Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 76(1) 2022 7–16

Topic Modelling for Discovering Themes in the Queries Raised at Farmers' Call Center

B.S. Yashavanth and P.D. Sreekanth

ICAR-National Academy of Agricultural Research Management, Hyderabad

Received 29 January 2022; Revised 04 April 2022; Accepted 04 May 2022

SUMMARY

Topic modelling has gained prominence in the recent years due to the availability and necessities for the analysis of large volumes of unstructured text data. In agriculture, a huge amount of text data is generated in kisan call centers in the form of queries raised by the farmers. This study attempts to use the Latent Dirichlet Allocation method of topic modelling to discover the hidden topics in the queries raised at kisan call centers of five south Indian states. Through exploratory text analysis, it was found that the most common terms appeared in the query texts are 'weather', 'management' and 'market'. The topic modelling lead to identification of 12 topics, out of which the topic 'pest management in paddy, cotton and chilli' reported the maximum number of queries.

Keywords: Topic models; Latent Dirichlet Allocation; Text analysis; Kisan call center.

1. INTRODUCTION

Textual databases are on the rise because of increased amount of information available on the Web (Gaazinoory *et al.*, 2013). These days, information from many of the public and private institutions and other organizations is stored electronically in textual databases (Feldman and Sanger, 2007). In today's digital era, the amount of information stored as texts in scientific journal databases, message boards and tweet archives, digitalized textbook collections, or newspaper archives is growing exponentially (Antons, 2016). Over the last decade, statistical machine learning algorithms have been developed to analyze distribution of words/ terms in text documents towards uncovering topics in a computer assisted, largely automated fashion (Blei, 2012).

Due to the increased application of ICT tools in agricultural sector, large amount of unstructured data including text data are available. Such unstructured text data needs to be analysed for its content using text mining techniques to capture the information hidden in them. Topic modeling is a text mining approach for automated content analysis which helps in identifying topic structures that are hidden in text corpora. Topic modeling is considered as an efficient alternative to the widely used manual (Biemans *et al.*, 2007, 2010; Durisin *et al.*, 2010) or computer-assisted (Guo, 2008) content analysis.

Latent Dirichlet Allocation (LDA) is the simplest (Blei *et al.*, 2003) and most popular topic modeling algorithm (Antons, 2016). The basic idea behind LDA is that the documents of a collection (called as corpus) are conceptualized as bundles of multiple topics and that each document present in the corpus is characterized by a particular topic distribution. Assigning topics to documents would result in a multinomial distribution of topics across articles showing the number of documents associated with each topic (Antons, 2016). For a single document, such a distribution would indicate which topic(s) the document address and which they do not address. The Dirichlet distribution refers to the corresponding probability distribution of this multinomial distribution. LDA assumes that topics and their word distributions existed before the creation

Corresponding author: B.S. Yashavanth
E-mail address: yashavanthbs@gmail.com

of the documents (Blei, 2012). With this understanding, the generative process of the statistical model of LDA tries to replicate the imaginary process through which the documents were generated.

In literature, the topic modeling methods have been effectively used for various research-oriented tasks such as multi-document summarization (Haghighi and Vanderwende, 2009), patient generated data on Reddit (Okon *et al.*, 2020), word sense discrimination (Brody and Lapata, 2009), sentiment analysis (Titov and McDonald, 2008), machine translation (Eidelman *et al.*, 2012), information retrieval (Wei and Croft, 2006), discourse analysis (Nguyen *et al.*, 2012), and image labeling (Fei-Fei and Perona, 2005), discovering the hot topics covered by the scientific journal (Griffiths and Steyvers, 2004), identifying important topics within news archives on the web (Kim and Oh, 2011). The topic models using LDA are successfully used for analysing consumer complaints (Bastani *et al.*, 2019), online accommodation reviews (Sutherland *et al.*, 2020) and investigating emerging trends in e-learning field (Gurcan *et al.*, 2021). In India, topic models have been successfully utilized for analysing text data from different fields. Maskeri *et al.*, (2008) used LDA based topic model for mining business topics in source code. Vamshi *et al.* (2018) used topic models for opinion mining and sentiment analysis of text reviews posted in web forums or social media site. Chauhan and Shah (2021) have surveyed the application of topic modeling using Latent Dirichlet Allocation over various domains. However, application of this relatively new technique in agricultural research is not reported barring a few (Biswas and Jain, 2018). This study is a pioneer effort in applying the concepts of text analytics techniques including topic models to analyse unstructured text data in agriculture. The study aims to identify the trending topics among the queries raised at Kisan Call Centers in South India.

2. MATERIALS AND METHODS

2.1 Topic Modeling

The topic modeling algorithms have been designed to discover the underlying set of topics of a given text corpus by analyzing the words in the texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time (Blei, 2012). The probabilistic topic models are a suite of algorithms with an aim to discover and annotate large archives of text documents with thematic information (Blei, 2012). These quantitative measures that are provided by these algorithms are useful in identifying the topics of single documents, investigating the document similarity and their thematic association and uncover temporal changes of the content of the collection (Griffiths and Steyvers, 2004). These statistical methods discover the themes that run through the texts, how those themes are connected to each other, and how they change over time (Blei, 2012). These algorithms are advantageous since they don't require prior categorization, labeling, or annotation of the documents or the collection (Blei, 2012; Blei, Ng, and Jordan, 2003).

2.2 Latent Dirichlet Allocation

The LDA algorithm is a generative probabilistic model for topic modeling which is based on collections of discrete data such as frequency and text corpora (Blei, 2012; Blei *et al.*, 2003). The basic idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is determined by a distribution over words (terms). The document generation process of LDA is as given in Fig. 1. In the generative process of LDA, for a corpus having D number of documents, let K be the number of topics and N_d be the number of words in the d^{th} document (d=1,2,…,D). Suppose, $W_{d,n}$ is the n^{th} word in the document d and $Z_{d,n}$ is the topic assignment to the word $W_{d,n}$. Let θ_d denote the topic proportions for the d^{th} document which follows Dirichlet distribution with parameter α; and φ_k denote the word distribution for the k^{th} topic which follows Dirichlet distribution with parameter β. The generative process of LDA can be denoted by the joint distribution of the random variables as follows:

$$p(w_d, z_d, \theta_d, \varphi \mid \alpha, \beta)$$
$$= \prod_{n=1}^{N_d} p\left(w_{d,n} \mid \varphi_{Z_{d,n}}\right) p\left(z_{d,n} \mid \theta_d\right) p\left(\theta_d \mid \alpha\right) p\left(\varphi \mid \beta\right)$$

The parameters of the LDA topic model, θ_d and φ_k, can be estimated using the Gibbs sampling method (Griffiths and Steyvers, 2004; Noel and Peterson, 2014).

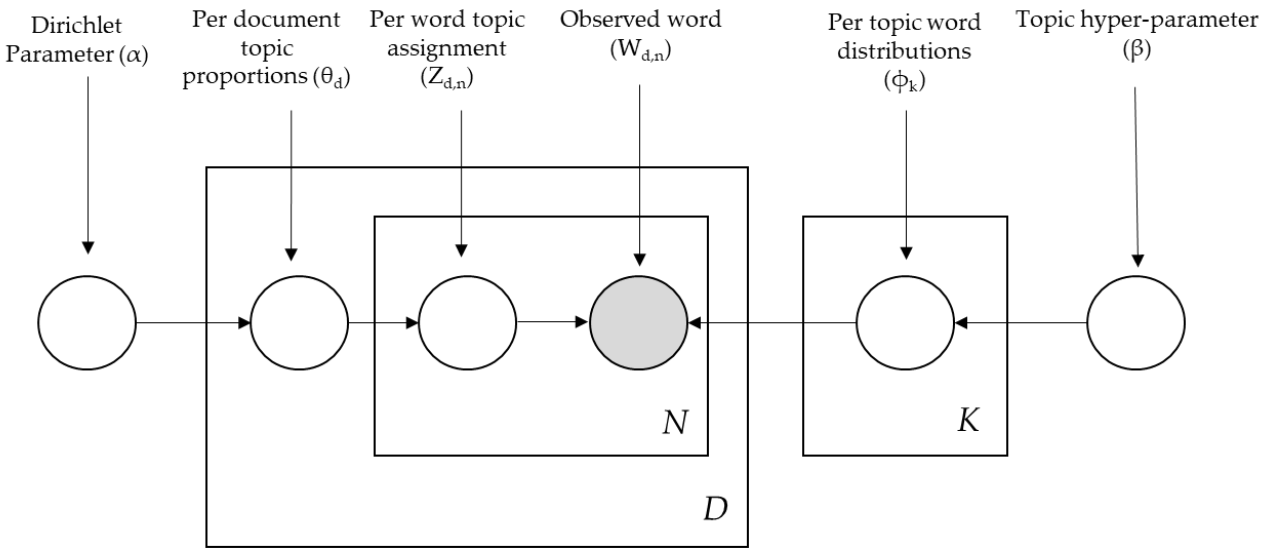

Fig. 1. The document generation process of LDA

2.3 Data

The data for this study has been extracted from the website www.data.gov.in, an online platform supporting the open data initiative of Government of India. This web portal provides single point access to datasets and documents published by various ministries, departments and organisations of India. We selected the data corresponding to the farmers' queries raised in Farmers' Call Center (Kisan Call Center, KCC). The KCC scheme was launched by GoI in 2004 in order to harness the potential of ICT in Agriculture. The main aim of the scheme is to answer farmers' queries on a telephone call in their own dialect. Under this, the farmers can call to Kisan Call Center and enquire about various queries/problems related to crops, seeds, fertilizers, agriculture commodity prices, pesticides, horticulture, veterinary, etc. These call centers are working in 14 different locations covering all the States and Union Territories. Replies to the farmers' queries are given in 22 local languages. The Farm Tele Advisors (FTAs) answer the call and simultaneously record the query details. These recorded data have been utilized in this study for discovering the pattern in the farmers' queries. The data for the year 2017 belonging to 120 districts of the 5 South Indian states viz., Andhra Pradesh, Karnataka, Kerala, Tamilnadu and Telangana were used. Since the data is available for every month for each district, we could obtain 1440 documents consisting queries raised in each month and in each district. All the data analyses were carried out using R Programming Language (R Core Team, 2020).

2.4 Text Cleaning

The query texts recorded by the FTAs consists technical terms regarding pests and diseases, crop management practices including sowing and harvesting and market information. However, most of the terms do not provide any meaningful information when alone (called as stop words). These stop words needs to be removed from the documents and only the most meaningful information needs to be kept for further analysis. Since LDA follows the bag of words assumption, every unique term in the target corpus is counted when assigning words to themes, and hence unnecessary or meaningless elements need to be removed before proceeding for topic modeling. In addition, to maintain the uniformity, all the terms need to be converted to their lower cases including acronyms (text normalization). So, the data cleaning consisted of removal of stop words, numeric values, non-alphabetic characters, punctuations and normalizing. We used the English stop word list available in the 'corpus' package in R.

2.5 Study framework

The LDA assumes that the order of the words does not matter, or the "bag of words" assumption (Blei, 2012). This assumption is satisfied in our study since the

individual query text make up our documents in which the order of the words does not make any difference. The other assumption of the LDA model is that the document order does not matter. This assumption becomes unrealistic when huge temporal collections of documents are analysed where topics change over time (Blei, 2012). The third assumption of the LDA model is that the number of topics in the data corpus is fixed and known apriori. Towards this, we used the Daveaud metrics which estimates the number of latent topics by maximizing the information divergence D (Jensen-Shannon divergence) between all pairs (k_i, k_j) of LDA's topics (Daveaud, 2014). Accordingly, the number of topics $\hat{K}$ estimated by this method is given by the following formula:

$$\hat{K} = \underset{K}{\boldsymbol{arg\,max}} \frac{1}{K(K-1)} \sum_{(k,k') \in T_K} D(k \| k')$$

where K is the number of topics given as a parameter to LDA, and T_K is the set of K topics modeled by LDA. The Jensen-Shannon divergence between all pairs of topics is formally written as:

$$D(k \| k') = \frac{1}{2} \sum_{\omega \in W_k \cap W_{k'}} P_{TM}(\omega | k) \log \frac{P_{TM}(\omega | k)}{P_{TM}(\omega | k')} \times \frac{1}{2} \sum_{\omega \in W_k \cap W_{k'}} P_{TM}(\omega | k') \log \frac{P_{TM}(\omega | k')}{P_{TM}(\omega | k)}$$

where $P_{TM}(\omega | k) = \varphi_{k,w}$.

3. RESULTS AND DISCUSSION

3.1 Exploratory text analysis

Prior to the topic modeling, an exploratory text analysis was carried out to understand the data. Overall, 5,710,093 queries were raised during the study year out of which majority of the queries were related to weather (> 60%). The number of queries raised varied between months, the monsoon months reporting the highest number of queries most of which were related to weather (> 70%). Therefore, we felt that the weather-related queries need to be removed from the text corpus for identifying themes appropriately. Thus, the study corpus (data excluding weather-related queries) consisted of 1,883,002 queries with 9,594,414 words which reduced to 6,476,095 words after removing stop words. The month-wise total and weather-related queries raised are given in Fig. 2.

The corpus had 17,754 unique words, out of which the words 'market', 'management' and 'rate' were found to be the unigrams appearing most in the text corpus. The top 20 unigrams appearing in the text corpus are given in Fig. 3. Among the 52,903 bigrams, 'market rate' was found to be appearing the highest number of times followed by 'plant protection' and 'nutrient management'. Fig. 4 gives the top 20

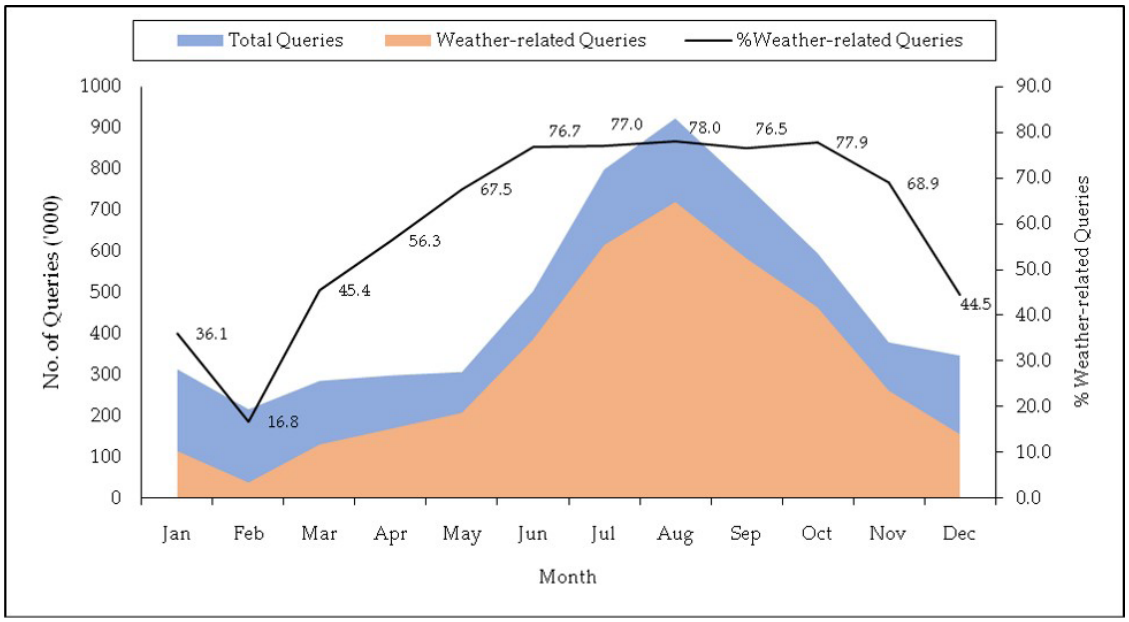

Fig. 2. Month-wise distribution of queries

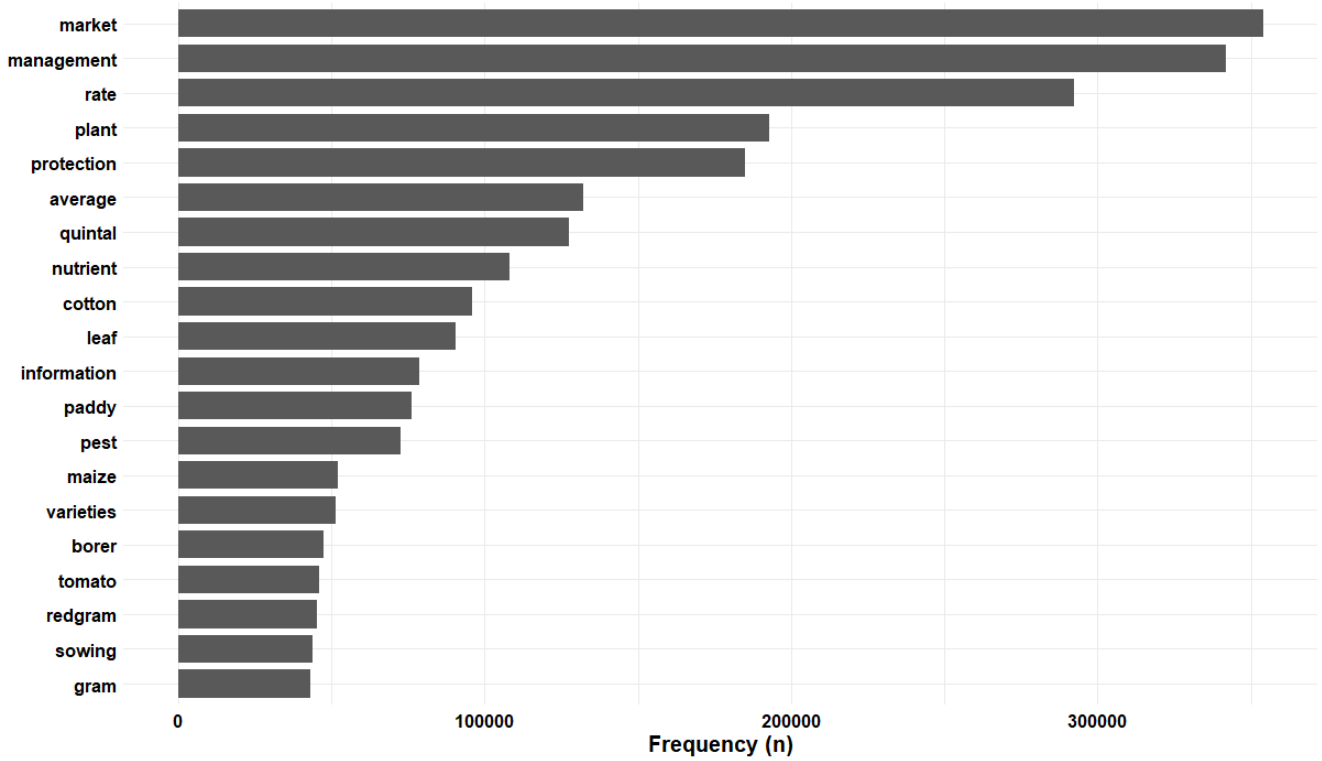

Fig. 3. Top 20 unigrams from the queries raised

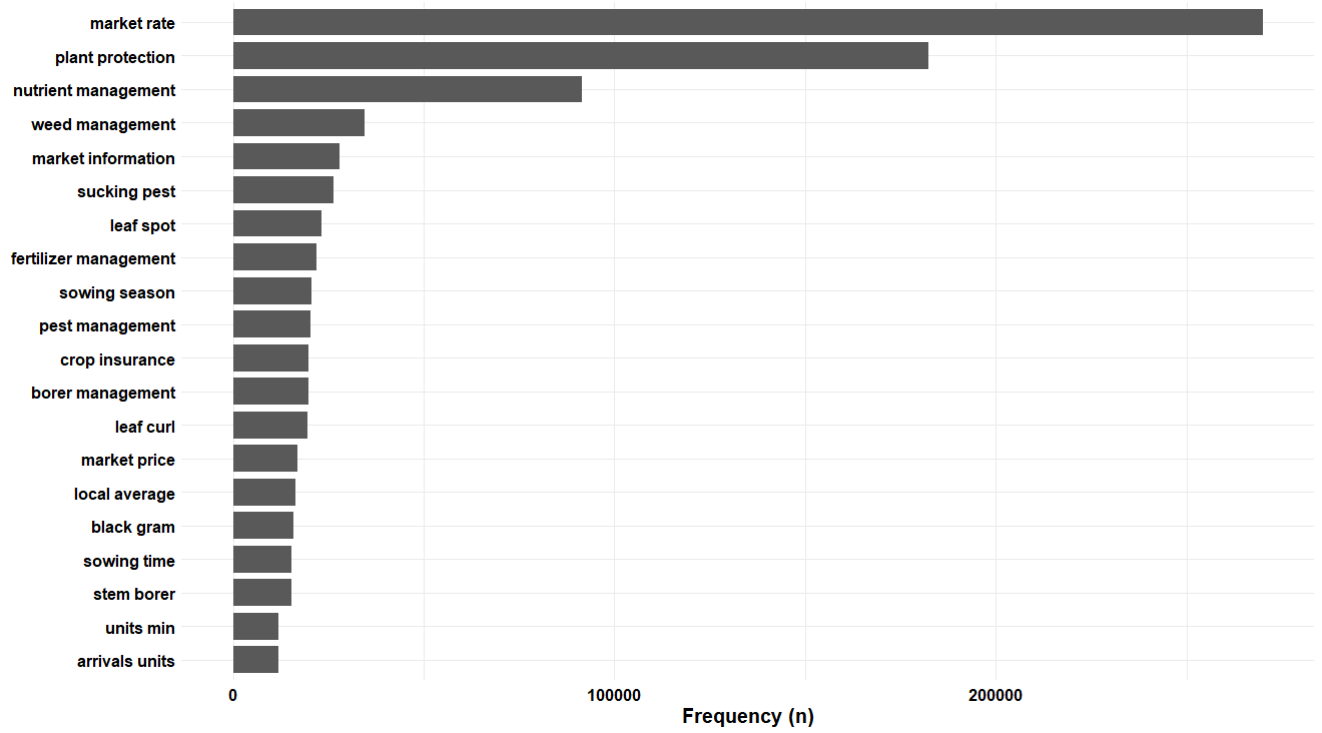

Fig. 4. Top 20 bigrams from the queries raised

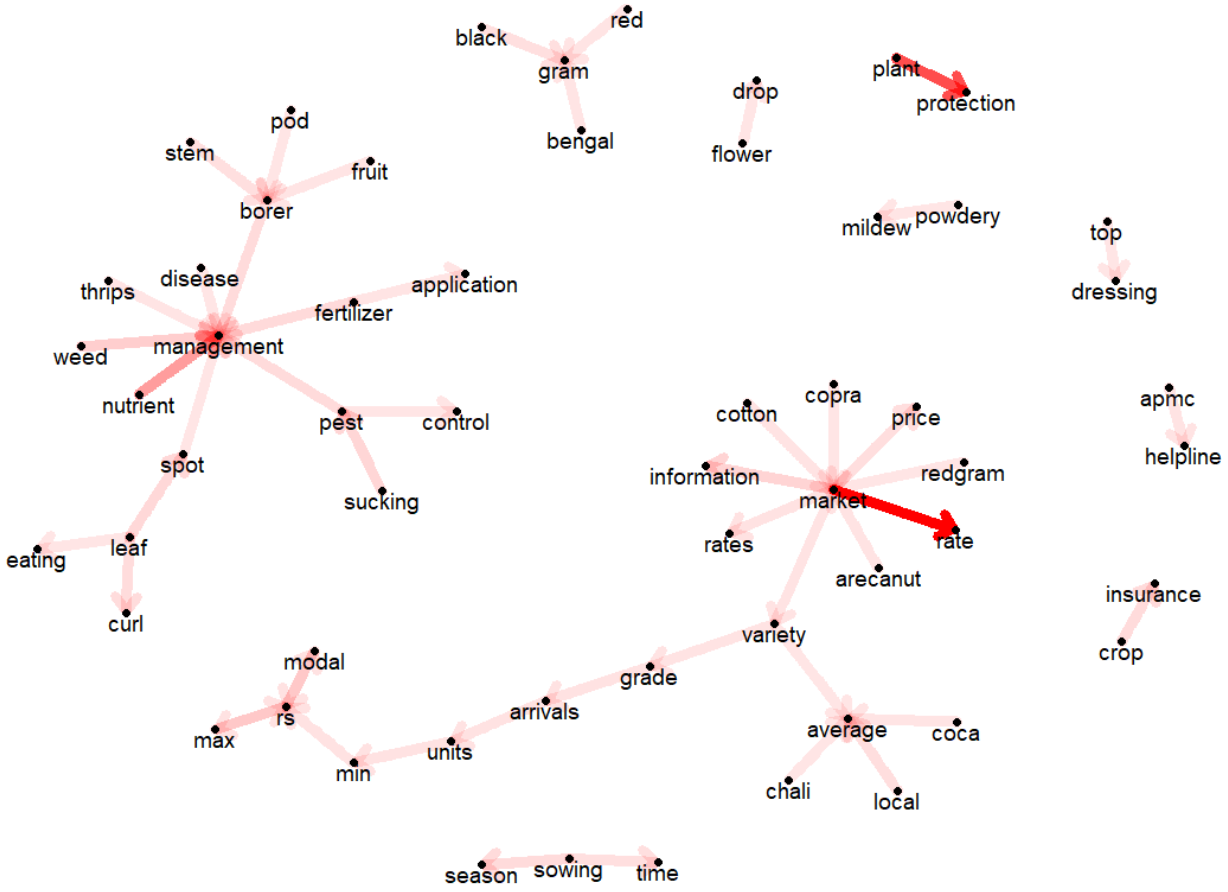

Fig. 5. Network of top 50 words most appearing in the queries

bigrams appearing in the text corpus. The network of top 50 words which commonly occur together (highest correlation between words) was plotted to investigate the relationship between them (Fig. 5). The dot indicates the preceding text and the arrow indicates succeeding text whereas the depth of the arrow indicates the correlation. It is observed that the word 'plant' commonly occurs with the word 'protection' with a correlation of r=0.96; the word 'market' with a set of words comprising 'rate' (r=0.806), 'price' (r=0.180), 'copra' (r=0.154), 'grade' (r=0.152) and 'arrivals' (r=0.150); and the word 'management' with the words 'nutrient' (r=0.428), 'weed' (r=0.258), 'paddy' (r=0.162), 'sucking' (r=0.143) and 'fertilizer' (r=0.134). This indicates that most of the queries raised were regarding plant protection measures, market related information and various crop management practices.

3.2 Topic models

The 1440 documents comprising the queries raised at 120 districts in 2017 are used for fitting the topic models using Latent Dirichlet Allocation algorithm. The LDA algorithm uses the Gibbs sampling method for developing the topic models. The estimation using Gibbs sampling requires specification of values for the parameters of the prior distributions. Griffiths and Steyvers (2004) suggest a value of 50/k for α and 0.1 for β. In addition, since the LDA requires that the optimum number of topics are known before modelling, the appropriate number of topics was identified using Daveaud metric (Daveaud, 2014). Deveaud metrics estimates the number of latent concepts of a corpus by maximizing the information divergence. The Deveaud metrics takes values from 0 to 1 and the number of topics (k) for which the Deveaud metrics is highest will be taken as the optimum number. In this study, the values of Deveaud metrics were calculated for number

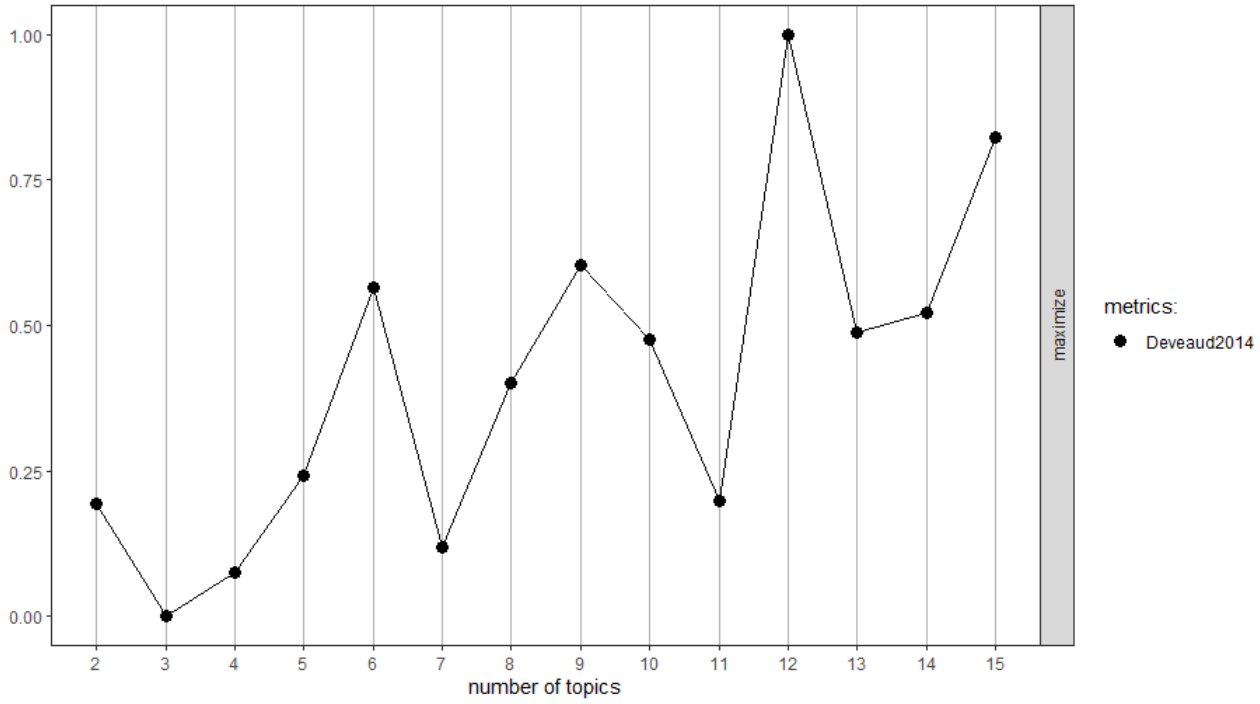

Fig. 6. Values of Deveaud metrics for different topic numbers

Table 1. Top 10 words appearing in the topics

Topic No	No. of documents (%)	Top 10 words	Topic interpretation
3	495 (35.4%)	management, paddy, cotton, chilli, leaf, sucking, borer, pest, nutrient, stem	Pest management in paddy, cotton and chilli
4	361 (25.8%)	management, paddy, fertilizer, season, market, top, foliar, sowing, gram, black	Fertilizer management
9	185 (13.2%)	management, attack, nutrient, pest, banana, tomato, contact, coconut, number, animal	Nutrient and pest management in banana, coconut and tomato
8	76 (5.4%)	management, plant, protection, nutrient, leaf, market, disease, tomato, rate, pest	Pest management in tomato
1	58 (4.1%)	management, nutrient, varieties, number, market, crop, rate, helpline, weed, disease	Nutrient management
12	56 (4.0%)	market, rate, management, plant, protection, maize, leaf, tomato, nutrient, helpline	Market information of maize and tomato
11	35 (2.5%)	market, rate, arecanut, copra, number, plant, protection, helpline, commodity, management	Market information of arecanut and copra
2	32 (2.3%)	protection, plant, market, rate, management, cotton, pest, nutrient, maize, onion	Plant protection in cotton, maize and onion
6	28 (2.0%)	plant, protection, management, cotton, pest, leaf, nutrient, market, redgram, rate	Plant protection in red gram
5	26 (1.9%)	market, crop, management, insurance, plant, protection, rate, nutrient, helpline, number	Market information
7	24 (1.7%)	market, rate, redgram, cotton, bengalgram, management, tur, plant, protection, commodity	Market information of cotton and pulses
10	24 (1.7%)	management, plant, protection, sugarcane, nutrient, maize, leaf, weed, borer, pest	Nutrient and pest management in sugarcane

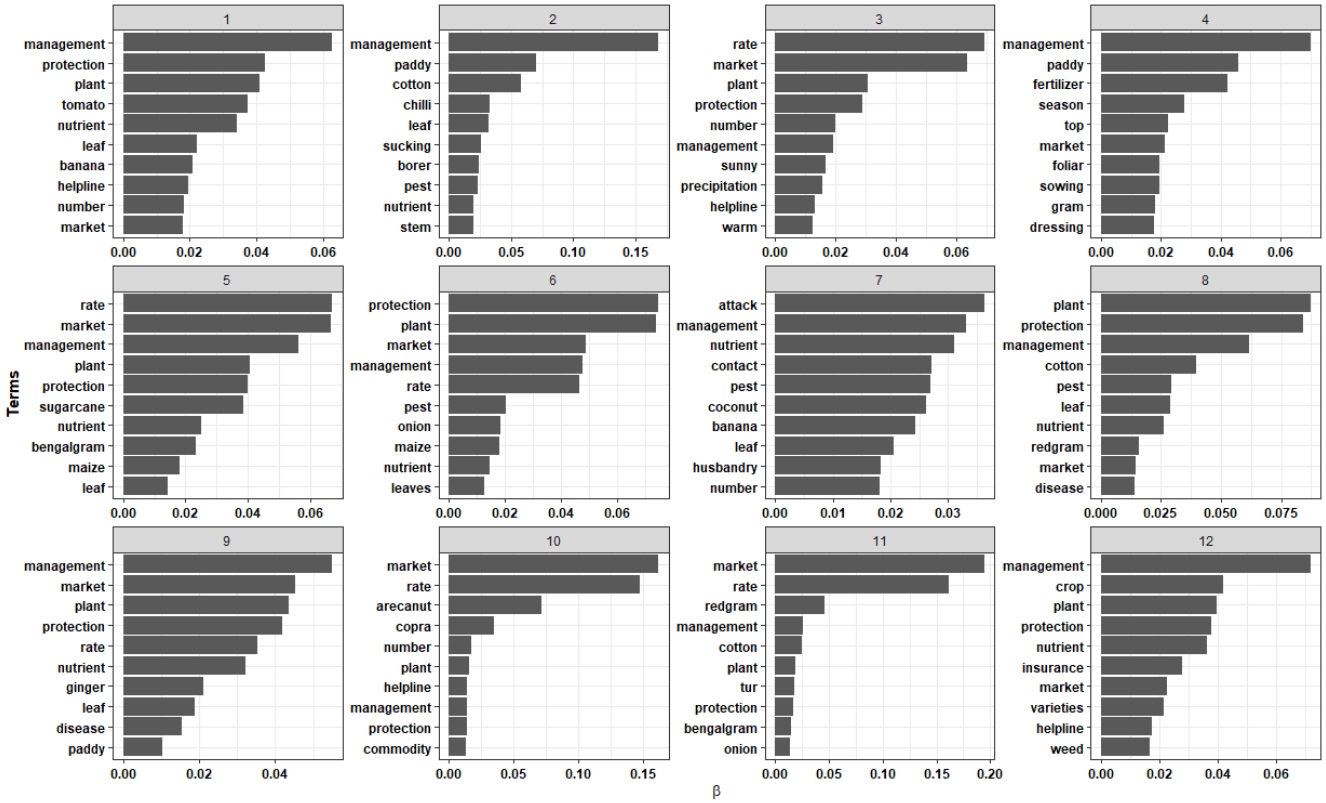

Fig. 7. Per topic per word probabilities for the top 10 words in each topic

of topics from k=1 to k=15. The Deveaud metrics value was found to be highest (close to 1, Fig. 6) at k=12 suggesting that the corpus has 12 distinguishable topics. Subsequently, the topic models were fit using the Gibbs sampling method. Among the 12 topics, the topic 3 consisted the highest number of documents (495 documents, 35.4%) followed by topic 4 (361 documents, 25.8%) and topic 9 (185 documents, 13.2%). The top ten words frequently appearing in these topics along with the topic labels are given in Table 1. A perusal of the topics and the words they comprise suggest that most of the documents belong to the 'pest management in paddy, cotton and chilli' topic. This is evident due to the fact that pests cause around > 25% and >18% yield loss in paddy and cotton respectively (Dhaliwal *et al.*, 2015). Also, since paddy and cotton are the important crops grown in the study area, the number of queries raised on these crops are more. The per topic per word probabilities, which gives the probability that the term belongs to that particular topic are given in Fig. 7. By looking at the figure, it can be concluded that the words 'management' and 'market' appear in most of the topics. However, there is a clear grouping with respect to the crops regard to which the information related to crop management and market rate are sought by the farmers. This kind of studies involving advanced machine learning techniques are helpful in identifying the most prominent problems faced by the farmers and reported at the kisan call centres. Results from such analysis are useful in taking policy decisions which addressing the needs of the farmers.

4. CONCLUSIONS

In this study, text analysis techniques have been employed to discover the latent topics in the queries raised at kisan call centers of five south Indian states for the year 2017. The exploratory analysis indicated that >60% of the queries raised were for weather related information. Among others, the most appeared terms are 'management' and 'market'. The Latent Dirichlet Allocation, a probabilistic topic model, was used to identify the hidden topics in the query text. The data corpus of 1440 documents comprising month-wise

and district-wise queries were found to be consisting 12 topics. Among them, the topic 'pest management in paddy, cotton and chilli' was found to be consisting highest number of documents (>35%). This study paves way for application of modern text analytics tools for analyzing unstructured text data in agriculture to obtain results that are helpful in making appropriate policy decisions.

ACKNOWLEDGEMENTS

Authors are grateful to the anonymous referees for their valuable comments which helped in improving the manuscript.

REFERENCES

Antons, D., Kleer, R. and Salge, T.O. (2016). Mapping the topic landscape of JPIM, 1984-2013: In search of hidden structures and development trajectories. *Journal of Product Innovation Management*, **33**, 726-749.

Bastani, K., Namavari, H. and Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints, *Expert Systems with Applications*, **127**, 256-271.

Biemans, W., Griffin A. and Moenaert, R. (2007). Twenty years of the Journal of Product Innovation Management: History, participants, and knowledge stock and flows. *Journal of Product Innovation Management,* **24(3)**, 193-213.

Biemans, W., Griffin, A. and Moenaert, R. (2010). In search of the classics: A study of the impact of JPIM papers from 1984 to 2003. *Journal of Product Innovation Management,* **27(4)**, 461-84.

Biswas, S., and Jain, R. (2018). Text document categorization using machine learning algorithm in agricultural domain. *Journal of the Indian Society of Agricultural Statistics*, **72(1)**, 61-69

Blei, D.M. (2012). Probabilistic topic models. *Communications of the ACM,* **55(4)**, 77–84.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research,* **3**, 993-1022.

Brody, S. and Lapata, M. (2009). Bayesian word sense induction. *In* Proceedings of the 12th Conferenceof the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 103-111.

Broniatowski, D.A. and Magee, C.L. (2017). The emergence and collapse of knowledge boundaries, *IEEE Trans. Eng. Manage.*, **64(3)**, 337-350.

Chauhan, U., and Shah, A. (2021). Topic modeling using latent Dirichlet allocation: A survey. *ACM Computing Surveys (CSUR)*, **54(7)**, 1-35.

Deveaud, R., SanJuan, E. and Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Doc. Numer*. **17**, 61-84

Dhaliwal, G.S., Jindal, V. an Mohindru, B. (2015). Crop losses due to insect pests: global and Indian scenario, *Indian Journal of Entomology*, **77(2)**, 165-168

Durisin, B., Calabretta, G. and Permeggiani, V. (2010). The intellectual structure of product innovation research. *Journal of Product Innovation Management,* **27(3)**, 437-51.

Eidelman, V., Boyd-Graber, J. and Resnik, P. (2012). Topic models for dynamic translation model adaptation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Vol. 2. Stroudsburg, PA, USA: Association for Computational Linguistics, 115-119.

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In Proceedings of the 10th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005). Los Alamitos, CA, USA: IEEE Computer Society, 524-531.

Feldman, R. and Sanger, J. (2007). The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press.

Ghazinoory, S., Ameri, F. and Farnoodi, S. (2013). An application of the text mining approach to select technology centers of excellence, *Technological Forecasting & Social Change,* **80**, 918-931.

Griffiths, T.L. and Steyvers, M. (2004). Finding scientific topics, Proc. Nat. Acad. Sci. **101**, 5228-5235.

Griffiths, T.L. and Steyvers, M. (2004). Finding Scientific Topics, Proceedings of the National Academy of Sciences of the United States of America, **101**, 5228-5235.

Guo, L. (2008) Perspective: An analysis of 22 years of research in *JPIM*. *Journal of Product Innovation Management* **25(3)**, 249-60.

Gurcan, F., Ozyurt, O., and Cagitay, N.E. (2021). Investigation of emerging trends in the e-learning field using latent dirichlet allocation. *International Review of Research in Open and Distributed Learning*, **22(2)**, 1-18.

Haghighi, A. and Vanderwende, A. (2009). Exploring content models for multi-document summarization. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 362-370, Boulder, Colorado

Kim, D., and Oh, A. (2011). Topic chains for understanding a news corpus. *In* International Conference on Intelligent Text Processing and Computational Linguistics (pp. 163-176). Springer, Berlin, Heidelberg.

Maskeri, G., Sarkar, S., and Heafield, K. (2008). Mining business topics in source code using latent dirichlet allocation. *In Proceedings of the 1st India software engineering conference* (pp. 113-120).

Nguyen, V.A., Boyd-Graber, J. and Resnik, P. (2012). SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers –Vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics, 78-87.

Noel, G.E., Peterson, G.L. (2014). Applicability of Latent Dirichlet Allocation to multi-disk search. *Digit. Investig*. **11(1)**, 43-56.

Okon, E., Rachakonda, V., Hong, H. J., Callison-Burch, C. and Lipoff, J. (2020). Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics. Journal of the American Academy of Dermatology, **83(3)**, 803-808.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Computing, Vienna, Austria. https://www.R-project.org/.

Sutherland, I., Sim, Y., Lee, S.K., Byun, J., and Kiatkawsin, K. (2020). Topic modeling of online accommodation reviews via latent Dirichlet allocation. *Sustainability*, **12(5)**, 1821.

Titov, I. and McDonald, R. (2008). A joint model of text and aspect ratings for sentiment summarization. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: Association for Computational Linguistics, 308-316.

Vamshi, K.B., Pandey, A.K. and Siva, K.A.P. (2018). Topic Model Based Opinion Mining and Sentiment Analysis, *International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-4, doi: 10.1109/ICCCI.2018.8441220.

Wei, X. and Croft, B. (2006). LDA-based document models for ad-hoc retrieval. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06). New York, NY, USA: ACM, 178-185.

Yang, L., Qiu, M., Gottipati, S. Zhu, F., Jiang, J., Sun, H. and Chen, Z. (2013). CQARank: jointly model topics and expertise in community question answering. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, 99-108.