Analyzing government health data to explore COVID-19 management strategies – the story of India's pandemic responsiveness

Satyaki Basu Sarbadhikary [1], **Soudeep Deb** [1], **Deepti Ganapathy** [1], **Rishideep Roy** [2]

[1] *Indian Institute of Management Bangalore, Bannerghatta Main Rd, Bangalore, KA 560076, India.*

[2] *University of Essex, Wivenhoe Park, Colchester, Essex CO4 3SQ, UK*

Abstract: Effective government communication is critical during public health emergencies, particularly in mobilising a population of over a billion, as seen in India during the COVID-19 pandemic. It plays a vital role in managing information flow and ensuring compliance with measures such as lockdowns. This paper proposes a novel statistical methodology to measure and monitor the impact of government communication during such crises. We introduce a network-based approach that leverages theme-based awareness materials from the Indian Government's open data repository to analyse the dissemination and effectiveness of communication strategies. Our methodology integrates a preferential attachment model to estimate the preferences of six thematic categories and a vector autoregressive (VAR) model to examine how sentiment scores are influenced by past values and external factors. Under suitable assumptions, we empirically demonstrate how these models provide insights into communication dynamics. Furthermore, we show that the proposed approach can guide targeted interventions to mitigate and manage future acute health events. The methodology is generalisable and adaptable to other high-risk scenarios, highlighting its utility for enhancing public health communication strategies globally.

Keywords and phrases: COVID-19 preparedness; Health communication; Network Models; Preferential attachment; Vector autoregression.

.

1. Introduction

In this paper we track government communication impacting 'Lockdown Measures' by analyzing themes that are distinct yet critical to managing the flow of information in the world's largest democracy. It is imperative to remember that India saw a rapid upsurge of COVID-19 pandemic, especially during the initial year, hence the Indian government's communication and response strategies became crucial for informing the public to remain calm yet determined to fight the pandemic, raise risk awareness and prompt the public to adopt effective measures to fight the pandemic (Ganapathy). A look at how Governments communicate suggests that from vote seeking to relation building, government communication has moved from tactical to strategic, often involving a plethora of options to communicate. This seeking and sharing of information has become a primary activity in the use of the Internet, second only to communicating. The content is not limited to popular information, but serious information which includes health and education.

It is interesting to note how data such as text, images, videos can help social science scholars make inferences about social interactions, and this affects society (Grimmer and Stewart, 2013). Impact of social media use during the global coronavirus pandemic can help understand the relationships between social media, government and society. In this paper, we seek to explore the use of newsletter on a digital platform as data for the Indian government to not only manage the spread of the pandemic, but to use this data to track citizens and collect information about cluster outbreaks (Ganapathy). We attempt to see how a physical event is deeply interwoven with digital newsletters and how this "Power dynamics" is played out on the platform. The Press in India is dependent on the information from the Ministry of Information and Broadcasting, Government of India. Given India's population and mass consumption of news media there are many avenues that the government utilizes to disseminate information to its citizens. We look at newsletters

published in the portal under the Ministry of Health and Family Welfare, Government of India dashboard [1] that was made public during the COVID-19 pandemic beginning 20 January 2020.

Our analysis captures diverse communication themes, including Awareness, Official Directives, Travel Advisory, etc., representing critical public health messaging strategies during the pandemic. Our study highlights the role of newsletter title as an important medium of communication, with varying word lengths reflecting the need to adapt messaging techniques. We find that lengthy and detailed communications give way to concise formats as the pandemic progressed and the public became more familiar with safety protocols, general awareness and vaccinations (Mheidly and Fares, 2020). In fact, every such communication effort during times of risk like the COVID-19 pandemic demands structured messaging and decisive continuity in the communication, both of which were present in India's response to the pandemic in the early stages (Ganapathy). Further, studies in the extant literature show that efforts to reduce impacts of the COVID-19 pandemic came from increased adherence to behavioral interventions, targeted messages that yielded effects in this direction (Ruggeri et al., 2024).

Our focus in this paper is to propose a sophisticated statistical framework to study the progression of communication during COVID-19. First, we take advantage of a preferential attachment model to identify how different communication themes evolved over time during the pandemic with reference to newsletters issued by the Indian Government. Subsequently, we posit that a vector auto regressive approach offers an excellent fit to the data and helps us understand how the themes' frequencies, contents as well as sentiments have impact on the roll-out of future newsletters. To the best of our knowledge, such a detailed analysis has not been conducted before.

For any government or agency that is dealing with a high-risk event such as a pandemic, our study has implications for several reasons. First, it offers the ability to make sense of a surge of communication by placing these into themes. Second, these themes indicate the relevance to the physical event happening on the ground, and demonstrate a response to mitigate the risk. Finally, when the authority that is handling the public health outbreak is making rapid transitions between themes, such as in the case of our analysis, when we see a significant increase in communications under Awareness and Citizens themes, we understand that this shift in preferences is shaped by informed decisions coming from the field, or in this case number of hospitalizations or deaths.

The rest of the paper is organized as follows. Section 2 is about the data and related exploratory analysis. Section 3 explains the methods and processes used to carry out our analysis. Section 4 enlists the results we have obtained from our analyses. Finally, in Section 5, we end this paper with some important concluding remarks and future scopes of our work.

2. Data and exploratory analysis

As alluded to above, we look at all public newsletters published in the portal under the Ministry of Health and Family Welfare (MoHFW), Government of India dashboard, starting from 20 January 2020 until 10 February 2023. There is a total of 407 such newsletters, published at different time-points. They have come in different formats, languages and are categorized into nine themes by MoHFW: *Awareness*, *Travel Advisory*, *Hospitals*, *Citizens*, *Employees*, *Inspirational*, *Official*, *Training*, *Psychosocial*. Each theme is related to announcements with information and guidelines pertaining to specific subjects. For example, the theme *Awareness* contains newsletters with information intended for raising awareness among general public during COVID-19. The theme *Citizens* is focused on citizenship behavior and restrictions. Similarly, the rest of the themes provide similar subject-specific information. In order to maintain consistency in our analysis, noting that the contents inside the newsletters may include information in various languages, we consider the main subject matters of the newsletters, all of which are in English. However, future research could steer

[1] https://covid19dashboard.mohfw.gov.in/

towards considering the entire content for better analysis, which will require more advanced methodology from the domain of natural language processing.

To address challenges posed by themes with varying frequencies, we employ a consolidation approach while cleaning and pre-processing the data. High-frequency themes, namely *Official*, *Travel Advisory*, *Hospitals*, *Citizens*, and *Awareness*, are maintained individually, while the three least frequent ones (*Psychosocial*, *Inspirational*, and *Training*) are aggregated into a broader category which we define as *Others*. This approach not only streamlines the analysis but ensures that the focus remains on dominant trends in public health communication without over-representing less frequent themes. Our refined dataset now serves as a comprehensive foundation to analyze the evolving communication dynamics and priorities during the COVID-19 pandemic. Figure 1 below presents a visualization of the dataset, emphasizing the relative frequency of the six primary communication themes observed during the COVID-19 pandemic. The themes with higher occurrence rates, i.e., *Awareness*, *Hospitals*, *Citizens*, *Official*, and *Travel Advisory*, are distinctly represented, facilitating a more straightforward analysis of public health messaging trends allowing for better interpretation of the dominant communication strategies and thus highlighting the focal points of information dissemination throughout the pandemic.

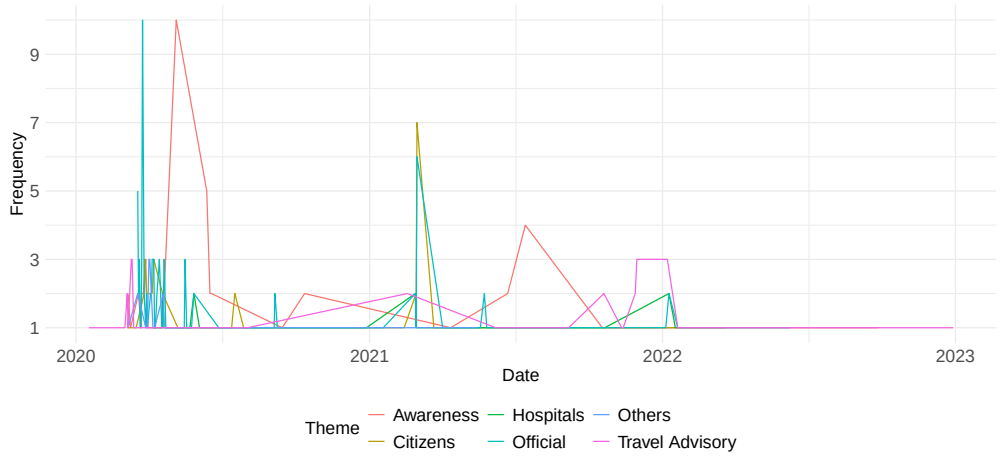

Fig 1: Frequency of public announcements in six themes over the entire time period.

From the graph, we observe that *Awareness* and *Official* are the most prominent themes, particularly during the early stages of the pandemic in 2020. Both themes experienced rapid increases in mentions, reflecting the urgency of public communication and official guidance during the onset of the COVID-19 crisis. After this initial surge, the frequencies of these themes drop significantly, though *Official* and *Awareness* related communications remain relatively steady compared to the rest. *Travel Advisory* announcements also show an initial spike, especially in early 2020, likely due to the travel restrictions and advisories issued worldwide. However, this theme tapers off quickly, reflecting stabilization as the restrictions became a more routine aspect of pandemic management. *Hospitals* is another important theme, indicating a strong concern with healthcare facilities, peaking in certain periods but remaining somewhat active throughout the timeline. The occurrences of these themes tend to decrease after the first year of the pandemic, with sporadic spikes occurring in response to significant pandemic events or health policy changes.

3. Methodology

The concepts of networks and vector autoregressive (VAR) models are fundamental to this study as they provide a nuanced framework for understanding the interconnections and dynamic interactions among public

health communication themes throughout the COVID-19 pandemic. By representing themes as nodes within a network, we can explore the intricate associations and flow of information, identifying patterns of thematic prominence, mutual reinforcement, and adaptability within a responsive communication strategy. In this regard, we rely on the concepts of preferential attachment models. In parallel, the VAR model helps in capturing the temporal inter-dependencies among themes providing insights into the interconnected nature of the arrival of messages. By treating all themes as interconnected, the vector-valued time series approach reveals how shifts in one theme affect others, offering a time-sensitive view of communication dynamics in response to pandemic developments. Below, in Section 3.1 we will walk through the details of the procedure undertaken using the preferential attachment model to identify how different communication themes evolved over time during the pandemic with reference to newsletters issued by the Indian Government. Then, in Sections 3.2 and 3.3, we present the methods to help us understand how the themes' frequencies, contents as well as the sentiments have impact on the roll-out of future newsletters respectively.

3.1. Preferential attachment model to study the flow of communication

Throughout this section and thereafter, to avoid confusion with the conventional terminology of the network theory, we shall use the terms "theme" and "topic" interchangeably. Our first objective is to determine the relative importance of the six topics and the likelihood of a new communication cropping up on any of these six topics as a function of time. The nature of the data naturally motivates a dynamic model, with a new topic coming at every new epoch. Our objective is to analyze these topics and identify the influence of each of them. Since the themes are interconnected, we use the preferential attachment (PA) model, which is a dynamic network approach, to determine this. However, it must be noted that one of the limitations of our data is that for newsletters released on the same day, the exact order of the communications is not determinable. To handle this issue, we show that our estimation process is not affected by the ordering of the data. This is guaranteed by the second theoretical result of this section.

The concept of PA has been used extensively to model growing networks with hubs. One of the most notable applications of it is the modeling of the World Wide Web, which developed from a single node (Giammatteo et al., 2010). Some other well-known applications of PA are found in studying the structure of Wikipedia (Capocci et al., 2006), the growth of referencing research materials via citation networks (Ming-Yang, Guang and Da Ren, 2010; Barabási, 2013), and in the field of strategic alliances through networks (Rossmannek and Rank, 2021). It gives a sophisticated framework for simulating and understanding the growth dynamics of networks where some nodes become increasingly prominent over time due to their connections. In this model, nodes with higher connectivity (equivalently, degree in the network literature) are more likely to attract new connections, a phenomenon often described as "the rich get richer" (Lyon and Mahmoud, 2020). This approach is grounded in the concept that popularity or relevance often reinforces itself, leading to network hubs that serve as key points of influence. The model assigns each node a probability of receiving new connections proportional to its existing degree. In each iteration, a new node is introduced and attaches itself to an existing node based on this probability, which reflects how frequently established nodes continue to attract new links. This probability function can be tuned to various forms, adapting the model to reflect different real-world network structures. By representing information dissemination or communication channels, the network offers a nuanced view of how influence, prominence, and connectivity evolve. This makes it highly relevant in studies like ours, where public health messaging themes emerge and gain prominence in response to an evolving pandemic. Applying the PA model enables us to capture the mechanisms by which certain themes, once established, continue to engage public attention and reinforce critical messaging in the collective consciousness over time.

For the main analysis, noting that the newsletters were published at irregular spaced discrete time points (say, at $t \in \mathscr{T}$), every time point as well as the main topic of the individual newsletter, represents the data we plan to use for that time point. Hence, every individual newsletter is denoted as a node. At every time

point, a new node gets attached to one of the themes via an edge if its theme already exists in the network. If not, this node initiates a new cluster for its theme. This process takes place with some probability, which we define below in the formal definition of the model. Before that, a simple visualization of the model in the context of our work is presented in Figure 2 for the reader's interest. There, it is assumed that the six themes have degrees (frequencies of newsletters) $1,2,2,3,1,3$ at a certain point of time. Then, the next newsletter, shown on the right hand side of the figure, can get attached to any of the six themes following the theory of PA network, that is, the probability of the next newsletter being in the theme *Travel Advisory* is proportional to $1/12$, the same for the theme *Awareness* is proportional to $2/12$, and so on.

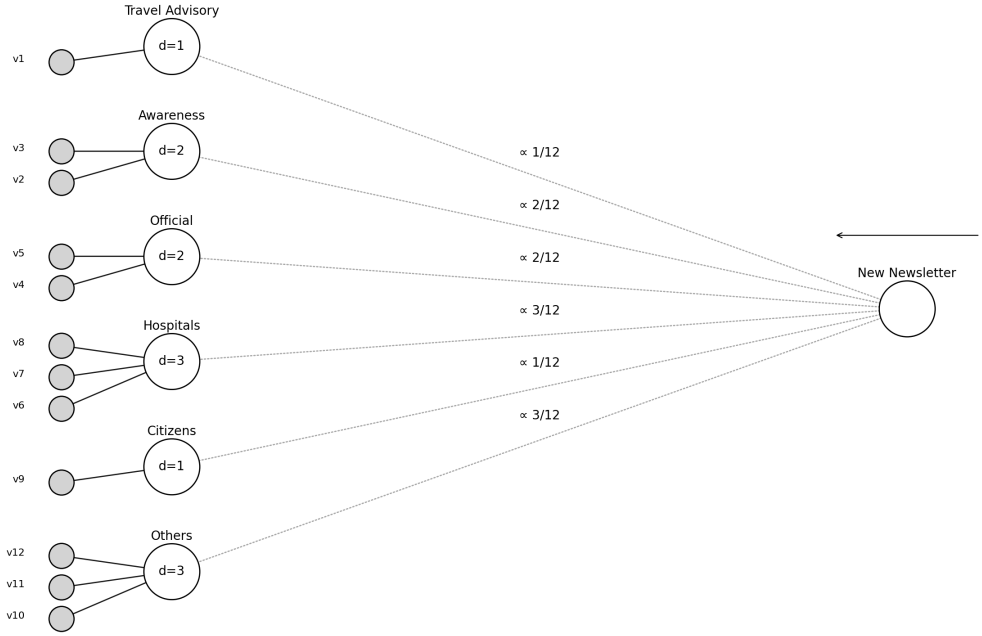

Fig 2: An illustration of the preferential Attachment network: a new newsletter can get attached to the six themes with probabilities proportional to the current frequencies of the themes.

To formally define our modeling framework, for $t \in \mathscr{T}$, let $\{v_1, v_2, \ldots, v_t\}$ represent the sequence of nodes that appear until that time point. We represent the degrees of these nodes interpreted as the frequency of revisiting a particular theme by $\{d_1^t, d_2^t, \ldots, d_t^t\}$ considering the evolution of the process until time point t. The next node v_{t+1} then can either join any of the previous nodes $v_k, k \leqslant t$, or can remain unattached to any previous node. We assume that the probability that it gets attached to v_k is proportional to $g\left(d_k^t\right)$ where $g(\cdot)$ is a function mapped from $\mathbb{N}$ to $\mathbb{R}_+$, i.e., from the set of natural numbers to the set of positive real numbers. Note that there are different functional forms that can be assumed for $g(\cdot)$. For example, in the previous illustration, we used the identity function $g(x) = x$, which has been a popular choice in the extant literature (see, e.g., Barabási and Albert, 1999). On the other hand, Móri (2002) considered the functional form $g(x) = x + \alpha$, for $\alpha > -1$. There has also been some work on the estimation of the functional form of $g(\cdot)$. For instance, Gao et al. (2017) considered a non-parametric approach, while Gao and van der Vaart (2021) developed a parametric method.

In this paper, we are specifically going to use an empirical estimate, along the lines of Gao et al. (2017), for the probabilities of a new node getting attached to any of the previous topics. This estimate is updated at every point in time for each of the different topics. The main result in this regard is that, uniformly over

the order of communications, the empirical estimate is consistent for the asymptotic probabilities of a new node getting attached to a topic. Before explaining the main theory, we find it prudent to emphasize that our work differs from a conventional PA model in one particular aspect, where the number of clusters, which also equals the total number of topics, is considered fixed. We shall denote this fixed number as K. Let $N_k(n)$ be the number of times a node (i.e., newsletter) is attached to the k^{th} cluster (i.e., topic), until the n^{th} time point. The empirical estimate of the attachment probability for every cluster is then given by

$$\hat{r}_k(n) = \frac{N(k)}{n}. \tag{1}$$

We also define r_k as the long-run probability of attachment under stable circumstances. Our main result is on the consistency of the estimator $\hat{r}_k(n)$.

Result 1. *For $k \in \{1,2,\ldots,n\}$, the empirical estimator of attachment probability for the k^{th} cluster, $\widehat{r_k}(n)$, converges to r_k, with probability 1, as $n \to \infty$.*

The above result discusses the theoretical properties of the empirical estimator of attachment probability given by (1), which changes with every new node in the network. The result, however, is proven for assuming that at every time-point, there is exactly one newsletter, whereas in our data, there is a possibility that on some days, there will be multiple newsletters. To make things more complicated, under those circumstances, the data does not mention the order in which the communications are published within a day. Interestingly, we can show that the result is not affected by the order in which the nodes come into the network within a time-point. Under certain assumptions on the mapping function, the model can indeed show the property of exchangeability, indicating that the model's behavior and associated estimation process are unaffected by the order of arrival of new nodes at a given time-point.

Result 2. *If the mapping function $g : \mathbb{N} \to \mathbb{R}^+$ satisfies the condition that $g(y+1) - g(y) = g(x+1) - g(x)$ for all $x, y \in \mathbb{N}$, then every new node coming into the network at a certain time-point can come in any order, and the formation of clusters is not affected by this.*

The proofs of both results are technical in nature and are provided in the Appendix. Their interpretation, however, is of the essence here. These results highlight how our model captures the evolving importance of public health communication themes while ensuring its reliability even when the exact timings of the messages are uncertain. Essentially, the first result shows that as we observe more communications over time, the probability of a theme being referenced or gaining attention stabilizes, giving us a clear picture of which themes consistently draw focus and by when the focus stabilized during the pandemic. The second result reassures us that the sequence in which these communications are sent or received during a particular time period does not affect the overall insights. This is especially important in our study, and guarantees that we can confidently analyze the relationships between themes and their prominence. Together, these results ensure that our model provides a reliable and robust framework for understanding how different themes in public health messaging interact and evolve over time.

3.2. Modeling empirical probability estimates using vector autoregression

Building upon our previous analysis of PA model in studying the evolution of communication, to deepen our understanding of the dynamic interactions among various COVID-19 related public communication themes, we extend our analysis by employing a vector autoregression (VAR) model (Hamilton, 2020). While the previous exploration with the PA model provided insights into how certain themes gained prominence over time, it primarily focused on the evolution of individual themes without addressing inter-dependencies. In contrast, the VAR model enables an integrated analysis of multiple time series as endogenous variables,

allowing us to capture the mutual influence among the six communication themes: *Awareness*, *Citizens*, *Hospitals*, *Official*, *Travel Advisory*, and *Others*. We also include two exogenous variables in this model. They are the numbers of COVID-19 cases and the number of deaths at the specific time-points, which help is understanding how the progression of the pandemic impacted the empirical importance and interconnectedness of the six themes.

To prepare the data for modeling, the estimated attachment probabilities $\hat{r}_k(n)$ are first transformed using an arcsin transformation given by

$$Y_k(t) = \arcsin\left(\sqrt{\hat{r}_k(t)}\right) \tag{2}$$

Such transformation is quite effective to stabilize the variance, and is commonly done when dealing with data related to sample proportions (see, for example, Deb, Roy and Das, 2024). Further, COVID-19 case and death numbers are log-transformed to manage the large variations in their scale, smoothing the data for better analysis. This is also a common practice in similar statistical studies related to COVID-19 datasets (Rawat and Deb, 2023). With these transformations taken into account, the structure of a VAR(p) model, where p denotes the number of lags, can be represented mathematically as:

$$\boldsymbol{Y}_t = \boldsymbol{c} + \sum_{i=1}^{p} \Phi_i \boldsymbol{Y}_{t-i} + \Gamma \boldsymbol{X}_t + \boldsymbol{\varepsilon}_t \tag{3}$$

In (3), $\boldsymbol{Y}_t$ is a vector of the six endogenous variables $Y_k(t)$, for $k = 1, 2, \ldots, 6$, observed at time t. In the context of this analysis, these variables correspond to the six communication themes that were prominent in public communications during the COVID-19 pandemic. Next, the vector $\boldsymbol{c}$ consists of 6 constants or intercept terms that capture the baseline levels for these communication themes, effectively setting a starting point for the system. The matrices Φ_i, for $i = 1, \ldots, p$, $(p = 1)$ are 6×6 coefficient matrices that quantify the influence of lagged values of the endogenous variables on their current values. The term $\boldsymbol{X}_t$ represents the design matrix of exogenous variables and Γ is a 6×2 matrix of parameters capturing the impact of these variables on $\boldsymbol{Y}_t$. Finally, $\boldsymbol{\varepsilon}_t$ is the random noise terms in the model.

Throughout this paper, we shall use $p = 1$ to ensure reliability of the model, as a higher lag order demands more data to perform the estimation correctly. This model is going to be vital to our study because it helps us understand how different themes in public health communication, like *Awareness* or *Hospitals*, interact and influence each other over time. Instead of looking at these themes in isolation, the model allows us to see the bigger picture – how they work together and adapt as the pandemic evolves. For example, a focus on healthcare resources might influence public awareness or change the tone of advisories, showing a ripple effect across communication efforts. By also considering external factors like the number of COVID-19 cases and deaths, the model provides insights into how these real-world events shape the way messages are communicated. This is important for understanding not just what was said, but why it was said at specific times and how it resonated with the public. Overall, the VAR model gives us a clearer understanding of the flow and adaptability of public health messages, which is essential for improving communication strategies in future crises. Note that a similar model framework will be used in the next segment of the analysis to understand the interaction and evolution of communication sentiment with respect to these themes over time in the course of the pandemic.

3.3. Analysis of newsletter topic

The topics of these newsletters are mediums of communication from the government to public. The previous analysis helps us in understanding the flow of communication themes during the testing times. To further understand the effectiveness of the government's communication strategy employed during the pandemic in information dissemination, we next look into two important aspects. First one is simply the number of

words in the subject matter of each newsletter, which offers an idea of how detailed the information was in different points of time. Hereafter, this metric will be called as content length.

Second, to capture the emotional tone of each newsletter, we conduct a sentiment analysis of the subject matter in every newsletter. To quantify this, we compute the sentiment scores for each newsletter across the given period of time, and connect it to the six themes. The sentiment score calculation initiates with the assignment of positive or negative scores to each word of the newsletter content based on a predefined lexicon, followed by the computation of the average of these scores based on the length of the content. For all calculations in this regard, we have utilized the R package `Syuzhet` (Jockers, 2015). It is imperative to note that sentiment analysis is quite common and effective in understanding health communication (Zunic, Corcoran and Spasic, 2020). Even in the context of COVID-19, multiple researchers have addressed various aspects of communication strategies during the pandemic (see, for example, de Las Heras-Pedrosa, Sánchez-Núñez and Peláez, 2020; Bulut and Poth, 2022).

After calculating the sentiment scores, as done in Section 3.2, we explore the dynamic interdependence of the sentiments on the topics of the newsletters, associated with the six themes and their evolution over time. To that end, we employ the VAR model of order 1, by treating each theme's sentiment score as an endogenous variable and confirmed COVID-19 cases as an exogenous variable. The intercept vector or the number of deaths are not utilized in this case, as the model showed poorer fit with them. Following the same notations as before, the model in this case can be represented as

$$\boldsymbol{Z}_t = \Theta_1 \boldsymbol{Z}_{t-1} + \Lambda C_t + \boldsymbol{\varepsilon}_t, \tag{4}$$

where $\boldsymbol{Z}_t$ is a vector representing the sentiment scores of the six communication themes at time t, and the matrix Θ_1 is the coefficient matrix that quantifies the effect of lagged sentiment scores from time $t-1$ on the current values at time t. Then, C_t is the number of COVID-19 cases and Λ is a 6×1 vector encompassing its impact on the sentiment scores. Finally, the term $\boldsymbol{\varepsilon}_t$ follows the similar definition and dimensions as discussed in the preceding section.

4. Results

4.1. Empirical estimates from the preferential attachment model

The results of the empirical probability estimates for the six key themes (*Awareness*, *Hospitals*, *Citizens*, *Official*, *Travel Advisory*, and *Others*) over time, beginning in early 2020 and extending to early 2023 is shown in Figure 3, where each line tracks the empirical attachment probability for the six themes and their evolution over the entire timeline (presented in the right panel). We also plot the same for the first six months (presented in the left panel) for having a better understanding of the initial days of the pandemic.

For the probabilities over the entire timeline, at the start, we can see that there is a sharp spike in the *Travel Advisory* theme, indicating a rapid increase in communications or updates related to travel restrictions during the early months of the COVID-19 pandemic. This peak quickly falls and stabilizes as travel restrictions became a norm. On the other hand, the themes *Awareness*, *Citizens*, *Hospitals*, and *Official* show an initial rise during the pandemic's onset and gradually stabilize with time. The *Official* theme sees consistent high relevance, reflecting the importance of official guidance and updates throughout the pandemic. Meanwhile, *Awareness* and *Citizens* also stabilize, indicating their sustained presence in communications. Turning attention to the other communications (given by the *Others* theme), which comprises less frequent categories like Employees, Psychosocial, Inspirational and Training, we see that the graph maintains a relatively low and stable occurrence after an initial rise, as these themes became less central to pandemic communications over time. Overall, we observe that all themes' attachment probabilities stabilized in the middle of 2022, which is in fact directly connected to the theory presented in Section 3.1.

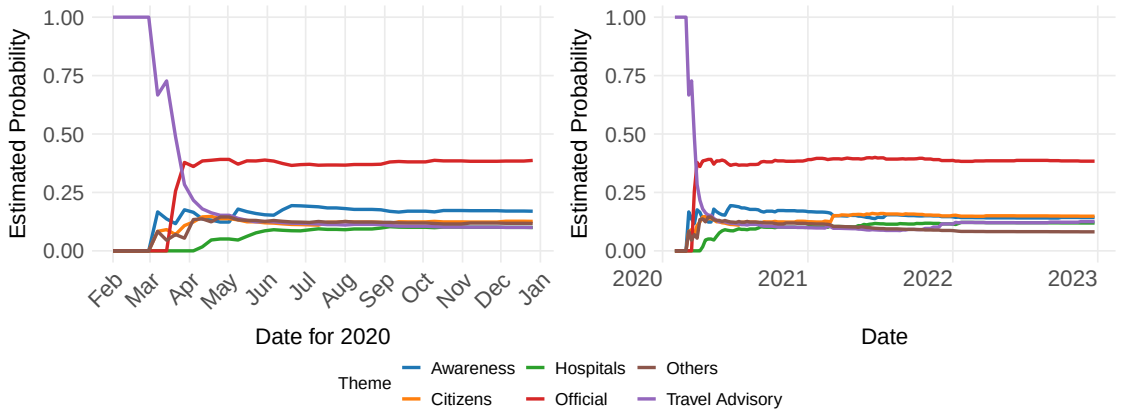

Fig 3: Visualization of the empirical attachment probability for the themes during the first six months of the pandemic (left panel), and over the entire timeline (right panel).

If we keep the focus only in the first six months, we see a lot of movement in the six themes. First, the dominance of the *Travel Advisory* theme is evident, with its probability starting near 1.0 before sharply declining as the pandemic progressed. As mentioned earlier, it reflects an initial prioritization of urgent travel restrictions and guidelines to contain the virus's spread. As time passed, themes like *Awareness*, *Citizens* and *Hospitals* gained prominence. The spike in *Awareness* theme is reflective of the efforts to educate the public about the virus, precautionary measures, and hygiene practices. Similarly, the increasing probability of the *Hospitals* theme highlights the growing emphasis on healthcare infrastructure, availability of medical resources, and updates on the preparedness of the healthcare system during this period. The probability of *Official* theme rose steadily from mid-March continuing till the beginning of April after which it remained relatively stable throughout. The initial spike suggests the increased role of government directives during that time encompassing key announcements, regulatory updates, and policy measures aimed at managing the pandemic's immediate impact.

4.2. Estimation of the VAR model for empirical estimates

As mentioned in Section 3.2, we implement the VAR model to study the dynamics of the estimated probabilities for the six themes, following equations (2) and (3). These results are summarized in Table 1. Each row of the table corresponds to a specific parameter. Overall, these parameters quantify how much one-unit change in the lagged values of the estimated attachment probabilities (with the arcsine transformation) in various communication themes or an exogenous variable (confirmed cases, COVID-19 deaths) affects the current level of the estimated attachment probabilities. The table includes key statistical measures, such as the estimate of the coefficient, its standard error (SE), t-statistic, and p-value, which together help determine the significance of each coefficient.

We observe that self-persistence is a strong feature across all themes except *Official*, with significant positive coefficients for the respective lagged terms. Specifically, the highest self-persistence is observed in the themes *Citizens* (estimated coefficient 0.6952, t-statistic 14.92, $p < 0.001$) and *Travel Advisory* (estimated coefficient 1.0226, t-statistic 13.06, $p < 0.001$). It suggests that once travel-related communications or information about citizenship behavior begin, they are likely to persist over time. Similarly, *Hospitals* (estimated coefficient 0.4536, t-statistic 11.48, $p < 0.001$), and *Awareness* (estimated coefficient 0.3366, t-statistic 4.2, $p < 0.001$) themes also demonstrate considerable self-sustainability, indicating that these themes tend to remain in focus once initiated suggesting a sustained communication strategy in these areas, likely driven by the evolving nature of pandemic management.

TABLE 1

Results of the VAR model for analyzing the time series of attachment probabilities of the six themes. Significance of the coefficients are indicated as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Response	**Regressor**	**Estimate**	**SE**	**t-statistic**	**p-value**
Awareness	Constant	-0.8657^{***}	0.1382	-6.2647	< 0.0001
	Awareness (lag 1)	0.3366^{***}	0.0801	4.2038	< 0.0001
	Citizens (lag 1)	-0.6251^{***}	0.0705	-8.8603	< 0.0001
	Hospitals (lag 1)	0.2637^{***}	0.0595	4.4350	< 0.0001
	Official (lag 1)	0.1393^{**}	0.0463	3.0056	0.0031
	Travel Advisory (lag 1)	-0.1051^{*}	0.0450	-2.3377	0.0208
	Others (lag 1)	-0.0268	0.0746	-0.3592	0.7200
	Confirmed Cases	0.0796^{***}	0.0126	6.3455	< 0.0001
	COVID-19 deaths	-0.1063^{***}	0.0162	-6.5548	< 0.0001
Citizens	Constant	-0.4633^{***}	0.0913	-5.0764	< 0.0001
	Citizens (lag 1)	0.6952^{***}	0.0466	14.9216	< 0.0001
	Awareness (lag 1)	-0.1226^{*}	0.0529	-2.3177	0.0219
	Hospitals (lag 1)	-0.0893^{*}	0.0393	-2.2733	0.0245
	Official (lag 1)	0.1456^{***}	0.0306	4.7567	< 0.0001
	Travel Advisory (lag 1)	0.0669^{*}	0.0297	2.2523	0.0258
	Others (lag 1)	0.0113	0.0493	0.2287	0.8194
	Confirmed Cases	0.0275^{**}	0.0083	3.3180	0.0012
	COVID-19 deaths	-0.0247^{*}	0.0107	-2.3074	0.0225
Hospitals	Constant	-0.2764^{**}	0.0918	-3.0094	0.0031
	Hospitals (lag 1)	0.4536^{***}	0.0395	11.4801	< 0.0001
	Awareness (lag 1)	-0.1010	0.0532	-1.8972	0.0598
	Citizens (lag 1)	0.1219^{*}	0.0469	2.6004	0.0103
	Official (lag 1)	-0.0758^{*}	0.0308	-2.4598	0.0151
	Travel Advisory (lag 1)	0.0745^{*}	0.0299	2.4932	0.0138
	Others (lag 1)	0.2788^{***}	0.0496	5.6239	< 0.0001
	Confirmed Cases	-0.0946^{***}	0.0083	-11.3455	< 0.0001
	COVID-19 deaths	0.1369^{***}	0.0108	12.6992	< 0.0001
Official	Constant	-2.1035^{***}	0.3253	-6.4668	< 0.0001
	Official (lag 1)	0.1505	0.1091	1.3793	0.1699
	Awareness (lag 1)	-0.5631^{**}	0.1885	-2.9879	0.0033
	Citizens (lag 1)	-0.4624^{**}	0.1660	-2.7850	0.0061
	Hospitals (lag 1)	-0.5716^{***}	0.1399	-4.0850	< 0.0001
	Travel Advisory (lag 1)	-0.3880^{***}	0.1058	-3.6661	< 0.0001
	Others (lag 1)	-0.3414	0.1756	-1.9443	0.0538
	Confirmed Cases	-0.1005^{***}	0.0295	-3.4004	0.0009
	COVID-19 deaths	0.1299^{***}	0.0382	3.4040	0.0009
Travel Advisory	Constant	1.4818^{***}	0.2407	6.1567	< 0.0001
	Travel Aadvisory (lag 1)	1.0226^{***}	0.0783	13.0574	< 0.0001
	Awareness (lag 1)	0.5791^{***}	0.1394	4.1528	< 0.0001
	Citizens (lag 1)	0.8240^{***}	0.1229	6.7068	< 0.0001
	Hospitals (lag 1)	-0.1183	0.1035	-1.1427	0.2551
	Official (lag 1)	-0.2160^{**}	0.0807	-2.6759	0.0083
	Others (lag 1)	0.2605^{*}	0.1299	2.0049	0.0469
	Confirmed Cases	-0.1990^{***}	0.0219	-9.1030	< 0.0001
	COVID-19 deaths	0.2453^{***}	0.0283	8.6820	< 0.0001
Others	Constant	-0.0418	0.1447	-0.2887	0.7732
	Others (lag 1)	0.4402^{***}	0.0781	5.6338	< 0.0001
	Awareness (lag 1)	0.2068^{*}	0.0839	2.4666	0.0148
	Citizens (lag 1)	-0.1053	0.0739	-1.4246	0.1565
	Hospitals (lag 1)	0.2121^{***}	0.0623	3.4059	0.0009
	Official (lag 1)	0.2659^{***}	0.0485	5.4780	< 0.0001
	Travel Advisory (lag 1)	-0.0603	0.0471	-1.2801	0.2026
	Confirmed Cases	0.0462^{***}	0.0131	3.5144	0.0006
	COVID-19 deaths	-0.0791^{***}	0.0170	-4.6587	< 0.0001

Cross-theme impacts are also notable in some cases. For example, *Citizens* communications are found to have a negative influence on *Awareness* in the following period (estimated coefficient -0.6251, with

$p < 0.001$), implying a potential trade-off between broad awareness campaigns and targeted messaging aimed at individuals. This suggests a shift in communication strategy, where information towards the *Awareness* theme is reduced and there is a focus towards the *Citizens* theme. Similarly, the *Official* theme has a significant positive impact on the *Citizens* theme (estimated coefficient 0.1456, with $p < 0.0001$), indicating that formal official announcements often trigger more engagement with citizens. Communication about *Hospitals* also positively impacts *Awareness* (estimated coefficient 0.2637, with $p < 0.001$) but reduces the emphasis on broader *Official* messaging (estimated coefficient -0.5716, with $p < 0.001$), reflecting a shift toward more urgent healthcare messaging when the pandemic conditions worsen. This trend is reinforced by the strong positive correlation between deaths and hospital communications (estimated coefficient 0.1369, $p < 0.001$), suggesting that rising fatalities necessitate more hospital-related updates. When cases increased, *Awareness* (estimated coefficient 0.0796, $p < 0.001$) and *Citizens* themes saw a surge in messages (estimated coefficient 0.0275, with $p = 0.0012$). However, while rising deaths saw an increase in hospital-related communications (estimated coefficient 0.1369, with $p < 0.0001$) and *Official* responses (estimated coefficient 0.13, with $p < 0.001$), they reduced *Awareness* efforts (estimated coefficient -0.1063, with $p < 0.001$).

4.3. Exploratory analysis of content length and sentiment of newsletters

We now move on to understand how detailed and focused the newsletters were during the pandemic. Recall that for our analysis, we look at the subject matter of the newsletters. The top panel in Figure 4 offers an analysis of the variation in the content length for all newsletters across different communication themes. We can see that from 2022 onward, the numbers reduced across three themes – *Awareness*, *Travel Advisory*, and *Official* – indicating more concise communication as the pandemic stabilized. This shift toward streamlined messaging aligned with vaccination campaigns, the gradual reopening of activities, and a decline in COVID-19 cases, suggesting that public awareness had improved and information needs became more specific. The sentiment scores as mentioned in Section 3 are also computed and presented in the bottom panels of the same figure. There, a higher positive score indicates optimistic or reassuring messages, while lower scores reflect more critical or urgent communication. For better understanding, we also present a snapshot of the public sentiment and the newsletter content length for the six themes in Table 2.

TABLE 2
A snippet of word count and sentiment score for randomly chosen six newsletter topics, one from each theme.

Date	Theme	Subject matter of newsletter	Sentiment score	Content length
2020-06-12	Awareness	Guidelines for Religious Places on preventive measures to contain spread of COVID-19	−0.4795	13
2020-07-13	Citizens	Fixation of rate for rt PCR Test for COVID-19 in respect of Central Services (Medical Attendance) beneficiaries	0.0361	17
2021-05-15	Others	Reimbursement of OPD medicines to CS (MA) beneficiaries- Special Sanction in view of COVID-19	0.0602	14
2021-06-07	Hospitals	Administration of Second Dose of Covishield Vaccine Prior to Prescribed Time Interval (after 28 days but before 84 days) to persons intending to undertake international travel for specific purposes	−0.2771	29
2022-01-05	Official	Revised guidelines for Home Isolation of mild or asymptomatic COVID-19 cases	−0.6386	11
2022-11-17	Travel Advisory	List of Countries Regions in respect of which primary vaccination schedule completion certificate is allowed to be considered (in context of guidelines for international arrivals updated on 2nd September 2022)	0.8072	30

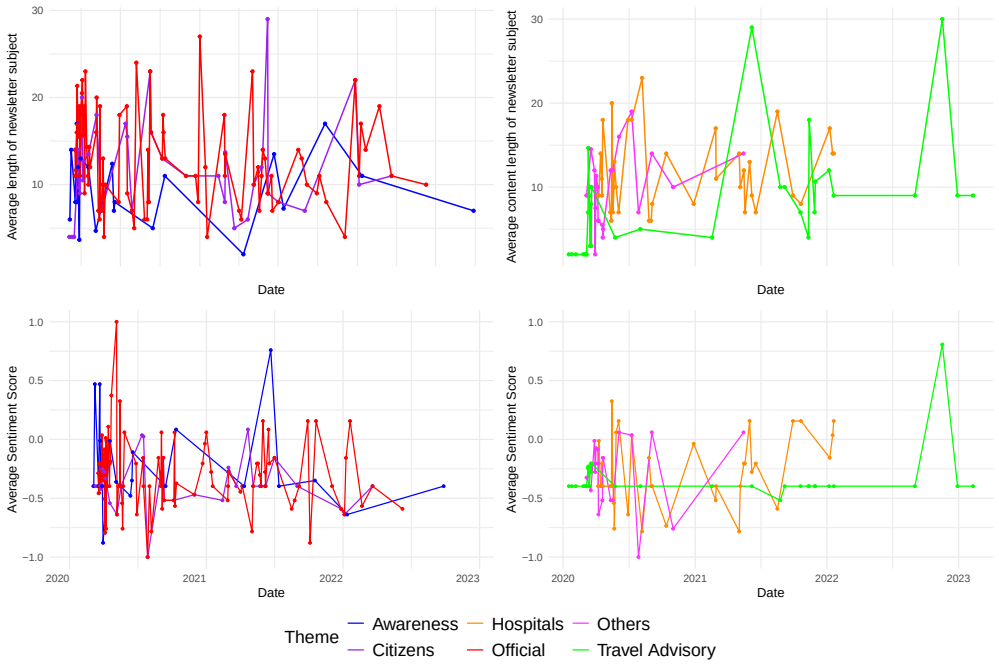

Fig 4: Average content length and sentiment scores for newsletters in the six themes. Left panels show the graphs for the themes *Awareness*, *Citizens* and *Official*; while the right panels show the same for the remaining three themes.

The negative sentiment in the *Hospitals* theme during this period corresponds to the overwhelming pressure on healthcare infrastructure, as India faced severe shortages of hospital beds, ventilators, PPE kits, and other essential medical supplies. These issues were highlighted by the Press Information Bureau (PIB) and NITI Aayog[2] in their socioeconomic reports. Communications under the *Citizens* theme attempted to address these challenges but were often perceived with anxiety, contributing to the observed negative sentiment. In contrast, the *Awareness* theme displayed relatively stable and neutral sentiment, primarily focusing on disseminating preventive measures, such as mask-wearing, social distancing, and hygiene practices. The *Official* theme, saw a rise in positive sentiment, reflecting organized efforts to manage the pandemic through vaccination drives and improved healthcare infrastructure. Announcements about vaccine availability, distribution logistics, and the gradual easing of restrictions contributed to this positive trend, as evidenced by updates from MoHFW and the Ministry of External Affairs (MEA)[3]. However, the *Hospitals* theme maintained a mixed sentiment during this period, particularly during the catastrophic second wave in mid-2021 driven by the Delta variant. The severe shortage of medical oxygen, ICU beds, and essential medicines led to a spike in negative sentiment, as reflected in media reports[4].

By 2022, as vaccination rates stabilized and infection rates decreased, communication themes shifted towards recovery, normalization, and resilience. During this phase, the *Travel Advisory* and *Official* themes demonstrated a distinctly positive trend, aligning with announcements of gradual reopening, easing of restrictions, and resumption of economic activities. This was corroborated by MEA advisories, which facilitated safe resumption of international and domestic travel. Finally, as the pandemic progressed, a positive sentiment in the *Others* category became more pronounced, particularly with psychosocial support programs and motivational campaigns that promote community resilience. This shift aligns with broader governmental

[2] https://www.niti.gov.in/
[3] https://www.mea.gov.in/
[4] https://www.ncdc.gov.in/

and non-governmental efforts to support mental health and provide skill-based training during the recovery phase, helping different communities adapt to new socioeconomic realities.

4.4. Estimation of VAR model for sentiment scores

As mentioned in Section 3.3, to understand the evolution of the sentiment of communication in the six themes with time, we utilize the VAR model on the time series of sentiment scores and the results are displayed in Table 3.

TABLE 3
Results of the VAR model for analyzing the time series of sentiment scores in the six themes. Significance of the coefficients are indicated as: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, $^{\cdot}$ $p < 0.1$.

Response	**Regressor**	**Estimate**	**SE**	**t-Score**	**p-value**
Awareness	Awareness (lag 1)	0.0288	0.0666	0.432	0.6660
	Citizens (lag 1)	−0.0142	0.0564	−0.252	0.8010
	Hospitals (lag 1)	−0.0334	0.0588	−0.568	0.5710
	Official (lag 1)	−0.0035	0.0388	−0.091	0.9270
	Travel Advisory (lag 1)	0.0135	0.0627	0.215	0.8300
	Others (lag 1)	0.0400	0.0753	0.532	0.5950
	Confirmed Cases	−0.0014	0.0009	−1.607	0.1090
Citizens	Citizens (lag 1)	−0.1048	0.0674	−1.554	0.1216
	Awareness (lag 1)	0.0140	0.0797	0.175	0.8612
	Hospitals (lag 1)	−0.0944	0.0704	−1.342	0.1810
	Official (lag 1)	0.0548	0.0464	1.181	0.2389
	Travel Advisory (lag 1)	0.0745	0.0750	0.994	0.3214
	Others (lag 1)	0.0488	0.0900	0.542	0.5880
	Confirmed Cases	-0.0041^{***}	0.0011	−3.800	0.0002
Hospitals	Hospitals (lag 1)	−0.0769	0.0674	−1.140	0.2553
	Awareness (lag 1)	−0.0114	0.0764	−0.149	0.8820
	Citizens (lag 1)	0.0189	0.0646	0.292	0.7705
	Official (lag 1)	−0.0213	0.0445	−0.479	0.6327
	Travel Advisory (lag 1)	0.0193	0.0718	0.269	0.7884
	Others (lag 1)	0.1391	0.0863	1.613	0.1081
	Confirmed Cases	-0.0034^{**}	0.0010	−3.269	0.0012
Official	Official (lag 1)	−0.0125	0.0544	−0.231	0.8179
	Awareness (lag 1)	−0.1174	0.0934	−1.258	0.2097
	Citizens (lag 1)	0.6060^{***}	0.0790	7.674	< 0.0001
	Hospitals (lag 1)	0.7847^{***}	0.0824	9.523	< 0.0001
	Travel Advisory (lag 1)	0.0483	0.0878	0.551	0.5825
	Others (lag 1)	0.0782	0.1055	0.741	0.4592
	Confirmed Cases	-0.0039^{**}	0.0013	−3.110	0.0021
Travel Advisory	Travel Advisory (lag 1)	0.2262^{***}	0.0650	3.479	0.0006
	Awareness (lag 1)	−0.0909	0.0691	−1.315	0.1899
	Citizens (lag 1)	0.0465	0.0585	0.796	0.4271
	Hospitals (lag 1)	−0.0331	0.0610	−0.543	0.5876
	Official (lag 1)	0.0539	0.0403	1.339	0.1818
	Others (lag 1)	0.0398	0.0781	0.510	0.6109
	Confirmed Cases	−0.0014	0.0009	−1.538	0.1255
Others	Others (lag 1)	0.0180	0.0633	0.284	0.7770
	Awareness (lag 1)	0.0770	0.0561	1.374	0.1710
	Citizens (lag 1)	0.0059	0.0474	0.124	0.9010
	Hospitals (lag 1)	0.0071	0.0495	0.144	0.8860
	Official (lag 1)	0.1609^{***}	0.0326	4.931	< 0.0001
	Travel Advisory (lag 1)	0.0046	0.0527	0.087	0.9310
	Confirmed Cases	0.0002	0.0008	0.304	0.7610

In contrast with the earlier VAR analysis, self-persistence is a notable feature in the sentiment score of only one themes, that is *Travel Advisory*, where prior sentiments strongly influenced subsequent communication (estimated coefficient 0.2262, t-statistic 3.48, $p < 0.001$). This finding suggests that the government

prioritized consistent and sustained messaging for travel restrictions, which were critical for public health during the initial phases of the pandemic. Furthermore, the absence of significant influence from other themes on *Travel Advisory* indicates that this sentiment evolved independently, reflecting its specialized and urgent nature. Communications in the *Citizens* and *Hospitals* theme demonstrate a unique relationship with the exogenous factor, the confirmed COVID-19 cases, which has a significant negative impact on both themes. This suggests that rising case numbers heightened public concerns or distress, shaping the sentiment associated with hospitals and citizen-focused messaging. In both of these cases, the lack of self-persistence or the dependence on other themes emphasize their reactive nature to real-world developments rather than continuity in sentiment over time.

Interestingly, the *Official* theme emerged as a central mediator in sentiment interactions. It was positively influenced by the sentiment of the communications related to the *Citizens* theme as well as the *Hospitals* theme, indicating that formal governmental communication was closely aligned with public concerns and healthcare-related sentiments. This alignment underscores the role of the *Official* theme as a platform for addressing societal priorities and institutional challenges. Moreover, the *Official* theme significantly influenced sentiments in the broader, less defined, other communications, demonstrating its anchoring role in the sentiment network. Finally, we notice that the sentiments associated with the newsletters in the *Awareness* theme are found to be not impacted by anything else, thereby underscoring its consistent pattern across the entire time horizon.

5. Discussion

A succinct summary of our study is warranted at this point. In this article, we have considered the communication from the Indian government during the COVID-19 pandemic in the form of newsletters' topics categorized primarily into six themes. In order to determine the relative importance of these themes and the likelihood of a new communication topic cropping up on any one of these six themes, we have used a dynamic preferential attachment network model by computing the empirical probability estimates for each theme throughout the period of study. With necessary theoretical guarantee, this help us in understanding when the communication patterns stabilized. We have also explored the lengths and sentiments of the newsletters corresponding to each theme as a function of time. Overall, in order to understand the temporal interactions among various themes and sentiment with time, we have separately implemented the VAR model on the estimated probabilities of the themes and sentiment scores for the same.

Our results show that initial communication efforts were heavily focused on creating broader *Awareness* campaigns, aimed at building a foundational understanding of preventive measures and mobilizing the public for collective action against the spread of the virus. This aligns with prior research, emphasizing the necessity of clear and consistent messaging during the initial stages of a public health crisis (Vaughan and Tinker, 2009). As the pandemic progressed, communication strategies adapted to emerging challenges, such as the strain on healthcare systems, vaccine roll-out complexities, and socioeconomic disruptions. Notably, themes related to *Hospitals* and *Citizens* remained prominent throughout, reflecting sustained efforts to address healthcare inadequacies and public distress due to socioeconomic adversities (Patel et al., 2020). The focus on health communication, even during vaccination phases, underscores the persistent pressure on healthcare systems and the importance of keeping people informed about public health protocols (Dubé et al., 2022). Similarly, continuous communication in the *Citizens* theme highlighted efforts to mitigate socioeconomic impacts, suggesting that addressing public welfare was a critical component of the overall communication strategy (Malecki, Keating and Safdar, 2021).

The use of the VAR model in our approach to watch these interactions and study how they are influenced by their own past values (lagged variables) and exogenous factors like COVID-19 statistics presents a unique opportunity for further research to assess the use of government communication during high-risk health outbreaks that can stall national progress if not addressed or communicated appropriately. Our sample is

unique to the Indian context, health systems, information flow mechanisms, and implementation techniques. This could pan out differently in different countries and regions, yet it can serve as a baseline to situate strong and persuasive health communication patterns to effectively communicate a need to adopt and change the behavior that the outbreak necessitates.

Our analysis during the early period of COVID-19 has important lessons on integrating public health with health-related policy. While this can help in effectively creating targeted interventions, there is the risk of information overload. This can create overlapping in themes and timelines that can blur the lines of information, creating an infodemic (see, for example, Briand et al., 2023), and causing the need for management of this infodemic. Though this is caused by the widespread use of social media for propagating misinformation, to address this, the World Health Organization (WHO) member states have drafted and negotiated a convention to manage this to create awareness, communication engagement, and health literacy which is crucial to combat this disinformation (Taguchi et al., 2023). Furthermore, as seen in our analysis, the shift in communication during vaccination drives from caution to optimism suggests that the government leveraged this as a means of restoring public confidence. This shift is aligned with the observed increase in positive sentiment scores in themes like *Awareness* and *Official* (Vilar-Lluch et al.). The consistent communication regarding vaccine efficacy, safety, and booster doses played a crucial role in sustaining public engagement and achieving high vaccination rates (Kappes et al., 2023). The Indian government's efforts in health communication, through these newsletters, demonstrates the importance of integrating multiple themes to ensure comprehensive coverage of all aspects of a public health crisis while continuously keeping track of the changing dynamics on the ground to communicate with clarity, brevity and care.

Let us end the paper with some potential future extensions of our study. It is important to recognize that a limitation of our analysis is that we relied only on official sources of information, namely the newsletters published on the government portal. The reason behind this is that amid the minefield of confusing, often misleading health communication, official newsletters such as these serve as spokes of the wheels of a continuously moving machinery holding the wheels firmly in place while steering public health as a priority. However, it will definitely be beneficial to extend the analysis further to study the communications via other channels, such as official social media outlets in X or Facebook (in similar lines of Jha, Lin and Savoia, 2016). From the perspective of the government, it will also be of interest to assess whether the communication strategies during the pandemic led to effective management and mitigation of the crisis. In that case, appropriate statistical modeling techniques will be required to capture the bidirectional relationship between the COVID-19 cases and the communication patterns.

References

Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371** 20120375.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512.

Briand, S., Hess, S., Nguyen, T. et al. (2023). Infodemic Management in the Twenty-First Century. In *Managing Infodemics in the 21st Century: Addressing New Public Health Challenges in the Information Ecosystem* (T. D. Purnat, T. Nguyen and S. Briand, eds.) 1 Springer, Cham (CH).

Bulut, O. and Poth, C. N. (2022). Rapid assessment of communication consistency: sentiment analysis of public health briefings during the COVID-19 pandemic. *AIMS Public Health* **9** 293.

Capocci, A., Servedio, V. D., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S. and Caldarelli, G. (2006). Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* **74** 036116.

DE LAS HERAS-PEDROSA, C., SÁNCHEZ-NÚÑEZ, P. and PELÁEZ, J. I. (2020). Sentiment analysis and emotion understanding during the COVID-19 pandemic in Spain and its impact on digital ecosystems. *International journal of environmental research and public health* **17** 5542.

DEB, S., ROY, R. and DAS, S. (2024). Forecasting elections from partial information using a Bayesian model for a multinomial sequence of data. *Journal of Forecasting*.

DUBÉ, È., LABBÉ, F., MALO, B. and PELLETIER, C. (2022). Public health communication during the COVID-19 pandemic: perspectives of communication specialists, healthcare professionals, and community members in Quebec, Canada. *Canadian Journal of Public Health* **113** 24–33.

GANAPATHY, D. Global Culture, Power, and Health Communication: India Fights Corona on the Battlefield of Social Media Platforms.

GAO, F. and VAN DER VAART, A. (2021). Statistical Inference in Parametric Preferential Attachment Trees. *arXiv preprint arXiv:2111.00832*.

GAO, F., VAN DER VAART, A., CASTRO, R. and VAN DER HOFSTAD, R. (2017). Consistent estimation in general sublinear preferential attachment trees. *Electronic Journal of Statistics* **11** 3979–3999.

GIAMMATTEO, P., DONATO, D., ZLATIĆ, V. and CALDARELLI, G. (2010). A PageRank-based preferential attachment model for the evolution of the World Wide Web. *Europhysics letters* **91** 18004.

GRIMMER, J. and STEWART, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* **21** 267–297.

HAMILTON, J. D. (2020). *Time series analysis*. Princeton university press.

JHA, A., LIN, L. and SAVOIA, E. (2016). The use of social media by state health departments in the US: analyzing health communication through Facebook. *Journal of community health* **41** 174–179.

JOCKERS, M. L. (2015). Syuzhet: Extract Sentiment and Plot Arcs from Text.

KAPPES, H. B., TOMA, M., BALU, R., BURNETT, R., CHEN, N., JOHNSON, R., LEIGHT, J., OMER, S. B., SAFRAN, E., STEFFEL, M. et al. (2023). Using communication to boost vaccination: Lessons for COVID-19 from evaluations of eight large-scale programs to promote routine vaccinations. *Behavioral Science & Policy* **9** 11–24.

LYON, M. R. and MAHMOUD, H. M. (2020). Trees grown under young-age preferential attachment. *Journal of Applied Probability* **57** 911–927.

MALECKI, K. M., KEATING, J. A. and SAFDAR, N. (2021). Crisis communication and public perception of COVID-19 risk in the era of social media. *Clinical infectious diseases* **72** 697–702.

MHEIDLY, N. and FARES, J. (2020). Leveraging media and health communication strategies to overcome the COVID-19 infodemic. *Journal of public health policy* **41** 410–420.

MING-YANG, W., GUANG, Y. and DA REN, Y. (2010). Measuring preferential attachment mechanism in citation network basing on sliding time windows. In *2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)* **2** V2–110. IEEE.

MÓRI, T. (2002). On random trees. *Studia Scientiarum Mathematicarum Hungarica* **39** 143–155.

PATEL, J. A., NIELSEN, F. B. H., BADIANI, A. A., ASSI, S., UNADKAT, V., PATEL, B., RAVINDRANE, R. and WARDLE, H. (2020). Poverty, inequality and COVID-19: the forgotten vulnerable. *Public health* **183** 110.

RAWAT, S. and DEB, S. (2023). A spatio-temporal statistical model to analyze COVID-19 spread in the USA. *Journal of Applied Statistics* **50** 2310–2329.

ROSSMANNEK, O. and RANK, O. N. (2021). Is it Really a Universal Phenomenon?-Preferential Attachment in Alliance Networks. *European Management Review* **18** 85–99.

RUGGERI, K., STOCK, F., HASLAM, S. A., CAPRARO, V., BOGGIO, P., ELLEMERS, N., CICHOCKA, A., DOUGLAS, K. M., RAND, D. G., VAN DER LINDEN, S. et al. (2024). A synthesis of evidence for policy from behavioural science during COVID-19. *Nature* **625** 134–147.

TAGUCHI, K., MATSOSO, P., DRIECE, R., DA SILVA NUNES, T., SOLIMAN, A. and TANGCHAROENSATHIEN, V. (2023). Effective infodemic management: A substantive article of the pandemic accord.

VAUGHAN, E. and TINKER, T. (2009). Effective health risk communication about pandemic influenza for vulnerable populations. *American journal of public health* **99** S324–S332.

VILAR-LLUCH, S., MCCLAUGHLIN, E., KNIGHT, D., ADOLPHS, S. and NICHELE, E. The language of vaccination campaigns during COVID-19.

ZUNIC, A., CORCORAN, P. and SPASIC, I. (2020). Sentiment analysis in health and well-being: systematic review. *JMIR medical informatics* **8** e16023.

Appendix A: Proofs

Proof of Result 1. The proof, just like the result, heavily relies on (Gao et al., 2017, Theorem 2). The result in there is on the number of nodes of degree k, whereas here we are interested in the degree of the k^{th} theme. Incorporating this modification, our empirical estimate is just a translation of the empirical estimate from their work, and can be worked out in a straightforward manner. □

Proof of Result 2. Suppose, till the time-point n, the degrees of the present nodes are $d_1^n, d_2^n, \ldots, d_K^n$. Also, assume that the next m nodes get attached to topics given by random variables $C_{n+1}, C_{n+2}, \ldots, C_{n+m}$. Then, the likelihood of this event is given by

$$P\left(C_{n+1}=c_1, C_{n+2}=c_2, \cdots, C_{n+m}=c_m\right) = \frac{g(d_{c_1}^n)}{\sum_{j=1}^K g(d_j^n)} \times \frac{g(d_{c_2}^{n+1})}{\sum_{j=1}^K g(d_j^{n+1})} \times \cdots \times \frac{g(d_{c_m}^{n+m-1})}{\sum_{j=1}^K g(d_j^{n+m-1})}. \tag{5}$$

We want to show that for any order, $\{c_{\pi(1)}, c_{\pi(2)}, \ldots, c_{\pi(m)}\}$,

$$P\left(C_{n+1}=c_1, C_{n+2}=c_2, \cdots, C_{n+m}=c_m\right) = P\left(C_{n+1}=c_{\pi(1)}, C_{n+2}=c_{\pi(2)}, \cdots, C_{n+m}=c_{\pi(m)}\right),$$

where $\pi(\cdot)$ is any permutation such that $\pi : \{1,2,\ldots,m\} \to \{1,2,\ldots,m\}$. To prove this, we shall show that the result holds for $m=2$, and the general result will follow from mathematical induction, which is straightforward. Thus, we essentially need to show that

$$P\left(C_{n+1}=c_1, C_{n+2}=c_2\right) = P\left(C_{n+1}=c_2, C_{n+2}=c_1\right).$$

If $c_1 = c_2$, then this follows trivially. Otherwise, we have

$$P\left(C_{n+1}=c_2, C_{n+2}=c_1\right) = \frac{g(d_{c_2}^n)}{\sum_{j=1}^K g(d_j^n)} \times \frac{g(d_{c_1}^n)}{\sum_{j=1, j\neq c_2}^K g(d_j^n) + g(d_{c_2}^n+1)}. \tag{6}$$

From the given condition on $g(\cdot)$, the term on the right hand side of (6) clearly equals $P(C_{n+1}=c_1, C_{n+2}=c_2)$, and that completes our proof. □