Samachar: Print News Media on Air Pollution in India

Karm Patel
VGEC, Chandkheda
Ahmedabad, India
karmpatel216@gmail.com

Rishiraj Adhikary
IIT Gandhinagar
Gandhinagar, India
rishiraj.a@iitgn.ac.in

Zeel B Patel
IIT Gandhinagar
Gandhinagar, India
patel_zeel@iitgn.ac.in

Nipun Batra
IIT Gandhinagar
Gandhinagar, India

Sarath Guttikunda
Urban Emissions
Delhi, India

ABSTRACT

Air pollution killed 1.67M people in India in 2019. Previous work has shown that accurate public perception can help people identify the health risks of air pollution and act accordingly. News media influence how the public defines a social problem. However, news media analysis on air pollution has been on a small scale and regional. In this work, we gauge print news media response to air pollution in India on a larger scale. We curated a dataset of 17.4K news articles on air pollution from two leading English daily newspapers spanning 11 years. We performed exploratory data analysis and topic modeling to reveal the news media response to air pollution. Our study shows that, although air pollution is a year-long problem in India, the news media limelight on the issue is periodic (temporal bias). News media prefer to focus on the air pollution issue of metropolitan cities rather than the cities which are worst hit by air pollution (geographical bias). Also, the air pollution source contributions discussed in news articles significantly deviate from the scientific studies. Finally, we analyze the challenges raised by our findings and suggest potential solutions as well as the policy implications of our work.

KEYWORDS

air pollution; text mining; nlp; news media; topic modeling

ACM Reference Format:
Karm Patel, Rishiraj Adhikary, Zeel B Patel, Nipun Batra, and Sarath Guttikunda. 2022. Samachar: Print News Media on Air Pollution in India. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '22), June 29-July 1, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3530190.3534812

1 INTRODUCTION

In 2019 alone, air pollution was responsible for 1.67M deaths in India (17.8% of total deaths) [67]. Ambient fine particulate matter ($PM_{2.5}$) is the most significant risk factor for premature death, shortening life expectancy at birth by 1.5 to 1.9 years [9]. Air pollution is not always visible, leading to an incorrect perception among people [5]. False perception could lead to increased exposure to air pollution and increased challenges in implementing mitigation strategies. For example, a study [80] has shown that people are less likely to go outdoors if the perceived pollution is high. Research has shown that timely and informed action against air pollution can help make informed decisions to reduce health risk [18, 69]. News media play a pivotal role in dissipating information and thereby shaping public perception [33, 36, 47, 77, 81]. News media influence the definition of a social problem and also set the agenda for policymakers to address the issues of national concern [14]. In this paper, we try to understand how print news media respond to air pollution in India.

Research has shown that stubble burning and meteorological factors result in poor air quality in the North Indian river plain, also known as the Indo-Gangetic (IG) plain [29, 40]. The IG plain covers 14 Indian states and many cities [87]. Any geographical bias in news media coverage could impact the behavioral measures and citizens' active involvement in mitigating air pollution. The success of any air pollution mitigation measure, both legislative and behavioral, to limit pollutant emissions requires acceptance by the citizens [51]. Our first research question **(RQ1)** tries to understand if newspapers have a geographical and temporal bias in reporting about air pollution. A study [15] in India showed that 90% of Indians across highly polluted cities do not understand the causes and effects of air pollution. One of the reasons stated in the report is the lack of evidence and science in media coverage about air pollution. Another study [43] conducted by a global public health organization says that most news about air pollution omits information about the significant health impact of air pollution. But these studies mainly looked at social media data across multiple geographies. We framed our second research question **(RQ2)** to understand if news media coverage around air pollution exhibits deviation from scientific evidence around sources and the impact of air pollution. We state both **(RQ1)** and **(RQ2)** as the following,

- **(RQ1) Does the news media coverage around air pollution exhibit a geographical and/or temporal bias?**
- **(RQ2) Does the news media coverage around air pollution exhibit deviation from scientific evidence around sources and impact of air pollution?**

We retrieved 3.17M articles from two news media houses[1] published over 2010 to 2021. We used an extensive list of queries related to air pollution followed by snowball sampling to extract 17.4K air

[1]The Times of India and The Hindu

COMPASS '22, June 29-July 1, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9347-8/22/06…$15.00
https://doi.org/10.1145/3530190.3534812

pollution related news articles. To address the first research question **(RQ1)**, we also curated $PM_{2.5}$ data for 88 cities from 2010 to 2021. Knowing that the cities in Indo-Gangetic (IG) plain [87] are highly polluted [29, 40], we divided (RQ1) into three sub-questions related to i) intensity of air pollution in IG plain throughout the year; ii) news media coverage of air quality in IG plain throughout the year; and iii) the cities getting highest attention from news media. We found that cities in IG plain are more polluted compared to other cities in India, but they get less news media attention. In contrast, news media focus more on the metropolitan cities, including, but not limited to, Delhi, Mumbai, Bengaluru, Chennai, Kolkata, Nagpur, Pune, and Hyderabad, but they are less polluted (excluding Delhi) compared to the cities in IG plain.

To address the second research question **(RQ2)**, we apply a topic modeling technique (Latent Dirichlet Allocation) to extract the topics of discussions on air pollution by news media. We classify the extracted topics into various categories to understand news media discussions on sources and effects of air pollution. We compare the source distribution found in news media with the ground truth from a scientific study [71]. We found that 'Vehicular emissions' contribute less air pollution but get the highest attention in news media. In contrast, 'Residential biomass' has a significant contribution but comparatively less attention.

We further discuss the policy implications of our study. Our findings suggest that air quality data should be published regularly in a widely accepted format to accelerate timely actions against air pollution. We then comment on potential criteria to decide the government budget for air pollution in different cities. After observing a significant dissimilarity between scientific and news media reported air pollution source contributions, we suggest news media refer to scientific studies to emphasize some sources over the others. We believe that news media can bring a significant impact by taking the right stand by mentioning several success stories of news media in resolving air pollution related problems.

Reproducibility and Dataset Release: Our work is fully reproducible, and we publish all the analyzed news articles with metadata in our project repository[2]. All the tables and figures in this paper are reproducible with the code shared in the same repository. We believe our news-articles dataset will be useful to other researchers working on news media discussion around air pollution in India.

Paper Organization: We discuss the related work in Section 2. Dataset information is presented in Section 3. We propose the research questions and our approaches to address them in Section 4. We evaluate our methodology and discuss the results and analysis in Section 5. We discuss the policy implications of our study in Section 6. We describe limitation and future work in Section 7 and conclude our work in Section 8.

2 RELATED WORK

We categorize the related work in three categories as the following.

1) Gauging public perception via social media: A recent study analyzed a large amount of Twitter data (1.2M tweets, 26.4K users) over four and a half years (Jan' 16 to Apr' 20) to gauge the public perception of air pollution [5]. The study found that public perception towards untested pollution mitigation strategies is largely supportive. Further, Twitter discussions on air pollution only peak up during winter when pollution is highly hazardous and visually apparent. Our study closely matches with [5] in terms of approach. In the past, a global public health organization [43], used social media data to conclude that there is lack of evidence and science around air pollution discussions. News media influence a large number of people compared to social media such as Twitter. For example, *Times of India* had 17.3M total readers in India till 2020 [64]. Thus, we focus on analyzing news articles to gauge the coverage and correctness of information they dissipate.

2) Media effect on public perception: Previous research has found that the public pays more attention to air pollution when its concerns are reflected in news media [83]. The content of news media can be amplified via social media. For example, a Chinese social media saw an outburst of air pollution related discussion in 2013. A study investigated and concluded that this was mainly due to $PM_{2.5}$ data published by US Embassy in China followed by an intense discussion in the news media [76].

3) News media on environmental problems: Murali et al. [56] manually analyzed 600 news articles (2019-20) related to 'climate change' covering four countries, including India. Authors find that air pollution discussions in news media are mostly driven by political influence in countries like India. They also emphasize the need for scientific evidence backing up news media. Another study investigates 'climate change' discussions from two Indian news media by curating 18.2K news articles covering the period of 20 years (1997-2016) [46]. They applied a topic modeling technique to describe topics of discussion. The study is related to our work in terms of dataset and methodology. However, our focus is on 'air pollution'. Olofsson et al. [65] collected 235 news articles published in the year 2011-12 from 5 media houses related to 'Delhi air pollution'. They analyzed causes and effects discussed by news media using automated text coding with 'AutoMap' software. They observed that news media presents 'transportation' as a leading cause and 'health' as a dominant effect of air pollution. A recent WHO study [58] highlights news media discussion on sources and solutions of air pollution, health risks, and air pollution policies in India by manually analyzing 500 random articles published in 2014-15 from 'Google News' and 'Meltwater'. The study also observed that news media focus more on Delhi and neglect other cities with poor air quality. Another study finds that news articles are tend to ignore the discussion on environmental causes such as air pollution while discussing asthma and emphasized that media can help in raising scientific knowledge about relationship between asthma and air pollution in people followed by policy changes on public interest [52].

Our Contribution: Previous studies are limited to either the short time span of articles or specific regions (such as Delhi, India). To the best of our knowledge, ours is the most extended study (2010-2021) of air pollution related news articles with 190 Indian cities analyzed. Existing literature does not analyze the ground truth pollution data at low granularity. We have collected and analyzed the ground truth air pollution data at a fine-grained level (daily) to check its correspondence with news articles. We also use advanced methods from data science compared to the manual investigation done in the majority of the previous work, which is laborious and does not scale well.

[2]https://github.com/karm-patel/Samachar-News-media-on-air-pollution

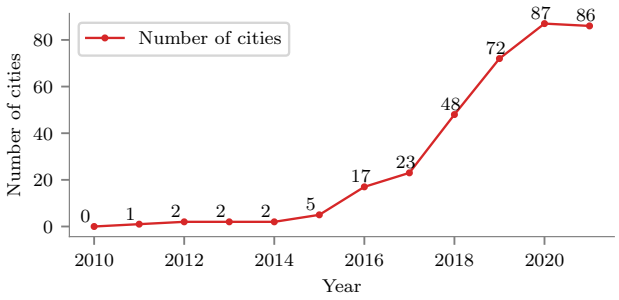

Figure 1: Number of cities for which $PM_{2.5}$ data is available each year. After 2015, number of cities steadily increase due to newly installed stations.

3 DATASET

We collect $PM_{2.5}$ and news articles data from various sources for our study. We elaborate on the details in the following sections.

3.1 $PM_{2.5}$ dataset

Air pollutants including, but not limited to, $PM_{2.5}$, PM_{10}, SO_2, NO_2, NO, CO, Ozone, and BTX are of major concern to air quality, but $PM_{2.5}$ exceeds the standards more often than the other pollutants [68] in India. Thus, we focus on $PM_{2.5}$ in our analysis. The $PM_{2.5}$ data before 2019 is curated directly from Central Pollution Control Board (CPCB) portal [27], and the data after 2019 is curated from CPCB sourced via OpenAQ [66]. We removed the missing data entries and filtered the erroneous data outside the measurement range (below 0 $\mu g/m^3$ and above 1000 $\mu g/m^3$) [5]. The data has 15 minutes of granularity. Since we need readings of cities for each day, we average across stations from the same cities and get a single reading for each city per day. The timespan of this data is variable for each city, depending upon the installation of air quality monitors. We have the longest timespan (2010-21) of data for Delhi. After 2015, there was a steady increase in the number of cities with data availability, as shown in Figure 1. Note that 66 out of 88 cities have only one station. Thus, data from those cities may not be representative of the entire city. However, it is a limitation that originates from the data source and not from our study.

3.2 News articles dataset

We select two English language Indian newspapers (*Times of India* (*TOI*) and *The Hindu*) considering their country wide reach, high readership, and well organized archives of news articles. *TOI* had 17.3M total readers and *The Hindu* had 8M total readers in India by the end of 2020 [64].

We scrapped 3.17M publicly available news articles from archives of these news media [10, 62] using Python libraries named 'requests' [73] , 'beautifulsoup' [12] and 'newspaper3k' [61]. We have a total of 17.4K Air Quality (AQ) related news articles, which comprises 11.8K articles from *TOI* and 5.6K articles from *The Hindu* from 2010 to 2021. Table 2 shows statistics related to news articles dataset. We have used the queries mentioned in Table 1 to filter the AQ-related articles. We choose these queries by i) discussing with AQ experts; and ii) snowball sampling using keywords as queries that appear in articles using other queries [16]. For example, while investigating articles retrieved by query 'air quality', we found that keywords '$PM_{2.5}$' and 'AQI' were frequently present in the articles. Thus, we include these two keywords in the queries. Our dataset also contains other meta information like *city, author, top image URL,* and *news category*. We have the *city* information for 84% articles, which is useful in air quality related analysis. Figure 2 shows the evolution of articles over time for *TOI* and *The Hindu*.

Verifying if the collected articles are about air quality: After the initial phase of data collection, we removed some queries which were adding noise in the data. We initially added 'carbon dioxide' and 'co2' as queries. Upon randomly checking 50 articles for these queries, we found that 44 articles are irrelevant with AQ. Thus, we narrowed down to 22 AQ-specific queries. To check the number of AQ-related articles in the dataset, we manually investigated several articles from the dataset. Due to the subjective nature of the process, two authors annotated 200 articles (AQ-related or not AQ-related) and found Cohen's kappa score equal to 94% (κ) [54] ($\kappa \geq 80\%$ is considered as almost perfect agreement [54]). We found 182 out of 200 articles labeled as AQ-related by both annotators. Thus, we can say that 91% of articles in our dataset are AQ-related. Remaining 9% articles contain AQ related words in a line or two, but they are not focused on AQ.

4 APPROACH

In this section, we describe our research questions and explain the approaches to address them.

4.1 RQ1: Does the news media coverage around air pollution exhibit a geographical and/or temporal bias?

4.1.1 ***Background:*** Previous studies have shown that air pollution is a year-long problem in Delhi, but its concern appears sporadically across social media [5]. Google trends plot on "air pollution" keyword search in Delhi also asserts the fact of sporadicity along with periodicity [6].

A study [35] shows that nearly 50% of the research literature focuses only on Delhi's air pollution. However, there is less attention towards the air pollution problem in other cities. To understand the air pollution intensity in various regions of India, we extract a satellite-derived district-wise annual average (2016) of $PM_{2.5}$ from a previous study [35] and plot a choropleth map as shown in Figure 3. $PM_{2.5}$ concentrations in Indo-Gangetic (IG) plain remains higher than the annual standard of India persistently due to emissions from primary sources of $PM_{2.5}$, unfavorable topography, and meteorology [29, 40]. On this ground, we break down RQ1 into three sub-questions:

Q1: *"Is air pollution a year-long problem in cities of Indo-Gangetic (IG) plain?"*

Q2: *"Do cities in IG plain get the news media attention in consonance with their pollution?"*

Q3: *"Which cities get high attention from the news media? Are these cities polluted throughout the year?"*

Queries
air pollution, airpollution, air quality, airquality, aqi, pm 2.5, pm2.5, pm10, pm 10, stubble burning, crop burning, ozone, air pollutants, sulphur dioxide, so2, carbon monoxide, smog, nitrogen dioxide, acid rain, odd even, oddeven, car emissions

Table 1: List of 22 queries used to filter air quality related news articles.

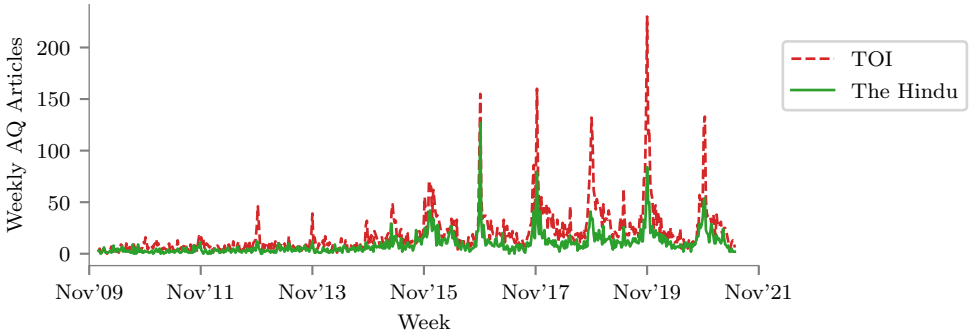

Figure 2: Articles from *TOI* and *The Hindu* follow similar trends indicating that news media attention is higher in October-November. News media had limited attention towards air pollution before 2015.

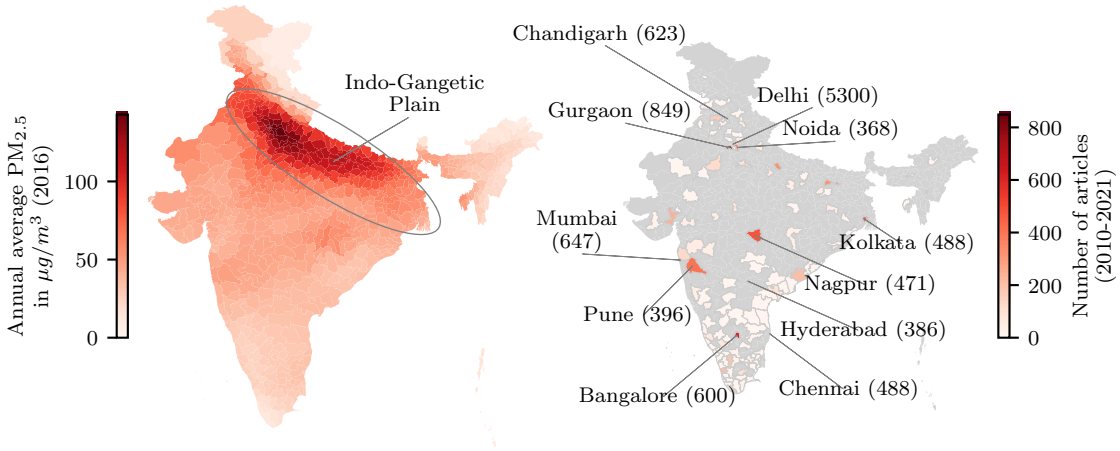

Figure 3: Indo-Gangetic plain is highly polluted compared to the rest of India (left map), but, only a few cities in IG plain get pertinent news media attention (right map). Most of the metro cities (annotated in right map) are relatively less polluted, yet they get more media attention than cities in IG plain.

4.1.2 ***Approach:*** We describe our approaches to solve each sub-question as the following,

Q1: We analyzed air pollution in Delhi along with at least one city from Punjab, Haryana, Uttar Pradesh, West Bengal, and Bihar (states in IG plain) as shown in Table 3. We compute the daily average of $PM_{2.5}$ for each of these cities over multiple years and compare it with the daily average $PM_{2.5}$ limit set by WHO ($25\mu g/m^3$) and India ($60\mu g/m^3$).

Q2: We calculate the articles count for the same cities analyzed in Q1 and compare them with air pollution levels to check if the number of articles follows air pollution levels. A previous study [5] shows that Twitter discussions around 'Delhi air pollution' are

News media	*Times of India*	*The Hindu*
Time-period	Jan'10 to May'21	Jan'10 to May'21
Total articles	1.96M	1.21M
Air quality articles	11,746	5,628
Cities	67	148
Percentage of articles with *city*	86%	80%

Table 2: Statistics on news articles dataset. 67% of air quality articles are from *TOI*. We have *city* metadata available for 86% and 80% articles in *TOI* and *The Hindu* respectively.

episodic. Similarly, we check if news media discussions on 'air pollution' follow similar trends in Delhi.

Q3: We extracted districts corresponding to each city present in our dataset using a Python library named 'geopy' [34]. Further, we plot the district-wise number of articles over the timespan Jan'2010 - May'2021. We find the most discussed cities from this plot, along with the number of articles. Finally, we check if these cities are polluted throughout the year by applying the same approach as mentioned in Q1.

4.2 RQ2: Does the news media coverage around air pollution exhibit deviation from scientific evidence around sources and impact of air pollution?

4.2.1 ***Background:*** Previous studies show that news media discussion play a vital role in shaping public perception [14, 33, 36, 47, 77, 81]. Government organizations like CPCB [27] and various previous studies [11, 35, 71] have acknowledged and published various sources and effects of air pollution. The citizens are usually unaware of such technical details impeding timely personal action to reduce the effect of air pollution on health. News media play an important role in dissipating air pollution related information with people. It is essential to reveal the topics discussed by news media to verify if the discussions in news articles are scientifically valid. For example, awareness about sources of air pollution will help the citizens and government take preventive actions towards reducing air pollution. Thus, news media should dissipate correct information about sources of air pollution. We state two sub-questions that would address our second research question as the following:

Q1: *"What are the topics discussed by news media? Are these topics pertinent with the ground reality of air pollution?"*

Q2: *"What are the major sources of air pollution? Are these sources discussed by news media?"*

4.2.2 ***Approach:*** Now, we describe our approaches to address the above mentioned research questions.

Q1: We apply an unsupervised machine learning technique known as topic modeling [17] to address Q1. In particular, we use the 'latent Dirichlet allocation' (LDA) technique, which is a popular technique among several topic modeling techniques [17]. As described in a previous literature [41], we feed the M news articles to the model, where each article has N number of words. The model returns z topics where each topic is a cluster of words. ψ is the probability distribution of words in a topic. θ is the probability distribution of topics per document. Concentration parameter α represents topic density per document, and β represents word density per topic. We feed our article corpus into the LDA and evaluate the output topics. The choice of hyperparameters, number of topics (z), and number of iterations of the algorithm (i) can be made using the topic coherence measure [8, 74], explained in Section 5.2.1.

Q2: Sources which contribute to air pollution have been described in a previous study [53] . We identified if the same sources are discussed proportionally in news articles by carefully curating specific queries. We focus this question on Delhi to narrow down the question and avoid diverse distribution of sources in various regions. We describe the details in Section 5.2.2.

City	Average $PM_{2.5}$ ($\mu g/m^3$)	India limit	WHO limit	Time span ($PM_{2.5}$ data)	Articles (2010-21)
Delhi	142	81%	98%	02/2010 - 06/2021	5300
Gurugram	116	71%	96%	01/2016 - 06/2021	857
Patna	119	69%	94%	10/2015 - 06/2021	317
Varanasi	114	69%	92%	09/2014 - 06/2021	51
Faridabad	120	68%	95%	05/2015 - 06/2021	51
Noida	114	66%	94%	07/2017 - 06/2021	369
Lucknow	111	66%	95%	03/2015 - 06/2021	320
Kanpur	113	63%	93%	05/2015 - 06/2021	56
Agra	103	61%	94%	05/2015 - 06/2021	88
Kolkata	64	39%	64%	04/2018 - 06/2021	488
Amritsar	57	32%	85%	02/2017 - 06/2021	44

Table 3: Percentage of time cities in IG plain violate daily $PM_{2.5}$ India limit (60 $\mu g/m^3$) and WHO limit (25 $\mu g/m^3$) for a given time span. These cities do not get enough news media attention in consonance with their pollution. For example, Delhi gets significant news media attention while 'Varanasi' and 'Faridabad' rarely get news media attention.

5 EVALUATION

In this section, we describe our experiments and results to address the research questions.

5.1 Addressing Research Question 1

As mentioned in Section 4.1 we framed three sub-questions to approach RQ1. We address each sub-question hereunder.

5.1.1 ***Is air pollution a year-long problem in cities of Indo-Gangetic (IG) plain?***

Table 3 shows the percentage of times eleven cities of IG plain exceed the WHO and India limit. We observed that nine out of eleven cities surpass the India limit more than 60% of the time and WHO limit more than 90% of time as shown in Table 3. Additionally, the days where the India limit is violated are spread across the year except a few months (Jun to Sep) for those cities as shown in Figure 4. For India, Jun to Sep is monsoon season and rains wash out the particles which effectively reduces the air pollution during these months [29]. **Hence, air pollution is a year-long problem for the majority of cities in the IG plain.**

5.1.2 ***Do cities in IG plain get the news media attention in consonance with their pollution?***

Figure 4: Violation of daily India limit of $PM_{2.5}$ level (60 $\mu g/m^3$) by eleven cities in the IG plain over the year. Most of these cities (except Amritsar and Kolkata) violated India's limit across the year except a few months (monsoon season); hence air pollution is a year-long problem for most cities in the IG plain. Note that the time-span of $PM_{2.5}$ data of all cities is from April 2018 to April 2021 (common time period based on the availability of $PM_{2.5}$ data of all cities) which is averaged out along same day to show the readings over the year.

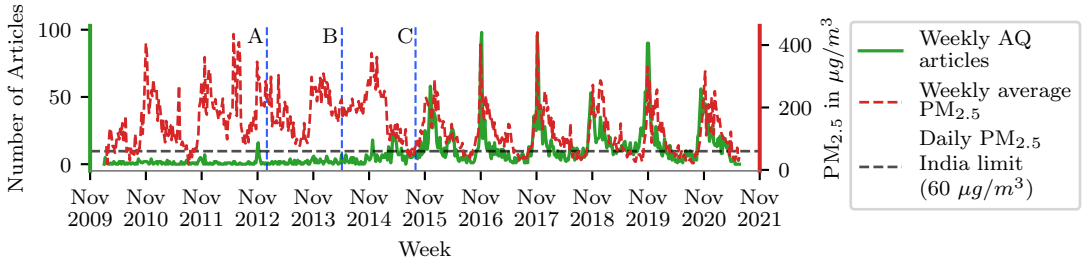

Figure 5: A: From Jan 2013 onwards, US embassy in India started posting air quality data online; B: In May 2014, Delhi became most polluted city in the world according WHO ambient air pollution guidelines; C: In Sept 2015, OpenAQ website made AQ data of India publicly available. This plot is comparison of AQ articles counts with $PM_{2.5}$ levels in Delhi over 11 years. News media attention is not persistent throughout the year, whereas air pollution is higher than India limit most of the time. The number of news articles increased comparatively after 2014 due to easily available air pollution data.

From Figure 3, we can infer that cities in IG plain are more polluted compared to cities in other regions; still, only a few of them get commensurable news media attention. To get the exact numbers, we calculate the count of articles published in eleven cities of IG plain for time span Jan'2010 to May'2021 as shown in Table 3. All cities except Kolkata and Amritsar experience high $PM_{2.5}$ exposure. However, Delhi gets significantly higher news media attention than other cities (36% of the total articles). Faridabad, Varanasi, Kanpur, and Agra are lacking adequate news media attention. Thus, **news media attention to cities in IG plain is not proportional to their pollution.**

News media attention on Delhi: Delhi violates the daily air pollution standards more often than any other city. Thus, it is necessary to check if news media discuss it throughout the year or not. We have $PM_{2.5}$ data for the longest time span (2010-2021) for Delhi as mentioned in Section 3.1. We plot time-series of weekly average $PM_{2.5}$ along with weekly AQ articles for Delhi in Figure 5. From this plot, we can derive two observations as the following:

1) News media attention increased after 2014: Before 2014, air pollution was high in Delhi, though news media discussions were comparatively low. This is because, a) Before 2013, the air pollution data of Delhi was mostly offline and hardly available to public. From 2013 onwards, US embassy in India started posting air pollution data of Delhi online [37], b) In May 2014, WHO updated its database of ambient air pollution in cities, which showed that Delhi surpassed Beijing in terms of $PM_{2.5}$ level and became most polluted city in the world [58], c) In Sept 2015, OpenAQ started to post air pollution data on its website which made it easier to access air pollution data from air quality stations pan India [30], and d) The number of AQ monitors increased drastically during the period of 2015-2018 [1] leading to availability of more data. The seasonal trend in $PM_{2.5}$ also becomes more stable and apparent after 2015 due to averaging effect of an increased number of stations (Figure 5). Thus, news

City	Average $PM_{2.5}$ ($\mu g/m^3$)	India limit	WHO limit	Time span ($PM_{2.5}$)	Articles (2010-21)
Delhi	142	**81%**	**98%**	02/2010 - 06/2021	5300
Gurugram	116	**71%**	**96%**	01/2016 - 06/2021	857
Mumbai	49	25%	64%	09/2014 - 06/2021	647
Bengaluru	36	11%	64%	03/2015 - 06/2021	538
Chennai	47	23%	81%	03/2015 - 06/2021	490
Kolkata	64	39%	64%	04/2018 - 06/2021	488
Nagpur	41	20%	58%	05/2011 - 06/2021	471
Pune	51	34%	76%	06/2015 - 06/2021	396
Hyderabad	47	29%	74%	03/2015 - 06/2021	386
Noida	114	**66%**	**94%**	07/2017 - 06/2021	369

Table 4: Ten most discussed cities in news media. All cities mentioned here except three cities (in bold) are less polluted than most cities in IG plain, but get higher news media attention.

media houses started to see a new storyline in air pollution and ramped up the coverage from 2014 onwards.

2) Periodic discussions after 2014: From 2015 onwards, news media attention has increased, but it is periodic and episodic, whereas air pollution is high throughout the year in Delhi. News articles drastically increase when Delhi is extremely polluted at the end of each year (October, November, and December). **Hence, 'air pollution' becomes a popular topic among news media when $PM_{2.5}$ level increases at the end of year. The popularity dies even though 81.28% of times Delhi's $PM_{2.5}$ level is above the India standard ($60\mu g/m^3$).**

5.1.3 *Which cities get high attention from the news media? Are these cities polluted throughout the year?*

Table 4 shows the cities discussed in news media along with the percentage of time these cities violate the minimum air quality standard. We observed that, except Delhi, Gurugram, and Noida, other cities are relatively clean in terms of $PM_{2.5}$ levels. Thus, **news media focus more on metro cities, although air pollution is not high compared to other cities in need of attention.**

5.2 Addressing Research Question 2

In this section we address the questions raised in Section 4.2.1.

5.2.1 *What are the topics discussed by news media? Are these topics pertinent with the ground reality of air pollution?*

To investigate the topics discussed in news media, we applied LDA as described in Section 4.2.2. We preprocessed each article to remove numbers, stop words, hyperlinks, and emails as they do not add any value in determining topics in LDA. Frequent words like 'Air' and 'Pollution' appear in almost all articles, eventually making a topic noisy. We removed words that appeared more than 80% of the time and less than 15% of the time [22] in the corpus. After preprocessing, we converted the entire corpus of 17.4K articles into a bag of words (BoW) representation and passed it to the LDA model. We tuned the number of iterations and number of topics (z) by calculating topic coherence [84]. We found optimal values of z to be 25 for 200 iterations. The value of the per-document topic distribution, α and per topic word distribution, β was as per the default settings in the gensim [70] library.

Visualizing Topics in LDA: Topics given by LDA are not directly interpretable by humans [23]. Thus, we use a web-based interactive visualization tool called *LDAvis* [78] to visualize the topic estimated by LDA. *LDAvis* provides a view of how topics differ from each other while allowing deep inspection of the terms most highly associated with each topic as shown in Figure 6.

Interpreting the LDA Visualisation: On the left of Figure 6, the topics are plotted as circles in the two-dimensional plane whose centers are determined by computing the Jensen–Shannon divergence [32] between topics, and then by using multidimensional scaling [26] to project the inter-topic distances onto two dimensions. Distance between two circles determines the similarity between two topics. For example, "poor air quality" and "weather" are close to each other revealing that air quality is highly influenced by weather conditions. The radius of a circle denotes the number of articles associated with a specific topic. For example, "Diwali" has lesser articles compared to "Climate Change". On the right side of Figure 6, the visualization depicts a horizontal bar-chart, whose bars represent the individual terms in a topic [25, 78]. We carefully evaluated the terms in each of the topics and assigned names to them. For example, the LDA algorithm returned words like 'climate', 'energy', 'change', and 'environmental' for a particular topic. All authors agreed to name this topic "Climate Change". Some of these topics and associated terms are mentioned in Table 5.

What are the topics of discussions in news media, and how do they evolve?: We showed in Section 5.1.2 that news media discussions spike up in October, November, and December. Here we reveal the topics of discussions. As mentioned in Section 4.2.2, LDA returns t ($t \leq z$) topics with some probability for each article. For example, LDA could return four topics for an article with probabilities 0.7 (health), 0.25 (Diwali), 0.04 (government), and 0.01 (weather). In this case, the article is most likely to be about 'health'. More the number of topics in an article, less likely the article would belong to one particular topic. We considered only those articles that belong to at least one topic with a probability of 0.5 or more. 43% articles belong to one topic with a confidence probability ≥ 0.5. We then plot topic frequency (most relevant topics) from 2015 to 2021 using a rolling window of 10 days as shown in Figure 7. We elaborate on these topics in detail by categorizing the topics in i) Event-specific and periodic discussions; ii) Event-specific and episodic discussions; and iii) Event agnostic discussions.

i) Event-specific and periodic discussions: These discussions contain the topics appearing periodically. A lot of news articles coinciding with the festival of **Diwali** appear during October and November. Burning firecrackers cause significant short-term degradation in air quality and harmful health effects [7, 48, 75]. Related terms such as 'green crackers' and 'green diwali' appear as a part of the discussions signifying that news media also focus on alternative ways of Diwali celebration. Other words like 'ban' and 'court' reflect the articles that discuss banning firecrackers and orders passed by the Supreme Court to act against crackers.

Stubble Burning is a phenomenon in which farmers burn their crop residues (stubble) to prepare their fields for the next season [20]. Previous studies [13, 28, 48, 85] show that the impact of 'stubble burning' on Delhi's air quality is higher than any other state in India

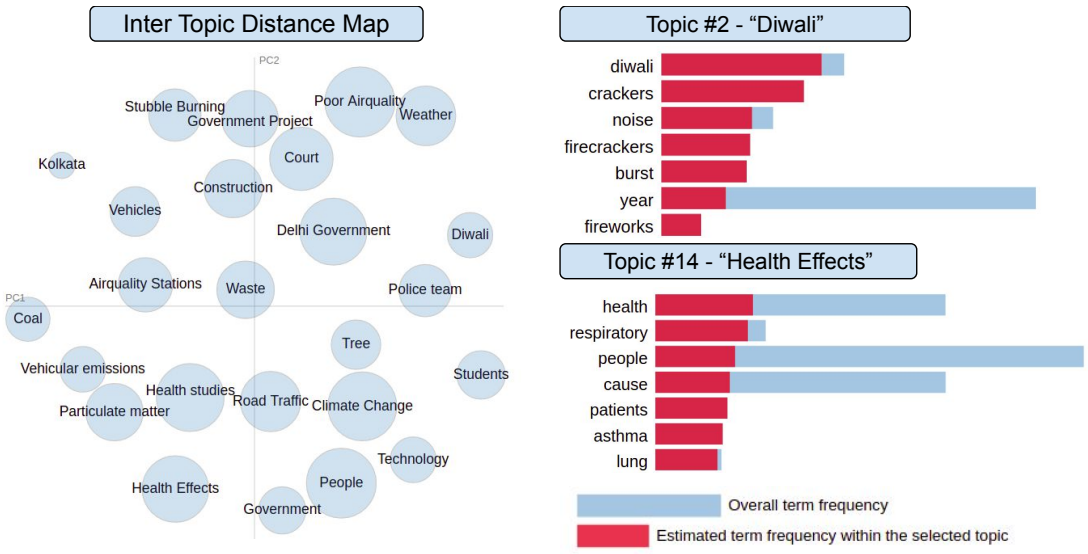

Figure 6: Inter topic distance map of 25 topics returned by LDA model. The radius of circles is proportional to the number of articles in a topic. Nearby topics (circles) are similar to each other. The blue bar on the right shows term frequency in the whole corpus, and the red bar shows term frequency in a particular topic. Interactive visualization is available here [39]. Note that Diwali is a festival in India which is celebrated by burning firecrackers.

Stubble burning	Weather	Delhi govt.	Poor air quality	Waste	Vehicular emissions	Climate Change	Construction	Students
burn	degrees	delhi	quality	waste	fuels	climate	construction	school
stubble	temp.	govt.	poor	garbage	power	energy	dust	awareness
farmers	celsius	minister	level	burn	diesel	change	control	children
punjab	rain	plan	wind	dump	emission	need	noida	campaign
crop	minimum	env.	delhi	residents	vehicles	env.	action	group
haryana	record	oddeven	category	municipal	petrol	global	board	class
paddy	maximum	scheme	severe	corporation	cars	research	measure	env.
farm	delhi	action	index	plastic	norms	countries	gurugram	event

Table 5: Most relevant topics and associated terms returned by LDA model. Topic names are manually created by authors based on the associated terms. Note that env. is environment and govt. is government.

due stubble burning in the neighboring states (Punjab and Haryana). Figure 7 shows that news media attention to 'stubble burning' increases during October-November. In the next paragraph, we discuss the stubble burning problem in detail.

Do news media report 'stubble burning' in accordance with open fire count in Punjab and Haryana?: To answer this question, we use open fire data (which is highly correlated with stubble burning in the regions of interest) from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor of NASA as done by previous studies [28, 42, 49, 79]. We extract the fire count data for Haryana and Punjab by manually selecting these regions in the GUI of NASA FIRM [72]. In the retrieved data, each fire count has a confidence value (C) varying between 1-100. To ensure fewer false detection, we filtered fire data having confidence of 'nominal' or 'high' ($C \geq 30$) [42, 49]. Finally, we get the daily count of 'open fires' from 2012 to 2020. Further, we compare the fire counts with the frequency of the articles on 'stubble burning'.

Stubble burning related articles: To identify the articles on 'stubble burning', we filtered the news articles for specific cities Delhi, Punjab, and Haryana, with keywords such as 'crop burning', 'stubble', 'burning crops', 'paddy burning', 'crop residue', and 'burning paddy'. These queries were selected by snowball sampling technique [16]. We found that 1280 articles are related to 'stubble burning' out of 6410 articles on Delhi, Punjab, and Haryana.

Figure 8 shows the comparison of open fire count and the number of articles on 'stubble burning'. We observed that: i) news media reporting on 'stubble burning' appeared significantly only after 2015, but, stubble burning occurred before 2015 also in Punjab and Haryana [28, 49]; ii) the fire count data reveals that fire incidents

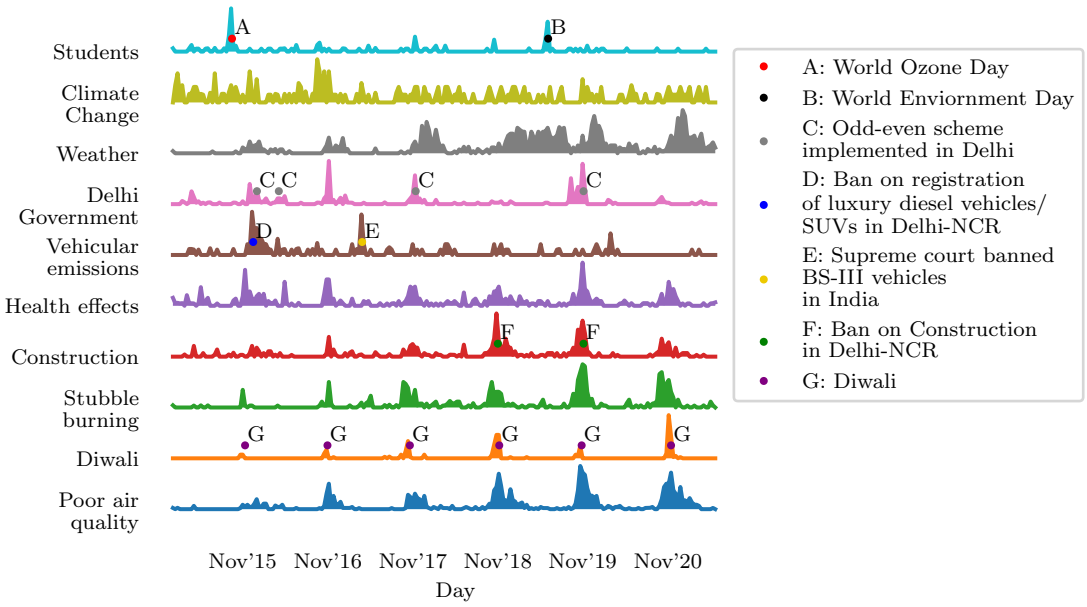

Figure 7: Topic evolution over time. Most of the topics are discussed at the end of the year (October-November) when $PM_{2.5}$ level is high. Some topics (such as 'Students' and 'Vehicular emissions') are influenced by specific events. Delhi-NCR stands for Delhi-National Capital Region which encompasses Delhi and several districts surrounding it from the states of Haryana, Uttar Pradesh and Rajasthan in India.

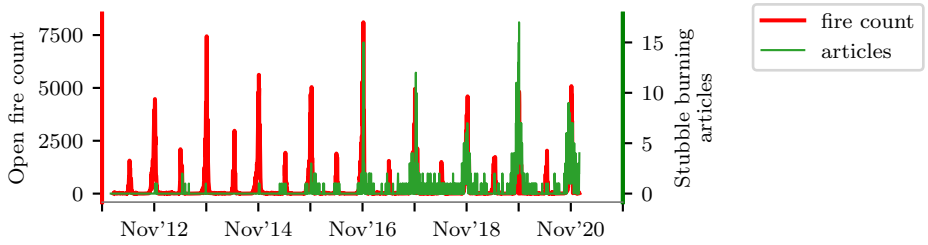

Figure 8: Open fire count data for 3 states (Punjab, Haryana, and Delhi). Stubble burning does not get enough media attention before 2016.

are significantly high during two time-periods every year (October-November and April-May). It indicates that 'stubble burning' happens two times a year, but news articles on 'stubble burning' appear only in October-November. We see the lack of news media attention on 'stubble burning' during the April-May period. This could happen because 'stubble burning' does not increase air pollution in Delhi-NCR regions due to weak northwesterly wind direction and high wind speed in Delhi-NCR compared to winters [44]. We further discuss the source contribution of 'Open burning' in Section 5.2.2. **ii) Event-specific and episodic discussions:** These discussions contain the topics which get highlighted after specific events related to air quality. i) **Vehicular emissions:** In October 2016, The local government in Delhi banned diesel vehicles of certain specifications for three months to control air pollution [50]. Again in March 2017, the Supreme Court banned the sale of BS-III standard vehicles in India [19]. ii) **Students:** Schools and colleges across India celebrate World Ozone Day on 16^{th} September to bring awareness towards air pollution. On 5^{th} June, 2019, 'air pollution' was the theme for world environment day [2]. iii) **Delhi government:** From Figure 7, we observe that this topic came in discussion when Delhi government implemented 'Odd-Even' scheme to curb air pollution in year 2016 (January and April), 2017 (November) and 2019 (November) [86]. **iii) Event agnostic but periodic discussions:** We observed some topics which are not event-specific but periodically discussed in October, November, and December. **i) Poor air quality:** This topic contains articles reporting about the effects of Diwali and Stubble burning by emphasizing the Air Quality Index (AQI). **ii) Construction:** news media discussed the 'Dust' effects from the construction

sites when the Supreme Court of India intervened for air pollution due to construction activities (Figure 7). 'Noida' and 'Gurugram' are highlighted in articles of 'construction' (Table 5). **iii) Health effects:** Health is also a year-long problem due to air pollution. However, discussions about health take place only at the end of each year (Figure 7). The term 'children' appears along with other health-related terms ('respiratory' , 'lungs', 'cancer', 'heart', and 'blood'), which indicates that children's health is discussed in news media as a major side effect of air pollution.
Climate change and weather conditions: Climate change is highlighted throughout the year in news media as shown in Figure 7. The discussions mainly focus on 'Hydro Fluoro Carbon (HFC)' and 'greenhouse emissions'. 'Weather' topic contains articles discussing the impact of weather on air quality.

5.2.2 ***What are the major sources of air pollution? Are these sources discussed by news media?*** There are five primary sources of air pollution as per the recent report by Health Effect Institute [53] (Table 6). The study also includes percentage source contribution to $PM_{2.5}$ for all states in India as per 2017. For this experiment, our objective was to analyze if news media discuss sources of air pollution in Delhi in proportion to the actual source contribution. We had 5.3K news articles (2010-21) in our corpus referring to Delhi's air pollution. We use the queries shown in Table 6, to identify the articles about sources of air pollution. We choose these queries by discussing with AQ experts and snowball sampling [16]. We found that 3285 out of 5.3K articles discussed the sources of Delhi's air pollution. One of the authors checked 100 random articles out of 3285 and found that 97 articles are relevant. Figure 9 shows the percentage source contribution to $PM_{2.5}$ in Delhi [53] , and the number of articles for each source.
We derive the following results from our analysis:

(i) 'Residential biomass' is a major source (32.3%) of $PM_{2.5}$ emissions, but it gets significantly less attention in the news articles (71 articles).
(ii) 'Vehicular emissions' related articles are the highest (1625 articles), but its contribution to $PM_{2.5}$ is 8%.
(iii) Percentage source contribution of 'Open burning' is 5.5%, but it gets high news media attention (1242 articles).
(iv) 'Industry and Energy production' is second highest source contributor to $PM_{2.5}$ (27.7%), but it has 754 articles which is lesser than articles of 'Vehicular emissions' and 'Open burning'.

The news media focuses more on 'Open-burning' and 'Dust' because the news coverage on air pollution is amplified post the harvest season in India. Post-harvest, the farmers burn the stubbles, which causes visible air pollution such as smog. Simultaneously, the focus also shifts towards 'Vehicular Emissions'. The general population of Delhi spends more than an hour in traffic [24]. As a result, perception also focuses on visible sources of pollution such as vehicular emissions during the same wave of air pollution discussion. This periodic wave of air pollution news and discussion completely misses out on non-visible sources of pollution such as 'residential cooking'.

6 DISCUSSION

Our work has shown a strong correlation between print media coverage and air pollution when the data become available. In Figure 5, we start seeing this correlation from 2015. As evident from the same figure, air pollution was a problem much before 2015, but we did not observe the same correlation before 2015 as air pollution data was not readily available. From January 2013, the United States Embassy in New Delhi started posting air quality data in its portal [37]. The OpenAQ website started posting air quality data on New Delhi from September 2015 [30]. Figure 5 shows that there is an increase in the volume of print news articles on air pollution post these landmarks due to the availability of data. The Central Pollution Control Board (CPCB) in India currently provides air quality data in Portable Document Format (PDF), making it difficult to parse. **We believe that air quality data should be made available sooner in a widely accepted format such as CSV to take timely actions against air pollution**. Change in mentality and public awareness about air pollution can only happen if enough timely information is available.

India's National Clean Air Programme (NCAP) [55] was launched in 2019 to meet clean air targets in 122 cities. The Union Government allocated a budget to all the 122 cities based on population. The budgetary allocation under NCAP for semi-urban cities in India like Bareily, Moradabad, Firozabad, all of which lie in the IG plain (Figure 3), is not enough to install even a single air quality monitor in these cities [21]. These cities have poor air quality (with yearly average $PM_{2.5}$ value of $114\mu g/m^3$, $121\mu g/m^3$ and $136\mu g/m^3$, respectively in the year 2016). Figure 3 shows that all rural areas in IG plain do not get English news media coverage and have an inadequate allocation from the NCAP. The Government should frame national policies to mitigate pollution in the rural areas of IG plain. **We believe that the NCAP budget allocation should be based on i) the cost of installing air pollution monitors for data availability and ii) should focus on data availability for rural regions**. Data availability complements in designing emission inventory through which sources of air pollution can be identified.

People tend to have little knowledge about the causes, evolution and sources of air pollutants [82]. Our research has shown that the contributions of various sources of air pollution discussed in newspaper media differ from scientific reality. Figure 9 shows that residential biomass is one of the key contributors to air pollution in Delhi, but the media hardly talks about it. Without proper awareness, people would not understand the significance of clean fuel. Furthermore, there is a geographic bias in print media reporting about air pollution. People perceive Delhi's air pollution as a problem because the media concentrated on the subject. This geographical bias in news media coverage makes people believe that Delhi's air pollution is a yardstick compared to other polluted cities. Fatehabad, a city in the Haryana state of India, had very poor air quality in 2020 [59] but people of Fatehabad may feel that the pollution in their town is not as harmful as Delhi because rarely there is a talk about it in the media. **The benchmark of judging the air quality should be based on its health effects**

Queries
Residential biomass: biomass fuels, wood burning, burning wood, firewood, cooking, household air pollution **Industry and Energy production:** coal, industry, industrial pollution, polluting industries, industrial emissions, energy sector, energy production, power plant, fossil fuel **Dust:** dust, dusts, construction, demolition **Vehicular emissions:** vehicles, vehicular emissions, petrol, diesel **Open burning:** open burning, garbage burning, burning garbage, waste burning, burning waste, crop burning, burning organic waste, stubble, burning crops, paddy burning, burning paddy

Table 6: Queries to identify news articles discussing the sources of air pollution.

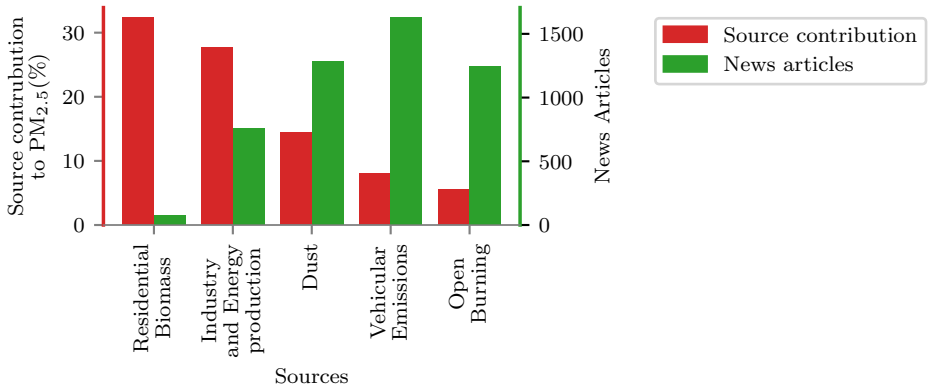

Figure 9: Comparison of source contributions to $PM_{2.5}$ as per report by Health Effect Institute [53] and related news articles for Delhi. Vehicular Emissions contribute less to the air pollution, but, is discussed highest times in news media articles. Residential Biomass contribute more but get less attention in news media.

and not on the air quality of another city. Many Indian cities suffer from air pollution, and communicating the health risk to the people via media is very important. **We believe that newspapers can play a pivotal role in dispersing scientifically valid knowledge of air pollution by curating articles that impact behavioral changes in people.** Furthermore, conversations on social media tend to reflect the narratives in traditional mass media like newspapers [57]. Newspapers articles amplified by social media have proven to be a strong tool in bringing policy changes [45]. Indian Express - a prominent media publication house, has successfully raised awareness of Delhi's air pollution [38]. Another broadcast news media named NDTV has successfully played the role of amplifying people's concern on a polluting industry; as a result, the industrial plant responsible for the pollution was shut down [60]. These examples show that print and digital media are paramount to raising awareness about air pollution and saving lives.

Finally, summarise our policy discussion points again:

- Data drives change. Air quality data should be made accessible by making it available on time in a widely accepted format such as CSV.
- Clean air budget allotment needs to be revised so that air quality monitors can be installed in rural areas. The new data on air pollution can be used to design emission inventory.
- The risk of air pollution should be communicated by raising awareness via news media.

7 LIMITATIONS AND FUTURE WORK

We now mention the limitations of the current work followed by potential future work to address them.

- Our study is limited to two English news media houses. Our rationale behind considering *The Hindu* and *Times of India* was, high readership and openly and easily available news-archive data. We could not consider similar other English dailies due to the lack of well-structured archives. Furthermore, in India, regional newspapers have high circulation [63]. We did not include regional newspapers in current work because of technical challenges. In the future, we can potentially apply language translation to analyze regional news media.
- Our focus in the current work was on print news media but due to low literacy (especially in some of the states spanning the IGP [3, 4]) , some people may not read the newspapers. Instead, they may get informed via broadcast media such as television news. Future work can look into the transcripts available for the broadcast media to understand the news media narrative around air pollution.
- In current work, we did not consider the positioning of the content in newspapers. For example, news on the front

page may have more chances of being read by people and thus have more impact compared to news on the non-front pages [31]. To do this, future work can use OCR techniques to parse online copy of the news-paper and classify the news based on their positions in the newspaper.

- In our experiments on understanding the scientific correctness of the sources and effects of air pollution, we only looked at the number of articles curated via handcrafted queries with respect to a source and effect. However, we do not consider tables and figures from such articles. Such tables and figures may contain include graphs or numbers around the source apportionment. To do such analysis, the future work needs to be able to read the images in the article and also parse the tables where such apportionment may be discussed.

8 CONCLUSION

Our work performs exploratory data analysis and topic modeling to study news media trends towards air pollution. We curated data of 17.4K news media articles from two English dailies and retrieved air pollution data for 88 cities in India spanning across 11 years. We found that news media articles on air pollution appear episodically, whereas air pollution is a year-long problem across several cities in India. Furthermore, the news articles are primarily focused on metropolitan cities, whereas the cities in the Indo-Gangetic plain bear the burns of air pollution. The causes of air pollution, as discussed in the media, deviate from the scientific reasons for air pollution. News media can act as an intermediator between the scientists and the government to find an amicable solution to air pollution.

ACKNOWLEDGMENTS

This research was supported by the Science and Engineering Research Board (SERB) project SRG/2020/001127 and Google Faculty Research Award. We would like to acknowledge the support from the Prime Minister's Research Fellowship (PMRF). Finally, we are grateful to the anonymous reviewers for their helpful reviews on revising the manuscript.

REFERENCES

[1] Delhi AQ Monitors Timesries 2009-19. https://urbanemissions.info/wp-content/uploads/images/2019-09-Delhi-AQ-Summary_1.jpg.
[2] World Environment Day 2019. https://www.unep.org/events/un-environment-event/world-environment-day-2019#:~:text=The%20theme%20for%202019%20is,worldwide%20efforts%20to%20address%20them..
[3] Bihar Literacy Rate 2022. https://www.indiacensus.net/states/bihar/literacy. (Accessed on 03/01/2022).
[4] Uttar Pradesh Literacy Rate 2022. https://www.indiacensus.net/states/uttar-pradesh/literacy. (Accessed on 03/01/2022).
[5] Rishiraj Adhikary, Zeel B Patel, Tanmay Srivastava, Nipun Batra, Mayank Singh, Udit Bhatia, and Sarath Guttikunda. 2021. Vartalaap: what drives# airquality discussions: politics, pollution or pseudo-science? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–29.
[6] Delhi air pollution Google Trends. https://trends.google.com/trends/explore?date=2010-01-01%202021-07-16&geo=IN-DL&q=air%20pollution.
[7] Balram Ambade. 2018. The air pollution during Diwali festival by the burning of fireworks in Jamshedpur city, India. *Urban climate* 26 (2018), 149–160.
[8] Tawfiq Ammari, Sarita Schoenebeck, and Daniel Romero. 2019. Self-declared throwaway accounts on Reddit: How platform affordances and shared norms enable parenting disclosure and support. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[9] Joshua S Apte, Michael Brauer, Aaron J Cohen, Majid Ezzati, and C Arden Pope III. 2018. Ambient PM2. 5 reduces global and regional life expectancy. *Environmental Science & Technology Letters* 5, 9 (2018), 546–551.
[10] The Hindu Archives. https://www.thehindu.com/archive/.
[11] Kalpana Balakrishnan, Sagnik Dey, Tarun Gupta, RS Dhaliwal, Michael Brauer, Aaron J Cohen, Jeffrey D Stanaway, Gufran Beig, Tushar K Joshi, Ashutosh N Aggarwal, et al. 2019. The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017. *The Lancet Planetary Health* 3, 1 (2019), e26–e39.
[12] beautifulsoup4 · PyPI. https://pypi.org/project/beautifulsoup4/.
[13] Gufran Beig, Saroj K Sahu, Vikas Singh, Suvarna Tikle, Sandeepan B Sobhana, Prashant Gargeva, K Ramakrishna, Aditi Rathod, and BS Murthy. 2020. Objective evaluation of stubble emission of North India and quantifying its impact on air quality of Delhi. *Science of The Total Environment* 709 (2020), 136126.
[14] Nandini Bhalla, Jane O'Boyle, and Dan Haun. 2019. Who is responsible for Delhi air pollution? Indian newspapers' framing of causes and solutions. *International Journal of Communication* 13 (2019), 24.
[15] Ramesh Bhushal. 2019. Asian Media Has Misled the Public on Air Pollution. https://thewire.in/environment/asian-media-misleads-public-on-air-pollution. [Online; accessed 20-February-2022].
[16] Patrick Biernacki and Dan Waldorf. 1981. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research* 10, 2 (1981), 141–163.
[17] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.
[18] Ilias Bougoudis, Konstantinos Demertzis, and Lazaros Iliadis. 2016. HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens. *Neural Computing and Applications* 27, 5 (2016), 1191–1206.
[19] registration of BS-III vehicles from April 1 Times of India BS-III Vehicles: SC bans sale. https://timesofindia.indiatimes.com/auto/miscellaneous/supreme-court-bans-sale-of-bs-iii-vehicles-from-april-1/articleshow/57891089.cms.
[20] Stubble burning Wikipedia. https://en.wikipedia.org/wiki/Stubble_burning.
[21] CarbonCopy and Respirer Living Sciences. 2022. NCAP Budget Tracker. https://ncaptracker.in/budget-dashboard/. [Online; accessed 14-February-2022].
[22] Youngchul Cha and Junghoo Cho. 2012. Social-network analysis using topic models. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 565–574.
[23] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*. 288–296.
[24] Capital chaos: Delhi's traffic has slowed down and doubled time spent on roads | Latest News Delhi Hindustan Times. https://www.hindustantimes.com/delhi/capital-chaos-delhi-s-traffic-has-slowed-down-and-doubled-time-spent-on-roads/story-ZTp1UviD50hOXvdZpGs8FN.html#:~:text=Today%2C%20a%20person%20travelling%20a,to%201.36%20hours%20in%202011.. (Accessed on 02/28/2022).
[25] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 443–452.
[26] Michael AA Cox and Trevor F Cox. 2008. Multidimensional scaling. In *Handbook of data visualization*. Springer, 315–347.
[27] CPCB. https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/caaqm-data-availability. .
[28] Daniel H Cusworth, Loretta J Mickley, Melissa P Sulprizio, Tianjia Liu, Miriam E Marlier, Ruth S DeFries, Sarath K Guttikunda, and Pawan Gupta. 2018. Quantifying the influence of agricultural fires in northwest India on urban air pollution in Delhi, India. *Environmental Research Letters* 13, 4 (2018), 044018.
[29] Sagnik Dey, Bhavesh Purohit, Palak Balyan, Kuldeep Dixit, Kunal Bali, Alok Kumar, Fahad Imam, Sourangsu Chowdhury, Dilip Ganguly, Prashant Gargava, et al. 2020. A Satellite-Based High-Resolution (1-km) Ambient PM2. 5 Database for India over Two Decades (2000–2019): Applications for Air Quality Management. *Remote Sensing* 12, 23 (2020), 3872.
[30] AWS S3 Explorer. https://openaq-fetches.s3.amazonaws.com/index.html. (Accessed on 02/25/2022).
[31] Anastassia Fedyk. 2018. *Front page news: The effect of news positioning on financial markets*. Technical Report. Working paper.
[32] Bent Fuglede and Flemming Topsoe. 2004. Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings*. IEEE, 31.
[33] Harold Gene Zucker. 1978. The variable nature of news media influence. *Annals of the International Communication Association* 2, 1 (1978), 225–240.
[34] PyPI geopy. https://pypi.org/project/geopy/.
[35] Sarath K Guttikunda, KA Nishadh, and Puja Jawahar. 2019. Air pollution knowledge assessments (APnA) for 20 Indian cities. *Urban Climate* 27 (2019), 124–141.
[36] Cody T Havard, Patrick Ferrucci, and Timothy D Ryan. 2021. Does messaging matter? Investigating the influence of media headlines on perceptions and attitudes of the in-group and out-group. *Journal of Marketing Communications* 27, 1 (2021), 20–30.
[37] Air Quality Data U.S. Embassy & Consulates in India. https://in.usembassy.gov/embassy-consulates/new-delhi/air-quality-data/. (Accessed on 02/25/2022).

[38] The Indian Express Indian Express journalists awarded for series on Delhi's air pollution | India News. https://indianexpress.com/article/india/india-news-india/indian-express-journalists-awarded-for-series-on-delhis-air-pollution/. (Accessed on 02/26/2022).

[39] LDA interactive visualization. https://karm-patel.github.io/Samachar-News-media-on-air-pollution/Research_Question_2/LDA_Visualization/LDA_topic_modelling_Visualization.html.

[40] Rajmal Jat and Bhola Ram Gurjar. 2021. Contribution of different source sectors and source regions of Indo-Gangetic Plain in India to PM2. 5 pollution and its short-term health impacts during peak polluted winter. *Atmospheric Pollution Research* 12, 4 (2021), 89–100.

[41] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78, 11 (2019), 15169–15211.

[42] Hiren T Jethva, Duli Chand, Omar Torres, Pawan Gupta, Alexei Lyapustin, and Falguni Patadia. 2018. Agricultural burning and air quality over northern India: a synergistic analysis using NASA's A-train satellite data and ground measurements. *Aerosol and Air Quality Research* 18, PNNL-SA-125481 (2018).

[43] Vital Strategies JH. 2019. Public Understanding of Air Quality and its Health Impact in South and Southeast Asia, 2015–2018. https://www.vitalstrategies.org/wp-content/uploads/import/2019/03/Hazy_Perceptions.pdf. [Online; accessed 20-February-2022].

[44] Leena Ajit Kaushal. 2020. Examining the policy-practice gap-The issue of crop burning induced Particulate Matter pollution in Northwest India. *Ecosystem Health and Sustainability* 6, 1 (2020), 1846460.

[45] Samuel Kay, Bo Zhao, and Daniel Sui. 2015. Can social media clear the air? A case study of the air pollution problem in Chinese cities. *The Professional Geographer* 67, 3 (2015), 351–363.

[46] Tobias R Keller, Valerie Hase, Jagadish Thaker, Daniela Mahl, and Mike S Schäfer. 2020. News media coverage of climate change in India 1997–2016: using automated content analysis to assess themes and topics. *Environmental Communication* 14, 2 (2020), 219–235.

[47] Gary King, Benjamin Schneer, and Ariel White. 2017. How the news media activate public expression and influence national agendas. *Science* 358, 6364 (2017), 776–780.

[48] LS Kurinji, Adeel Khan, and Tanushree Ganguly. 2021. Bending Delhi's Air Pollution Curve. (2021).

[49] Kurinji LS. 2019. *Alternative Methods to Monitor Air Pollution: A Study of Crop Residue Burning in Punjab.* Council on Energy, Environment and Water, New Delhi.

[50] NCR The Hindu Luxury diesel vehicles banned in Delhi. https://www.thehindu.com/news/cities/Delhi/sc-bans-registration-of-diesel-suvs-in-delhi-till-march/article7995819.ece.

[51] Michela Maione, Elisabetta Mocca, Kristina Eisfeld, Yuri Kazepov, and Sandro Fuzzi. 2021. Public perception of air pollution sources across Europe. *Ambio* 50, 6 (2021), 1150–1158.

[52] Brian Mayer. 2012. 'Relax and take a deep breath': Print media coverage of asthma and air pollution in the United States. *Social Science & Medicine* 75, 5 (2012), 892–900. https://doi.org/10.1016/j.socscimed.2012.04.024

[53] Erin E. McDuffie, Randall V. Martin, Joseph V. Spadaro, Richard Burnett, Steven J. Smith, Patrick O'Rourke, Melanie S. Hammer, Aaron van Donkelaar, Liam Bindle, Viral Shah, Lyatt Jaeglé, Gan Luo, Fangqun Yu, Jamiu A. Adeniran, Jintai Lin, and Michael Brauer. 2021. Source sector and fuel contributions to ambient PM2.5 and attributable mortality across multiple spatial scales. *Nature Communications* 12, 1 (14 Jun 2021), 3594. https://doi.org/10.1038/s41467-021-23853-y

[54] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[55] Government of India Ministry of Housing & Urban Affairs. 2022. Long-Term, Time-Bound, National Level Strategy to Tackle Air Pollution-National Clean Air Programme (NCAP). https://pib.gov.in/PressReleasePage.aspx?PRID=1655203. [Online; accessed 14-February-2022].

[56] Ranjini Murali, Aishwarya Kuwar, and Harini Nagendra. 2021. Who's responsible for climate change? Untangling threads of media discussions in India, Nigeria, Australia, and the USA. *Climatic Change* 164, 3 (2021), 1–20.

[57] Nandita Murukutla, Namrata Kumar, and Sandra Mullin. 2019. A review of media effects: Implications for media coverage of air pollution and cancer. *Annals of Cancer Epidemiology* 3, 3 (2019).

[58] Nandita Murukutla, Nalin S Negi, Pallavi Puri, Sandra Mullin, and Lesley Onyon. 2017. Online media coverage of air pollution risks and current policies in India: a content analysis. *WHO South-East Asia journal of public health* 6, 2 (2017), 41–50.

[59] Sukanya Nair. 2021. Air quality worse in smaller towns in Indo-Gangetic Plains compared to bigger cities, says CSE's latest analysis. https://www.cseindia.org/air-quality-worse-in-smaller-towns-in-indo-gangetic-plains-compared-to-bigger-cities-10614. [Online; accessed 26-February-2022].

[60] Bengaluru NDTV Impact: Top Court To Look At Air Pollution In Whitefield. https://www.ndtv.com/video/shows/trending-10/ndtv-impact-top-court-to-look-at-air-pollution-in-whitefield-bengaluru-496214. (Accessed on 02/26/2022).

[61] newspaper3k · PyPI. https://pypi.org/project/newspaper3k/.

[62] Times of India Archives. https://timesofindia.indiatimes.com/archive.cms.

[63] List of newspapers in India by readership Wikipedia. https://en.wikipedia.org/wiki/List_of_newspapers_in_India_by_readership.

[64] Readership of The Hindu and TOI. https://www.thehindu.com/news/national/the-hindu-is-indias-fastest-growing-english-daily-fourth-time-in-a-row/article31565845.ece.

[65] Kristin L. Olofsson, C. Weible, Tanya Heikkila, and J. Martel. 2018. Using Non-profit Narratives and News Media Framing to Depict Air Pollution in Delhi, India. *Environmental Communication* 12 (2018), 956 – 972.

[66] OpenAQ. https://openaq.org/#/.

[67] Anamika Pandey, Michael Brauer, Maureen L Cropper, Kalpana Balakrishnan, Prashant Mathur, Sagnik Dey, Burak Turkgulu, G Anil Kumar, Mukesh Khare, Gufran Beig, et al. 2021. Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019. *The Lancet Planetary Health* 5, 1 (2021), e25–e38.

[68] Pallavi Pant, Raj M Lal, Sarath K Guttikunda, Armistead G Russell, Ajay S Nagpure, Anu Ramaswami, and Richard E Peltier. 2019. Monitoring particulate matter in India: recent trends and future outlook. *Air Quality, Atmosphere & Health* 12, 1 (2019), 45–58.

[69] Khaiwal Ravindra, Maninder Kaur Sidhu, Suman Mor, Siby John, and Saumyadipta Pyne. 2016. Air pollution in India: bridging the gap between science and policy. *Journal of Hazardous, Toxic, and Radioactive Waste* 20, 4 (2016), A4015003.

[70] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[71] SPECIAL REPO. 2018. Burden of disease attributable to major air pollution sources in India. (2018).

[72] NASA FIRMS Creating Archive Download Request. https://firms.modaps.eosdis.nasa.gov/download/create.php.

[73] Python library Requests. https://pypi.org/project/requests/.

[74] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*. 399–408.

[75] M Sateesh, VK Soni, and PVS Raju. 2018. Effect of diwali firecrackers on air quality and aerosol optical properties over mega city (Delhi) in India. *Earth Systems and Environment* 2, 2 (2018), 293–304.

[76] Julian Schwabe. 2018. The Impact of Severe Air Pollution in January 2013 in Beijing on Sustained Elevation of Public Environmental Concern. *European Journal of East Asian Studies* (2018).

[77] Elizabeth A Shanahan, Mark K McBeth, and Paul L Hathaway. 2011. Narrative policy framework: The influence of media policy narratives on public opinion. *Politics & Policy* 39, 3 (2011), 373–400.

[78] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 63–70.

[79] Dineshkumar Singh, Jayantrao Mohite, Suryakant Sawant, and Srinivasu Pappula. 2020. Monitoring and Analysis of Viirs Fire Events Data Over Indian States of Punjab and Haryana. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 4538–4541.

[80] Torsten Skov, Torben Cordtz, Lilli Kirkeskov Jensen, Peter Saugman, Kirsten Schmidt, and Peter Theilade. 1991. Modifications of health behaviour in response to air pollution notifications in Copenhagen. *Social Science & Medicine* 33, 5 (1991), 621–626.

[81] Michael D Slater. 2007. Reinforcing spirals: The mutual influence of media selectivity and media effects and their impact on individual behavior and social identity. *Communication theory* 17, 3 (2007), 281–303.

[82] Kirsty Smallbone. 2010. Individuals' interpretation of air quality information.

[83] Tongxin Sun and Bu Zhong. 2020. A tale of four cities: A semantic analysis comparing the newspaper coverage of air pollution in Hong Kong, London, Pittsburgh, and Tianjin from 2014 to 2017. *Newspaper Research Journal* 41, 1 (2020), 37–52. https://doi.org/10.1177/0739532919873438 arXiv:https://doi.org/10.1177/0739532919873438

[84] Shaheen Syed and Marco Spruit. 2017. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE, 165–174.

[85] Krishna Vadrevu, E. Ellicott, K.V.S. Badarinath, and E. Vermote. 2011. MODIS derived fire characteristics and aerosol optical depth variations during the agricultural residue burning season, north India. *Environmental pollution (Barking, Essex : 1987)* 159 (03 2011), 1560–9. https://doi.org/10.1016/j.envpol.2011.03.001

[86] Benefits of Delhi Odd-Even Scheme | Business Standard What is Odd-Even Scheme, Origin. https://www.business-standard.com/about/what-is-odd-even-scheme.

[87] Indo-Gangetic Plain Wikipedia. https://en.wikipedia.org/wiki/Indo-Gangetic_Plain.