INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Textual Mining- Evaluation of Mann Ki Baat Repository

Mounika Kandukuri[1], **V V Haragopal**[2]*

1 INSPIRE SRF, Department of Statistics, University College of Science, Osmania University, Hyderabad, Telangana, India
2 Professor of Statistics, Department of Mathematics, BITS-Pilani, Hyderabad Campus, Hyderabad, Telangana, India

OPEN ACCESS

Received: 09-06-2020

Accepted: 14-07-2020

Published: 31-07-2020

Editor: Dr. Natarajan Gajendran

Citation: Kandukuri M, Haragopal VV (2020) Textual Mining- Evaluation of Mann Ki Baat Repository. Indian Journal of Science and Technology 13(27): 2711-2719. https://doi.org/10.17485/IJST/v13i27.879

* **Corresponding author**.
V V Haragopal
Professor of Statistics, Department of Mathematics, BITS-Pilani, Hyderabad Campus, Hyderabad, Telangana, India
haragopal_vajjha@yahoo.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

Abstract

Background and Objective: As computers and the Internet are broadly utilized in nearly every region, numerous computerized text data is produced each day. It becomes a fundamental task to explore and effectively search such massive data. The main aim of the present study is to emphasize the recurrence of topics and identifying main ideas from a popular monthly addressing radio program Mann Ki Baat by using topic modeling technique. **Data and Method**: The present study utilizes the unstructured data of Mann ki Baat from January 2020 to March 2020, collected from the PMINDIA website. This program was initiated by the Honorable Prime Minister of India, Mr. Narendra Modi. This examination uses a popular technique Topic modeling based on LDA (Latent Dirichlet Allocation). **Findings:** The results show that the method automatically extracts the main ideas and issues discussed. Besides it provides information about the most likely topics and themes discussed in each month that left an impact on people and helped in raising awareness. **Novelty:** This is a first study of the application of popular technique topic modelling on Mann ki Baat. Further, this is the first attempt to extract the ideas discussed in a social campaign using a statistical model.

Keywords: Unstructured data; preprocessing; topic modelling; latent dirichlet allocation (LDA); mann ki baat

1 Introduction

Day by day, there is a considerable increment in the advancement of digitalization, bringing about an enormous collection of a massive volume of textual data as digital libraries, news stories blog sites, online networking sites like Twitter, Facebook, and so on[1]. The vast amount of data stored in an unstructured format cannot directly be used for further processing by computers[2]. Consequently, there is a requirement for the adaption of specific pre-processing methods and algorithms to extract the hidden patterns from text data. Text Mining is the process of deriving new, previously unknown and interesting information from text data. There are various text mining approaches include topic tracking, topic extraction, summarization, and keyword extraction techniques[3].

Among these approaches, generally in text mining traditional techniques are used for searching information's or expressions. Topic modelling is different from a rule-based text mining approach and also provides a brief and quick overview of the main content consisting of a few words without reading the whole data[4]. It is an unsupervised approach that distributes the entire document automatically into a various number of topics to summarize, organize and understand the hidden ideas from an extensive collection of text documents in less time[3,5]. The significance of topic modelling is to determine the pattern of word use and to link records that share a similar pattern[6]. It presents a brief overview of the text to the readers more efficiently than document summarization[4]. This model illustrates topics by assuming that each document was formed through some generative process. There are various types of topic models existing in the literature. Still, among them, Latent Dirichlet Allocation (LDA) is the most powerful and proven to be very successful when applied to the massive content of text data[7]. To extract the most relevant information from high dimensional data in a concise manner by eliminating all the redundant information quickly and easily.

Literature suggests that several techniques and algorithms can be used to extract hidden information from a collection of documents. Some of the studies have experimented with similar ideas and also indicated that there are various representations of topic models and they are remarkably utilizing this strategy[5,6,8–10]. The primary topic model extensions have also been developed, entailing topic models for images and text[11,12], author-topic models[13], author-role -topic models[14], and hidden-Markov topic models for segregating semantic and syntactic topics[10]. In a recent study, sentiment analysis was used to study the impact of the Man ki Baat on Indian citizens[15]. This present study helps to emphasize the recurrence of topics and to identify the main ideas of Mann Ki Baat.

Mann Ki Baat is a popular social Initiative and communication revolution program pioneered by the honorable prime minister of India, Mr.Narendra Modi[16]. This ubiquitous monthly radio address program gained popularity not only because of its host but also due to its quality in the view of bringing awareness, spreading knowledge and also enabling people to know about various essential government initiatives and their importance or discussing the incidents happening across the globe and about its consequences on the masses clearly[17]. Given its stupendous success, there has been ample curiosity to know the recurrence of the topics and standard terms used in his monthly address to discover the various issues focussed and the reasons it became a sensation in every section of society within and outside India.

In this study, we used an approach to identify the topics discussed in Mann Ki Baat, which gives the basic hints to recognize main ideas, themes and their implementation and also to identify the issues that mass of people expecting to speak. The method associates with Topic modelling based on LDA to analyze the episodes of Mann Ki Baat 2020 and to extract the set of topics discussed in the campaign over time.

2 Materials and methods

This section includes techniques like: Data collection, Pre-processing, topic modelling with LDA.

2.1 Data collection

Data collection and preparation is a significant phase for any substance investigation as the nature of the crude information has a high impact on the quality of the examination. In general, the organizations of leading political candidates generate official speech transcripts and make them generously available on their official websites. In similar, the present study data was downloaded from the PmIndia.gov.in website (https://www.pmindia.gov.in/en/mann-ki-baat/). It consists of Mann Ki Baat episodes written speeches. For this analysis, we downloaded three latest episodes from January to March of 2020. We considered only these three latest episodes to see the reaction of Indian PM on Covid-19 and other major issues covered or not.

The developed framework is illustrated in Figure 1. Our analysis includes three steps Pre-processing, Topic Modelling and Post-processing (where the topic model LDA is used). We utilized the LDA technique in the topicmodels package and its dependencies in R version 3.4.3 (R Development Core Team, Vienna, Austria) to evaluate the results.

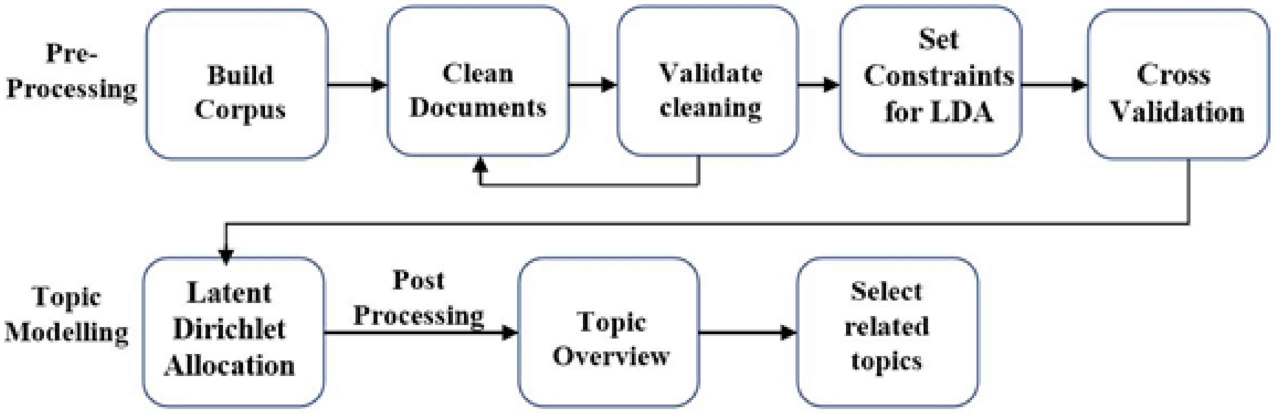

Fig 1. Process overview of Topic Modelling for Mann ki Baat

2.2 Pre-Processing

The text data collected may consist of a lot of noise and maybe in a variety of forms from individual words to multiple paragraphs consisting of special characters like punctuation marks and numbers etc. Hence, it is necessary to clean the data for extracting exact hidden information. Pre-processing is a method of resolving such issues, and it involves the removal of numeric values, removal of stop words, lower case conversion, stemming and tokenization[18,19]. Performing the pre-processing method on high dimensional text documents becomes a difficult task. Hence, this can be handled by constructing a corpus object available in tm package in R.

install. Packages (”*tm*”)

docs<-Corpus(DirSource(”path to your folder”))

The algorithmic approach of pre-processing steps is explained in Figure 2 .

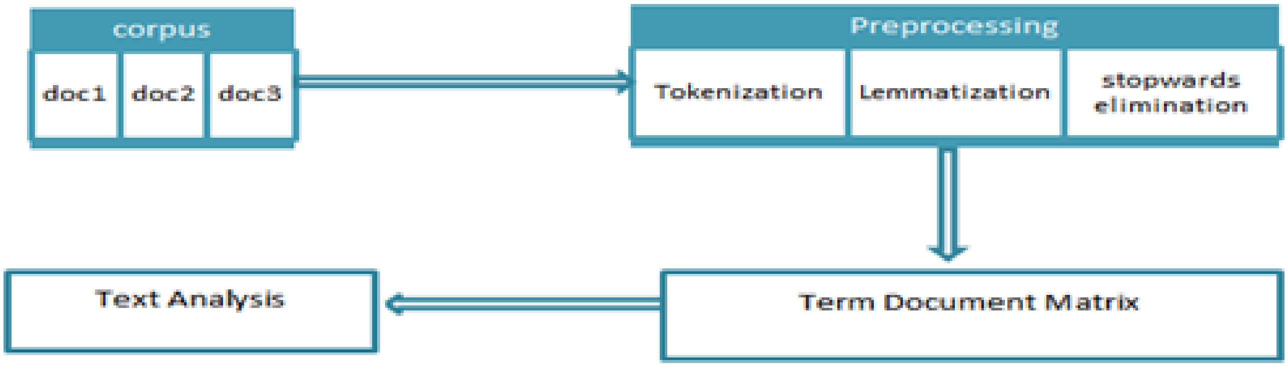

Fig 2. Pre-processing Steps of text data

2.3 Topic modelling

The theory of topic modelling comprises entities namely words, documents and corpora—some definitions of the topic model including[11,20].

1. It is a type of statistical model that determines the conceptual topics that occur in the collection of documents.
2. It is determined as probabilistic topic models, which refer to a set of statistical algorithms that disclose the hidden themes in extensive document collection and also to organize, summarize the vast amount of text data.
3. It is recurrently used techniques to discover the hidden semantic structure of text without labels. The topic consists of a set of words that would appear together in the document more or less frequently

All topic models are based on the same underlying assumption that First, Each document consists of a mixture of topics. Secondly, each topic consists of a collection of words[20].

The general framework of topic modelling is illustrated in Figure 3.

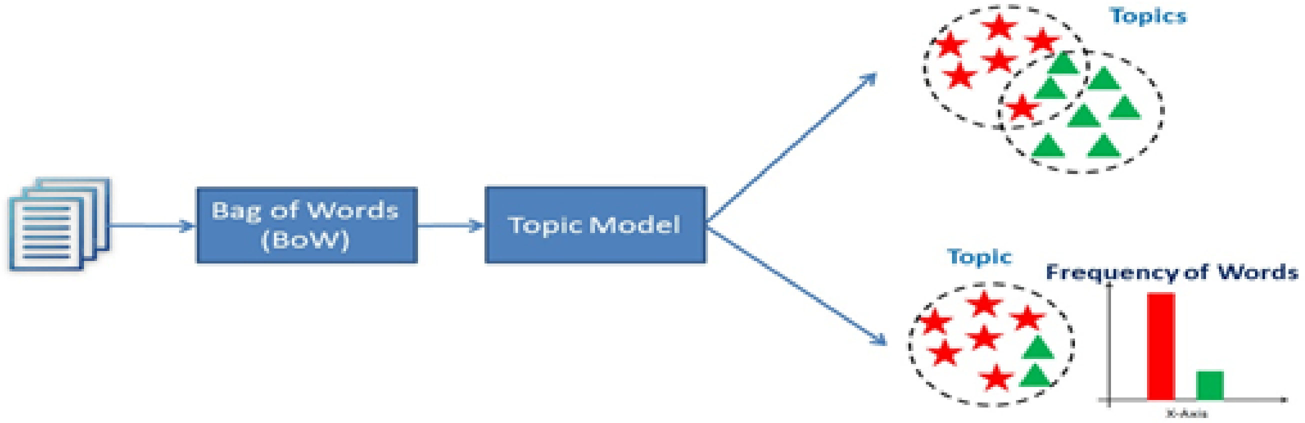

Fig 3. A Topic modelling framework

Latent Dirichlet Allocation(LDA)

Latent Dirichlet Allocation (LDA) was proposed by Blei in 2003[11]. This procedure considers the document as a mixture of topics. It is a Generative probabilistic model in which topic probability distribution over a word and the word with similar probability values lies under one topic[20–22]. Most of the studies used LDA as a preferred method for topic modeling since it works at the document level comprises of many topics. LDA is termed as a Bayesian probabilistic latent semantic analyzing method in which each item of accumulation is modeled as a finite mixture over an underlying set of topics[11]. The distribution used to attain the distribution of per-document topics is known as the Dirichlet distribution[23]. In the process, the outcomes from Dirichlet are used to allot the terms in the document to different topics.

Then LDA assumes the following generative process for a Corpus D comprising of M documents, with document d having N_d words (d $\in$ 1,2..M)[11,24,25].

1. Choose a multinomial distribution φ_t for topic t (t $\in$ (1….K)) from Dirichlet distribution with parameter β.
2. Choose a multinomial distribution θ_d for document d (d $\in$ (1….M)) from a Dirichlet distribution with parameter α.
3. For a word W_n (n ε (1,2… N_d)) in document d.

   (a) Select a topic Z_n from θ_d

(b) Select a word W_n from φ_{zn}

The Graphical model representation of LDA is, as shown in Figure 4 :

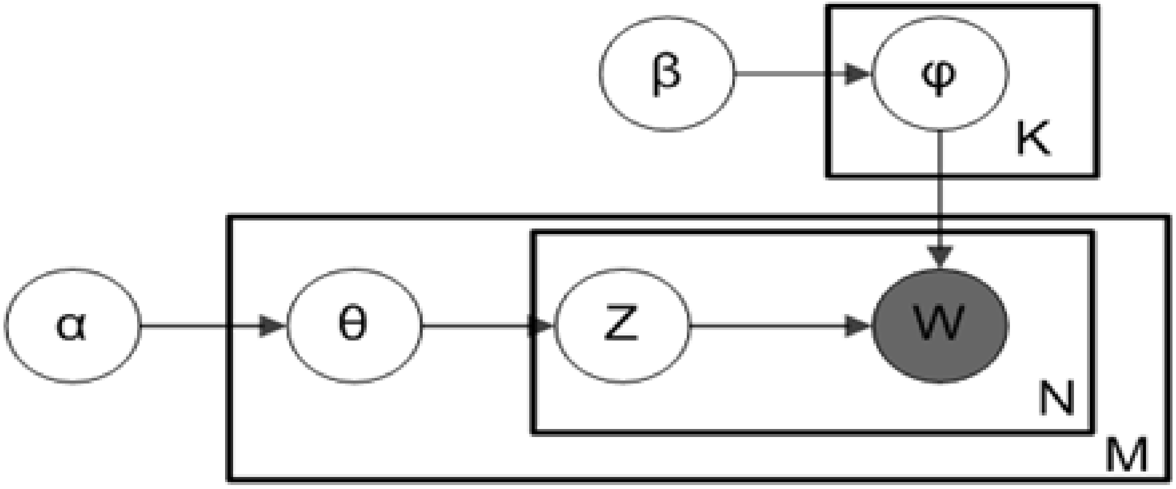

Fig 4. Graphical Model representation of LDA

Where

K- Number of topics; N-Number of words in the document

M-Number of documents to analyse; α- Dirichlet -Prior concentration parameter of the per-document topic distribution

φ(k)-Word distribute on, for topic K; θ(i)-topic distribution for topic i

Z(i,j)- topic assignment for W(i,j) ; W(i,j)-jth word in the ith document

φ and θ are Dirichlet distribution, Z, and W are multinomial distribution

In this study, we performed pre-processing steps, the process to perform LDA model for the corpus to determine the topic models requires the loading of topicmodels packages and its dependencies in R. after that, an optimal number of topics (K) is determined using a simple harmonic mean method from Martins work to group the documents based on topics. An LDA model of a corpus of 3 episodes returned after 1000 iterations of Gibbs sampling with K=24 topics (from the simple harmonic mean method), and Dirichlet hyper parameters β=0.1 and α=50/k producing each document topic distribution(10). The topic distribution created signifies the relation between topic and document.

3 Results and Discussion

The present study analysed is for three episodes of the Mann Ki Baat campaign by using topic modelling and results in the form of values and visuals based on data are presented in this section. In this initial analysis, a step was to load corpus of 3 episodes into the R environment and perform pre-processing steps like lower case conversion, removal of numeric, punctuation, and so on to clean data. Therefore, to apply the LDA model to determine the topic models, the necessary step is to determine the optimal number of topics. This is approximated using a simple harmonic mean method from Martins work and the experimental results are illustrated in Figure 5 :

As the optimal number of topics is determined, then further step is to run the LDA model on the documents. The total execution time of 3 episodes for 24 topics took nearly half an hour on a standard system. An outcome of the model is a three by 24 matrix of topic probabilities. The document to each topic with maximum probability was done in MS excel. Some topic probabilities values of documents have closer values, which indicate the similarity of topics besides the whole of 24 topics, only the topics with the highest probability were selected. The document probability values are presented in Table 1 and the plot of the document probabilities values is shown in Figure 6. From this Figure 6 topics with the highest probabilities are depicted usually by the top node in the graph plotted and are presented in Table 2.

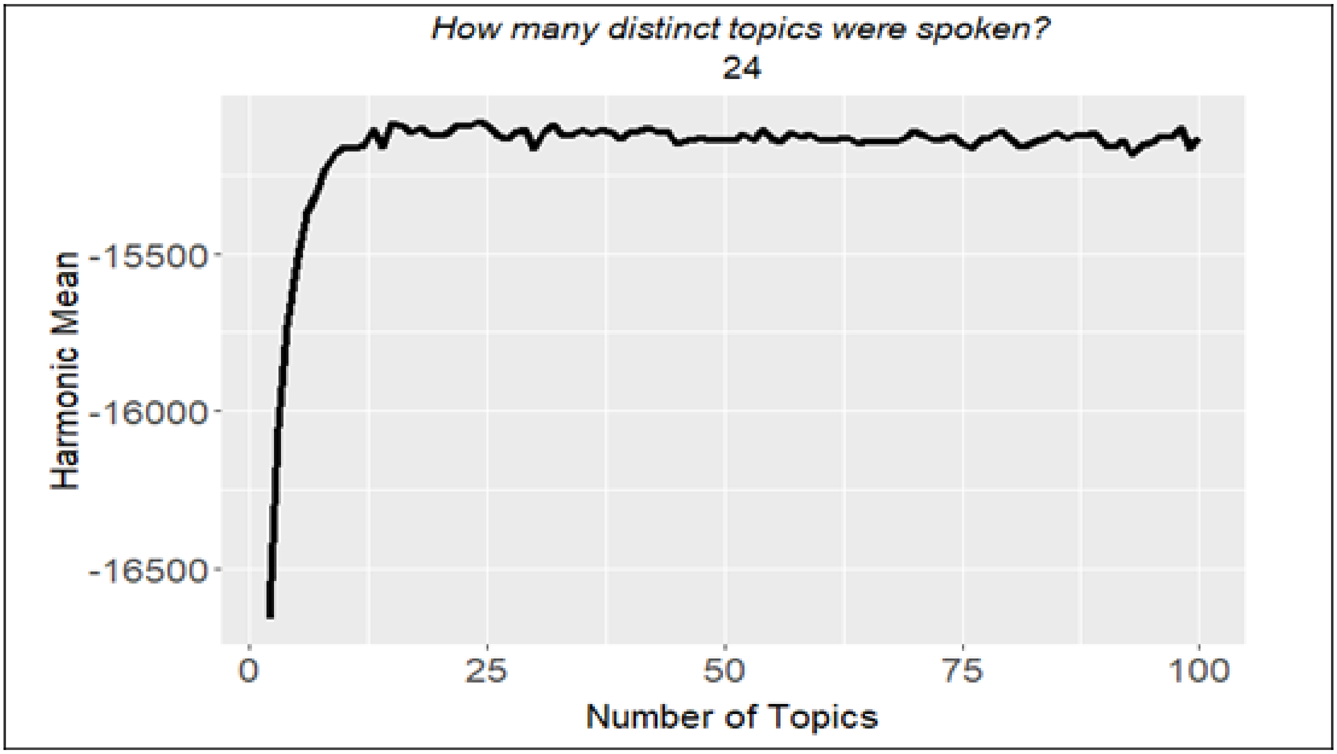

Fig 5. Optimal number of topics spoken in Mann Ki Baat 3 episodes of Jan, Feb, Mar2020

Table 1. Document Topic Probability values of 3 episodes of Mann ki Baat 2020

Month	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**
Jan	0.0224	0.0065	0.0691	0.0128	0.0022	0.0542	0.1307	0.0107
Feb	0.0158	0.2557	0.0171	0.0460	0.0080	0.0145	0.0027	0.0132
Mar	0.0072	0.0048	0.0060	0.0025	0.9174	0.0025	0.0025	0.0048
Month	**Topic 9**	**Topic 10**	**Topic 11**	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**
Jan	0.0139	0.0075	0.0595	0.0160	0.0457	0.0638	0.0203	0.0203
Feb	0.0447	0.1639	0.0040	0.0381	0.0158	0.0067	0.0119	0.0538
Mar	0.0036	0.0025	0.0025	0.0048	0.0036	0.0036	0.0025	0.0025
Month	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**	**Topic 21**	**Topic 22**	**Topic 23**	**Topic 24**
Jan	0.0532	0.0139	0.0096	0.0075	0.0595	0.2835	0.0107	0.0065
Feb	0.0211	0.0316	0.0499	0.0971	0.0054	0.0040	0.0316	0.0473
Mar	0.0025	0.0036	0.0048	0.0036	0.0036	0.0036	0.0025	0.0025

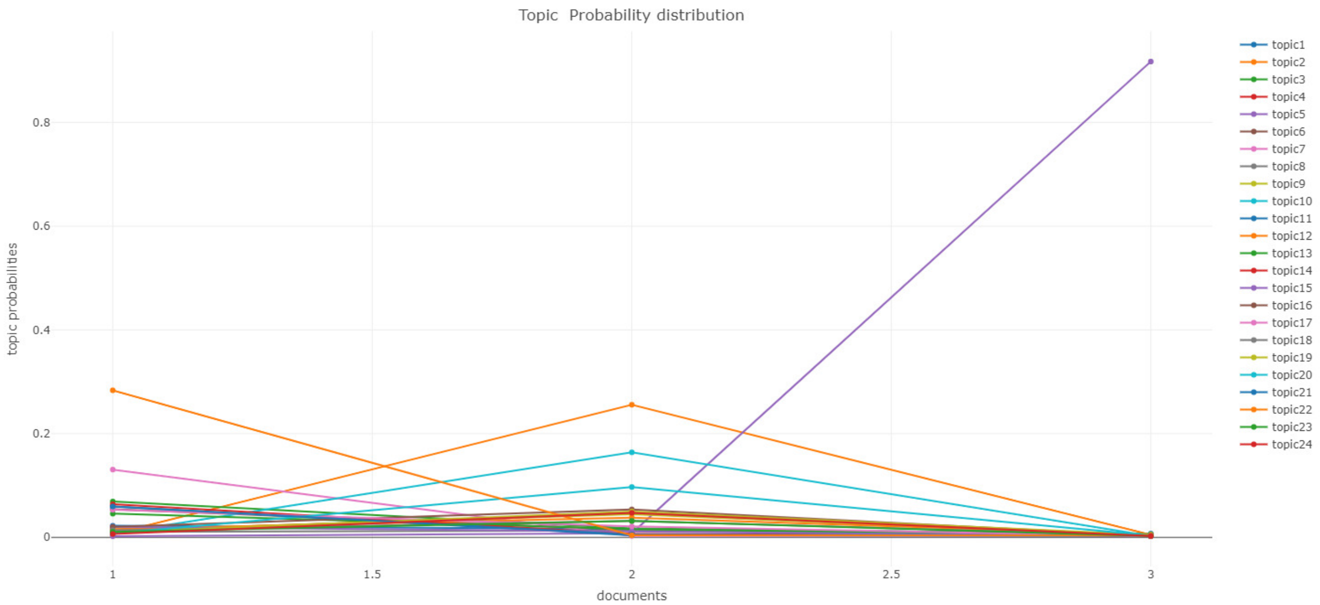

Fig 6. Document topic probability graph of 3 episodes of MannKi Baat 2020

Table 2 clearly shows that there are three topics with the highest probability values- topic 22, topic 2 and topic 5 in each month respectively, which shows the close association between documents and each topic, Table 3 shows the terms in each topic with 20 terms each.

Table 2. Topic and spoken probability for three episodes of Mann Ki Baat in 2020

Month	(Topic, Probability)
January	(Topic22,0.283528)
February	(Topic 2,0.255679)
March	(Topic 5,0.917356)

Table 3. Topic and terms of 3 episodes of Mann Ki Baat 2020

January	February	March
Topic 22	**Topic 2**	**Topic 5**
water	villag	doctor
khelo	opportun	famili
games	women	modi
agreement	amma	yes
particip	haat	corona
stori	recent	sir
congratul	salman	quarantin
game	student	patient
step	age	virus
campaign	hunar	hospit
contribut	pride	happen
endeavour	craft	servic
gift	kamya	chang
brureang	yuvika	battl
children	busi	ashok
famili	children	gupta
issu	daughter	nitesh
lake	technolog	nurs

Continued on next page

Table 3 continued

public	employ	namast
sport	fish	feel

4 Conclusion

Topic modelling plays a significant role in text mining in various fields. The fundamental commitment of our work is to explore documents comprising more than one topic in a fully programmed way. From these three episodes of the text data, we would like to quantify legitimately how this method specifically extracting the topics from sentences or sections. The work concisely coveys and also clearly shows that topic modelling with the LDA algorithm helped us greatly to derive topic models from the corpus. The results show that honourable PM in the Mann Ki Baat program addressed the nation covering only the generalized notions of various topics of the country; also, about various vital initiatives undertaken by the government. Topics such as social life, public life, lifestyle, cleanliness, Covid-19 and the environmental conversation had been addressed, which helped to spread positivity among the people and left an impact on them. The analysis also shows that there could be more focus on topics such as employment opportunities for the youth, economy, details about the GDP and development schemes. In future work, we would like to apply this strategy to more challenging textual data, inspect the pattern structure, and find the connection between the words that represent topics at a granular level.

Acknowledgments

The authors gratefully acknowledge the financial and fellowship support granted by the Department of Science and Technology (DST-INSPIRE (Innovation in Science and Pursuit for Inspired Research)), New Delhi, India for carrying out the study. The authors also wish to thank Dr.S.A.Jyothi Rani from Department of Statistics, Osmania University and Mokalla Thirupathi Reddy from National Institute of Nutrition (NIN) for encouragement and support.

Author Disclosures

There is no conflict of interest associated with this work.

Author contribution: VHG designed the study, KM undertook data collection and carried out analysis,

VHG verified analysis and encouraged to investigate and supervised the findings of this study. All authors discussed the results and contributed to the final manuscript.

References

1) Wang J, Geng X, Gao K. Study on topic evolution based on text mining. In: and others, editor. Proc - 5th Int Conf Fuzzy Syst Knowl Discov FSKD;vol. 2. 2008;p. 509–513.
2) Patel FN, Soni NR. Text mining: A Brief survey. *Int J Adv Comput Res*. 2012;2:243–248.
3) Tong Z, Zhang H. A Text Mining Research Based on LDA Topic Modelling. In: and others, editor. The Sixth International Conference on Computer Science, Engineering and Information Technology. 2016. Available from: https//doi.org/10.5121/csit.2016.60616.
4) Sajid A, Jan S, Shah IA. Automatic Topic Modeling for Single Document Short Texts. In: and others, editor. Proceedings - 2017 International Conference on Frontiers of Information Technology. 2017. Available from: https//doi.org/10.1109/FIT.2017.00020.
5) Dredze M, Wallach HM, Puller D. Generating summary keywords for emails using topics. In: and others, editor. International Conference on Intelligent User Interfaces, Proceedings IUI. 2008. Available from: https//doi.org/10.1145/1378773.1378800.
6) Lau JH, Newman D, Karimi S, et al. Latent Dirichlet allocation. In: and others, editor. Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference. 2010.
7) Mazarura J, Waal AD, Kanfer F. Topic Modelling for Short Text. 2015. Available from: http://hdl.handle.net/2263/50694.
8) Dimaggio P, Nag M, Blei D. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*. 2013. Available from: https://doi.org/10.1016/j.poetic.2013.08.004.
9) Mao XL, Ming ZY, Zha ZJ. Automatic labeling hierarchical topics. In: and others, editor. ACM International Conference Proceeding Series. . 2012.
10) Griffiths TL, Steyvers M, Blei DM. Integrating topics and syntax. *Advances in Neural Information Processing Systems*. 2005.

11) Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003. Available from: https://doi.org/10.1016/b978-0-12-411519-4.00006-9.

12) Blei DM, Jordan MI. Modeling Annotated Data. *SIGIR Forum*. 2003.

13) Rosen-Zvi M, Chemudugunta C, Griffiths T, Smyth P, Steyvers M. Learning author-topic models from text corpora. *ACM Transactions on Information Systems*. 2010;28(1):1–38. Available from: https://dx.doi.org/10.1145/1658377.1658381.

14) Mccallum A, Corrada-Emmanuel A, Wang X. Topic and role discovery in social networks. In: and others, editor. IJCAI International Joint Conference on Artificial Intelligence. 2005.

15) Garg K. Sentiment analysis of Indian PM's "Mann Ki Baat". *Int J Inf Technol*. 2020;12. Available from: https//doi.org/10.1007/s41870-019-00324-8.

16) Upadhyay S, Upadhyay N. Investigating Prime Minister Narendra Modi's Usage of Pathos in the Cyber-Physical Society – A Case of Public Relations Campaign. *Procedia Computer Science*. 2019;162:400–404. Available from: https://dx.doi.org/10.1016/j.procs.2019.12.003.

17) Saxena S, Ki M, Baat. Radio as a Medium of Communication by the Indian Premier, Narendra Modi. . *Asian Polit Policy*. 2016. Available from: https://doi.org/10.1111/aspp.12267.

18) Leopold E, Kindermann J. Text categorization with support vector machines. How to represent texts in input space? *Mach Learn*. 2002;46:423–444. Available from: https://doi.org/10.1023/A:1012491419635.

19) Munková D, Munk M, Vozár M. Data Pre-processing Evaluation for Text Mining: Transaction/Sequence Model. In: and others, editor. Procedia Computer Science;vol. 18. Elsevier BV. 2013;p. 1198–1207. Available from: https://dx.doi.org/10.1016/j.procs.2013.05.286.

20) Liu L, Tang L, Dong W. An overview of topic modeling and its current applications in bioinformatics. *Springerplus*. 2016;5. Available from: https://doi.org/10.1186/s40064-016-3252-8.

21) Blei D, Carin L, Dunson D, et al. Probabilistic Topic Models. In: and others, editor. IEEE Signal Processing Magazine;vol. 27 of 6. ;p. 55–65. Available from: https://doi.org/10.1109/MSP.2010.938079.

22) Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*. 2019;78(11):15169–15211. Available from: https://dx.doi.org/10.1007/s11042-018-6894-4.

23) Moubayed NA, McGough S, Hasan BAS, B. Beyond the topics: how deep learning can improve the discriminability of probabilistic topic modelling. *PeerJ Computer Science*. 2020.

24) Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*. 2015;16(S13). Available from: https://dx.doi.org/10.1186/1471-2105-16-s13-s8.

25) Hu DJ. . Available from: http://cseweb.ucsd.edu/~dhu/docs/research_exam09.pdf.