Acta Scientific COMPUTER SCIENCES

Volume 5 Issue 9 September 2023

Research Article

Evaluating Text Preprocessing Methods for Discovering Quality Topics to Improve the Information Retrieval Mechanism

Lakshmi Sonkusale[1], Krishna Kumar Chaturvedi[2]*, Anu Sharma[2], Shashi Bhushan Lal[2], Mohammad Samir Farooqi[3], Achal Lama[4], Dwijesh Chandra Mishra[4], Pratibha Joshi[5], Murari Kumar[1]

[1]Ph.D. Scholar, The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, India

[2]Principal Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

[3]Senior Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

[4]Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

[5]Scientist, ICAR-Indian Agricultural Research Institute, New Delhi, India

***Corresponding Author:** Krishna Kumar Chaturvedi, Principal Scientist, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India.

Received: July 24, 2023

Published: August 14, 2023

© All rights are reserved by **Krishna Kumar Chaturvedi., *et al.***

Abstract

Topic discovery is the innovation towards extracting the underlying semantic structure from large collection of unstructured text. It is a convenient way to analyze unclassified text into topic clusters that can be utilized in classification of documents. A topic contains a set of words that frequently occurs together and defines the complete text into specific category. Topic discovery can group words with similar meaning and distinguish between uses of words with multiple meaning. It is an important and challenging task useful in information retrieval process. This paper discusses different preprocessing methods of text mining by using Latent Dirichlet Allocation (LDA) in determining number of topics. This will help in developing new computational methods to identify topics from text dataset. The LDA is a statistical modelling approach to analyse unclassified text into useful topics. In this study, the effect of text preprocessing methods on collected research articles for obtaining quality topics by applying grid search method for hyperparameters optimization are explored and evaluated using coherence score and topic score. The study suggests that preprocessing affects the number of topics and quality of these topics. The findings of the study will help in enhancing the information retrieval mechanism based of the identified topics and also useful in recommending related research articles to the researchers.

Keywords: Topic Model; Hyperparameters; Topic Discovery; Latent Dirichlet Allocation (LDA); Grid Search

Introduction

The collection and generation of data on various aspects is continuously carried out by researchers from experimental labs, fields, surveys, questionnaires, literature, and other sources. The researchers share this data with the research community in the form of research publications, reports, data, etc., which are stored in publication repositories. To obtain relevant knowledge and useful insights from text data, machine learning techniques need to be used to mine these repositories. However, the unstructured nature of text data makes this process cumbersome. Unstructured textual data is generated on a daily basis in many areas such as social media, news articles, performance reports, books, literature, etc., and the rapid growth of digitization poses many challenges to researchers in terms of storage, access, and use of important and useful information.

In machine learning and natural language processing, a topic model is a type of statistical and machine learning technique to discover the "topics" from the collection of documents. Topic modelling is one of the widely used machine learning techniques to find semantic relationships from text documents [1,10,14]. Since last decade, topic modelling has become popular among the users of machine learning and natural language processing community for handling large amount of unstructured text and annotating these texts with suitable themes, called as topic clusters [11]. It involves the collection of research articles or documents, identifying features as words and phrases, and then generating word clusters [15]. It is a technique for discovering semantic relationships in these documents. The results of the topic modelling can be utilized for many aspects of text mining including text summarization, sentiment analysis, document classification, dataset exploration, and information retrieval. Different methods for topic modelling: Latent Semantic Analysis (LSA) or Latent Semantic Indexing (LSI) [5], Probabilistic Latent Semantic Analysis (pLSA) [7], and Latent Dirichlet Allocation (LDA) [4].

LDA is a most popular in topic modelling field [4]. This is unsupervised machine learning algorithm that identifies latent topics from collection of text documents. It has potential to generate topics from text with respect to associated subject domain. It has three hyperparameters namely *K* (number of topics), α (document-topic density) and β (topic-word density). These hyperparameters play a major role in determining the topic clusters from the given data. Along with, the optimal combination of hyperparameters helps in determining good quality of topic clusters. Currently, there is hardly any approach available that specifies and defines the optimal number of topics [2].

Topic discovery in text datasets is a challenging endeavour. By the grid search hyperparameters can be obtained but it requires a scale within which the topics can be generated. The best combination of hyperparameters will help in discovering the topics in given dataset. These topics will help individuals to automatically organize, understand, search and summarize large volume of text documents into interpretable categories [9]. Selection of hyperparameters can be affect the model performance as well as topic characteristic. A study was conducted to investigate the suitability of using LDA for topic modelling with text representation techniques such as Bag-of-Words (BOW) and Term Frequency - Inverse Document Frequency (TF-IDF). According to the study's findings, TF-IDF outperformed BOW in performance measures (coherence score and perplexity). Additionally, TF-IDF has been employed in applying LDA for topic modelling in various fields such as Social media, bioinformatics, and communication research [8,12,17,20].

The paper aims to evaluate text preprocessing methods by using grid search method for hyperparameters optimization in obtaining the quality topics. Three text preprocessing methods were applied on published research articles related to the Crop Science domain for evaluation. In addition, a quality topics were also determined based on optimum combination of hyperparameters of LDA based modelling.

This paper is organized into five sections. Section 2 describes LDA technique and section 3 mentions description of data and methods used in the study. Section 4 discusses results and finally concludes the study in conclusion section. The present paper also mentions the limitation and future scope.

Materials and Methods

This study was conducted to develop topic models in Crop Science domain to find topics in research articles with three different preprocessing methods. Natural language toolkit (NLTK) was used for tokenization, lemmatization and stop words removal as a part of preprocessing. Three combinations of preprocessing have been used in the study. A robust and scalable open source library Gensim, used to build dictionary, model training and model development. Google Colab used for web scraping of research articles using beautiful soup technology. The entire methodology is depicted in figure 1 shows the steps to conduct the study. The evaluation has been done by using performance metrics.

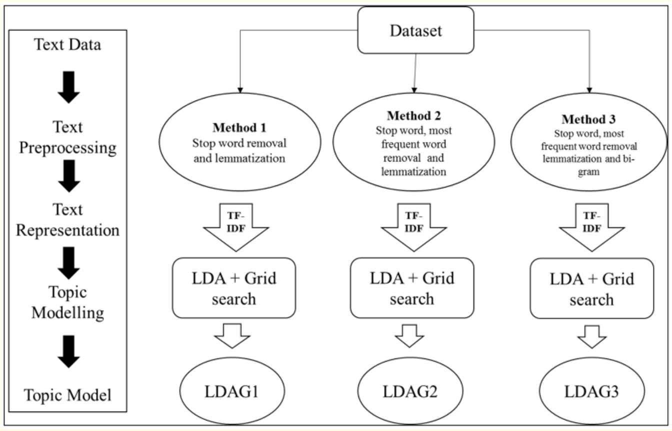


Figure 1: Workflow of topic discovery.

Dataset

The data were obtained from the KRISHI Portal's publication repository (krishi.icar.gov.in). The Indian Council of Agricultural Research (ICAR) has taken the initiative to provide its knowledge resources to all stakeholders at one place. The portal is being created as a centralized data repository system of ICAR consisting of Technology, Data generated through Experiments/Surveys/Observational studies, Geo-spatial data, Publications, Learning Resources etc. The data were collected through web scraping from publication archive of the Krishi portal (krishi.icar.gov.in) by custom searches covering the period from 2000 to 2022. In figure 2, the flow of the data gathering procedure is succinctly described. The dataset contains 2371 research articles, consisting of titles and abstracts.

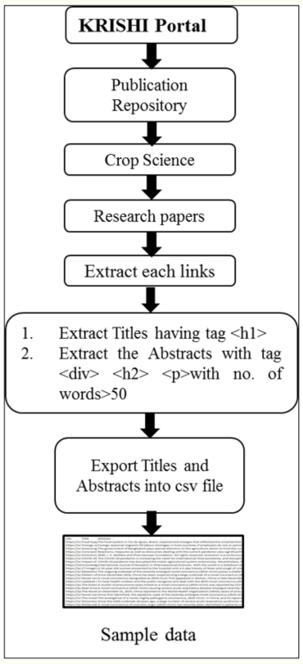

KRISHI Portal
Publication Repository
Crop Science
Research papers
Extract each links
1. Extract Titles having tag <h1>
2. Extract the Abstracts with tag <div> <h2> <p>with no. of words>50
Export Titles and Abstracts into csv file
Sample data

Figure 2: Flow of text data extraction from KRISHI.

Preprocessing methods and TF-IDF

In order to examine the effects of preprocessing on dataset and topics, three preprocessing methods were prepared and applied them sequentially on the text dataset. In text analysis, features are crucial since they influence the model's performance. Preprocessing methods were used to extract the key elements from the text, after performing the preprocessing, the dataset size was reduced. Dictionary was constructed only with distinct features, each of these features is represented by unique id. While using dictionary, user may identify unique feature in a dataset.

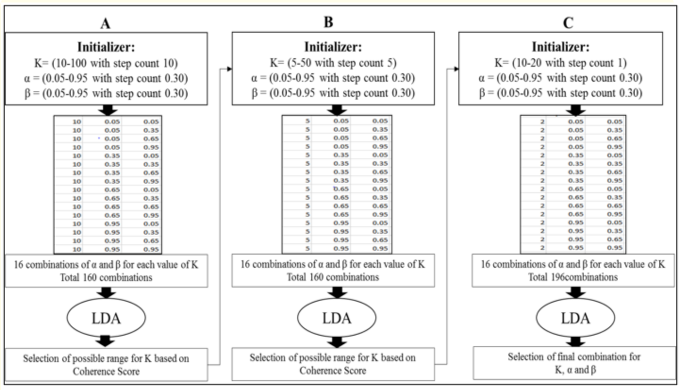
The standard process of preprocessing involves punctuation, special characters and language specific documents, tokenization, stop words removal, lemmatization etc. At the first, textual data were cleaned through removal of punctuation, digits, lower casing and the deletion of any word composed of a single character and tokenization. In this study, three methods of preprocessing have considered and mentioned below (Method 1, Method 2 and Method 3). In which, lemmatization is a process to find root word from the various forms of the words like a word experience considered for the words experience, disrupt from disruption, disrupted and disrupting, and many more. The bi-gram used to convert two continuous and meaningful words as a single word like supply and chain as supply-chain, small and scale as small-scale etc. After preprocessing, a vocabulary was prepared in all the three methods and each term of the vocabulary was assigned with unique number. The lemmatization is a process to find root word from the various forms of the words like a word experience considered for the words experience, disrupt from disruption, disrupted and disrupting, and many more. The bi-gram used to convert two continuous and meaningful words as a single word like supply and chain as supply-chain, small and scale as small-scale etc. After preprocessing, a vocabulary was prepared in all the three methods and each term of the vocabulary was assigned with unique number.

- Method 1: Stop words removal + Lemmatization
- Method 2: Stop words removal + Remove most frequent words + Lemmatization
- Method 3: Stop words removal + Remove most frequent words + Lemmatization + Bi-gram

To understand the text, it is essentially required to convert the text into numeric form using meaningful representation. The conversion of features into values for model development, TF-IDF was used. The measurement TF-IDF is a product of TF and IDF to neutralize the length of the document as well as number of documents and it is an easy measure to determine the importance of a term in the set of documents [13].

Grid search for LDA hyperparameters optimization

The ideal combination of the hyperparameters (K, α and β) are necessary for the LDA. The coherence score will be used to determine the best combination of these parameters, and used to assess the model too. The interaction of these variables influences how interpretable the identified topics [18]. This method needs a potential set (Potential set: the range of K in which the topics can be discovered) of hyperparameters, used random search to find a potential range of K, which was 10–20 in Crop Science dataset. For obtaining the potential set for K, vary the values of K from 10 to 100 with step count 10 and later vary the values from 5 to 50 with step count 5. The selected values for α that range from 0.05 to 0.95 with a step of 0.30 and the same ranges are employed for β. The optimal combination of these characteristics was chosen using the grid search method and assessed using the coherence score. The complete process of hyperparameters optimization is shown in figure 3. Finally, the LDA models were developed namely LDAG1, LDAG2 and LDAG3 with the optimized hyperparameters using three preprocessing methods respectively.


A
B
C
Initializer:
K= (10-100 with step count 10)
α = (0.05-0.95 with step count 0.30)
β = (0.05-0.95 with step count 0.30)
Initializer:
K= (5-50 with step count 5)
α = (0.05-0.95 with step count 0.30)
β = (0.05-0.95 with step count 0.30)
Initializer:
K= (10-20 with step count 1)
α = (0.05-0.95 with step count 0.30)
β = (0.05-0.95 with step count 0.30)
16 combinations of α and β for each value of K
Total 160 combinations
16 combinations of α and β for each value of K
Total 160 combinations
16 combinations of α and β for each value of K
Total 196combinations
LDA
LDA
LDA
Selection of possible range for K based on
Coherence Score
Selection of possible range for K based on
Coherence Score
Selection of final combination for
K, α and β

10	0.05	0.05
10	0.05	0.35
10	0.05	0.65
10	0.05	0.95
10	0.35	0.05
10	0.35	0.35
10	0.35	0.65
10	0.35	0.95
10	0.65	0.05
10	0.65	0.35
10	0.65	0.65
10	0.65	0.95
10	0.95	0.05
10	0.95	0.35
10	0.95	0.65
10	0.95	0.95

5	0.05	0.05
5	0.05	0.35
5	0.05	0.65
5	0.05	0.95
5	0.35	0.05
5	0.35	0.35
5	0.35	0.65
5	0.35	0.95
5	0.65	0.05
5	0.65	0.35
5	0.65	0.65
5	0.65	0.95
5	0.95	0.05
5	0.95	0.35
5	0.95	0.65
5	0.95	0.95

2	0.05	0.05
2	0.05	0.35
2	0.05	0.65
2	0.05	0.95
2	0.35	0.05
2	0.35	0.35
2	0.35	0.65
2	0.35	0.95
2	0.65	0.05
2	0.65	0.35
2	0.65	0.65
2	0.65	0.95
2	0.95	0.05
2	0.95	0.35
2	0.95	0.65
2	0.95	0.95

Figure 3: Process of Hyperparameter Optimization.

Evaluation metrics

To evaluate the performance of the models coherence score was used. CV is the popular coherence metrics and based on co-occurrences of words by calculating the normalized pointwise mutual information (NPMI) and cosine similarity [16,19]. As per literature, it is observed that CV score is good in human interpretation of topics [6,19] and it positively supports lemmatization [3]. In addition, topic coherence metric was used for assessing how well a topic is supported by a text set, it measures the degree of semantic similarity between high-scoring words within the same topic. It is shown as a number that represents the topics interpretability and is used to assess the topics quality. Topic coherence calculation process for a topic is shown in figure 4. This process consists of four step: Segmentation, Probability Calculation, Confirmation Measure and Aggregation. Through this process, topic coherence was achieved for each of the topics with the top ten features.

Results and Discussion

This study was carried out to compare the preprocessing methods in identifying the no. of topics in the given dataset in section 3.1. Alongwith, comparative study among grid search based topic models have been done in section 3.2 and section 3.3 deals with the

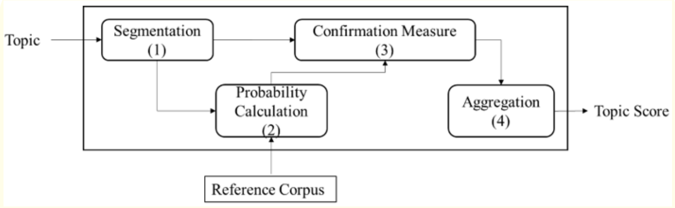


Figure 4: Topic coherence calculation process.

topic quality based on identified no. of topics and also compare the topics quality based on topic coherence.

Preprocessing effect and K

Table 1 provides the characteristics discovered both before and after preprocessing. After applying these preprocessing methods on the dataset, achieved three different results: total features, unique feature and *K* using grid search. Table 1 depicts the value of *K* obtained using three different preprocessing methods in Crop Science dataset. Number of topics for each preprocessing methods are 10, 11 and 10 respectively.

Here, the dataset was used for applying three preprocessing methods and analysed the effects on the dataset and number of topic. Different preprocessing methods reduces the size of dataset and also affected the number of topics in the dataset (Table 1).

Table 1: Identification of *K* using grid search approach in dataset.

Preprocessing	Features		Vocabulary size	K
	Before preprocessing	After preprocessing		
Method 1	531877	316024	17053	10
Method 2	531877	304320	17044	11
Method 3	531877	288395	17758	10

Grid search based topic models

The preprocessed text has been utilised to find the three LDA hyperparameters *K*, α and β. Moreover, applied LDA model on three preprocessed text to developed three different model of the LDA based on grid search, named LDAG1, LDAG2, and LDAG3.

Table 2 describe the comparative study among grid search based topic models in the Crop Science domain with the respective preprocessing methods. This comparison was made based on performance metrics i.e. coherence score to compare the performance of developed topic models. Table 2 contains three topic model (LDAG1, LDAG2 and LDAG3) based on grid search method for obtaining topics. In topic model 1, 2 and 3 indicate the preprocessing way through which dataset was preprocessed and then train the base model on optimized topics using grid search.

It is clear that the LDAG3 (Table 2) having highest coherence score (0.626) among three. On the basis of coherence score, LDAG3 performed better than remaining two. LDAG1 and LDAG3 have similar topics as well as same values for alpha and beta but only preprocessing methods are different. It shows the preprocessing may affect the model performance.

Table 2: Comparison based on three preprocessing methods.

Topic Model	Optimized Topics (K)	α	β	CV
LDAG1	10	0.05	0.65	0.448
LDAG2	11	0.35	0.65	0.607
LDAG3	10	0.05	0.65	0.626

Topic quality assessment based on Topic coherence

The topic models of LDA i.e. LDAG1, LDAG2 and LDAG3 are used to generate topic clusters of the dataset (crop science), after selecting them on the basis of coherence scores. The topic cluster of each created model of LDA in given dataset shown in Table 3, 4 and 5. Each topic represent the top ten features of the topic and topic coherence metric assesses is how well a topic is supported by a text set. It is shown as a number that represents the topics interpretability and is used to assess the topics quality.

The threshold value for topic coherence is set to 0.45 to assess the quality of topics. The LDAG1 (Table 3) have 60% of topics with higher score of topic coherence than the threshold value and LDAG2 (Table 4) have 36.36% topics with higher value of topic coherence while LDAG3 (Table 5) with only 50%. This shows that the LDAG1 approach is significantly good over the LDA1 and LDA3 based on topic coherence values.

Although the LDAG3 topic model performed well, it did not generate high-quality topics based on coherence. Conversely, LDAG1 did not perform as well, but it produced better topics in terms of topic coherence. These findings demonstrate that preprocessing techniques have an impact on both the model's performance and the quality of the topics it generates.

Table 3: Generated topics of LDAG1 topic model.

S. No.	Features	Topic coherence
1	Aphid tocopherol sbs biochar dhan irgc msp phytic emission sls	0.60
2	Bph mirna planthopper pantnagar extractant ammi osmotic groundwater smd arima	0.57
3	Spike drip woman dormancy fertigation lea inm pearl machine capsule	0.56
4	Marker line gene trait genetic genotype isolate cluster resistance analysis	0.50
5	Dominance module bidi rustica hp catch dap quizalofop capsule dnak	0.47
6	Rust ann lysine tryptophan ncrna reflectance azolla ll neural inversion	0.47
7	Endophyte solanacearum cajan cajanus ralstonia rme bpr spv dss xenorhabdus	0.44
8	Tillage cropping microbiology humanity pant department corn science kanchan sri	0.42
9	Soil rice tobacco stress crop yield system growth leaf increase	0.42
10	University mealybug dsr chew basic psb tribal solenopsis cassava pbnd	0.41

Table 4: Generated topics of LDAG2 topic model.

S. No.	Features	Topic coherence
1	Eb aggregate dsr partellus scylv μg false kumar photoperiod	0.60
2	Corn sweet sri vermicompost psii groundwater chloride smw mpa msp	0.59
3	Qpcr fortify snp constituent pbnd grassland phenolic tannin bag antioxidative	0.52
4	Biochar inm lysine tribal shatter thermotolerant altitude biofortified ashwagandha biofortifie	0.52
5	Export pantnagar pant starch mobile market phytic ann smut supply	0.43
6	Phytoplasma sbs mealybug watershed elevate epigenetic solenopsis bmp lipase phenacoccus	0.41
7	Tillage fatty oleic nue inhibitor drip tocopherol sandy trypsin loam	0.38
8	Bph rust planthopper incognita lugen nilaparvata stal adult chilli bunch	0.38
9	Peanut performance bacteria kernel potassium intervention pesticide antioxidant swarna nitrate	0.36
10	Rdf college irgc flooding gca psb combiner may zea sca	0.36
11	Gene genotype seed marker grain tobacco line stress trait genetic	0.36

Table 5: Generated topics of LDAG3 topic model.

S. No.	Features	Topic coherence
1	Trichoderma rust export inbred tocopherol colony virulence tribal sisal watershed	0.61
2	Phytoplasma snp carotene folder cspa pepper computer ysb pet pepper_capsicum	0.51
3	Dsr drr_dhan finger_millet ncrna sbs bmp shrimp atrazine biofortified acceptability	0.48
4	Cms pigeonpea biodiesel prediction introgression rf restorer scsmv productive_tiller maintainer	0.47
5	Biofilm incognita primary_spike arima dpc pa ghg_emission ptr autoregressive move_average	0.45
6	Biochar somaclone shatter kbsh karyotype robustum pm hs officinarum spontaneum	0.41
7	Phytic_acid gs rg metal nonallelic contaminate maymv poultry_manure ps fm	0.32
8	Gene genotype seed grain trait line genetic isolate crop marker	0.30
9	Npk submergence hill micronutrient rainfall tillage stem zn chickpea iron	0.28
10	Technology_pantnagar pant_university science_humanity department_microbiology college_basic uttarakhand mobile heavy_metal fortify mealybug	0.22

Conclusion

This article examines the use of the LDA model in topic modelling and compares three different preprocessing methods to determine the number of topics in a selected group of research articles. Preprocessing methods include text cleaning, tokenization, stop word removal, lemmatization and Bi-gram which are used to prepare the text data for analysis. Our study revealed that different preprocessing methods have a significant impact on the K value and affect the performance and quality of the generated topics. LDAG3 outperformed the other models in terms of performance, while LDAG1 produced higher quality topics. This finding highlights the importance of selecting appropriate preprocessing techniques to improve the performance and quality of topic models. In the future, our study aims to include a diverse set of research articles from different sources to explore new approaches to topic discovery. With a broader range of data, one can further investigate the effectiveness of different preprocessing methods and develop new techniques to improve topic modelling performance.

Acknowledgements

First author acknowledges The Graduate School, Indian Agricultural Research Institute, New Delhi-110012, India. Authors also acknowledge the support received from ICAR - Indian Agricultural Research Institute, New Delhi and ICAR- Indian Agricultural Statistics Research Institute, New Delhi.

Conflict of Interest

Authors declare that they have no conflict of interest.

Bibliography

1. Barde BV and Bainwad AM. "An overview of topic modeling methods and tools". In Proceedings of the International Conference on Intelligent Computing and Control Systems (2017): 745-750. IEEE.

2. Baumer Eric PS., *et al*. "Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?". *Journal of the Association for Information Science and Technology* 68.6 (2017): 1397-1410.

3. Bellaouar S., *et al*. "Topic modeling: Comparison of LSA and LDA on scientific publications". *2021 4th International Conference on Data Storage and Data Engineering* (2021): 59-64.

4. Blei D M and Jordan M I. "Modeling annotated data". In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2003): 127-134.

5. Deerwester S., *et al*. "Indexing by latent semantic analysis". *Journal of the American Society for Information Science* 41.6 (1990): 391-407.

6. Gupta R K., *et al*. "Prediction of Research Trends using LDA based Topic Modeling". *Global Transitions Proceedings* 3.1 (2022): 298-304.

7. Hofmann T. "Probabilistic latent semantic indexing". In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1999): 50-57.

8. Hong L and Davison BD. "Empirical study of topic modeling in twitter". In Proceedings of the first workshop on social media analytics (2010): 80-88.

9. Hurtado J L., *et al*. "Topic discovery and future trend forecasting for texts". *Journal of Big Data* 3.1 (2016): 1-21.

10. Jelodar H., *et al*. "Latent Dirichl*et al*location (LDA) and topic modeling: models, applications, a survey"*. Multimedia Tools and Applications* 78.11 (2019): 15169-15211.

11. Kherwa P and Bansal P. "Topic modeling: a comprehensive review". *EAI Endorsed Transactions on Scalable Information Systems* 7.24 (2019).

12. Lee N., *et al*. "Combining TF-IDF and LDA to generate flexible communication for recommendation services by a humanoid robot". *Multimedia Tools and Applications* 77.4 (2018): 5043-5058.

13. Mimno D., *et al*. "Optimizing semantic coherence in topic models". In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (2011): 262-272.

14. Murakami A., *et al*. "What is this corpus about?' using topic modelling to explore a specialised corpus". *Corpora* 12.2 (2017): 243-277.

15. Purver M., *et al*. "Unsupervised topic modelling for multi-party spoken discourse". In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (2006): 17-24.

16. Order M., *et al*. "Exploring the space of topic coherence measures". In proceedings of the 8th ACM International Conference on Web Search and Data Mining (2015): 399-408.

17. Sonkusale, L., *et al*. "Exploring the Applicability of Topic Modeling in SARS-CoV-2 Literature and Impact on Agriculture". *Indian Research Journal of Extension Education* 22.4 (2022): 48-56.

18. Steyvers M and Griffiths T. "Probabilistic topic models". In Handbook of latent semantic analysis (2007): 439-460.

19. Syed S and Spruit M. "Full-text or abstract? examining topic coherence scores using latent dirichlet allocation". In proceedings of the IEEE International Conference on Data Science and Advanced Analytics, (2017): 165-174.

20. Zhao W., *et al*. "A heuristic approach to determine an appropriate number of topics in topic modeling". *BMC Bioinformatics* 16.13 (2015): 1-10.