Vartalaap: What Drives #AirQuality Discussions: Politics, Pollution or Pseudo-science?

RISHIRAJ ADHIKARY, IIT Gandhinagar, India
ZEEL B PATEL, IIT Gandhinagar, India
TANMAY SRIVASTAVA, IIT Gandhinagar, India
NIPUN BATRA, IIT Gandhinagar, India
MAYANK SINGH, IIT Gandhinagar, India
UDIT BHATIA, IIT Gandhinagar, India
SARATH GUTTIKUNDA, UrbanEmissions.info, India

Air pollution is a global challenge for cities across the globe. Understanding the public perception of air pollution can help policymakers engage better with the public and appropriately introduce policies. Accurate public perception can also help people to identify the health risks of air pollution and act accordingly. Unfortunately, current techniques for determining perception are not scalable: it involves surveying few hundred people with questionnaire-based surveys. Using the advances in natural language processing (NLP), we propose a more scalable solution called *Vartalaap* to gauge public perception of air pollution via the microblogging social network Twitter. We curated a dataset of more than 1.2M tweets discussing Delhi-specific air pollution. We find that (unfortunately) the public is supportive of unproven mitigation strategies to reduce pollution, thus risking their health due to a false sense of security. We also find that air quality is a year-long problem, but the discussions are not proportional to the level of pollution and spike up when pollution is more *visible*. The information required by *Vartalaap* is publicly available and, as such, it can be immediately applied to study different societal issues across the world.

CCS Concepts: • **Human-centered computing → Social and Behavioral Science**.

Additional Key Words and Phrases: social media; air pollution; perception

ACM Reference Format:
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?. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 96 (April 2021), 29 pages. https://doi.org/10.1145/3449170

1 INTRODUCTION

Ambient fine particulate ($PM_{2.5}$) is the most significant risk factor for premature death, shortening life expectancy at birth by 1.5 to 1.9 years [7]. 91% of the world's population lives in areas where air pollution exceeds safety limits[1]. 99% of the people in countries like India, Pakistan, Nepal, and Bangladesh experience ambient exposures of $PM_{2.5}$ exceeding 75 $\mu g/m^3$ to 100 $\mu g/m^3$ [8]. India is among the top 13 countries with the highest number of death (82 per million in 2016) attributable to

[1]https://www.who.int/health-topics/air-pollution Last accessed: 8 October 2020

Authors' addresses: Rishiraj Adhikary, IIT Gandhinagar, India; Zeel B Patel, IIT Gandhinagar, India; Tanmay Srivastava, IIT Gandhinagar, India; Nipun Batra, IIT Gandhinagar, India; Mayank Singh, IIT Gandhinagar, India; Udit Bhatia, IIT Gandhinagar, India; Sarath Guttikunda, UrbanEmissions.info, New Delhi, India.

2573-0142/2021/4-ART96 $15.00
https://doi.org/10.1145/3449170

ambient air pollution[2]. Previous studies show that timely and informed action against air pollution can help policymakers make informed decisions to reduce health risk [15, 49]. The successful implementation of mitigation strategies depends on public participation and response. Both public participation and response are highly dependent on the public's perception of air pollution and its associated health effects. For instance, studies show that people are less likely to go outdoors when perceived pollution is high [61]. Understanding public perception in social media is an active research area [24, 36].

Understanding the public perception of air pollution and its associated health effects can help policymakers engage with the public, appropriately educate and introduce mitigation strategies, understand public participation and pain points, and estimate the odds of success for a mitigation strategy. Further, in an ideal case, the threat perception from air pollution should be proportional to the severity of air pollution [2, 3]. Effective communication and education strategies can potentially help the public understand the health impacts of air pollution better. Thus, accurately understanding perception can inform policymakers and potentially save millions of lives [22, 28, 68].

Existing techniques, like questionnaire-based surveys, are not scalable to understand air pollution perception. They do not incorporate temporal reaction changes in opinions and involve a few, tens to a hundred participants only. In this paper, we leverage the advances in natural language processing (NLP) and the volume, variety, and velocity of data [51] to propose a scalable approach called *Vartalaap*[3] to gauge public perception of air pollution via the microblogging platform Twitter.

We focus our study on air pollution in the capital city of Delhi, India. We choose Delhi as our testbed because i) We have local expertise about air pollution in Delhi; ii) Delhi is often called the most polluted city in the world; iii) Delhi has a dense population, and thus potentially millions of people are at health risk; iv) Various mitigation strategies have been proposed so far in Delhi, and people extensively discuss these strategies in social media. In this work, we have collected more than 1.2M tweets reflecting Twitter discussions on Delhi's air pollution from more than 26K individual users based on 34 search queries and hashtags on air pollution (January 2016 to March 2020).

This paper broadly discusses four intertwined research questions

(1) What is the public sentiment towards different mitigation strategies?
(2) Does air pollution 'cause' Twitter discussions in proportion to pollution levels?
(3) Who are the air quality protagonists, and how do they influence the discussion?
(4) What are the set of topics discussed, and how do the topics evolve?

First, we evaluate the sentiment associated with untested mitigation strategies on air pollution. Untested mitigation strategies can give the public a false sense of security besides wasting taxpayers' money. We manually labelled a set of 1523 tweets representing two particular air polluting mitigation strategies. We also trained a sentiment classifier and demonstrate its state-of-the-art performance gain. Overall, we found that sentiment towards both the strategies is supportive, albeit unscientific.

Most of the tweets on mitigation strategies are event-driven, which let us presume a relationship between pollution levels and Twitter discussion. Thus we propose our second research question i.e, whether air pollution is a year-long problem in Delhi and does it 'Granger-cause' [30] Twitter

[2]https://www.who.int/gho/phe/outdoor_air_pollution/burden/en/ Last accessed: 8 October 2020
[3]Vartalaap means conversation in Hindi

discussion? We first show that air pollution remains above the World Health Organisation (WHO) standard for more than 90% of the time. To verify 'Granger-causality', we ensure that the two time series ($PM_{2.5}$ and number of tweets) are stationary. We found that $PM_{2.5}$ 'Granger-causes' Twitter discussion only in winter months while pollution is above WHO limits throughout the year, thus harming health. This finding suggests the need for a sustained conversation throughout the year.

Having understood the awry discussion throughout the year, we apply topic modelling to understand the discussed topics and their evolution. Topic modelling helps to understand if the public has a good understanding of causes, implications, and remedies for air pollution. We categorise the observed topics into i) event specific such as short-term mitigation strategies, fog v/s smog, world health day, climate change, and; ii) event agnostic such as lung health, and road traffic. A significant amount of air quality discussion comes from event-specific topics. The discussion is limited to a few months and a small set of users. Lastly, we address our third research question, i.e. is the air quality discussion dominated by a small group of users who follow power-law characteristics?

Our research can significantly improve the scalability of air quality perception (and similar societal issues) studies by mining social media data. Our work on air pollution perception is over an extended time, involves 1.2M tweets from 26.4K users, and hence more extensive in scale compared to questionnaire-based surveys. To the best of our knowledge, this is the first large scale study in terms of dataset and analysis performed to understand the air pollution perception. Our study can be replicated to any other part of the world, and our techniques can be adapted for real-time information via a dashboard to policymakers.

Reproducibility and Dataset release: We believe that our work is fully reproducible, and we publish 'Tweet IDs' of tweets used in our work adhering to the Twitter ToS (Terms of Service) in our project repository[4]. Archiver tools such as DocNow's twarc[5] can be used to archive the tweets' metadata from 'Tweet IDs'. All our tables and figures are reproducible with the code shared in the same URL.

We now discuss the paper organisation. Section 2 describes the related work. Section 3 explains the dataset curation process. Section 4 motivates all the research questions and lays down the approach to validating them. Section 5 evaluates the research questions and Section 5.3 evaluates the topic model. We discuss the limitation and future direction of our work in Section 6 and conclude in Section 7.

2 RELATED WORK

We can categorise the related work in air quality perception into two categories: i) Questionnaire-based surveys, and ii) Social media sentiment and topic analysis. We now describe the perception studies based on questionnaire-based surveys.

2.1 Perception studies using questionnaire-based surveys

Traditionally, researchers have used questionnaire surveys to collect data on perceived air quality level and public displeasure with air pollution. Egondi et al. [28] performed a cross-sectional study of 5,317 individuals aged 35+ years in Nairobi in the year 2013. The study established levels and associations between perceived pollution and health risk perception among slum residents. They

[4]https://github.com/rishi-a/Vartalaap
[5]https://github.com/DocNow/twarc Last accessed: 8 October 2020

found a mismatch between air pollution and its perceived levels. A similar mismatch was reported by Peng et al. [44]. They used social survey data from over 5000 respondents and statistical data from the Ministry of Environment Protection of China. In Canada, Atari et al. [9] conducted a similar survey and observed a significant correlation between odour annoyance scores and modelled ambient pollution. In another study, Semenza et al. [56] revealed that only ~ 10-15% of one-third of the participants claimed to have reduced outdoor activities during high pollution as per government advisory. Thus, the advisory may not lead to behaviour change as much as air quality perception does. Bickerstaff et al. [13] interviewed members of the public and demonstrated the need to understand air pollution perception if the objectives of environmental are to be achieved. On the policy front, Huang et al. [34] conducted surveys on three cities of China and revealed a gap in China's policy objective and public acceptable risk level from air pollution.

Questionnaire-based surveys have focused more on finding a relationship between actual air quality and its perceived levels. There is a need to study the causal analysis between air pollution and its perception. There is also a need to understand what changes public opinion. In contrast, our work scales to a much larger population and is not limited to a one-time survey.

2.2 Social media for sentiment analysis

Social media is now used as the largest ubiquitous sensor. It has been used to understand public sentiments on climate change [1, 6, 23], air quality [27, 31, 63], and finding correlation between air quality index (AQI) and volume of social media messages [37]. Abbar et al. [1] analysed conversations and sensed public awareness of climate change. They examined the discourse in 36K tweets which talk about climate change. In another work, An et al. [6] attempts to understand whether Twitter data mining can complement and supplement insights about climate change perception. They showed how Twitter data could be used to illustrate the change in opinions over time after specific events. Dahal et al. [23] performed topic modelling and sentiment analysis on geotagged tweets to conclude that climate change discussion on Twitter concerning USA residents is less focused on policy-related topics than other countries. Deteriorating air quality is a primary concern for countries like India and China. In China, Tao et al. [63] collected 27,500 comments from a Chinese microblog regarding the air quality of primary tourist destinations in China. Results indicated that tourists' perceptions of air quality were mainly positive, and they had a poor air pollution crisis awareness. Dong et al. [27] used data from another Chinese microblog and explored the relationship between the actual level of air pollution and residents concern about air pollution. They found out that residents perceived the deprivation of air quality and expressed their interest in air pollution within a day after the pollution level rose. There is a need to extend such studies to one of the most polluted and densely populated cities of the world, i.e. Delhi (India).

In India, Basu et al. [11] tried to understand the public perception of a Government policy targeting air pollution mitigation. Gurajala et al. [31] collected Twitter data for nearly two years and analysed these data for three major cities, namely, Paris, London and New Delhi. They identified three hashtags that best determine people response to air quality and also conclude that health-related discussions spike up when air quality indices deteriorate. Furthermore, topic modelling analysis revealed topics associated with sporadic air quality events, such as fireworks during festivals. Pramanik et.al. [46] tailored an algorithm to identify the users who actively tweet about air pollution events in Delhi, India. They curated a dataset of 166K tweets using six keywords and hashtags.

The studies on climate change and air quality perception have so far been small in scale either concerning data collection or for the data being finally analysed.

3 DATASET

The current study curates rich time-stamped datasets from two different sources. The first dataset comprises air quality real-time readings. The second dataset contains tweets discussing the air quality and its significant repercussions. Table 1 shows an overview.

3.1 The $PM_{2.5}$ Dataset

The $PM_{2.5}$ dataset comprises near real-time information on particulate matter of size less than 2.5 microns in diameter from the majority of locations in Delhi. The data is curated from Central Pollution Control Board[6], India sourced via OpenAQ[7]. OpenAQ posts raw files on $PM_{2.5}$ values. The data is available at every 10 minutes interval from different stations across the city. We define Delhi $PM_{2.5}$ data as a mean aggregate of the data from all the stations. Since there can be deviations in $PM_{2.5}$ readings across different locations in the city, we averaged across locations to create a single representative reading for the city. We removed data points which were missing and outside the measurement range[6] of $0\ \mu g/m^3 \leq PM_{2.5} < 1000\ \mu g/m^3$. After data cleaning, we took the average of all $PM_{2.5}$ values on a single day. Thus, we have a $PM_{2.5}$ value for every day from 2016 to 2019 for the entire city of Delhi.

3.2 The Tweet Dataset

We collect publicly available Twitter data using a Python library named 'GetOldTweets-python'[8]. We have 1.2 Million tweets on Delhi air pollution from 26.4K unique users. We use the queries mentioned in Table 2, which includes keywords and hashtags. We chose these specific queries by: i) consulting with air quality experts; ii) studying trending queries; iii) initial exploratory analysis; iv) studying hashtags and text used by a few sets of accounts that exclusively talk about air quality, and v) Snowball sampling, i.e. using keywords that appear in the tweets retrieved from other queries. For example, while investigating tweets retrieved using the query 'delhi air quality', we found that the queries 'delhi smog' and 'delhi fog' were interchangeably used to discuss Delhi's air pollution. The data set also contains meta-information, including username, follower count, and likes.

Verifying if the collected data is about Delhi air pollution: We found that there were several irrelevant tweets whenever we used a more general query. For example, the query 'air pollution' returned tweets specific to China, Germany, UK and Delhi. To choose Delhi relevant tweets, we changed our query ('delhi air pollution') and re-evaluated 200 random tweets from that query. We iterated this procedure for other queries as well. We used lookahead and look behind regular expressions in conjunction with these queries. We finally curated a list of 34 Delhi specific queries. Finally, we manually investigated 1000 random tweets from the final 1.2 M tweets (curated using 34 queries) and found 2.4% of them to be irrelevant to Delhi air pollution. Most of the unrelated tweets were advertisements which used air pollution-related hashtag.

4 APPROACH

In this section we state our research questions and explain our approach to address them.

[6]https://cpcb.nic.in/ Last accessed: 8 October 2020

[7]https://openaq.org Last accessed: 8 October 2020

[8]https://github.com/Jefferson-Henrique/GetOldTweets-python Last accessed: 8 October 2020 (such web-crawling is allowed under Twitter ToS adhering to terms and conditions)

$PM_{2.5}$	Time-period	Jan'16 to Dec'19
	Locations	13
	Granularity level	1 Hour
Twitter	Time-period	Jan'16 to Apr'20
	Total Tweets	1,252,999
	User Profiles	2,645,61

Table 1. Salient statistics of the curated datasets.

Queries
delhiagainstpollution, delhi against pollution, delhipollution, delhi pollution, smogtower, smog tower, delhi air, delhiair, delhi air emergency, delhiairemergency, delhichokes, delhi chokes, delhi air quality, delhiairquality, delhi smog, delhismog, oddeven, odd even, delhi fog, delhifog, air lodhi garden, air sarojini nagar, air chandni chowk, air gurgaon, delhincr, air ncr, air noida, air punjab, air haryana, air vehicle delhi, air road delhi, air school delhi, air children delhi, stubble burning delhi

Table 2. List of 34 queries used to identify Delhi based tweets

4.1 RQ1: What is the public sentiment associated with untested mitigation strategies?

4.1.1 Background. Governments across the world have proposed and implemented several strategies for reducing air pollution. In Germany renewable energy investment and expansion of public transport is being promoted[9]. The Netherland government has industrial emissions policy to combat air pollution[10]. Various mitigation strategies are also being adopted in India. These strategies include, but, are not limited to: i) using higher-grade fuel for vehicles[11]; ii) cutting emissions from power plants[12]. However, some of the proposed and implemented strategies have not yet been proven to work. In fact, several studies [20, 43] suggest that these strategies will not help reduce air pollution. One such highly debated strategy is installing outdoor air purifiers called "Smog Tower" [17]. These "Smog Towers" have been piloted in China in 2018[13]. In India, a "Smog Tower" was installed by a member of parliament and inaugurated in the first week of January 2020. A recent paper outlines the basics of atmospheric sciences to show that installing "Smog Towers" is an unrealistic solution [32]. Such solutions may give the public a false sense of security and harm their health besides not reducing air pollution. Experts suggest that the best strategy is to cut emissions from the sources such as household [19] and industrial emissions [33].

[9]https://www.unenvironment.org/resources/policy-and-strategy/air-quality-policies-germany Last accessed: 8 October 2020

[10]https://ec.europa.eu/environment/air/pdf/reduction_napcp/NL%20final%20NAPCP%201Apr19%20EN.pdf Last accessed: 8 October 2020

[11]https://www.thehindubusinessline.com/economy/policy/smoggy-delhi-to-get-bsvi-fuel-from-april-oil-ministry/article9961764.ece Last accessed: 8 October 2020

[12]https://www.thehindu.com/news/national/other-states/rajghat-badarpur-coal-plants-to-close/article7950494.ece Last accessed: 8 October 2020

[13]https://www.businessinsider.com/china-builds-worlds-biggest-air-purifier-2018-12?IR=T Last accessed: 8 October 2020

Another rigorously discussed solution is a vehicle rationing scheme called "Odd-Even". Vehicles having number plates ending with even numbers are allowed to operate on even dates, while those with odd numbers are allowed on odd dates. Two-wheelers, women, and private vehicles carrying school children are exempted in the scheme. A study by Chowdhury et al. [20] shows that the average reduction in $PM_{2.5}$ concentration was about 4-6% during Jan-2016 implementation, which is within the uncertainty range of satellite-based estimates. Thus, the jury is still out on the efficacy of the "Odd-Even" scheme. Therefore, it is important to know the public perception of such solutions. We try to analyse the same in the subsequent sections.

We chose these two mitigation strategies for our analysis because i) "Odd-Even" was implemented four times in the last four and half years in Delhi and has received significant attention on social media. ii) Although it is now proven [32], we knew that "Smog Tower" is a pseudo-scientific solution while working closely with air quality experts. Hence, we wanted to understand how people perceive their effectiveness. Other mitigation strategies including but not limited to restricting stubble burning [42], regulating industrial pollution [58], and increasing the number of electric vehicles [66] do not gather much social media attention or out of the actionable region of the local government.

4.1.2 Problem Statement. Our goal is to measure the public sentiment towards the untested mitigation strategies by classifying the pertinent tweets into three classes: i) supportive ii) unsupportive, and iii) neutral. By supportive, we mean supportive towards untested mitigation strategies and vice versa.

4.1.3 Approach. We require an accurate and diverse language model to represent the semantics of the tweet text. Bidirectional Encoder Representation from Transformers (BERT) [26] is a state-of-the-art language model. BERT contains bidirectional representations of words in all layers on the unlabelled text as shown in Figure 1. However, BERT is trained on Wikipedia data and retraining it directly on social-media data is a challenging task due to resource requirements. The Wikipedia documents are different from Twitter tweets in the following ways,

- *Subjective nature of Twitter data* differs completely from the objective nature of Wikipedia documents. Certain tweets can show different sentiments by changing the context.
- *Use of slang and informal language* is common on Twitter. However, Wikipedia uses a more formal and structured language.
- *Length of documents* in Wikipedia is larger than tweet-length limited to 280 characters. Thus, contextual continuation is lengthier in Wikipedia documents compared to Twitter tweets.
- *Hashtags and mentions* are specific features of Twitter and not visible in Wikipedia documents.
- *Tweets are more conversational* in nature, and a significant proportion of tweets are replies.

Adding to that, BERT models can not be used directly for sentiment predictions as they are trained on unlabeled datasets to learn the linguistic features automatically. We fine-tune the weights of its hidden layers for specific tasks (sentence classification task in our case). In this variant, we use our tweet dataset to fine-tune the BERT layers as described in Figure 2. During the process, all the attention layers try to learn a representation with a specific context. In the end, we have a fully connected softmax layer to output the prediction probability of each of the three classes (supportive, unsupportive, and neutral).

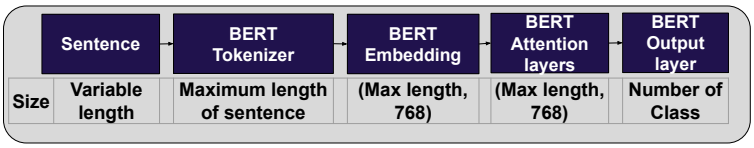


Fig. 1. Simplified architecture of language model BERT.

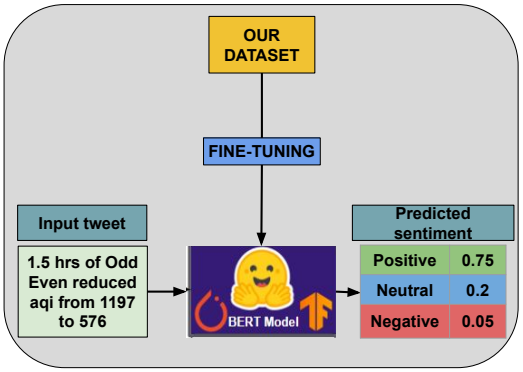


Fig. 2. Representation of BERT fine-tuned model

4.2 RQ2: Is Air pollution an year-long problem and does it 'Granger-cause' Twitter discussions?

We believe that events like visible smog or policy announcement bring about an episodic concern among people about air pollution. Episodic discussions are ignored as they are short-lived. An ongoing concern about air pollution in social media contributes to political and social enthusiasm for the enforcement of mitigation policies. We hypothesise that air pollution is a year-long problem, but it does not 'Granger-cause' twitter discussion [30].

There are two steps required in addressing this research question. First, we need to validate that air pollution is a year-long problem. We will use the WHO specified categorisation of $PM_{2.5}$ values by their severity (shown in Table 3). We hypothesise that air pollution would be a year-long problem as per the WHO norms. For validating the same, we will be "aggregating" the hourly $PM_{2.5}$ values into 24 hours. We then visualise the $PM_{2.5}$ values and compare it with the WHO standard and Indian standard of air quality.

Next, we need to validate that Air pollution ($PM_{2.5}$) does not 'Granger-causes' Twitter discussion as follows.

(1) We would aggregate the the total number of tweets for each day from 2016 to 2019. We also have $PM_{2.5}$ value for each day as explained in Section 3.
(2) We would check if both these time series are mean and variance stationary. If, not we apply the difference operation $y_t - y_{t-1}$ and make them stationary. We perform Kwiatkowski–Phillips –Schmidt–Shin (KPSS) statistical test [39] to confirm that data is stationary.
(3) Next, we compute the rolling 'Granger-causality' [64] value of $PM_{2.5}$ causes Twitter discussion over each month of all the years.
(4) We next compute the p value statistic and check if the magnitude is significant.
(5) For our hypothesis to be true, we would expect, p to be less than 0.05.

$PM_{2.5}$ Level	Basis for the selected level
35	These levels are associated with about a 15% higher long-term mortality risk relative to the AQG
25	In addition to other health benefits, these levels lower the risk of premature mortality by approximately 6%
15	In addition to other health benefits, these levels reduce the mortality risk by approximately 6%
10	These are the lowest levels at which total, cardiopulmonary and lung cancer mortality has been shown to increase with more than 95% confidence in response to long-term exposure to PM

Table 3. WHO air quality guidelines and interim targets for particulate matter: annual mean concentrations.

4.3 Topic Modelling

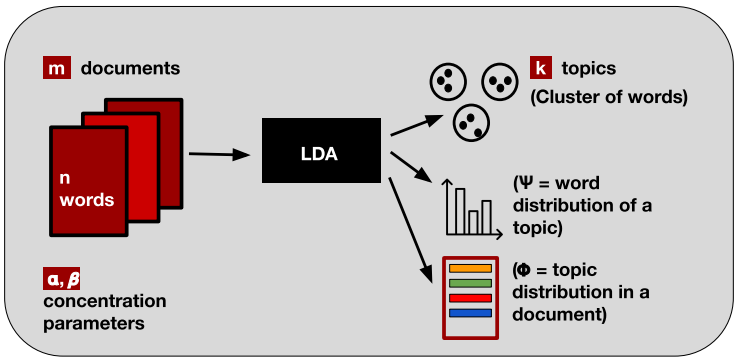


Fig. 3. A representative diagram of LDA for topic modelling.

Previously we have talked about sentiment towards mitigation strategies and 'Granger-causality' between $PM_{2.5}$ and twitter discussion. It is also important to reveal the topics of air pollution discussion on social media to understand the public's awareness and perception. An important aspect of air pollution is to understand i) what topics of air pollution does each tweet represent and ii) how do the most common topics evolve. It is practically impossible to look into each tweet and assign a topic label due to the dataset size. We can apply an unsupervised machine learning technique known as topic modelling [14] to determine the topics for each tweet. Suppose, the entire tweet corpus consists of k topics, then each of the tweets will belong to all the k topics with some probability score. The topics produced per tweet are a probabilistic mixture of words. A good topic model would provide topics that are human-interpretable and are distinct from each other.

Latent Dirichlet Allocation (LDA): Among several topic modelling techniques, LDA is a popular topic modelling technique created by Blei et al. [14]. We feed the m tweets (documents), where each tweet has n number of words and the number of topics (k) is feed into the LDA model. The

model returns k topics where each topic is a cluster of words. ψ is the probability distribution of words in a topic. ϕ is the probability distribution of topics per document. Concentration parameter α represent topic density per document, and β represent word density per topic. We feed our tweet corpus into the LDA and evaluate the output topics. Figure 3 demonstrates the process, and more details can be found in the appendix. The choice of initialisation parameters: number of topics (k) and number of iterations (i) can be made using topic coherence measure [5, 52], explained in the experimental settings in Section 5.3.

4.4 RQ3: Is the air quality discussion dominated by a small set of users who follow power-law characteristics?

Efforts to inform the scientific basis of air pollution can influence opinion and thereby influence policy-making. An endorser's Twitter post could positively impact attitude towards air pollution mitigation effort [47]. When it comes to air pollution in Delhi, we hypothesise that a small set of users dominate Twitter posts [38] and the number of tweets follow a Pareto distribution. To address this research question, we aggregate the total number of tweets by every user and sort them in decreasing order. We then try to validate if the probability distribution of the number of tweets by each user follow the form $p(x) \propto x^{-\alpha}$, where x is the user number.

5 EVALUATION

We now evaluate our hypotheses with various experiments and analyse them in the subsequent sections.

5.1 Addressing Research Question 1

"*RQ1: What is the public sentiment associated with untested mitigation strategies?*".

5.1.1 Dataset filtering and labelling. We test the sentiment towards two untested mitigation strategies "Smog Tower" and "Odd-Even". We discussed these mitigation strategies in Section 4.1.1. Thus, we need to filter the dataset for "Smog Tower" and "Odd-Even" tweets, and annotate the individual tweets. We use the same strategy used in Section 3.2 to ensure that there is least number of false positives or false negatives present in the filtered dataset. There is a subset of tweets (around 20%) in regional languages in our dataset. For this analysis, we have considered only the English tweets as the BERT model is pre-trained on English text only. We used Textblob[14] for detecting the language of each tweet in our dataset [48]. After removing non-English tweets, we have 516 and 80,343 tweets for "Smog Tower" and "Odd-Even", respectively. We use the BERT model to account for scalability in the future despite a lower number of tweets on "Smog Tower". For "Odd-Even" we have enough tweet to gauge the perception. For annotation, we consider all 516 tweets for "Smog Tower" and randomly sampled 1100 tweets over time for "Odd Even". Multiple tweets with the same text do not add any value to the sentiment detection task. Thus, we remove duplicate tweets.

Three authors annotated these tweets after referring to experts in air quality analysis. We annotated each tweet into one of the three classes: Supportive, Neutral, or Unsupportive, signifying their nature of support for these mitigation strategies. In these datasets, we have less than 2% tweets that are difficult to annotate because of subjectivity, satire or multiple sentiments present in a single tweet. The Cohen's kappa score[15] , which is a statistic that measures inter-annotator agreement, was 0.98 and 0.97 for "Smog Tower" and "Odd-Even", respectively. A Cohen's kappa score greater than 0.9 is sufficiently good value for confirming inter-annotator agreement[41]. We finally annotated

[14]https://textblob.readthedocs.io/en/dev/ Date accessed: 8 October 2020
[15]https://en.wikipedia.org/wiki/Cohen%27s_kappa Last accessed: 8 October 2020

Sentiment	Mitigation strategies	
	Odd-Even	Smog Tower
Supportive	238	285
Neutral	668	82
Unsupportive	187	63
Total	1093	430

Table 4. Class distribution for sentiment towards untested mitigation strategies in our labelled data set.

430 and 1093 unique tweets for "Smog Tower" and "Odd-Even", respectively. We found that for each tweet at least two annotators have given the same labels thus we get the final labels via majority voting among all annotators. We prepossessed the tweets to remove URLs. We do not remove hashtags as hashtags can present important contextual cues in tweets. As an example, *"Say NO to #FossilFuel. Adopt #OddEven , #Solarand #OrganicFarming all over the world"*. We observe that without hashtags the preceding sentence would be incomplete. Another example is a sarcastic tweet, *"What's to fear? We have #SmogTower now. #SmogTowerSoGood #Sarcasm #MyRightToBreathe"*. We can notice that '#Sarcasm' and '#SmogTowerSoGood' hashtags indicate the nature of tweet, which is sarcastic. The class distribution for each untested mitigation strategy in our labelled dataset is shown in Table 4.

5.1.2 ***Baseline Approaches.*** We compare the performance of our approach against three baseline approaches. We use preprocessed dataset discussed in section 5.1.1 for each of the following baselines:

- **Naive Bayes Classifier:** Naive Bayes Classifier is a generative classifier commonly used as a baseline in sentiment analysis [54, 55, 65]. The first step is to tokenise the preprocessed tweets. Once we have the tokens, we create a bag of words[16]. Each token in this representation has a certain probability of belonging to one of the three classes. We ignore the order of occurrence of tokens and focus only on the number of occurrences of the tokens in this baseline. We maximise P($t|c$) for token t with respect to class c using Bayes' theorem, $P(t|c) = \frac{P(c|t)P(t)}{P(c)}$.
- **BiLSTMs (Bidirectional Long Short Term Memory):** Sequential neural network models have made great strides in the semantic composition methods for sentiment analysis [40, 67]. In recent years, LSTMs have overcome the problem of vanishing gradients in RNNs. BiLSTMs connects two hidden LSTM layers to the same output, such that one of the LSTM layers is on the input sequence and other on the reverse of the input sequence. BiLSTM helps in understanding the context concerning both preceding and succeeding tokens.
- **Classifiers with BERT embeddings**: In this baseline approach, the pre-trained BERT model generates the sentence embeddings for the tweets. The feature vector serves as input for different classifiers shown in Table 9, which predict the tweet's sentiment[17].

5.1.3 ***Evaluation Metric.*** We use macro averaged F1 score as the evaluation metric based on its usage in prior work [10, 45]. The F1 score takes both precision and recalls into account, is computed by the following equation, $F1\ score = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$. The macro averaged F1 score tells about the overall performance of the baselines and BERT twice fine-tuned for a multi-class classification, is given by: $\frac{1}{n}\sum_{n=1}^{N} F_n$, where F_n is the F1 score for n^{th} class and N is total number of classes.

[16]https://en.wikipedia.org/wiki/Bag-of-words_model Last accessed: 8 October 2020

[17]http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/ Last accessed: 8 October 2020

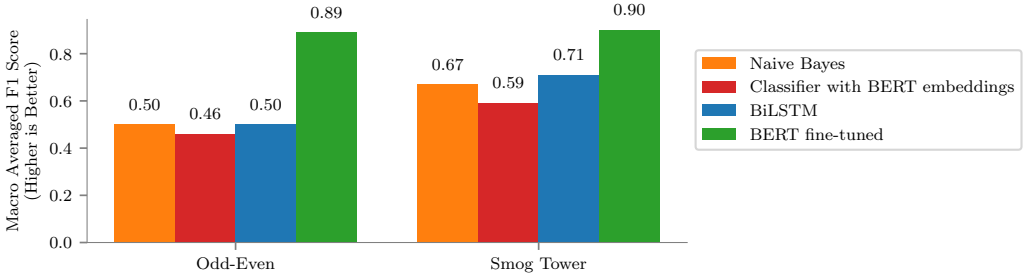


Fig. 4. BERT fine-tuned model performs with a better macro averaged F1 score than baseline approaches across both mitigation strategies under evaluation.

5.1.4 ***Experimental Setup.*** In this section, we describe the training and hyperparameter tuning of baseline approaches and our model. For all of our models, we split our annotated dataset into five stratified folds[18]. We used nested cross-validation[19] for tuning hyperparameters via grid search. We select the best performing set of hyperparameters with the highest macro averaged F1 score obtained. We have used the following set of hyperparameters for the baselines and our approach:

- Naive Bayes: We tune smoothing parameter α, which is known as pseudo-count and makes sure that probability is never zero. Another hyperparameter *fit_prior* is also tuned, which determines whether to learn class prior probabilities or not. We used the following hyperparameter space : $\alpha \in \{0.01, 0.05, 0.1, 0.5, 1\}$ and fit_prior $\in$ {True, False}
- Bidirectional LSTMs: We have used two Bidirectional LSTM layers followed by a dense layer as output. We tune the *embedding size, the number of neurons in the second hidden layer, epochs* and the *batch size*. We have the number of neurons in the first hidden layer to a constant 32. The following hyperparameter space was used: embedding size $\in \{32, 64, 128\}$, neurons $\in \{8, 16, 32, 64\}$, batch size $\in \{16, 32, 64\}$ and epochs $\in \{20, 30, 50, 100\}$
- Classifiers with BERT embeddings: We use pre-trained BERT to get embeddings for each tweet from the annotated dataset. These embeddings serve as input for classifiers. We tune the hyperparameter of the classifiers, logistic regression, SVC (Support Vector Classifier), and a neural network. The hyperparameters for these classifiers is given in Table 9 (Appendix).
- BERT fine-tuned: We only tune the *epochs* and *batch size* for the BERT keeping the dropout rate as 0.3 as suggested by previous literature [25] on the annotated dataset. The following hyperparameter space was used based on [25]: batch size $\in \{16, 32\}$ and 10 epochs with early-stopping enabled based on validation loss.

5.1.5 ***Results.*** Using BERT fine-tuned, we achieve 0.90 and 0.89 macro averaged F1 score on "Smog Tower" and "Odd-Even" datasets, respectively, which is better than other state of the art methods [10, 57]. Figure 4 shows the comparison among all approaches. The models with BERT outperform other approaches as BERT utilises the bidirectional training of transformer, a popular attention model, for language modelling in contrast to previous methods which look at text sequence from either left or right. While BERT is similar to LSTMs in bidirectional training but BERT has a sufficiently robust language model. Naive Bayes does not account for sequential nature. The classifiers in Table 9 with BERT embeddings does not achieve comparable performance compared to our approach which goes through end-to-end learning.

[18]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html Last accessed: 8 October 2020

[19]https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html Last accessed: 8 October 2020

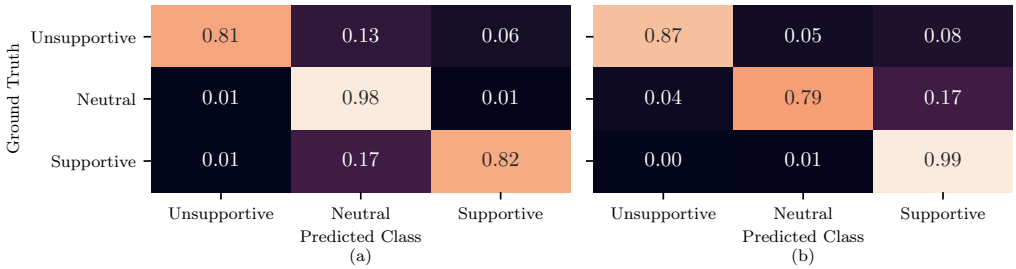


Fig. 5. Confusion matrices for untested mitigation strategies: (a) BERT fine-tuned for "Odd-Even"; (b) BERT fine-tuned for "Smog Tower"

ID	Tweet	Ground Truth	Sentiment Predicted
1	day 7: #hatersgonnahate but our #smogtower doing its job. fabricated news about rains and wind bringing relief. tomorrow, when weather reverses, even then smog tower will stand. #smogtowersogood that we are breathing. in your face , haters #myrighttobreathe #Saracsm	-1	1
2	I came to know about smog tower in delhi. appreciate work of gautam gambhir with some private investors in making this possible. this 20ft tall filter is not the ultimate solution but yes it will work for 10 to 15 % in a locality. nice step. #airpollution	-1	1
3	Wow clean Air in Delhi... Reason.. #OddEven no.. Thank you Nature for cleaning Delhi Air so that we can breath now.. @ArvindKejriwal sir please stop this odd even business and do something to stop stubble burning in adjoining states.. People will appreciate your efforts.. TYpic.twitter.com/ILtw1QQYio	-1	1
4	@VijayGoelBJP I can make out from a distant place that #OddEven is a small contribution to deal with a large problem. Now, Sir, you and your colleagues are making it a Centre versus State issue caring little about public health of millions of people https://twitter.com/ANI/status/1191248747302252545	1	-1

Table 5. Tweets from the "Odd-Even" and "Smog Tower" dataset. These tweets can be tough to classify correctly because of their ambiguity, subjectivity, and presence of multiple sentiment in same tweet.

While our results outperform the SOTA baseline, there are several reasons behind the imperfect sentiment classification. First, there is instance of irony in the tweet to show disapproval for a mitigation strategy like the first tweet in Table 5. There are also instances of multiple sentiments present in a single tweet like the second and third tweet in Table 5. The second tweet appreciates the concern of policymaker towards air pollution levels but is not in support of the mitigation strategy. Similarly, the third tweet has a supportive sentiment towards natural phenomenon mitigating the pollution levels but opposes the mitigation scheme started by the Government. Such challenges are not unique to our data set and known in the NLP literature [35].

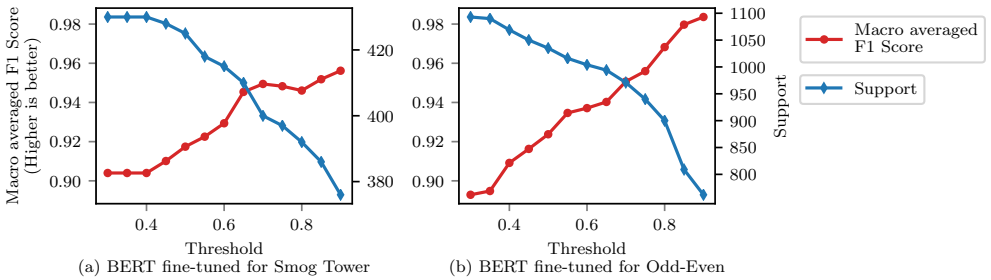


Fig. 6. Variation of Macro averaged F1 score and support with *probability of predicted class (X-axis)*. At 0.6 confidence, BERT fine-tuned model's macro averaged F1 score increases from 0.9 to 0.93 and from 0.89 to 0.94 with dropping support of 3% and 8% for "Smog Tower" and "Odd-Even", respectively.

5.1.6 ***Analysis.*** We now use our approach for sentiment classification on the larger unlabelled dataset and analyse the sentiment. However, before classification, we analyse the sensitivity of the model's macro averaged F1 score to the confidence in the predictions. We have a softmax output layer, which gives a probability score for each class (supportive, unsupportive, neutral). We take the class having the highest probability into account for a particular tweet. We now apply a threshold on the probability of a predicted class, that is we consider only those samples for evaluation for which the probability of predicted class is above a certain value. We do so to predict only when our model is confident. This may lead to a reduction in predicted samples, but, the predictions will likely be more accurate. We look at the trade-off between increasing threshold and corresponding support in Figure 6. We find that at 0.6 confidence, BERT fine-tuned model's macro averaged F1 score increases from 0.9 to 0.93 with 3% drop in support for "Smog Tower". For "Odd-Even", at 0.6 confidence, macro averaged F1 score increases from 0.89 to 0.94 with 8% drop in support. We do further analysis on unlabaled data with 0.6 confidence as there is significant increase in macro averaged f1 score.

We use BERT fine-tuned model to obtain sentiment around "Odd-Even". We use a class probability threshold of 0.6 and are able to predict 78, 597 out of 80, 343 samples. We then apply a 60 day moving average to smoothen the predictions over time, as shown in Figure 7. Between 2016 and 2019, "Odd-Even" scheme was implemented multiple times. We indicate these instances by vertical lines with 'A' tag in Figure 7. We found that for most of the timeline under consideration, the percentage distribution of supportive tweets is comparable to unsupportive. Before every implementation of "Odd-Even", the percentage of supportive tweets rises.

There are some driving events towards increment in unsupportive sentiment, as shown in Figure 7. Nearly six months later the second implementation of "Odd-Even", the National Green Tribunal of India (NGT) reported that "Odd-Even" did not help in reducing the air pollution[20]. The sentiment of people became unsupportive for some time after this news. In February 2018, the Delhi Chief Minister (CM) said that "Odd-Even" scheme is not a permanent solution to Delhi pollution[21]. After the CM's statement, there is a huge rise in unsupportive sentiment. Six months later, in November 2018, the Supreme Court of India allowed two-wheelers to be exempted from the future implementation of "Odd-Even" to reduce the burden on the public transport system. Twitter discussion reflecting the negative sentiment became more prominent (Figure 7) after the Supreme Court's

[20]https://indianexpress.com/article/cities/delhi/ngt-aap-odd-even-scheme-delhi-3089838/ Last accessed: 8 October 2020
[21]https://twitter.com/timesofindia/status/962369979424432129 Last accessed: 8 October 2020

Tweet	Associated sentiment
It's about time you guys move ahead of #OddEven and #publichealth emergency.. Do something concrete. Arrest culprits, control industrial pollution, stop crop burning Even trying to get an artificial rain will b more impact full What ur doing is just public gimmick	Unsupportive
I care for my family's health. I care for Delhi's health I care for every child's health Ipledge to reduce pollution I promise to follow #OddEven Do you ??? Those who care forDelhi's Health please share your msgs. Friends, #LetsUniteAgainstPollution	Supportive
Delhi CM Says Odd-Even Scheme Can Be Extended #Delhi #OddEven #arvindkejriwal #DelhiCM#oddevenscheme #WednesdayWisdom #DelhiAirPollution #viralbake www.viralbake.com/delhi-cm-says-odd-even-scheme-can-be-extended	Neutral

Table 6. Table showing tweets with different sentiments for "Odd-Even" Scheme.

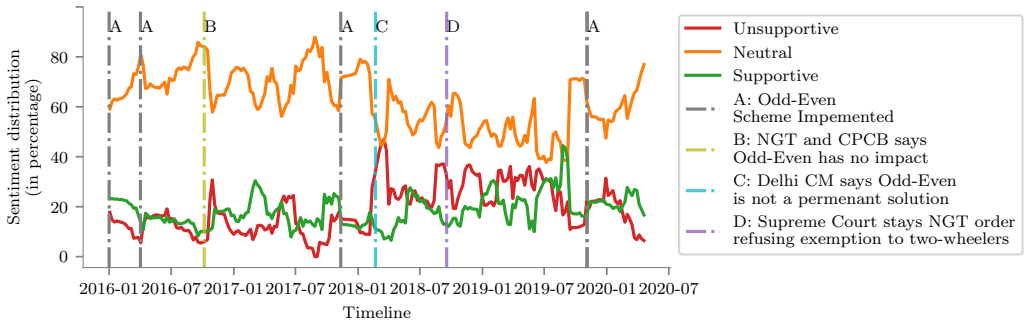


Fig. 7. Evolution of sentiment around "Odd-Even" scheme over time. The vertical lines with 'A' tag, signify the instances when scheme was implemented, January of 2016 and November of 2017 and 2019. 'B' and 'C' are two driving events for change in public sentiments.

verdict.

We now analyse the sentiment of the people for "Smog Tower". For this analysis, the timeline is from September 2016 to April 2020. We have annotated the entire set of tweets, and we use ground truth for analysis. In the future, when there will be more discussion around "Smog Tower", then our model could be used for predicting sentiment and further analysis.

Due to the lesser number of tweets on "Smog Tower", we show the sentiment over time in Figure 8 for the cumulative tweets. We observe from the plot that at any point in time, the supportive sentiment associated with the "Smog Tower" is more than unsupportive. There are five key events related to the "Smog Tower", which are shown by vertical lines in the plot. People in Delhi exchanged Twitter interactions about "Smog Tower" for the first time when a print media tweeted about its installation in Netherlands[22]. There was a significant increase in the number of supportive

[22]http://atlasofthefuture.org/project/smog-free-tower/ Last accessed: 8 October 2020

Tweet	Associated Sentiment
Smog towers are white elephants that will not fix Delhi's air pollution crisis. #Delhiairpollution #cleanair #Mission808080	Unsupportive
Intent combined with action will always yield results! Sharing initial readings of the prototype air purifier installed last week!	Supportive
On lines of China, Delhi may get anti-smog tower by next winter. China has recently installed an anti-smog tower that is 100m in height and cleans up to about 75M m^3 air per day.	Neutral

Table 7. Table showing tweets with different sentiments for "Smog Tower". The supportive tweet was by local politician had nearly 25,500 likes and 4,000 retweets while the unsupportive tweet was by the organisation which conducted study on "Smog Tower" had 61 likes and 48 retweets. The neutral tweets is by a leading media house in India with 108 likes and 24 retweets.

and neutral sentiment after a smog tower was installed in Beijing, China. People became highly supportive after India got the first working prototype. Supportive sentiments skyrocketed after a local politician's tweet went viral, which was in favour of smog tower. The sentiment towards "Smog Tower" continues to remain positive. Positive sentiment could be attributed to the higher social outreach of local politicians being vocal in favour of "Smog Tower". Famous politicians usually have a much higher follower count as compared to twitter handles of think-tanks who tweet on debunking the myth about air pollution. Examples of few tweets regarding the "Smog Tower" are given in Table 7. While there are few exceptions, the conclusion holds that the overall public sentiment towards untested mitigation strategies is supportive, giving a false sense of security.

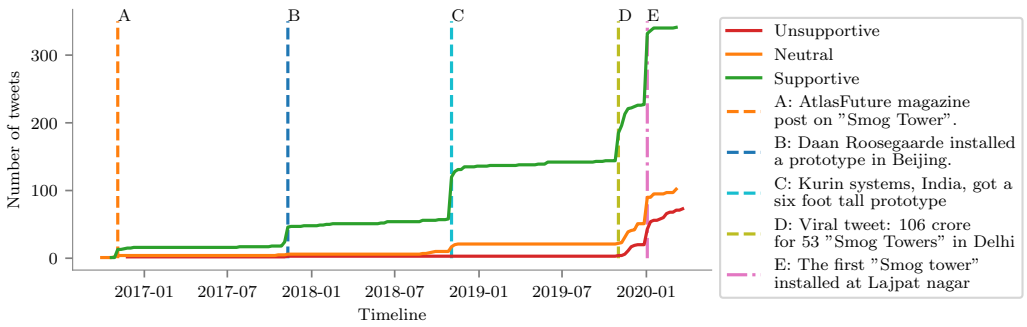


Fig. 8. Cumulative sum of supportive, neutral and unsupportive tweets for "Smog Tower" over time. The rate of increase in number of supportive and neutral tweets is significantly greater than unsupportive tweets.

5.2 Addressing Research Question 2

"*RQ2: Is Air pollution an year-long problem and does it 'Granger-cause' Twitter discussions?*".

5.2.1 Experiment 1. To answer if air pollution is a year-long problem, we plot the daily mean of $PM_{2.5}$ for the territory of Delhi, India, as shown in Figure 9. The data of $PM_{2.5}$ was from January

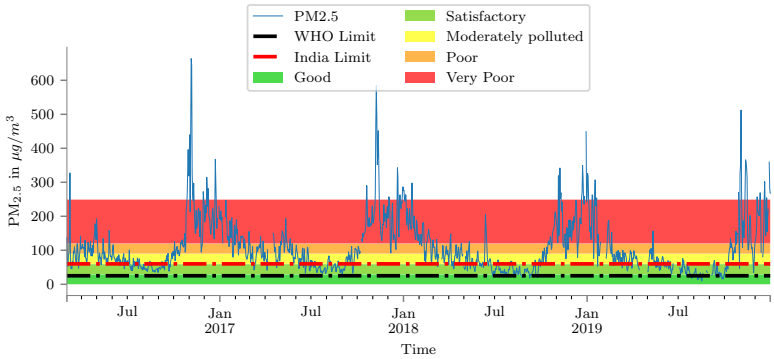


Fig. 9. Yearly $PM_{2.5}$ Level In Delhi. Colours corresponds to Indian air quality standards. The WHO and India limit on $PM_{2.5}$ is also shown.

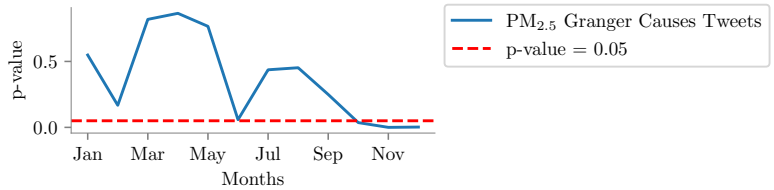


Fig. 10. $PM_{2.5}$ 'Granger-cause' twitter discussion only for the month of October, November and December.

2016 to December 2019. It was observed that 92% of the time, $PM_{2.5}$ levels breached the World Health Organisation (WHO) threshold on $PM_{2.5}$, which is 25 $\mu g/m^3$. It violated the Indian standard (60 $\mu g/m^3$) 66.5% of the time. It is thus, imperative that air pollution is a year-long problem in Delhi. The colour-coding in Figure 9 refers to different intervals of Indian air quality standard from good to very poor.

5.2.2 Experiment 2. Now, we evaluate if $PM_{2.5}$ 'Granger-causes' Twitter discussions. For this test, we follow the steps outlined in previous literature [64] and as mentioned in Section 4.2. One critical step is to make the time-series stationery by taking the difference operation as $y_t - y_{t-1}$. If the p-value of the 'Granger-causality' test is less then the significance level (0.05), then the corresponding time series ($PM_{2.5}$) causes the other time-series (Number of tweets).

Result: We found out that, for most months of the year (January to September), $PM_{2.5}$ does not 'Granger-cause' Twitter discussion on air pollution (Figure 10). During October, November and December 'Granger-causality' holds and it is at this time that air pollution becomes visible as smog. Consequently, the public reacts to it via social media. Public's reaction represented as the density of discussion in Figure 13 also shows a similar result. Almost all the discussion about air pollution happens during the end of a year (winter months) when pollution becomes visible as smog. While evaluating the evolution of '*Delhi Smog*' topic over time, we observe that its density is high only during the winter months. The correlation between the total number of tweets (on October, November and December) and the number of tweets under the topic '*Delhi Smog*' is 0.91. This seasonal increase in $PM_{2.5}$ value (Figure 11) becomes extremely hazardous, and public health

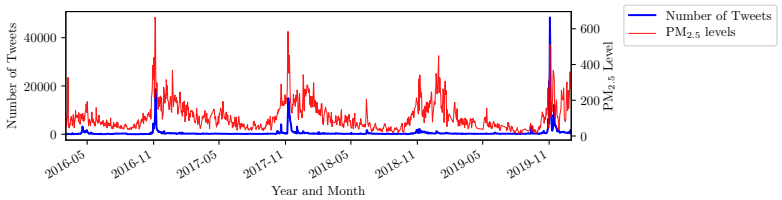


Fig. 11. The frequency of tweet and $PM_{2.5}$ value for the year 2016 to 2019. There were specific episodes when both $PM_{2.5}$ and frequency of tweet rise rapidly.

goes for a toss when nothing can be done to control emissions. While air pollution is a year-long problem (Figure 9) in Delhi but discussions are sparse and episodic. The next Section will reveal one of the primary reason for high pollution during winter months.

World Environment Day: We observe in Figure 10 that 'Granger-causality' approaches the significance level between May and July. On analysis, we found that the volume of tweets on air pollution increases abruptly on June 5th 2016, i.e. World Environment Day. Besides this event, we observe the topic '*Plant Tree*' in Figure 12, which evolved in the first week of June 2016.

5.3 Evaluating The Topic Model

We use LDA to observe the topics about air quality discussion. We feed the entire tweet corpus (1.2M) to the LDA model. We explained about LDA in Section 4.3. We pre-processed our tweet corpus by removing punctuation, hyperlinks and other non-ASCII characters. A frequently occurring word might appear in all the topic, whereas words that appear in a few tweets might add up as noise to an otherwise good topic. Thus, as a common practice we removed words that appeared more than 80% of the time, and less than 15% of the time (Threshold Noise Filtering [16]). Post text cleaning, we converted the pre-processed tweet corpus into a Bag-of-words (BoW) representation where each tweet is the one-hot encoding of the whole corpus.

We feed the Bag of Word (BoW) representation in the LDA model. The most critical parameter of LDA is the number of topics (k). A small value of k could potentially merge distinct topics into one. A large value of k could result in several topics that are individually confusing. To find the optimal value of k, we calculated the topic coherence [52] using different topic models by varying k [5]. The optimal value of k is where the topic coherence score is the highest.

The result of finding the optimal topic model is in Figure 15 (in Appendix). We found that the optimal value of k is 25 with 100 iterations of the algorithm. The value of the per-document topic distribution, α and per topic word distribution, β was as per the default settings in the gensim [50] library.

Visualising Topics in LDA: Topics inferred by LDA are not always easily interpretable by humans [18]. The visualisation makes it easy to interpret the topics. Therefore, we used a web-based interactive visualisation tool called *LDAvis* [59] that visualises the topic estimated by LDA. *LDAvis* provides a view of how topics differ from each other while at the same time allowing for deep inspection of the terms most highly associated with each topic. On the left of Figure 12, the topics are plotted as circles in the two-dimensional plane whose centres are determined by computing the distance between topics, and then by using multidimensional scaling to project the inter-topic distances onto two dimensions. The right panel of the visualisation depicts a horizontal bar-chart

whose bars represent the individual terms that are the most useful for interpreting the selected topic on the left [21, 59].

Interpreting the LDA Visualisation in Figure 12: The circles represent a topic. The radius of the circle depicts the number of terms in a topic. For example, the topic "Rain" has lesser number of terms then the topic "Diwali". Term frequency in a topic is lesser or equal to the term frequency in the corpus. It also means that the number of tweets on "Rain" is lesser than the number of tweets on "Diwali" (Diwali is a festival in India where firecrackers are burnt). Closer the circles to each other, greater the similarity of the topics those respective circles represent. For example, the topic "Government" and "Citizen Air" are more similar (have intersecting terms) than any other topic. For each topic, the histogram on the right side lists the top 10 (by term frequency in a topic) most relevant terms. For example, the terms like "burn","punjab" (a state in India),"haryana" (another state in India), "stubble" etc , appears entirely in the topic of "Stubble Burning".

Our result of the LDA model was evaluated by 'human in the loop' [62] as per the LDA algorithm (in Appendix A). We observed Delhi air pollution-related topics and associated terms while performing LDA visualisation on our tweet corpus (Figure 12). We explain some of these topics and their related terms. **"Odd-Even"** is an air pollution mitigation strategy implemented in Delhi. We have explained about "Odd-Even" in detail in Section 4.1.1. One of the associated terms with the topic "Odd-Even", is "arvind, kejriwal" - the first and the last name of the Chief Minister of Delhi. These terms reflect that the tweets on 'Odd-Even' had the mention of the Chief Minister.

A topic closer to "Odd-Even" is "public transport" with associated terms like "metro, bus, public, increase". Discussion around transport medium like metro train and bus increased among the public due to increased dependency of people on public transport. In Figure 13 we observe that the density of discussion around "Odd-Even" was highest in the year 2016 and lowest in the year 2018. The Delhi Government implemented the "Odd-Even" policy for the first time in the year 2016. In 2018, "Odd-Even" was not implemented.

Another topic in Figure 12 is "Stubble Burning" which is the primary cause of air pollution in Delhi due to "crop" burning by "farmers" in "punjab" and "haryana" - two neighbouring states of Delhi. Terms like "smog, delhichokes, smoke, cigarette" are associated with smog in Delhi ("Delhi Smog"). As evident from Figure 13, the discussion around "smog/fog" happens only during the winter months when pollution becomes visible. Most of the topics identified have a corresponding word in the word cloud in Figure 14, which shows that discussions around such topics are galore during air pollution in Delhi. Besides, the presence of media news articles on the same topics validates our selection of topics [12].

Table 8 shows some other topics and ten associated terms. Some tweets from the topic "Mumbai", were not directly related to Delhi air pollution. Example of one such tweet is , "*Hey Delhi! Look at the Mumbai sky. Pretty hazy too. It's not all pollution but it's not clean sea air either as it should be.No escape!*"[Retrieved on 01 September 2020]. All tweets in the topic of "Hindi Words" were in hindi language.

5.3.1 Analysis. We premised from topic visualisation that topics like "Odd-Even" and "Stubble Burning" are event specific. We now wanted to understand other topics which are event agnostic. Figure 13 shows different topics along y-axis and time along x-axis. The y-axis shows the likelihood that the topic of discussion occurs on a particular day. The dashed vertical line are event markers.

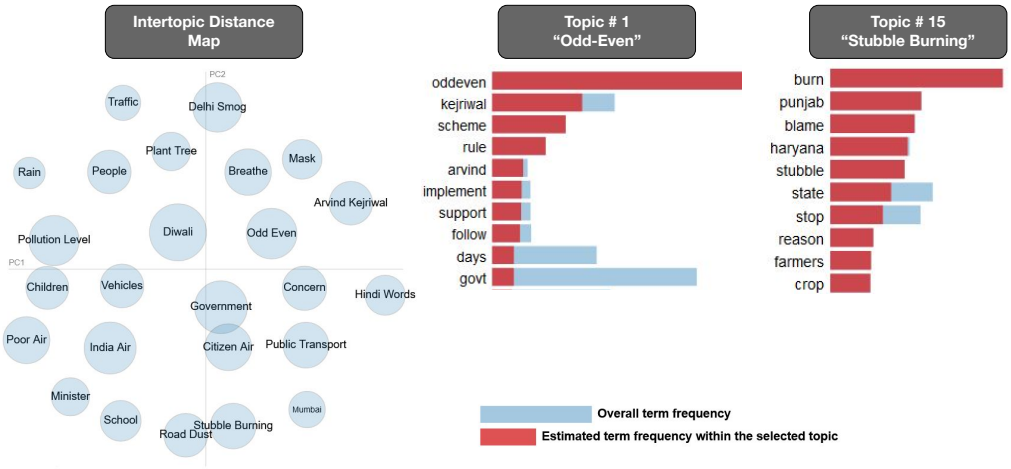


Fig. 12. Abridged form of the LDA visualisation. Two important topics, "Odd-Even" and "Stubble Burning" with associated terms are shown.

Topics	Traffic	Public Transport	Delhi Smog	People	Stubble Burning	Diwali	Hindi Words
Terms	swachhbharat	transport	delhismog	narendramodi	burn	burst	nahi
	police	metro	delhichokes	pmoindia	punjab	celebrate	mein
	cleandelhi	long	delhiairquality	problem	blame	hell	kuch
	dtptraffic	travel	delhibachao	solve	haryana	fireworks	liye
	mycleanindia	term	smogindelhi	mlkhattar	stubble	festival	raha
	jam	parali	lahore	captamarinder	state	marathon	hain
	pics	cycle	cigarettes	amitshah	stop	f***	rahe
	stick	short	delhiweather	cmohry	reason	kill	bhai
	hike	charge	smell	mohfwindia	farmers	judge	dilli
	encroachment	bus	sky	zeenews	crop	shameful	hoga

Table 8. Some interesting topics and associated terms as returned by LDA.**"Diwali"** is an Indian festival which involves bursting of crackers. We observed that all **Hindi Words** were segregated on a single topic.

We observed that discussion on the health effects of air pollution takes place throughout the year. Other topics of discussion are event specific like people talk about smog when it is visible. We elaborate on these topics and events by categorising them into i) Event Specific Topics, and ii) Event Agnostic Topics.

Event Specific Topics: These are topics on which Twitter discussion spurts on specific events. For example, i) **Odd Even**: "Odd-Even" was a Government policy to curb air pollution emitting from vehicles. On November 2017 and 2019, the "Odd-Even" again came to force. We observe that air pollution discussions on the subject of "Odd-Even" raised only during the announcement of the scheme, and on the implementation phase and not otherwise. ii) **Fog/Smog**: Air pollution discussion on the subject of fog/smog raised three times from 2016 to 2019. The first discussions were on November 30, 2016, when the smog was visibly apparent. Some users on Twitter confused

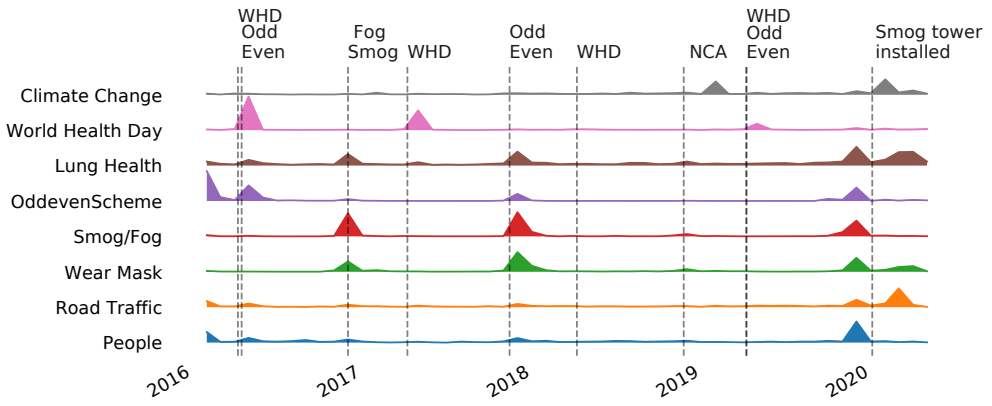


Fig. 13. Topics (in Y-axis) evolution over time. Peaks represents discussion density. Vertical lines represent events, i)"Odd-Even", ii) "World Health Day (**WHD**)", iii)"Fog/Smog" iv) "National Climate Assessment (**NCA**)", and v) "Smog Tower Installation".

smog with fog causing debate on the topic. The topic again came to light when visible smog on November 2017 prompted the Government to reimplement the "Odd-Even" policy. The third spurt in the discussion about smog/fog was a few days before the "Smog Tower" installation in Delhi. iii) **Climate Change**: Air pollution discussion on climate change occurred after the National Climate Assessment (NCA, US), 2018. The discussions reoccurred after the "Smog Tower" installation in Delhi. iv) **World Health Day:** Every year on April 5th, we observe a rise on air pollution discussion on World Health day. On June 5th,2016, we found a similar trend on World Environment Day.

Event Agnostic Topics: The density of tweets per day on two particular topics, namely, i) **Lung Health** and ii) **People**, was higher as compared to other event-specific topics. The topic of "Lung Health" comprises twitter discussion on Chronic obstructive pulmonary disease (COPD), cough, respiratory illness and even cancer. The topic of "People" includes names of cabinet ministers both at national and state level of politics. Twitter users often mention heads of states to voice concern about air pollution.

"Smog Tower" Installation In Delhi: The Indian capital territory of Delhi got its first "Smog Tower" inaugurated by an Indian Parliamentarian on January 3rd, 2020. We observe from Figure 13 that during the same period, discussion on topics like "Climate Change", "Lung Health", "Smog/Fog", "Road Traffic" and "People" rose significantly. Discussion on air quality happens more during events related to air pollution. Events like "Odd-Even" and "Smog Tower Installation" are perfect scenarios to gauge the public sentiment about mitigation strategies.

What do people talk about "Odd-Even" policy?: To understand the topic of discussion on "Odd-Even", we performed topic modelling using LDA on all the tweets that belong to "Odd-Even". We found the following relevant topics: i) "Women Exemption.": Tweets on this topic concentrated on the debate of exempting women drivers from the "Odd-Even" policy; ii) "Uber Price Surge.": During the odd even period of November 2019, people vented out their frustration of price rise in cab aggregators service. Uber is one of the leading cab aggregators in Delhi; iii) "Court.": The supreme court of India has pulled up Delhi government on the hardship faced by people due to the vehicle rationing policy and has questioned the effectiveness of the "Odd-Even" policy. It can be

concluded that beside the debates on Twitter about success or failure of "Odd-Even" policy, other concerns of the public were also highlighted. Our technique could show these concerns accurately which could be of interest to stakeholders.

5.4 Addressing Research Question 3

"*RQ3: Is the air quality discussion dominated by a small set of users who follow power-law characteristics?*".

5.4.1 Experiment and Analysis. Our objective is to confirm if small set of users dominate the air pollution discussion on Twitter and if we can identify these users. Figure 16 (in Appendix) confirms that the degree distribution of our data follow power-law which is indicative of the existence of fewer users in the network with higher levels of interactions, and many other users with less interaction [53]. The definition of protagonist is subjective. Previous work [46] defines a protagonist by a user's retweet, and followers. But $PM_{2.5}$ does not 'Granger-cause' tweets on any month of the year (except). Thus, we are not certain if the timeframe chosen by previous researchers is adequate to define the protagonists.

To add more objectivity to our analysis, we selected air pollution protagonist as the top 5% of users who have the maximum number of tweets and has more than 1 million followers. 13,229 users are in the top 5% of users with most tweets, and 124 of them has more than 1 million followers. Out of these 124 users, 52.42% are electronic/print/social news media, 4.84% are journalist, 5.65% are either political parties or politicians, and the remaining 37.1% comprises of celebrities (actors and sportsperson), data aggregators, and Non Government Organisations (NGO). We conclude that influential users with a high number of followers can bias opinion on air pollution mitigation strategy but they choose not to do so. Our analysis shows that a lot of celebrities including actors spoke about Delhi Air pollution only on November 3rd and November 4th 2019 when India played Bangladesh in a cricket match which grabbed International attention. Researchers talk about air pollution throughout the year, but they are more confined to their research domain which is not surprising. They do not appear in the top 5% of users who tweet about air pollution in our dataset.

6 LIMITATIONS AND FUTURE WORK

Our work has the following limitations

(1) Ours is the first large scale study in terms of analysis on air pollution perception. Due to domain expertise of our investigators, we limit our study for Delhi, India. This study can be extended to any other part of the world. In the future, we plan to extend our analysis to multiple geographies.
(2) In our current analysis, we do not do a detailed dive into the individual users and their characteristics. In our preliminary analysis of users' 'Bio', we are able to extract the age for 20% of users, gender for 1% of the users and occupation for 35% of users. In the future, we plan to augment our Twitter based work with a carefully planned user study on a subset of the users. The objectives of such a study would be to: i) acquire and relate user characteristics with their perception; ii) verify the perceived perception via social media and the perception via user study; and most importantly study what shapes perception with regards to air pollution.
(3) All of our code is publicly available and reproducible. In the future, we plan to work with regional authorities and develop a real-time dashboard to monitor and analyse social media data pertinent to understanding air pollution.

(4) In this work, we labelled 1523 tweets in total pertaining to sentiment classification towards untested mitigation strategy. Such annotation is human labour intensive. We plan to study the sentiment towards multiple such strategies in the future. Thus, in the future, we plan to leverage techniques such as transfer learning and active learning to minimise annotation cost. The key idea would be to "transfer" the knowledge from strategies with labelled data to strategies without labelled data, and to "actively" query or annotate to maximise accuracy using minimum number of annotations.
(5) In the current work, we have analysed only the text data. However, 30% of tweets associated media (images and videos) such as Figure 17 (in the Appendix). Such images can help us understand the data behind the individual user's perception. We plan to work on such data extraction from images in the future.

7 CONCLUSION

Relatively expensive, time and cost-prohibitive survey techniques were traditionally used to assess the opinions and responses of the citizens towards the issues of broader societal interests. Our work leverges the ubiquity of social media and advances in time-series analysis and NLP, to study perception towards societal issues. From the domain perspective, our work shows some ground realities as the public perception is largely supportive of untested strategies. Further, the public discussions peak up only during winter months when the air pollution is hazardous. We believe that our approach has the potential for immediate impact at scale for different societal issues.

8 ACKNOWLEDGEMENT

We would like to thank Professor Anirban Dasgupta of the Indian Institute of Technology (IIT) Gandhinagar (India), for his constructive feedback during this work. We are thankful to ACM PhD Clinic[23] mentor, Professor Ponnurangam Kumaraguru of Indraprastha Institute of Information Technology (IIIT) Delhi (India), for his valuable feedback. Finally, we are also grateful to the anonymous reviewers for their helpful advice on revising the manuscript.

REFERENCES

[1] Sofiane Abbar, Tahar Zanouda, Laure Berti-Equille, and Javier Borge-Holthoefer. 2016. Using twitter to understand public interest in climate change: The case of qatar. In *Tenth International AAAI Conference on Web and Social Media*.
[2] Rishiraj Adhikary and Nipun Batra. 2020. Computational tools for understanding air pollution. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 199–203.
[3] Rishiraj Adhikary and Nipun Batra. 2020. Do we breathe the same air?. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 1–4.
[4] Jeff Alstott and Dietmar Plenz Bullmore. 2014. powerlaw: a Python package for analysis of heavy-tailed distributions. *PloS one* 9, 1 (2014).
[5] 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.
[6] Xiaoran An, Auroop R Ganguly, Yi Fang, Steven B Scyphers, Ann M Hunter, and Jennifer G Dy. 2014. Tracking climate change opinions from twitter data. In *Workshop on Data Science for Social Good*.
[7] 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.
[8] Joshua S Apte and Pallavi Pant. 2019. Toward cleaner air for a billion Indians. *Proceedings of the National Academy of Sciences* 116, 22 (2019), 10614–10616.

[23] https://india.acm.org/research/phd-clinic

[9] Dominic Odwa Atari, Isaac N Luginaah, and Karen Fung. 2009. The relationship between odour annoyance scores and modelled ambient air pollution in Sarnia,"Chemical Valley", Ontario. *International Journal of Environmental Research and Public Health* 6, 10 (2009), 2655–2675.

[10] Noureddine Azzouza, Karima Akli-Astouati, and Roliana Ibrahim. 2020. TwitterBERT: Framework for Twitter Sentiment Analysis Based on Pre-trained Language Model Representations. In *Emerging Trends in Intelligent Computing and Informatics*, Faisal Saeed, Fathey Mohammed, and Nadhmi Gazem (Eds.). Springer International Publishing, Cham, 428–437.

[11] Arnab Jana Rounaq Basu, Aparup Khatua, and Saptarshi Ghosh. 2017. Harnessing Twitter Data for Analyzing Public Reactions to Transportation Policies: Evidences from the Odd-Even Policy in Delhi, India. *Proceedings of the Eastern Asia Society For Transportation Studies (EASTS)* (2017).

[12] Eric PS Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410.

[13] Karen Bickerstaff and Gordon Walker. 2001. Public understandings of air pollution: the 'localisation'of environmental risk. *Global Environmental Change* 11, 2 (2001), 133–145.

[14] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[15] 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.

[16] 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.

[17] K Chandramouli, N Pannirselvam, D Vijaya Kumar, Sagar Reddy Avuthu, and V Anitha. 2019. A STUDY ON SMOG FILTERING TOWER. *Journal of Advanced Cement & Concrete Technology* 2, 1, 2 (2019).

[18] 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.

[19] Sourangsu Chowdhury, Sagnik Dey, Sarath Guttikunda, Ajay Pillarisetti, Kirk R Smith, and Larry Di Girolamo. 2019. Indian annual ambient air quality standard is achievable by completely mitigating emissions from household sources. *Proceedings of the National Academy of Sciences* 116, 22 (2019), 10711–10716.

[20] Sourangsu Chowdhury, Sagnik Dey, Sachchida Nand Tripathi, Gufran Beig, Amit Kumar Mishra, and Sumit Sharma. 2017. "Traffic intervention" policy fails to mitigate air pollution in megacity Delhi. *Environmental science & policy* 74 (2017), 8–13.

[21] 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.

[22] Anna-Sara Claeson, Edvard Lidén, Maria Nordin, and Steven Nordin. 2013. The role of perceived pollution and health risk perception in annoyance and health symptoms: a population-based study of odorous air pollution. *International archives of occupational and environmental health* 86, 3 (2013), 367–374.

[23] Biraj Dahal, Sathish AP Kumar, and Zhenlong Li. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining* 9, 1 (2019), 24.

[24] Michael A DeVito, Jeremy Birnholtz, and Jeffery T Hancock. 2017. Platforms, people, and perception: Using affordances to understand self-presentation on social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 740–754.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[27] Daxin Dong, Xiaowei Xu, Wen Xu, and Junye Xie. 2019. The Relationship Between the Actual Level of Air Pollution and Residents' Concern about Air Pollution: Evidence from Shanghai, China. *International Journal of Environmental Research and Public Health* 16, 23 (2019), 4784.

[28] Thaddaeus Egondi, Catherine Kyobutungi, Nawi Ng, Kanyiva Muindi, Samuel Oti, Steven Van de Vijver, Remare Ettarh, and Joacim Rocklöv. 2013. Community perceptions of air pollution and related health risks in Nairobi slums. *International journal of environmental research and public health* 10, 10 (2013), 4851–4868.

[29] Colin S. Gillespie. 2015. Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software* 64, 2 (2015), 1–16. http://www.jstatsoft.org/v64/i02/

[30] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* (1969), 424–438.

[31] Supraja Gurajala, Suresh Dhaniyala, and Jeanna N Matthews. 2019. Understanding public response to air quality using tweet analysis. *Social Media+ Society* 5, 3 (2019), 2056305119867656.

[32] Sarath Guttikunda and Puja Jawahar. 2020. Can We Vacuum Our Air Pollution Problem Using Smog Towers? *Atmosphere* 11, 9 (2020), 922.

[33] Sarath K Guttikunda, Pallavi Pant, KA Nishadh, and Puja Jawahar. 2019. Particulate Matter Source Contributions for Raipur-Durg-Bhilai Region of Chhattisgarh, India. *Aerosol and Air Quality Research* 19, 3 (2019), 528–540+.

[34] Lei Huang, Chao Rao, Tsering Jan van der Kuijp, Jun Bi, and Yang Liu. 2017. A comparison of individual exposure, perception, and acceptable levels of PM2. 5 with air pollution policy objectives in China. *Environmental research* 157 (2017), 78–86.

[35] Doaa Mohey El-Din Mohamed Hussein. 2016. A survey on sentiment analysis challenges. https://www.sciencedirect.com/science/article/pii/S1018363916300071

[36] Aarti Israni, Sheena Erete, and Che L Smith. 2017. Snitches, Trolls, and Social Norms: Unpacking Perceptions of Social Media Use for Crime Prevention. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1193–1209.

[37] Wei Jiang, Yandong Wang, Ming-Hsiang Tsou, and Xiaokang Fu. 2015. Using social media to detect outdoor air pollution and monitor air quality index (AQI): a geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese Twitter). *PloS one* 10, 10 (2015), e0141185.

[38] B Kahng, I Yang, H Jeong, and A-L Barabási. 2004. Emergence of power-law behaviors in online auctions. In *The Application of Econophysics*. Springer, 204–209.

[39] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, Yongcheol Shin, et al. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of econometrics* 54, 1-3 (1992), 159–178.

[40] F. Long, K. Zhou, and W. Ou. 2019. Sentiment Analysis of Text Based on Bidirectional LSTM With Multi-Head Attention. *IEEE Access* 7 (2019), 141960–141969.

[41] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

[42] Susheel K Mittal, Nirankar Singh, Ravinder Agarwal, Amit Awasthi, and Prabhat K Gupta. 2009. Ambient air quality during wheat and rice crop stubble burning episodes in Patiala. *Atmospheric Environment* 43, 2 (2009), 238–244.

[43] Dinesh Mohan, Geetam Tiwari, Rahul Goel, and Paranjyoti Lahkar. 2017. Evaluation of odd–even day traffic restriction experiments in Delhi, India. *Transportation Research Record* 2627, 1 (2017), 9–16.

[44] Minggang Peng, Hui Zhang, Richard D Evans, Xiaohui Zhong, and Kun Yang. 2019. Actual air pollution, environmental transparency, and the perception of air pollution in China. *The Journal of Environment & Development* 28, 1 (2019), 78–105.

[45] H. T. Phan, V. C. Tran, N. T. Nguyen, and D. Hwang. 2020. Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model. *IEEE Access* 8 (2020), 14630–14641.

[46] Prithviraj Pramanik, Tamal Mondal, Subrata Nandi, and Mousumi Saha. 2020. AirCalypse: Can Twitter Help in Urban Air Quality Measurement and Who are the Influential Users?. In *Companion Proceedings of the Web Conference 2020*. 540–545.

[47] Melissa Pujazon-Zazik and M Jane Park. 2010. To tweet, or not to tweet: gender differences and potential positive and negative health outcomes of adolescents' social internet use. *American journal of men's health* 4, 1 (2010), 77–85.

[48] Manish Rana and Mohammad Atique. 2019. Language Translation: Enhancing Bi-Lingual Machine Translation Approach Using Python. *i-Manager's Journal on Computer Science* 7, 2 (2019), 36.

[49] 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.

[50] 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.

[51] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature* 566, 7743 (2019), 195–204.

[52] 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.

[53] Arif Mohaimin Sadri, Samiul Hasan, Satish V Ukkusuri, and Juan Esteban Suarez Lopez. 2018. Analysis of social interaction network properties and growth on Twitter. *Social Network Analysis and Mining* 8, 1 (2018), 56.

[54] K. Sarkar and M. Bhowmick. 2017. Sentiment polarity detection in bengali tweets using multinomial Naïve Bayes and support vector machines. In *2017 IEEE Calcutta Conference (CALCON)*. 31–36.

[55] Kamal Sarkar and Saikat Chakraborty. 2015. A Sentiment Analysis System for Indian Language Tweets. In *Mining Intelligence and Knowledge Exploration*, Rajendra Prasath, Anil Kumar Vuppala, and T. Kathirvalavakumar (Eds.). Springer International Publishing, Cham, 694–702.

[56] Jan C Semenza, Daniel J Wilson, Jeremy Parra, Brian D Bontempo, Melissa Hart, David J Sailor, and Linda A George. 2008. Public perception and behavior change in relationship to hot weather and air pollution. *Environmental research* 107, 3 (2008), 401–411.

[57] Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter Sentiment Analysis with Deep Convolutional Neural Networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 959–962. https://doi.org/10.1145/2766462.2767830

[58] Kaushik K Shandilya, Mukesh Khare, and Akhilendra Bhushan Gupta. 2007. Suspended particulate matter distribution in rural-industrial Satna and in urban-industrial South Delhi. *Environmental monitoring and assessment* 128, 1-3 (2007), 431–445.

[59] 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.

[60] Jabrinder Singh, Naveen Singhal, Shailey Singhal, Madhu Sharma, Shilpi Agarwal, and Shefali Arora. 2018. Environmental implications of rice and wheat stubble burning in north-western states of India. In *Advances in health and environment safety*. Springer, 47–55.

[61] 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.

[62] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *23rd International Conference on Intelligent User Interfaces*. 293–304.

[63] Yuguo Tao, Feng Zhang, Chunyun Shi, and Yun Chen. 2019. Social Media Data-Based Sentiment Analysis of Tourists' Air Quality Perceptions. *Sustainability* 11, 18 (2019), 5070.

[64] Hiro Y Toda and Taku Yamamoto. 1995. Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics* 66, 1-2 (1995), 225–250.

[65] Alexandre Trilla and Francesc Alıas. 2012. Sentiment analysis of Twitter messages based on multinomial Naive Bayes. *Comput. Surv* 34 (2012), 1–47.

[66] Rachana Vidhi and Prasanna Shrivastava. 2018. A review of electric vehicle lifecycle emissions and policy recommendations to increase EV penetration in India. *Energies* 11, 3 (2018), 483.

[67] Yang Wang, Shi Feng, Daling Wang, Yifei Zhang, and Ge Yu. 2016. Context-Aware Chinese Microblog Sentiment Classification with Bidirectional LSTM. In *Web Technologies and Applications*, Feifei Li, Kyuseok Shim, Kai Zheng, and Guanfeng Liu (Eds.). Springer International Publishing, Cham, 594–606.

[68] Xiaojing Zhang, Pawel Wargocki, Zhiwei Lian, and Camilla Thyregod. 2017. Effects of exposure to carbon dioxide and bioeffluents on perceived air quality, self-assessed acute health symptoms, and cognitive performance. *Indoor air* 27, 1 (2017), 47–64.

A APPENDIX

In this supplementary document, we provide additional details for our paper.

A.1 Dataset visualisation

We visualise the word-cloud of our dataset in Figure 14 and observe phrases such as 'Stubble Burning', 'Odd-Even' and 'Public Health'. These phrases become important when we evaluate the result of topic modelling in Section 5.3.

Fig. 14. World cloud from the 1.2 Million tweets representing Delhi air pollution. We observer words like, **Stubble Burning** is a significant cause of pollution in Delhi. Farmers in the neighbouring states burn stubble to prepare for next sowing season. $PM_{2.5}$ and PM_{10} pollutants are released in a large amount during the process. Winds often bring these pollutants to Delhi [60]. **Public Health** is a major area of concern as there is increasing scientific and anecdotal evidence of the severe health impacts of air pollution. **Odd Even scheme** is the vehicle rationing mitigation strategy by Delhi Government implemented from time to time to reduce air pollution.

A.2 LDA and coherence score

(1) Initialise Hyperparameters. We initialise the following hyperparameters
- Number of topics, k
- Number of iteration of the algorithm, i
- Set Concentration parameters, α, β

(2) Initialise topics assignment randomly.
- Each word in each document is assigned a random topic

(3) **Iterate** until convergence
- For each word in each document:
  - Resample topic for word, given all other words and their current topic assignments.
  - Update frequency of words, ψ
  - Update distribution of topics, ϕ

(4) Evaluate Model
- **Human in the loop:** For each trained topic, we take the first ten words and substitute a term of one of them with another randomly chosen term (intruder). Now we check if one of our investigators can reliably tell which term (word) is an intruder in a topic. If so, the trained topic is good, if not, the topic has no discernable theme.

Figure 15 shows number of topics (k) and corresponding coherence score. The optimal number of topic is where the coherence score is the highest.

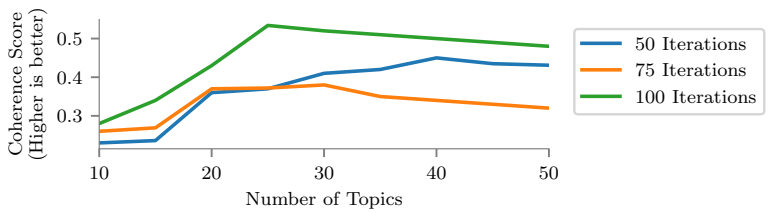


Fig. 15. **The optimal number of topics is 25**.The coherence score for 100 iterations with 25 topics is the highest.

Classifier List			
Logistic Regression	**SVM with RBF kernel**	**SVM with polynomial kernel**	**Neural Network**
C: {0.01,0.1,10,100} penalty: {ℓ_1,ℓ_2}	C: {0.01,0.1,10,100} kernel coefficient: {0.01,0.001,0.0001}	C: {0.01,0.1,10,100} kernel coefficient: {0.01,0.001,0.0001} polynomial degree: {3,5,7,9}	hidden layer neurons: {8,16,32} batch size: {16,32,64} epochs: {20,30,50}

Table 9. Different classifiers used and their corresponding hyperparameters with space. C is the is the inverse of regularisation strength.

A.3 Hyperparameter space

Table 9 shows the different classifiers using embeddings from BERT and BERT fine-tuned on sentiment140, along with their corresponding hyperparameter space.

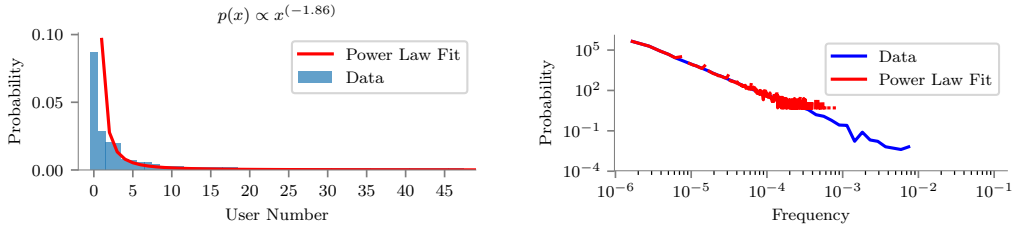
A.4 Power law fit

We used [29] to verify how likely is it that our data comes from either of power-law distribution or log-normal distribution. We use Vuong's test statistic with the null hypothesis that "H0: Both distributions are equally far from the true distribution." and "H1: One of the test distributions is closer to the true distribution". While fitting the data, we found the α value to be -1.86 and p-value of 0.08 thus rejecting "H0". Figure 16 shows the power law fit and the goodness of fit of our data (Distribution of tweets per user)

A.5 Tweet with an associated image

Figure 17 shows an instance where user tweeted an image related of air quality at a specific location.

Received June 2020; revised October 2020; accepted December 2020



(a) (b)

Fig. 16. **a)** The number of tweets by user follow a power-law distribution. **b)** The log-log plot for power-law.'The best fit power law may only cover a portion of the distribution's tail' [4].

Fig. 17. An example image associated with a tweet related to air pollution.