Samachar: Print News Media on Air Pollution in India

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

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

Nipun Batra
IIT Gandhinagar
Gandhinagar, India

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

Sarat 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 curate 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:

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 houses¹ 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
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\(^3\). 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 analysis of 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.

\(^3\)https://github.com/karm-patel/Samachar-News-media-on-air-pollution
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 BTEX are of major concern to air quality, but PM$_{2.5}$ exceed 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 (2010-2021) 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 scraped 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 ‘AQ’ 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.

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.

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% ($\kappa$) [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, 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, odd even, car emissions

<table>
<thead>
<tr>
<th>Week</th>
<th>TOI</th>
<th>The Hindu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov’09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov’11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov’13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov’15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov’17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov’19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nov’21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 analyze 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 (25µg/m$^3$) and India (60µg/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
We find the most discussed cities from this plot, where each article has a concentration parameter \( \alpha \) distribution of words in a topic. Concentration parameter \( \theta \) is the probability distribution of topics per document. Concentration parameter \( \beta \) represents word density per document, and \( \beta \) represents word density per topic. We feed our article corpus into the LDA and evaluate the output 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.

### 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 µg/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
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 the year. The popularity dies even though 81.28% of times Delhi’s PM$_{2.5}$ level is above the India standard (60µg/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, $\alpha$ and per topic word distribution, $\beta$ 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.

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.

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

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 ≥ 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
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 16th September to bring awareness towards air pollution. On 5th 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

---

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.
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 Bareilly, 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 114µg/m$^3$, 121µg/m$^3$, and 136µg/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.](image)

...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
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 intermediary between the scientists and the government to find an amicable solution to air pollution.

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 intermediary 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

[22] 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-delhis-traffic-has-slowed-down-and-doubled-time-spent-on-roads/story-2Tp1Uvd5b6OhOvXeZpGis8FN.html#:~:text=Today%20s%20person%20traveling%20at,02%20hours%20at%2002:01.
[29] Sarath K Guttikunda, KA Nishadh, and Puja Jawahar. 2019. Air pollution knowledge and awareness of the sources and effects of air pollution, we only looked to parse online copy of the newspaper and classify the news pages [31]. To do this, future work needs 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 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.
Samachar: Print News Media on Air Pollution in India

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

The Indian Express Indian Express journalists awarded for series on Delhi’s air pollution | India News https://indianexpress.com/article/india/india-news/india-print-indian-express-journalists-awarded-for-series-on-delhis-air-pollution/

(Accessed on 02/26/2022)


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-


https://www.pyjournal.com/project/newspaper3k/.


https://www.india.com/newspaper3k.