDA Model.



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
Term1	training	pattern	neuron	node	control	image	vector	unit	state
Term2	classifier	rule	cell	circuit	signal	feature	distribution	cell	action
Term3	class	unit	synaptic	vector	task	cell	variable	layer	policy
Term4	classification	representation	response	tree	noise	object	approximation	hidden_unit	step
Term5	feature	feature	spike	current	stimulus	visual	linear	activation	control
Term6	trained	layer	pattern	chip	response	pixel	probability	response	reinforcement_learning
Term7	word	task	circuit	voltage	target	face	class	activity	task
Term8	training_set	memory	stimulus	structure	controller	region	matrix	connection	sequence
Term9	recognition	object	activity	source	dynamic	view	equation	motion	optimal
Term10	test	structure	connection	graph	trajectory	response	estimate	net	transition
Term11	net	hidden_unit	current	map	word	local	sample	direction	dynamic
Term12	task	activation	firing	bit	subject	location	bound	word	agent
Term13	experiment	training	neural	representation	visual	orientation	let	stimulus	memory
Term14	hmm	learn	synapsis	analog	change	filter	gaussian	structure	probability
Term15	mlp	category	chip	level	movement	representation	optimal	local	reward
Term16	error_rate	role	layer	signal	frequency	recognition	theory	architecture	environment
Term17	layer	net	synapse	code	position	motion	solution	recurrent	machine
Term18	size	vector	threshold	processor	neural	map	noise	sequence	goal
Term19	architecture	trained	signal	solution	motion	texture	consider	training	mdp
Term20	hidden_unit	connectionist	dynamic	design	feedback	edge	density	receptive_field	equation

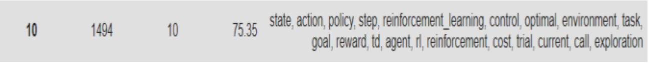
Fig. 11. Topic – word matrix for NMF model.

Example 1:



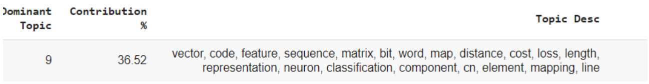
```
The effect of eligibility traces on finding optimal memoryless
policies in partially observable Markov decision processes
John Loch
Department of Computer Science
University of Colorado
Boulder, CO 80309-0430
l och cs.colorado.edu
Abstract
Agents acting in the real world are confronted with the problem of
making good decisions with limited knowledge of the environment.
Partially observable Markov decision processes (POMDPs) model
decision problems in which an agent tries to maximize its reward in the
face of limited sensor feedback. Recent work has shown empirically that
a reinforcement learning (RL) algorithm called Sarsa( .) can efficiently
find optimal memoryless policies, which map current observations to
actions, for POMDP problems (Loch and Singh 1998). The Sarsa() .)
algorithm uses a form of short-term memory called an eligibility trace,
which distributes temporally delayed rewards to observation-action
pairs which lead up to the reward. This paper explores the
```

Fig. 12. Sample document 1.



10	1494	10	75.35	state, action, policy, step, reinforcement_learning, control, optimal, environment, task, goal, reward, td, agent, rl, reinforcement, cost, trial, current, call, exploration

Fig. 13. Predicted topic number for given document using LDA.

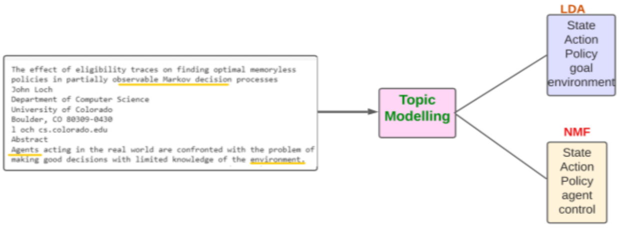


Dominant Topic	Contribution %	Topic Desc
9	36.52	vector, code, feature, sequence, matrix, bit, word, map, distance, cost, loss, length, representation, neuron, classification, component, cn, element, mapping, line

Fig. 14. Predicted topic number for given document using NMF.

Fig 13, fig 14, represents the dominant topic to that document , how much percent the document is related to that topic and the top 10 words related to that topic.

The Fig 12, document is similar to the domain of Reinforcement Learning by looking the document description. LDA predict the given document as topic 10 which are state, action, and agent. Which are exactly matching to the given document, coming to the NMF it predicts the given document as topic 9 and words for that topic are state, action, policy and step. NMF also correctly predicts the topic. From this we can conclude that LDA with optimal topic 22 and NMF with topics 9 can perform well for this dataset. Both models are producing similar and correct outputs.

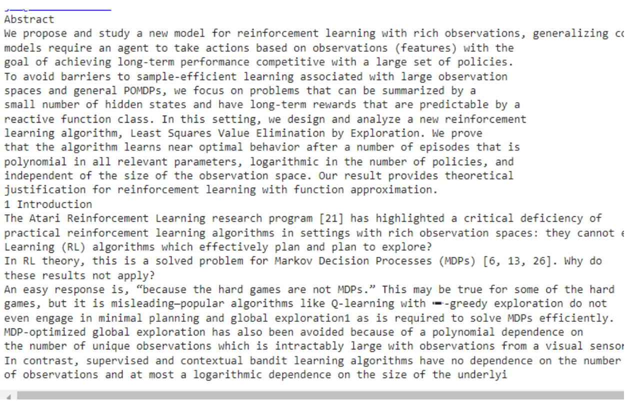


```
The effect of eligibility traces on finding optimal memoryless
policies in partially observable Markov decision processes
John Loch
Department of Computer Science
University of Colorado
Boulder, CO 80309-0430
l och cs.colorado.edu
Abstract
Agents acting in the real world are confronted with the problem of
making good decisions with limited knowledge of the environment.
```

Fig. 15. Diagrammatic representation of example 1

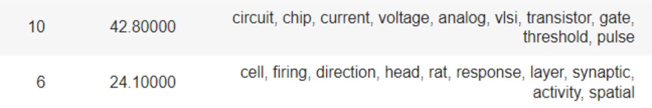
Fig 15, represents the how topic modelling algorithms classify documents into topics. First preprocessing is done after preprocessing only main words related to document will be there then topic model algorithms cluster the words which have closer relationship between the words. If any new document want to classify models will compare the words in the documents with different topics words weights and which topic has highest probability it will assign to that topic.

Example 2:



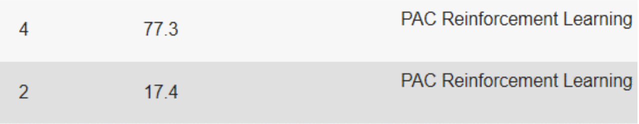
```
Abstract
We propose and study a new model for reinforcement learning with rich observations, generalizing cc
models require an agent to take actions based on observations (features) with the
goal of achieving long-term performance competitive with a large set of policies.
To avoid barriers to sample-efficient learning associated with large observation
spaces and general POMDPs, we focus on problems that can be summarized by a
small number of hidden states and have long-term rewards that are predictable by a
reactive function class. In this setting, we design and analyze a new reinforcement
learning algorithm, Least Squares Value Elimination by Exploration. We prove
that the algorithm learns near optimal behavior after a number of episodes that is
polynomial in all relevant parameters, logarithmic in the number of policies, and
independent of the size of the observation space. Our result provides theoretical
justification for reinforcement learning with function approximation.
1 Introduction
The Atari Reinforcement Learning research program [21] has highlighted a critical deficiency of
practical reinforcement learning algorithms in settings with rich observation spaces: they cannot e
Learning (RL) algorithms which effectively plan and plan to explore?
In RL theory, this is a solved problem for Markov Decision Processes (MDPs) [6, 13, 26]. Why do
these results not apply?
An easy response is, “because the hard games are not MDPs.” This may be true for some of the hard
games, but it is misleading—popular algorithms like Q-learning with ▬-greedy exploration do not
even engage in minimal planning and global exploration1 as is required to solve MDPs efficiently.
MDP-optimized global exploration has also been avoided because of a polynomial dependence on
the number of unique observations which is intractably large with observations from a visual sensor
In contrast, supervised and contextual bandit learning algorithms have no dependence on the number
of observations and at most a logarithmic dependence on the size of the underlyi
```

Fig. 16. Sample Document 2



10	42.80000	circuit, chip, current, voltage, analog, vlsi, transistor, gate, threshold, pulse
6	24.10000	cell, firing, direction, head, rat, response, layer, synaptic, activity, spatial

Fig. 17. Predicted topic number for given document using LDA.



4	77.3	PAC Reinforcement Learning
2	17.4	PAC Reinforcement Learning

Fig. 18. Predicted topic number for given document using NMF.

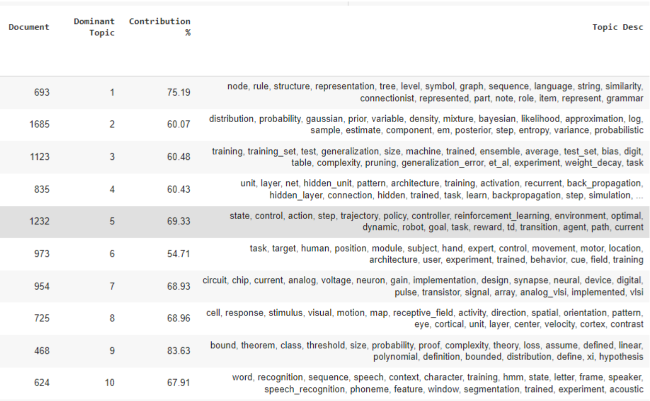
Fig 17, 18 shows the top two dominant topics for the given sample document 2 and its contribution to that topic.

The sample document 2 talks about the one of the type in machine learning method which is reinforcement learning. The words in the document similar to the topic 10 in LDA and topic 9 in NMF model. While coming to categorization LDA model classify as topic 10 which is similar to the words in document but NMF model predicts the topic as 4 which is completely not similar to that topic.

Here the above two examples are taken on similar topic which is reinforcement learning but different document , as we compare the results LDA predicts correctly in two cases but NMF predicts only once. For this research papers categorization LDA model can perform better than NMF. Finally from this comparison we conclude that if the document vocabulary size is less than NMF performs well and if document vocabulary is high then LDA performs well.

VI. RESULTS

The experiment was carried out on a sample of 1740 papers from a specific database containing realistic textual data and entire text in several study fields. Each paper consists of a title, author's name, abstract of the paper, introduction, relevant work, methods and material, results, conclusion and discussion, and references. Cleaning the text of letters and numbers, eliminating stop signs and special symbols, removing stop words, and re-configuring data for the representation of unstructured data in a bag of words were all done as part of the pre-processing activities. We have used two different yet widely used generative topic modeling approaches namely LDA and NMF to train the data. We further test our model by checking our respective model's performance on 4 different articles and mapping them to the respective dominant topic. We further manually evaluate which of these two models performs better with our database. On evaluating LDA performs better than NMF on our dataset mapping the articles to more accurate topics.



Document	Dominant Topic	Contribution %	Topic Desc
693	1	75.19	node, rule, structure, representation, tree, level, symbol, graph, sequence, language, string, similarity, connectionist, represented, part, note, role, item, represent, grammar
1685	2	60.07	distribution, probability, gaussian, prior, variable, density, mixture, bayesian, likelihood, approximation, log, sample, estimate, component, em, posterior, step, entropy, variance, probabilistic
1123	3	60.48	training, training_set, test, generalization, size, machine, trained, ensemble, average, test_set, bias, digit, table, complexity, pruning, generalization_error, et_al, experiment, weight_decay, task
835	4	60.43	unit, layer, net, hidden_unit, pattern, architecture, training, activation, recurrent, back_propagation, hidden_layer, connection, hidden, trained, task, learn, backpropagation, step, simulation, ...
1232	5	69.33	state, control, action, step, trajectory, policy, controller, reinforcement_learning, environment, optimal, dynamic, robot, goal, task, reward, td, transition, agent, path, current
973	6	54.71	task, target, human, position, module, subject, hand, expert, control, movement, motor, location, architecture, user, experiment, trained, behavior, cue, field, training
954	7	68.93	circuit, chip, current, analog, voltage, neuron, gain, implementation, design, synapse, neural, device, digital, pulse, transistor, signal, array, analog_vlsi, implemented, vlsi
725	8	68.96	cell, response, stimulus, visual, motion, map, receptive_field, activity, direction, spatial, orientation, pattern, eye, cortical, unit, layer, center, velocity, cortex, contrast
468	9	83.63	bound, theorem, class, threshold, size, probability, proof, complexity, theory, loss, assume, defined, linear, polynomial, definition, bounded, distribution, define, xi, hypothesis
624	10	67.91	word, recognition, sequence, speech, context, character, training, hmm, state, letter, frame, speaker, speech_recognition, phoneme, feature, window, segmentation, trained, experiment, acoustic

Fig. 19. Research Papers per Topic based on Dominance for10 topics for LDA

Fig. 19. Consists of the Dominant topic, Document, Contribution percentage, Topic description, and paper. The above table 2 tells that among 10 dominant topics which documents have the highest percent similarity to that topic, from this we can conclude that if any random document is similar to topic 10 then the document is similar domain to the dominant topic 10 paper, from this we can group the similar documents into one cluster so, that it can help in for searching similar domain papers.

VII. CONCLUSION

In this paper, we attempted to classify the obtained data according to its prominent topics using two topic modeling techniques (LDA and NMF) and compared the results. We provided a research papers classification to effectively perform paper categorization, which is necessary to give users a quick and efficient search for the research papers they want. The suggested method uses a bag of words for NMF and LDA aims to calculate the importance of each publication and uses the same technique to group papers with similar subjects. As a result, it will produce accurate classification results for users' relevant papers. Topic modeling methods, as we've observed, are simpler and more efficient in summarizing and also searching compared to other techniques. Despite the fact that there are various subject modeling methods accessible, LDA is the simplest and most efficient topic modelling technique.

REFERENCES

[1] Owa, D. (2021) Identification of Topics from Scientific Papers through Topic Modeling. Open Journal of Applied Sciences, 11, 541-548. doi: 10.4236/ojapps.2021.104038.

[2] Kim, SW., Gil, JM. Research paper classification systems based on TF-IDF and LDA schemes. Hum. Cent. Comput. Inf. Sci. 9, 30 (2019).

[3] Blei et al., Latent dirichlet allocation Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[4] Bravo-Alcobendas D, Sorzano COS (2009) Clustering of biomedical scientific papers. In: 2009 IEEE Int. symp. on intelligent signal processing. pp 205–209

[5] Mahesh Korlapati, Tejaswi Ravipati, Abhilash Kumar Jha, Kolla Bhanu Prakash (2019) Categorizing Research Papers By Topics Using Latent Dirichlet Allocation Model IJSTR VOLUME 8, ISSUE 12, DECEMBER 2019 ISSN 2277-8616.

[6] A. Ritter, S. Clark and O. Etzioni, "Named entity recognition in tweets: An experimental study", Proc. Conf. Empirical Methods Nat. lang. Process, pp. 1524-1534, 2011.

[7] Gui Y, Gao G, Li R, Yang X (2012) Hierarchical text classification for news articles based-on named entities. In: Proc. of int. conf. on advanced data mining and applications. pp 318–329.

[8] Taheriyan M (2011) Subject classification of research papers based on interrelationships analysis. In: ACM proc. of the 2011 workshop on knowledge discovery, modeling and simulation. pp 39–44

[9] Nanba H, Kando N, Okumura M (2011) Classification of research papers using citation links and citation types: towards automatic review article generation. Adv Classif Res Online 11(1):117–134

[10] Blei, D. (2012) Probabilistic Topic Models. Communications of the ACM, 55, 77-84.https://doi.org/10.1145/2133806.2133826

[11] Blei, D., NG, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3, 993-1022.

[12] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., et al. (2019) Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. Multimedia Tools and Applications, 78, 15169-15211.

[13] Belford, M., Namee, Mb., and Greene, D., (January 2018), “Stability of Topic Modeling via Matrix Factorization’, Expert Systems with Applications.

[14] Alghamdi R., Alfalqi K., (2015), “A Survey of Topic Modeling in Text Mining”, (IJACSA) International journal of Advanced Computer Science and Applications, 6(1).

[15] Barman P.C., Nadeem Iqbal, and Soo-Young Lee, (2006), “Non-negative Matrix Factorization Based Text Mining: Feature Extraction and Classification”, © Springer-Verlag Berlin Heidelberg, 703 – 712.