Received : 28.05.2022 | Accepted : 24.07.2022 | Online published : 01.10.2022
https://doi.org/10.54986/irjee/2022/oct_dec/48-56
IRJEE SOCIETY OF EXTENSION EDUCATION

RESEARCH ARTICLE

Exploring the Applicability of Topic Modeling in SARS-CoV-2 Literature and Impact on Agriculture

Lakshmi Sonkusale[1], K.K. Chaturvedi[2], S.B. Lal[3], M. S. Farooqi[4], Anu Sharma[5], Pratibha Joshi[6], Achal Lama[7] and D.C. Mishra[8]

1. Ph.D. Scholar,
2,3 & 5. Pr. Scientist,
4&7. Scientist,
8. Sr. Scientist,
ICAR-IASRI, New Delhi,
6. Scientist,
ICAR-IARI, New Delhi, India
Corresponding author e-mail : kk.chaturvedi@icar.gov.in

ABSTRACT

For the last two years, countries around the globe have been suffering and severely affected by the Covid-19 pandemic due to the novel coronavirus. Researchers from various disciplines are conducting research and publishing number of articles related to this virus and its effects. Furthermore, articles related to Covid-19 are being continuously published in the form of research papers, popular articles, blogs, surveys, short stories etc. These possess useful information and this information can be processed to infer important knowledge by applying text mining techniques. The Latent Dirichlet Allocation (LDA) technique provides an efficient way to analyse unclassified text into useful sets of terms, called topics. LDA can group terms with similar semantic meaning into topics called "themes". A theme is a group of terms that frequently appear together. The objective of the present study is to explore the applicability of topic modeling in identifying the hidden themes or topics by using published research articles related to Covid-19 and agriculture through Google scholar. After pre-processing of titles and abstracts, two approaches namely LDA with Bag of Words (LDAB) and LDA with Term Frequency-Inverse Document Frequency (LDAT) were applied to find the hidden themes. There are thirteen and seven topics are identified by applying LDAB and LDAT respectively. These identified topics comprised with different set of words or features will play an important role in developing the information retrieval system for specific search related to agricultural production, supply chain mechanism in agriculture, health and agri-tourism.

***Key words**: Topic modeling; Covid-19; Latent dirichlet allocation; Machine learning; Text Analytics.*

In late November 2019, a highly infectious disease outbreak due to Covid-19 was started from Wuhan, China. It is caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and the entire world is severely affected by this virus (*Feng et al., 2020*). The cases of Covid-19 are being continuously increasing worldwide, with confirmed cases in Southeast Asia, the United States of America, and several European countries. In January 2020, the World Health Organization (WHO) agreed that the disease Covid-19 meets the criteria of a public health emergency of international concern (*Duarte et al., 2020*), and coronavirus disease was officially declared as a global pandemic in March 2020 (*Cucinotta & Vanelli 2020*). The unusual outbreak of novel SARS-CoV-2 is one of the most serious global threats to public health. Currently, a large number of researchers and specialists have made their best possible efforts to develop vaccines and other remedial measures to mitigate its further spread through active support of health and medical community (*Cheng et al., 2020*). The entire medical and health industries are concerned with its long-term implications on human. Important areas of human are economy, industry,

global markets, agriculture, health care, etc. which are severely affected in this pandemic (*Kumar & Nayar, 2020*). Agriculture, a primary contributing sector of the economy, plays a crucial role in the economic growth of country because it serves the human through providing them food, income, and employment opportunities. The agriculture in our country is also severely affected in this pandemic due to migration of farm workers, affected timely application of plant protection practices, harvesting and post-harvesting operations etc. A number of case studies related to this pandemic have been published in leading scientific journals (*Alga et al., 2020*; *Cheng et al., 2020*; *Feng et al., 2020*; *Holshue et al., 2020; Kumar & Nayar, 2020*) to address the issues related to mutational changes in virus behaviour and its symptoms, clinical diagnosis, food habits, mitigation and management strategies.

Text analytics or text mining is defined as "the process of finding useful or interesting patterns, models, directions, trends, or rules from unstructured text (*Nahm & Mooney, 2002*). It refers to the discovery of knowledge and to find actionable insights from text data. The main tasks of text analytics are text classification, text clustering, text summarization, sentiment analysis, entity extraction and recognition, similarity analysis and topic modeling etc. (*Sarkar, 2016*). The topic modeling is one of the most powerful techniques for finding relationships that exists in text data through text analytics. Various methods for topic modeling namely, Latent Semantic Analysis or Latent Semantic Indexing (LSA or LSI), Probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) have been applied in different subject domains like social network, software engineering, crime science, geographical, political science, medical/biomedical, linguistic science (*Jelodar et al., 2019*). Topic modeling is a content-based analysis technique to find the “hidden” themes from selected literature (*Maier, 2018*). It builds a topic per document model and terms per topic model in the form of dirichlet distributions. The present study aims to discover topics or themes from published literature related to Covid-19 and agriculture and addresses following research questions:

Research Question 1: What are relevant terms in identification of topics or themes?

Research Question 2: What are the important topics selected based on evaluation measures?

This paper is organized into four sections, including an introduction and a conclusion. Section 2 provides the data and methods used in the study. Section 3 discusses the analytical results, and Section 4 concludes the study and mentions the scope for future research.

METHODOLOGY

LDA is a topic modelling technique for discovering latent topics or themes from collected literature based on selected domain. The steps required in conducted the experiment is briefly summarized as workflow (fig. 1). The first step deals with the *creation of the corpus.* A corpus is a collection of d documents, denoted by $D = \{w_1, w_2... w_d\}$. A document is a set of n terms, denoted by $d = (w_1, w_2....w_n)$, where w_n is the n^{th} term in the document. A term is the basic measurement unit in text mining and set of terms are used in creating the vocabulary V.

In the first step, the documents were collected from Google Scholar using the keywords "Covid-19" and "Agriculture" during January 2020 to November 2020. The corpus consists with 994 research articles**.** Fig. 2 shows the trend of research papers related to these keywords in the year 2020. These articles are published in various journals such as World Development, Outlook on Agriculture, International Journal of Research in Pharmaceutical Sciences, Health, Risk & Society, Human Vaccines & Immunotherapeutics, Molecular Medicine Reports, Emerging Microbes & Infections, Journal of Pure and Applied Microbiology, Ocular Immunology and Inflammation. The second step deals with *corpus pre-processing. The* text needs to be converted into numeric form to make it suitable for analysis. The pre-processing of this text is essentially requiring corpus cleaning, tokenization, stop word removal and lemmatization. Python libraries such as natural language processing toolkit (NLTK), *pandas*, *numpy*, *spacy*, *re*, etc. were used and a workflow was created for pre-processing of these documents. The third step deals with *topic modeling and text representation*, with two approaches of text representation i.e. LDA with BOW (LDAB) and LDA with TF-IDF (LDAT). BOW, a text is represented as the bag of its terms for each document in corpus (*Baeza-Yates & Ribeiro-Neto, 1999*). TF-IDF is weighting scheme for a term in each document (*Salton & Buckley, 1988*). LDA is a topic modeling technique for extracting topics from

text corpus developed by (*Blei et al., 2003*). LDA assumes the following generative process for each document d in a corpus D:

Step 1: Choose N ~ Poisson (ξ) where N stands for number of terms/words

Step 2: Choose θ ~ Dir (α) where θ stands for distribution of topics over documents

Step 3: For each of the N terms w_n:

Choose a topic z_n ~ Multinomial (θ)

Choose a term w_n from p (w_n |z_n, β), a multinomial probability conditioned on the topic z_n

Step 1 : Creation of the Corpus

Data Sources → Searching Criteria →Text data corpus

↓

Step 2 : Corpus Pre-processing

↓

Corpus Cleaning → Tokenization → Stop Word Removal → Lemmatization

↓

Step 3 : Topic Modelling and Text Representation

↓

BOW + LDA TF-IDF + LDA

↓

Step 4 : Model Evaluation based on Perplexity and Coherence

↓

Step 5 : Topic Evaluation

Fig. 1. Workflow of the study

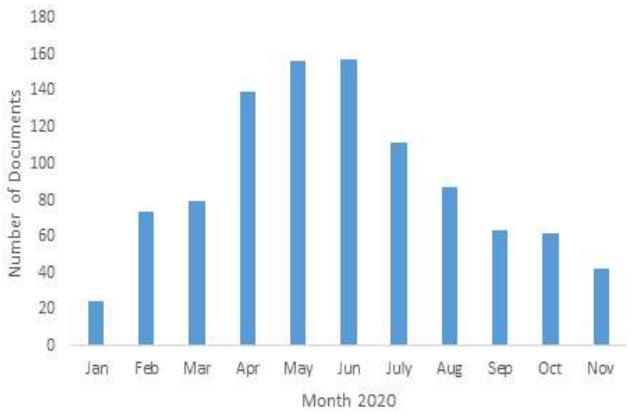

Fig. 2. Published research articles in year 2020

The LDA requires the optimal combination of parameters namely k, α and β. The optimum combination of these parameters will be chosen by using coherence score and model will be evaluated by perplexity measure. The combination of these parameters affects the interpretability of the identified topics (*Steyvers, & Griffiths, 2007*). The values for k range from 2 to 14 with step of 1 and values for α vary from 0.05 to 0.95 with step of 0.30. Similar ranges are used for β as well. The grid search method was used to select the best combination of these parameters and evaluated by coherence score.

The fourth step deals with the *model evaluation* using perplexity measure that estimates how well a model produced for the major part of the corpus by predicting a held-out smaller portion of the documents. The best model will be determined based on the lower value of perplexity. The last step deals with *topic evaluation* and these topics will be evaluated and examined through human judgement that will help in eliminating meaningless terms from the topics.

RESULTS AND DISCUSSION

The results of the study have been discussed in four parts: pre-processing, LDA topic modeling, comparative study between LDAB and LDAT and the impact of Covid-19 in agricultural domain.

Pre-processing : Pre-processing was used to clean and tokenize the title and abstracts of the selected documents by applying lemmatization and stop word removal. Prior the pre-processing, there were 187 average number of terms per document, 24 minimum number of terms, 559 maximum number of terms in each document 559 and the total features were 1,86,235. Top 50 features of the preprocessed data are shown in fig. 3. After the pre-processing of 994 documents, there were 107.25 average number of terms per document, 15 minimum number of terms, 335 maximum number of terms in each document and the total features were 1,06,613. There were Approximately 80,000 features were removed during pre-processing. Top 50 common features were shown in fig. 4. These features/terms are the relevant terms and useful in identification of topics or themes.

Topic modeling using LDA : LDA techniques were applied by using Bag of Words and Term Frequency by Inverse Document Frequency (TF*IDF) as LDAB and LDAT respectively. The fig. 5 and fig. 6 are showing the coherence scores by varying values of k, α and β for LDAB and LDAT respectively.

Table 1 shows the best value of coherence score with respect to best combination of K, α and β. From this Table 1, the highest coherence score i.e., 0.625

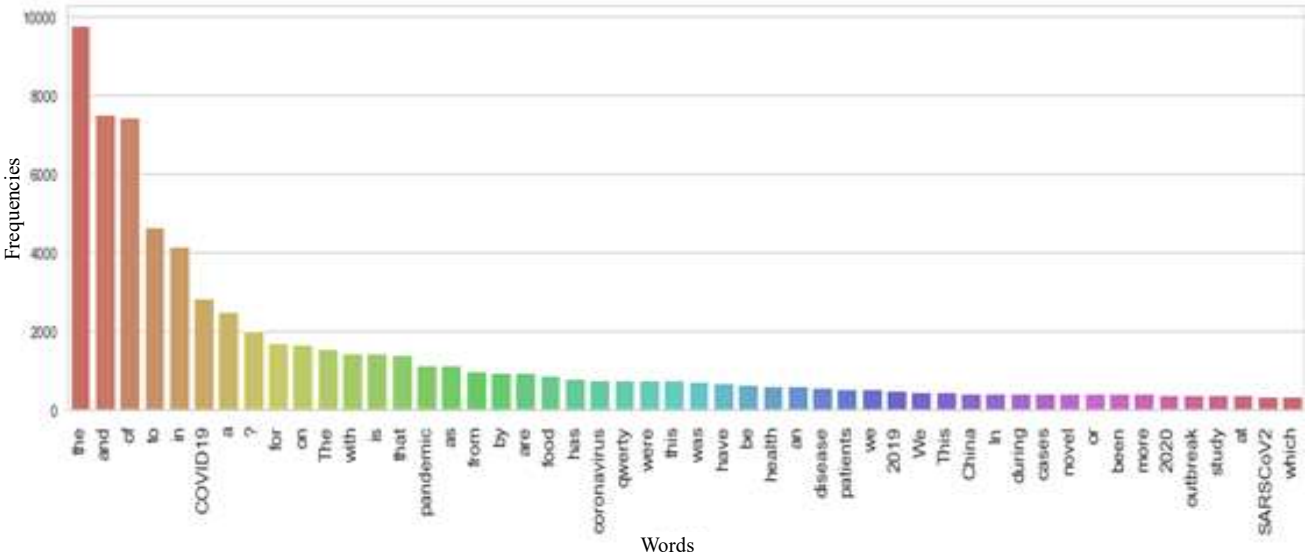

Fig. 3 Top term frequencies before pre-processing

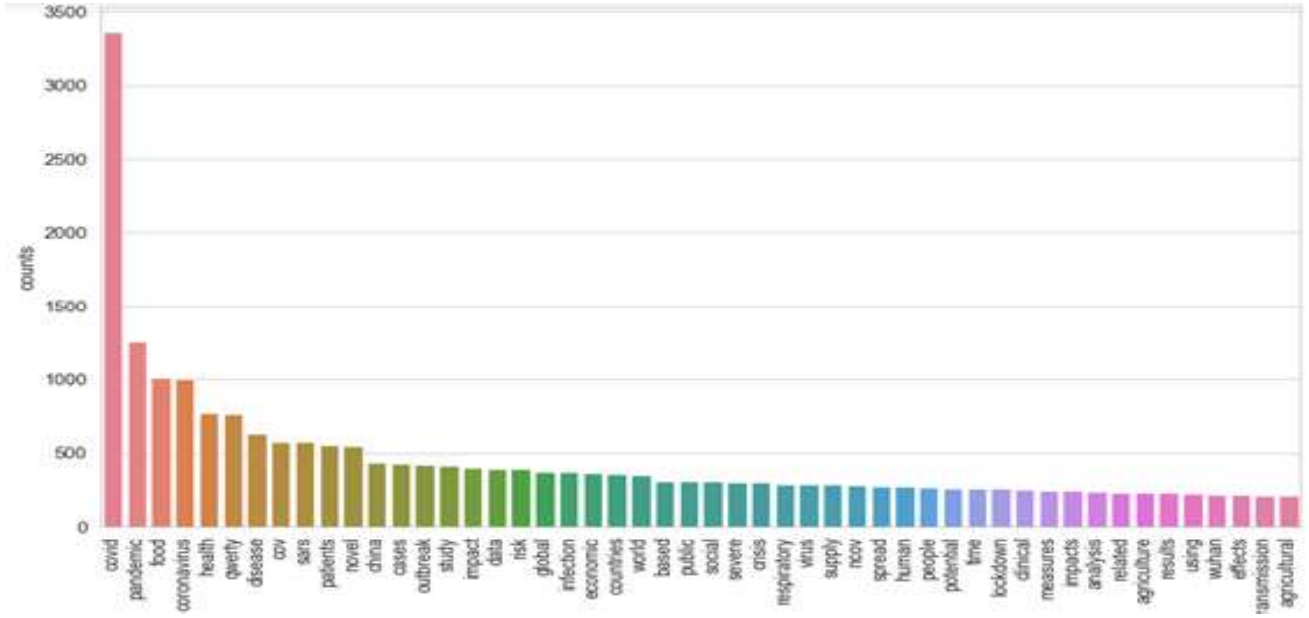

Fig. 4. Top term frequencies after pre-processing

was obtained for 13 number of topics with 0.95 for α and 0.95 for β for LDAB but the highest coherence score i.e., 0.849 was obtained on seven number of topics with 0.05 for α and 0.95 for β for LDAT. Thus, it is clear that more interpretable topics are obtained with high coherence score as compared to number of topics with low coherence score (*Mimno et al., 2011*; *Maier et al 2018*).

Similarly, the perplexity is another important measure to get more interpretable, non-overlapped, and meaningful topics with respect to higher perplexity value (*Ghosh & Guha, 2013*; *Jacobi et al., 2015*; *Mimno et al., 2018*). However, there is increase in number of topics and this will lead to over-fitting

Table 1. Combination of parameters with respect to number of topics identified

No. of Topics	LDAB			LDAT		
K	α	β	Coherence Score	α	β	Coherence Score
2	0.95	0.35	0.545	0.65	0.35	0.719
3	0.35	0.95	0.539	0.65	0.65	0.713
4	0.05	0.35	0.548	0.35	0.65	0.782
5	0.95	0.65	0.512	0.05	0.95	0.808
6	0.35	0.65	0.488	0.35	0.65	0.739
7	0.65	0.95	0.549	0.05	0.95	0.849
8	0.65	0.95	0.528	0.05	0.65	0.743
9	0.95	0.95	0.550	0.05	0.95	0.758
10	0.95	0.95	0.590	0.35	0.95	0.793
11	0.95	0.95	0.607	0.35	0.95	0.700
12	0.95	0.95	0.609	0.35	0.95	0.721
13	0.95	0.95	0.625	0.35	0.95	0.739

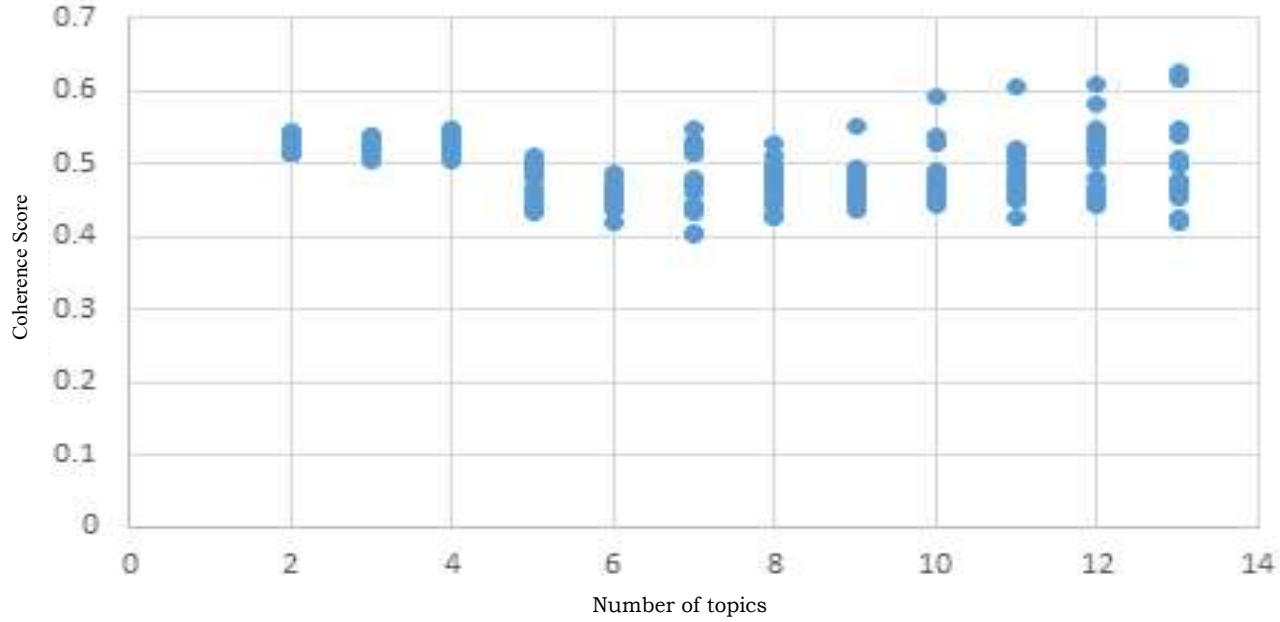

Fig. 5 Coherence scores for different number of topics in LDAB

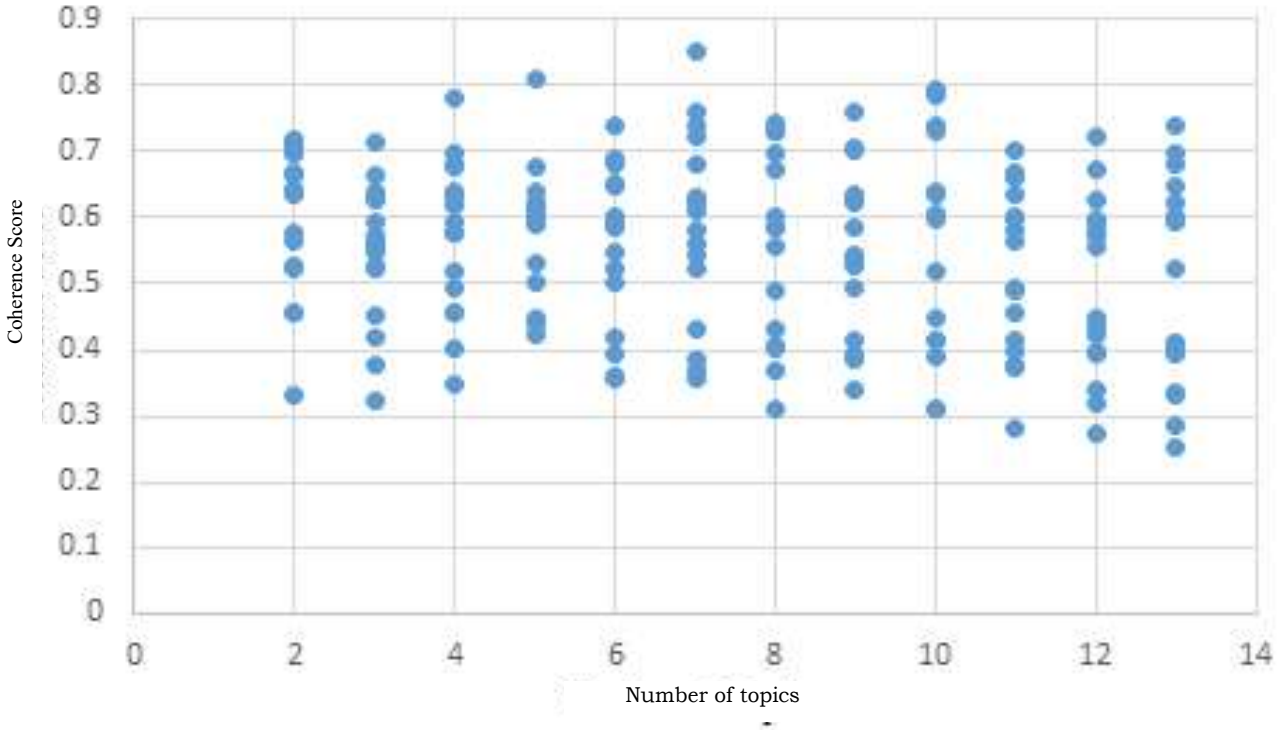

Fig. 6. Coherence scores for different number of topics in LDAT

of the model (*Hidayatullah & Ma'arif, 2017*). In this, we obtained the perplexity value as -10.4 for LDAT and -7.52 for LDAB which are relevant under the considered dataset. Topics obtained by applying LDAB and LDAT are shown in fig. 7 and Fig. 8 respectively with respect to word frequency and weights for individual terms identified in particular theme i.e., topic number. As LDA is a mixture model, the same words can belong to more than one topic too. Some of these important terms identified by applying LDAB are covid, pandemic, impact, food, consumer, tourism, behavior, economic, policy, clinical, human, worker, lockdown, health, hospital, severe, research and response while in LDAT, the identified terms are

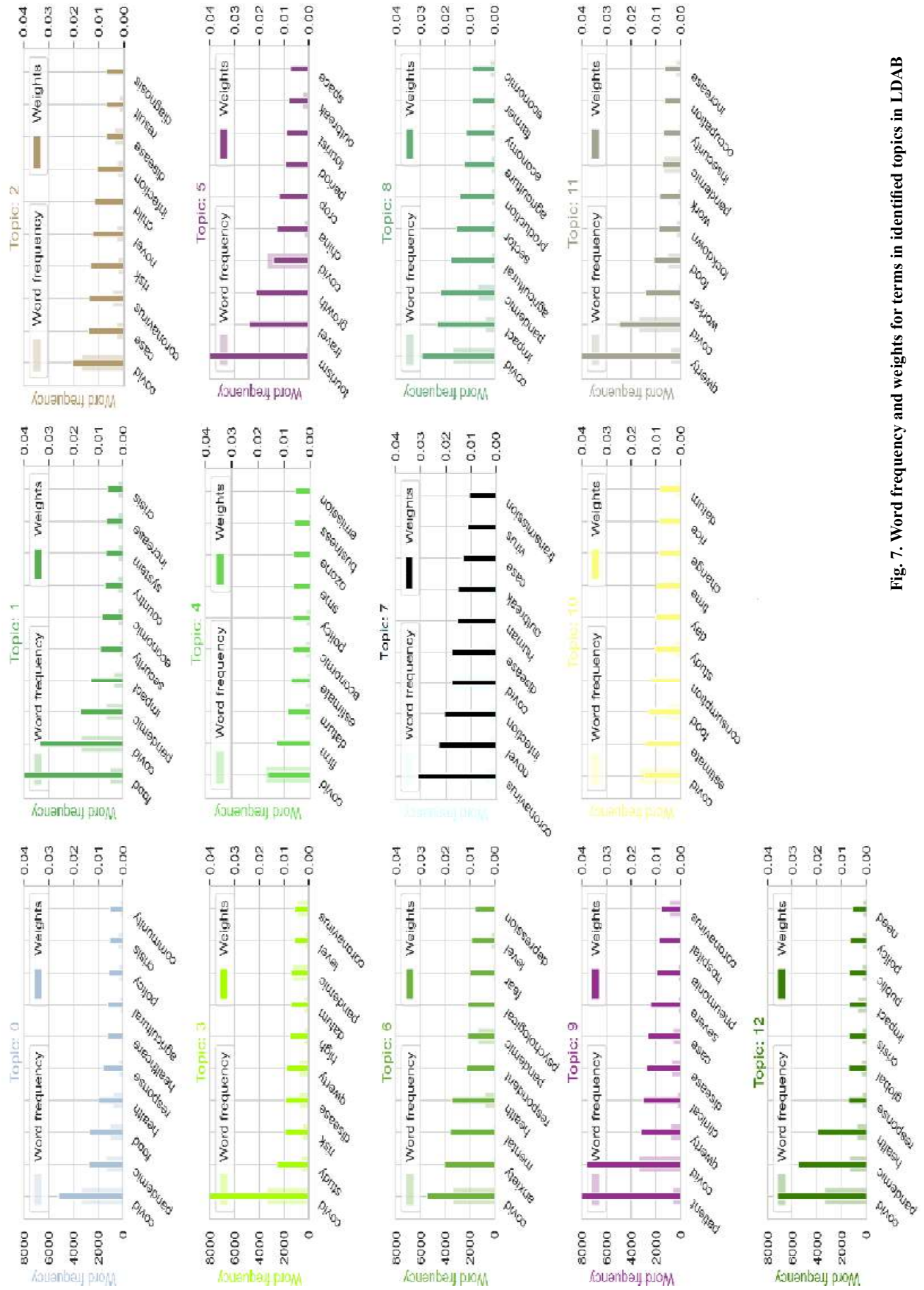

Fig. 7. Word frequency and weights for terms in identified topics in LDAB

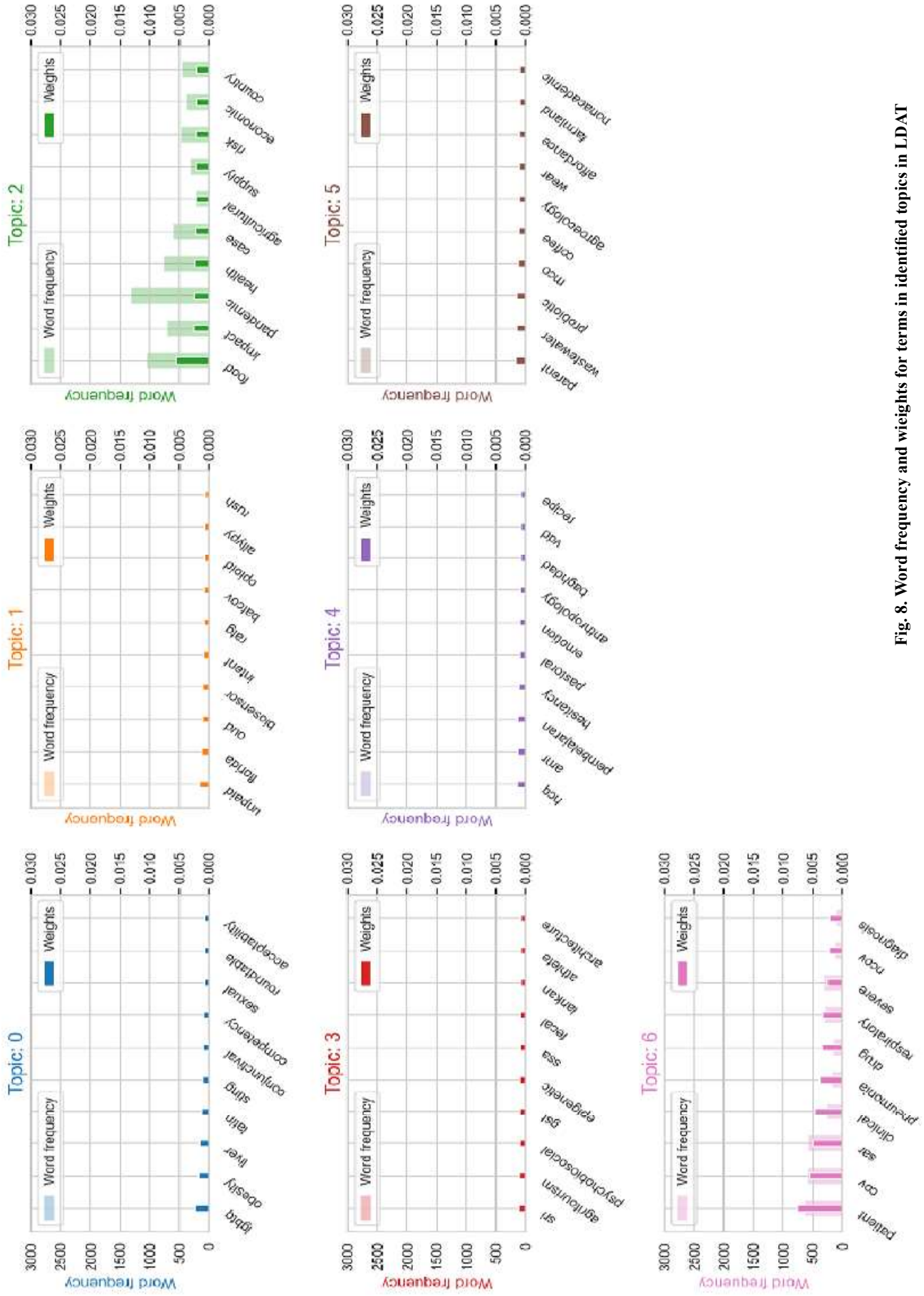

Fig. 8. Word frequency and wieights for terms in identified topics in LDAT

food, impact, pandemic, agriculture, cov, coronavirus, sar and patient. The weights of the individual terms in LDAB are high as compared to word frequency that resulted in large number of topics. However, it is also noted that the weightage of individual terms or words are almost similar across multiple topics identified during the analysis through LDAT that results the smaller number of topics.

Impact of covid pandemic in agriculture : It is noted that the spread of Covid-19 affected various sectors and also agriculture. Some of these domains of agriculture are production, applying package of practices, availability of required farm input, harvesting and post harvesting operations, marketing, supply chain and delivery of produce to the end users (*Aday & Aday, 2020*; *Cariappa et al., 2021*; *Kalogiannidis & Melfou, 2020*; *Mouloudj et al., 2020*). Additionally, there were shortage of inputs like seeds, fertilizer and irrigation devices due to lockdown and lack of production and supply from industry. The production system is also severely affected migration of labor and fear of coming out from the houses (*Shirsath et al., 2020*). There was difficulty faced by producers in ensuring the safe transportation of produce from villages to mandis that reduced their agricultural income (*Rawal et al., 2020*). This study reveals that there are multiple dimensions that largely affects the production system, food consumption, supply chain service providers, lack of mobility of the people and affected their livelihood. Thus, the severity of covid-19 pandemic largely affected many areas of agriculture production and supply chain system.

There were certain limitations to this study and only title and abstract related to keywords "Covid-19" and "Agriculture" were considered during data collection through Google Scholar. However, the search criteria can be further extended with other sub-domains of agriculture such as agronomic practices, plant protection measures, harvesting and post harvesting measures and biotechnological including various omics technologies. Further, other literature databases such as Scopus, Web of Science, PubMed, Semantic Scholar etc. may also be considered for better insights and global scope of topic identification related to agriculture. However, the complete articles will also able to provide the insights related to different aspect of data generation and analysis aspects of material and methods but additionally introduce the computational complexities in the system. This study explored the application of LDA techniques for topic modeling in agriculture. These generated topics will help to develop the customized search criteria for fine tune the literature search as well as focus on current context of research in the academic and industrial interest.

CONCLUSION

Text analytics is playing a major role in identifying different topics to help in developing the information retrieval mechanism for finding relevant documents from the plethora of repositories. In this study, we explore the applicability of LDA techniques with two text representation techniques namely BOW and TF-IDF. Further, there are thirteen and seven topics were identified by mining the literature from Google Scholar after applying keywords "Covid-19" and "Agriculture". These identified topics or themes related to agriculture affected many domains of agriculture that are identified during the study. These topics are production system, food consumption, severity of health, agri-tourism etc. Moreover, the relevant and effective topics are found in LDAT as compared to LDAB due to non-overlapping of words or terms in the identified topics. These finding have been evaluated by coherence score and perplexity measure. This study can be further explored to find the relationships between different terms found in various identified themes or topics. In future, the experiment will be conducted to collect more diverse literature from more than one source and effort will also be made to develop pipeline by providing options for parameter settings and incorporate the findings in developing context specific search engine for agriculture.

CONFLICTS OF INTEREST

The authors have no conflicts of interest.

REFERENCES

Aday, S. and Aday, MS. (2020). Impact of COVID-19 on the food supply chain. *Food Quality and Safety*, **4** (4) : 167-180.

Alga, A.; Eriksson, O. and Nordberg, M. (2020). Analysis of scientific publications during the early phase of the COVID-19 pandemic: topic modeling study. *J . Medical Internet Res.*, **22**(11). https://96/21559

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern information retrieval, 1st Edition, Addison Wesley, New York.

Blei, D. M.; Ng, A.Y. and Jordan, M.I. (2003). Latent dirichlet allocation. *J. Machine Learning Res.*, **3**(Jan), 993-1022.

Cariappa, A.A.; Acharya, K.K.; Adhav, C.A.; Sendhil, R. and Ramasundaram, P. (2021). Impact of COVID-19 on the indian agricultural system: A 10-point strategy for post-pandemic recovery. *Outlook on Agri.*, **50**(1) : 26-33.

Cheng, X.; Cao, Q. and Liao, S.S. (2020). An overview of literature on COVID-19, MERS and SARS: Using text mining and latent dirichlet allocation. *J. Info. Sci.*, https://doi.4

Cucinotta, D. and Vanelli, M. (2020). WHO declares COVID-19 a pandemic. *Acta Bio Medica: Atenei Parmensis*, **91**(1) : 157-160.

Duarte, R.; Furtado, I.; Sousa, L. and Carvalho, C.F.A. (2020). The 2019 novel coronavirus (2019-nCoV): novel virus, old challenges. *Acta Medica Portuguesa*, **33**(3) : 155-157.

Feng, W.; Zong, W.; Wang, F. and Ju, S. (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): a review. *Molecular Cancer*, **19**(1) : 1-14.

Ghosh, D. and Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Info. System. *Cartography and Geographic Info. Sci.*, **40**(2) : 90-102.

Hidayatullah, A.F. and Ma'arif, M.R. (2017). Road traffic topic modeling on Twitter using latent dirichlet allocation. In proceedings of IEEE International Conference on S*ustainable Info. Engg. and Tech.*, 47-52.

Holshue, M. L.; DeBolt, C.; Lindquist, S.; Lofy, K.H.; Wiesman, J.; Bruce, H. and Pillai, S.K. (2020). First case of 2019 novel coronavirus in the United States. *New England J. Medicine.* https://doi: 10.1056/ NEJMoa2001191

Jacobi, C.; Van Atteveldt, W. and Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, **4**(1) : 89-106.

Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y. and Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, **78**(11) : 15169-15211.

Kalogiannidis, S. and Melfou, K. (2020). Issues and opportunities for agriculture sector during global pandemic. *Intl. J. Eco., Busi. and Mngt. Res.*, **4** (12) : 204-211.

Kumar, A. and Nayar, K.R. (2020). COVID 19 and its mental health consequences. *J. Mental Health*, **30**(1) : 1-2.

Maier, D.; Waldherr, A.; Miltner, P.; Wiedemann, G.; Niekler, A.; Keinert, A. and Schmid-Petri, H. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Comm. Methods and Measures*, **12** : 93-118.

Mimno, D.; Wallach, H.; Talley, E.; Leenders, M. and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* 262-272.

Mouloudj, K.; Bouarar, A.C. and Fechit, H. (2020). The impact of COVID-19 pandemic on food security. *Les Cahiers du CREAD*, **36**(3) : 159-184.

Nahm, U. Y. and Mooney, R. J. (2002). Text mining with information extraction. In *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 60-67.

Rawal, V.; Kumar, M.; Verma, A. and Pais, J. (2020). COVID-19 lockdown: Impact on agriculture and rural economy. *Society for Social and Eco.Res.*, **13**(2) : 34.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Info. Processing & Mngt.*, **24** (5) : 513-523.

Sarkar, D. (2016). Text Analytics with python. Apress, New York, USA.

Shirsath, P.B.; Jat, M.L.; McDonald, A. J.; Srivastava, A. K; Craufurd, P.; Rana, D. S. and Braun, H. (2020). Agricultural labor, COVID-19, and potential implications for food security and air quality in the breadbasket of India. *Agril. Systems*, **185** :102954.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis*, **427** (7) : 424-440.

• • • • •