VOLUME V
ISBN No.: 978-81-953602-8-4
Computational Science

NALLAMUTHU GOUNDER MAHALINGAM COLLEGE

An Autonomous Institution, Affiliated to Bharathiar University, An ISO 9001:2015 Certified Institution,

Pollachi-642001

SUPPORTED BY

ABT Industries Ltd.,
Dairy Division

PROCEEDING

One day International Conference

EMERGING TRENDS IN SCIENCE AND TECHNOLOGY (ETIST-2021)

27th October 2021

Jointly Organized by

Department of Biological Science, Physical Science and Computational Science

NALLAMUTHU GOUNDER MAHALINGAM COLLEGE

An Autonomous Institution, Affiliated to Bharathiar University

An ISO 9001:2015 Certified Institution, Pollachi-642001.

Estd. 1957

Proceeding of the

One day International Conference on

EMERGING TRENDS IN SCIENCE AND TECHNOLOGY (ETIST-2021)

27th October 2021

Jointly Organized by

Department of Biological Science, Physical Science and Computational Science

ISBN No: 978-81-953602-8-4

978- 81- 953602- 8- 4

Nallamuthu Gounder Mahalingam College

An Autonomous Institution, Affiliated to Bharathiar University

An ISO 9001:2015 Certified Institution, 90 Palghat Road, Pollachi-642001.

www.ngmc.org

ABOUT THE INSTITUTION

A nations's growth is in proportion to education and intelligence spread among the masses. Having this idealistic vision, two great philanthropists late. S.P. Nallamuthu Gounder and Late. Arutchelver Padmabhushan Dr.N.Mahalingam formed an organization called Pollachi Kalvi Kazhagam, which started NGM College in 1957, to impart holistic education with an objective to cater to the higher educational needs of those who wish to aspire for excellence in knowledge and values. The College has achieved greater academic distinctions with the introduction of autonomous system from the academic year 1987-88. The college has been Re-Accredited by NAAC and it is ISO 9001 : 2015 Certified Institution. The total student strength is around 6000. Having celebrated its Diamond Jubilee in 2017, the college has blossomed into a premier Post-Graduate and Research Institution, offering 26 UG, 12 PG, 13 M.Phil and 10 Ph.D Programmes, apart from Diploma and Certificate Courses. The college has been ranked within Top 100 (72nd Rank) in India by NIRF 2021.

ABOUT CONFERENCE

The International conference on “Emerging Trends in Science and Technology (ETIST-2021)” is being jointly organized by Departments of Biological Science, Physical Science and Computational Science - Nallamuthu Gounder Mahalingam College, Pollachi along with ISTE, CSI, IETE, IEE & RIYASA LABS on 27th OCT 2021. The Conference will provide common platform for faculties, research scholars, industrialists to exchange and discus the innovative ideas and will promote to work in interdisciplinary mode.

EDITORIAL BOARD

 ISBN No.: 978-81-953602-8-4

LIST OF ARTICLES

A Brief Survey on Topic Modeling Techniques

T.Rajalakshmi [1] – V.Srividhya [2] – E.Ramadevi [3]

ABSTRACT: This survey aims to provide a comprehensive overview of the present research on Topic modeling and interesting issue in the text mining domains. The usual topic models contains word co-occurrences gather the hidden semantic structure from an amount of documents. A topic was made up of a group of words which appear together. It can combine words with similar meanings and distinguish between different definitions of the same phrase. Natural language processing (NLP) is used in text mining to enable machines to understand and process human language automatically. These algorithms support in the creation of new methods for searching, browsing, and summarizing massive text archives. The field of topic modeling can be divided into two groups, according to this paper. The first one explores to the topic model methods group, which includes four methods that can be used. This paper has been presenting a survey on readily accessible topic modelling techniques of Latent semantic analysis (LSA), Probabilistic latent semantic analysis (PLSA), Latent Dirichlet Allocation (LDA) and Non Negative Matrix Factorization (NMF).The survey highlights on classification of supervised and unsupervised algorithms. At last it will be concluded that the combination of text classifiers and topic models.

Keywords: TOPIC MODEL, LSA, PLSA, LDA, NMF**.**

1. INTRODUCTION

Topic modeling is an unsupervised machine learning method that can search a group of documents, recognize word and expression patterns within them, and automatically gather word groups and corresponding phrases. So, topic models are a concept which can be used to deal with documents that are combination of topics, with a subject being such a possibility allocation over terms [1]. In other words, the topic model is a document generative model. To

T.Rajalakshmi *[1], Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India.. E-mail: ramatj1994@gmail.com
V.Srividhya *[2], Department of Computer Science, NGM College, Coimbatore, Tamil Nadu, India. E-mail: vidhyavasu@gmail.com

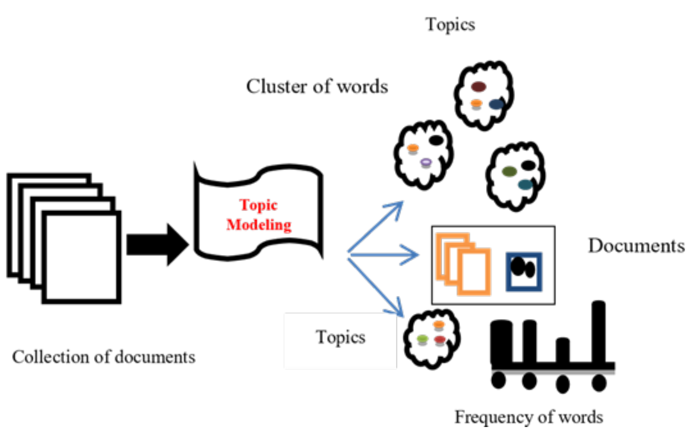
obtain topic models, various methods are used. It is the method most extensively used text mining technique in Natural Language Processing for extracting information from text documents. The algorithm is similar toward numerical data dimensionality reduction techniques. In Figure.1 it represents the basic methods of topic modeling.

It can be considered as being ability to extract desired characteristics from a bag of terms. This step was crucial as in NLP; every word in the quantity was recognized as a characteristic. Thus, attribute reduction allows focusing on the correct content but instead of wasting time going through all of the text in the data. Let us avoid mathematics for a better understanding of the concepts.

Topic modeling helps to develop techniques for organizing, comprehending, and summarizing vast amounts of textual data. It aids in:

- Categorizing latent topical variations that occur in the collection
- Annotating information with these subjects in opinion.
- To coordinate, check, and review documents, uses these annotations.

Figure 1: The basic methods of topic modeling



NLP is changing approach those who examines and interact with language-based data by training machines to make intellect of speech and text, and execute particular task like summarization, translation, extraction and classification. It finds topics by employing a probabilistic structure to gather subjects within the data based on the words found in the documents [2]. Topic modeling is adaptable method for organizing an unstructured collection of text documents. It can be used to help organize and understand large amounts of text data by automating processes of sifting through it. Once key topics have been identified, create a text document. This paper is organized as follows. Section II provides a literature review of the paper. Section III explains about the methods of topic modeling. Section IV explains the scope and challenges of topic modeling. Section V comparative study on two categories. Finally, it is concluded with text classifier and topic model in this paper.

2. LITERATURE REVIEW

***Scott Deerwester et.al,* (1990) [3]** introduced a novel technique for indexing and retrieving information automatically. The authors solved issues with accessible retrieval approaches for matching query terms to document words. The objective of LSA is to mine semantic relationships among terms in documents by analyzing correlation consequences with different values of truncated semantic space.

***Pooja Kherwa* (2017) [7]** proposed a technique for examining an element of text for arithmetical calculation among the documents. LSA utilized in various natural language processing research papers to discover associations of terms of semantic spaces, with consumer reservation, where LSA is used as a key performance for rising applications.

***Wood et.al.,* (2017) [8]** introduced a semi-supervised Latent Dirichlet Allocation (LDA) model, Source-LDA, which includes preceding information point the subject modeling procedure progress together the excellence of resultant topic labeling and topics. The author's proficient with incorporating obtainable labeled information sources representing recognized a topic model. These information resources were interpreted to allocation and utilized the hyper parameters of Dirichlet created distribution above words. Their approach guaranteed that the subject deduction procedure was reliable with obtainable knowledge, and concurrently, permits for detection of novel topics. The consequences demonstrate an enhanced subject creation and increased correctness in topic classification when evaluated to those achieved using assorted labeling approaches based off LDA.

***M. Shams and A. Baraani-Dastjerdi* (2017) [9]** discovered more accurate features integrates co-occurrence associations as previous field information into the Latent Dirichlet Allocation (LDA) subject form. In their technique, initially, beginning features were created based on LDA. Then, in a routine technique, earlier information was mined routinely as of co-incidence association's related features of applicable subjects. Lastly, the mined knowledge is integrated into the LDA form.

***Y. Liang, Y. Liu, C. Chen, and Z. Jiang* (2018) [10]** studied an interesting issue called Topic-Sensitive Content Extraction (TSCE). TSCE tries to reduce contents that are applicable toward instances subject features tinted by consumers from a particular text in a known text group. To deal with TSCE, authors proposed a novel mixture topic model which incorporates dissimilar structures in both context and subject space. It centers on recognizing contents linked an exacting subject feature as of every document. By forming ascent documents through phrase profiles for situation forming and by leveraging limited total differences among likelihood distributions above words in together context modeling and topic modeling, it has an improved captured the attributes of different language patterns.

***Yang et.al.*, (2018) [11]** presented a topic recognition is the technique that removes precious hot topics from media stream data. It is the tool to assist to resolve the crisis of excess information. The topic positive correctness of cluster technique is extremely small. The authors proposed single topic discovery technique based on Key Graph to progress the positive correctness, and took testing evaluated with baseline technique on corpus marked by graduate students. In the consequence, the positive correctness of Key Graph technique reaches 88.48 percent with huge development. The result established the effectiveness of their proposed technique.

Hoffman (2017) [12] PLSI may be a statistical model proposed by for comparing two modes and evaluating data co-occurrence. In accordance with this model, each word in an extremely document is sampled from a mix of multinomial distributions that can be interpreted as topics, and proportions resembling mixture weights are sampled from a separate multinomial distribution for each document.

***Feng Xue et.al.*, (2020) [13]** proposed a knowledge priors- and max-margin-based topic model for multi-modal social event analysis, called the KGE-MMSLDA, in which characteristic demonstration and knowledge priors are together learned. Their model has three major merits above the present techniques: (1) incorporates information from outside facts based into a combined subject model in which the multi-modal information and max-margin classifier are developed to raise the amount of occurrence images obtained. (2) The reduced information priors from over 74,000 web documents. (3) A high-dimension multi-modal dataset was composed and has been released publicly for event topic mining and classification research.

3. METHODS OF TOPIC MODELING

This section provides a description of three major methods of topic modeling as follows:

A. Latent Semantic Analysis (LSA)

Latent semantic indexing (also known as Latent Semantic Analysis) is a method for examining a collection of documents to find statistical co-occurrences of terms that appear together, and then collects data about the topics of these words and documents [3] [4]. By statistically analyzing the terms that cooccur with it in a text, LSI can predict which significance a word represents. Foltz, Peter [5] examining the method of relationships among a group of documents and the words they enclosed by creating a position of models significant to the documents and terms is recognized as latent semantic analysis (LSA).

LSA believed that words with related significances will become visible in comparable text. The SVD is also used in Latent semantic analysis (LSA) to reorganize the data. SVD is a method that obtainable a matrix to regulate additionally calculates all vector space decreases Furthermore, the damnation will be intended in vector space and prearranged from most to slightest important.

A huge portion of text is transformed into a matrix enclosing word calculations for each document (instances characterize distinctive words and features represent every document), and an arithmetical method recognized as the SVD is used to decrease amount of instances while preserving parallel structure of features. The cosine of the position among the 2 vectors created by every two features is then used to evaluate documents. Standards closer to one point out documents that are parallel, while values closer to zero indicate documents that are different [12].

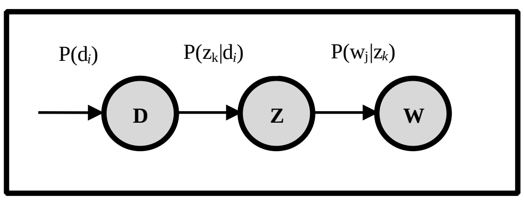
B. Probabilistic Latent Semantic Analysis (PLSA)

Liu [6] described the correlated topic model and latent dirichlet allocation, which were originally designed for statistical text modeling of large document collections, have recently become an dynamic research subject for annotation and multimedia demonstration in together computer visualization and model recognition. Another benefit of such models is that they automatically learn topics without the need for labeled training data. However, the success of these models is based on a flawed assumption. An arithmetical latent aspect model for co-occurrence data, probabilistic latent semantic analysis (PLSA) correlates an unobserved class attribute with every examination, an examination being the occurrence of a word in an exacting text.

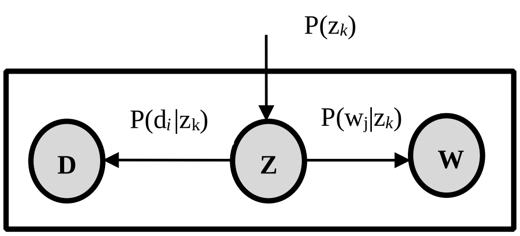
In Figure.2 it shows the graphical representation of this approach is an improvement on the PLSA, as well as asymmetric and symmetric formulations.

Figure.2: The graphical representation of PLSA

(a) Asymmetric Representation



(b) Symmetric Representation



The PLSA method was created to improve the LSA approach as well as to handle several additional difficulties that LSA was unable to handle. Many real-world applications of PLSA contain been valuable, containing recommender systems of computer vision. PLSA, on the other hand, endures from overfitting problems because the amount of parameters expands linearly with the amount of documents.

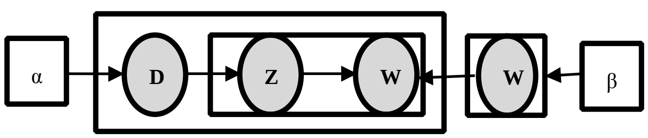
C. **Latent Dirichlet Allocation (LDA)**

The statistical and graphical model of Latent Dirichlet Allocation is used to discover associations among multiple documents in a corpus. The maximum likelihood estimation from the entire amount of text is obtained using the Variation Exception Maximization (VEM) algorithm. This method is typically solved by selecting the top tiny words from a bag of words. The sentence, however, is devoid of semantics. Each text represented by a probabilistic distribution of topics, and each topic described by a probabilistic distribution of terms, under this model. As a result, can get a much better understanding of how the topics are linked. The most widely used topic modeling method is LDA, which is an unsupervised generative probabilistic method for modeling a corpus. Each document is assumed to be interpreted as a probabilistic distribution over latent topic in LDA. Documents in LDA have many subjects. To delete punctuation and stop words during text pre-processing (such as, “if” “the” or “on” which contain little topical content). As a result, each text is interpreted as a series of corpus-wide topics.

i. Retrieval’s accuracy
ii. Inferring hidden topic structure

The literature on the application of subject models was discovered and examined in depth. In figure.3 it shows the graphical representation of LDA. According to the kinds of forms and the similarity among the theory of document-topic word categorized the related studies and provided an outlook on the use of topic models for developing various applications.

Figure 3: The graphical methods of LDA



D. Non Negative Matrix Factorization (NMF)

NMF technique ensures the non-negative fundamentals of the factorized matrices. Regard as the document-term matrix attained after extracting the stop words from a corpus. The term-topic matrix and therefore the topic-document matrix are two matrices which will be factored out of the matrix. Matrix factorization is accomplished employing a range of optimization models. Hierarchical Alternating Least Square be earlier and improved thanks for performing NMF. Here the factorization occurs by updating one column at a time while keeping the opposite columns as constant.

4. TOPIC MODELING APPLICATIONS

Topic modeling can be used in various applications of research in the last two decades since its inception such as Scientific Research, Bioinformatics, Social Network, and Software Engineering and in many other domains. TM methods can be supervised, unsupervised, or semi-supervised, and they will use unstructured data. They can also be used in various fields, including health, agriculture, education, e-commerce, social network opinion analysis, and transportation. In a set of text, such as documents, short text, chats, Twitter and Facebook messages, user comments on news sites, blogs, and emails, TM can be used to discover latent abstract topics.

5. SCOPE AND CHALLENGES IN TOPIC MODELING

Topic models have proved to be an effective technique for quantitative text analysis. Subject models can be much more useful than simple word searches or dictionary-based techniques, depending on the requirements. Topic models provide the best results when used on data that is not short, such as tweets, and has a uniform composition. Similarly, topic models have a several valid limitations.

- Topic modeling can be used to obtain semantic relationships between words in graph-based models.
- It can be used to quickly determine the topic of a paper or book using text summarization.
- It can be used to address applicant prejudice in exam assessments. It also makes it easy for users to get their results and saves them a lot of time.
- It will improve customer service by recognizing and adapting to the keyword the customer is searching for. Customers will trust you more because you provided them with the assist they desirable at the appropriate time and not including causing them any inconvenience. As a result, customer loyalty rises quickly, and the company's reputation grows as well.

The implementation of topic modeling raises numerous research challenges. One key dispute in topic modeling is to expand a quick explanation for indexing large dimensional data, which is an essential to construct high scale application data.

- Labeling a multinomial topic model
- Handling huge amount of documents
- Building huge amount of topics
- Scalability
- Difference of the illustration of a topic model and a label.
- Accurately interpreting meaning of each topic.

6. COMPARISON ANALYSIS

This paper aims to gather and consider papers that deal with Topic modeling methods. The aim is not to assume a constraint reviews, but quite to provide a broad state-of-the-art view on these related fields. Several previous methods have been projected to assist feature reductions, classification, which has mentioned in a body of literature that is spread over a wide variety of applications.

Table 1: SUMMARY TABLE FOR SURVEY OF TOPIC MODELING BASED ON FEATURE EXTRACTION AND CLASSIFICATION TECHNIQUES

S_N0	Reference	ALGORITHM	KEY-IDEA	TECHNIQUES	PERFORMANCE
1	**[8]**	Source- Latent Dirichlet allocation (LDA) algorithm.	To progress equally excellence of the resultant topics and of the topic labeling.	Knowledge source selection and Bijective Mapping.	25.7 percent of accuracy and confusion value of 1119.9.
2	**[9]**	Enriched LDA (ELDA), Hub and Bridge Extraction.	Topic modeling, Latent Dirichlet Allocation (LDA), Sentiment analysis, Aspect extraction, and Co-occurrence relations.	Aspect extraction techniques.	ELDA technique is around 10 percent and 18 percent improved than the LTM technique.
3	**[10]**	Generalized ExpectationMaximization (GEM) algorithm.	Probabilistic topic modeling, Topicsensitive content, Topic network and Semisupervised.	Three-stage framework and Context modeling with biased topic network.	The consequences about 20 percent and 10 percent superior than the finest consequences of normal recall and regular precision reported from LDA and PLSA correspondingly.
4	**[14]**	The Learning Algorithm for Regularized Non-negative Matrix Factorization Topic model (TRNMF)	Topic model, short texts, document correlation, word embedding, nonnegative matrix factorization and regularization.	Feature learning methods and Graph-based word embedding methods.	Average classification accuracy on three datasets (TMNews - 63.7%, Snippet - 65.7% and Twitter - 76.0%) with different number of topics.

5	**[15]**	Machine learning algorithms.	Text Mining Algorithms and Challenging Topics and Learning Analytics.	Deep Learning (DL) techniques.	The 90 percent accuracy of educational tasks topic labeling.

7. CONCLUSION

This survey article discussed two categories that can be classified as topic modeling. The general concept of four topic modeling methods, covering Latent semantic analysis (LSA), Probabilistic latent semantic analysis (PLSA), Latent Dirichlet allocation (LDA), and Non-negative Matrix Factorization (NMF) model, was discussed in the first category. It also discussed how these four methods differed in terms of qualities, limits, and theoretical bases. The paper does not go into great depth about each of these techniques. It simply gives a highlevel overview of these issues and how they relate to topic modeling. In the second category, paper has discussed the algorithm between supervised and unsupervised methods of topic models. It is clear from the discussion that no topic model or classification model can be considered a universal model for any application. Different classification algorithms perform differently depending on the dataset.

REFERENCES

[1] Blei, D.M., and Lafferty, J. D., “Dynamic Topic Models”, Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[2] Steyvers, M., and Griffiths, T., “Probabilistic topic models”. In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (eds), Latent Semantic Analysis: A Road to Meaning. Laurence Erlbaum, 2007.

[3] Liu, S., Xia, C., and Jiang, X., “Efficient Probabilistic Latent
Semantic Analysis with Sparsity Control”, IEEE International Conference on Data Mining, 2010, 905-910.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K.Landauer and
R. Harshman, “Indexing by latent semantic analysis,” Journal of the American Society for Information Science, vol. 41, no. 6, pp. 391-407, 1990.

[5] Foltz, Peter, “Latent Semantic Analysis for Text-Based Research”, Behavior Research Methods. 28. 197-202, 1996.

[6] Liu, S., Xia, C., and Jiang, X., “Efficient Probabilistic Latent
Semantic Analysis with Sparsity Control”, IEEE International Conference on Data Mining, 2010, 905-910.

[7] Pooja Kherwa, and Poonam Bansal,-Latent Semantic Analysis: An Approach toUnderstand Semantic of Text, International Conference International Conference on Current Trends in Computer, Electrical, Electronics and Communication (2017).

[8] J. Wood, P. Tan, W. Wang, and C. Arnold, ``Source-LDA: Enhancing probabilistic topic models using prior knowledge sources," in Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE), Apr. 2017, pp. 411-422.

[9] M. Shams and A. Baraani-Dastjerdi, ``Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word cooccurrence analysis for aspect extraction," Expert Syst. Appl., vol. 80, pp. 136-146, Sep. 2017.

[10] Y. Liang, Y. Liu, C. Chen, and Z. Jiang, ``Extracting topicsensitive content from textual documentsA hybrid topic model approach," Eng. Appl. Artif. Intell., vol. 70, pp. 81-91, Apr. 2018.

[11] S. Yang, Q. Sun, H. Zhou, Z. Gong, Y. Zhou, and J. Huang, ``A topic detection method based on KeyGraph and community partition," in Proc. Int. Conf. Comput. Artif. Intell. ICCAI, 2018, pp. 30-34.

[12] T. Hofmann, “Probabilistic Latent Semantic Indexing,” ACM SIGIR Forum, vol. 51, no. 2, p. 8, 2017.

[13] Thomas Hofmann, “Probabilistic Latent Semantic Indexing”, ACM SIGIR Forum 51(2):211-218, August 2017.

[14] Feng Xue, Richang Hong, Xiangnan He, Jianwei Wang, Shengsheng Qian, and Changsheng Xu, “Knowledge-Based Topic Model for Multi-Modal Social Event Analysis”, IEEE Transactions on Multimedia, VOL. 22, NO. 8, AUGUST 2020.

[15] Feng Yi, Bo Jiang, and Jianjun Wu, “Topic Modeling for Short Texts via Word Embedding and Document Correlation”, IEEE Access, Feburary 2020.

[16] Lourdes Araujo, Fernando López-Ostenero, Juan Martínez-Romo and Laura Plaza, “Deep-Learning Approach to Educational Text Mining and Application to the Analysis of Topics' Difficulty”, IEEE Access, December 2020.