

A Transformer Based Model to Detect Environmental Related Distress on Online Social Media

Utkarsh Sharma¹, Prateek Pandey¹ and Shishir Kumar²

¹Department of Computer Science and Engineering, Jaypee University of Engineering and Technology, Guna, India

²Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India

Article history

Received: 10-11-2025

Revised: 17-05-2026

Accepted: 20-08-2026

Corresponding Author:

Utkarsh Sharma

Department of Computer

Science and Engineering,

Jaypee University of

Engineering and Technology,

Guna, India

Email: utkarsh_shar@yahoo.co.in

Abstract: Online social media has been recently employed as the source of data for the detection of traffic-related activities and to control the flow of traffic in smart cities all over the world. Till now an ample amount of research work has been done in identification of traffic-related incidents with the help of online social media platforms (mostly Twitter). Along with the problem of traffic the society face the problem of increase in pollution and its related health hazards. Online social media proved to be effective in correctly mapping the sentiments of the people suffering from the problem of pollution with the exact measure of the harmful substances in the air. In this work, we presented a novel approach for analysing the environment-related distress on online social media related to pollution using BERT based transformer. In this paper we analysed the tweets and classified them into two classes: i) Pollution specific, ii) Location Specific by using the machine learning algorithms. From the location-specific tweets about the pollution, we then identify the location with serious pollution-related issues. The model was trained and evaluated on a dataset of 21,549 geotagged tweets collected from Delhi over a one-year period (January 2019–January 2020). The proposed BERT-based classifier achieved an accuracy of 96.5% and an F1-score of 96%, outperforming baseline machine learning algorithms and existing BERT-based benchmarks. Unlike prior BERT-based sentiment models that focus on general positive/negative polarity, our approach is specifically fine-tuned to detect pollution-related environmental distress and integrates geographic information extraction to identify urban pollution hotspots. This dual classification into Pollution Specific (PS) and Location Specific (LS) categories enables direct, actionable outputs for urban planners and traffic management authorities. The findings carry direct implications for smart city governance, enabling authorities to identify pollution hotspots and design data-driven interventions including traffic rerouting and targeted public health alerts.

Keywords: Air Pollution, Tweet Classification, Text Mining, Online Social Media, Traffic Control

Introduction

Traffic and traffic-related problems are now a major concern of almost every megacity in this world. It causes a great loss of health and wealth to almost every nation in this world. As a report, due to traffic congestions, an American on average loses 97 hours or \$1348 in a year (INRIX, 2018). As of this increase in health and economic hazards caused by traffic problems, a new dimension of study has emerged as an Intelligent transportation system (Picone et al., 2015). Traffic analysis has become a critical

component in implementing smart city concepts, aiming to identify key problematic locations and improve urban air quality. The basic need for traffic scheduling is to provide a smooth transit of vehicles around the various pathways of a city.

In this paper we presented a model for analysing the sentiments of the people gathered from online social media platforms for the factor of pollution, as we all know that pollution is the most severe problem that all of the major cities are facing. For this study, ‘environmental distress’ is defined as expressions of negative affect

specifically tied to pollution and environmental degradation, distinct from generic positive/negative sentiment analysis. In India, most of the population in all the states were living with air having a concentration of PM_{2.5} pollutant at a level way above the limit recommended by national ambient air quality standards of India (Balakrishnan et al., 2019). According to Myllyvirta and Sivalingam (2016) the total mortality rate in India due to pollution is the highest in the world which is 1.2 million. This study addresses the pressing issue of air pollution caused by traffic congestion by proposing a model that leverages Environmental Distress Classification from online social media platforms to assess pollution levels. The objective is to identify key problematic locations based on public sentiment and pollution-related health concerns. This study is grounded in theories of environmental affect and public discourse, framing pollution-related social media posts as expressions of collective distress. Drawing on Stacy Alaimo's concept of trans-corporeality, which highlights the entanglement of human bodies and environmental conditions, we conceptualize 'environmental distress' as both a lived and mediated phenomenon. Similarly, the idea of citizen sensing (Gabrys, 2016) informs our view of Twitter users as active participants generating grassroots environmental data. These posts form what (Papacharissi, 2015) calls affective publics – networked publics mobilized through shared feelings and digital interactions. This framework positions our technical approach within a broader socio-political context, linking computational detection with environmental humanities perspectives. This research aims to answer the following question: How can Environmental Distress Classification of public concerns help identify pollution hotspots for targeted intervention?

This research question aligns with our objective of identifying high-priority urban areas affected by pollution based on real-time public sentiment. We operationalize this by first classifying tweets as pollution-related distress and then mapping those that mention specific locations. A high frequency of such tweets from a location signals a pollution hotspot, thereby directly linking online public concern to actionable city planning decisions.

Curb on pollution measure of a particular area is the need of the current era because pollution not only affects the people in transit, but it also affects the health of the person nearby the traffic, and it may include elderly people and children also. As we all know that air pollution affects mostly our children and old age people, so we have to ultimately take a step and create some plan to help them cope with this situation lest it is too late. According to Heydon and Chakraborty (2020), people are concerned about the negative impact on the health of their wards due to air pollution, their study involves a qualitative survey of 45 individuals across England regarding the influence of air pollution on their children. They used a monitoring

device to monitor the air pollution level encountered on the streets and the people indulged in the survey carried those devices. A similar study was conducted by Ashmore et al. (2000) about the effect of carbon monoxide on school children. Although in their study they considered the comparison of the students who are going to school by either car or by walk. Still, their study shows that using a car as a mode of commute increases the CO level as compared to the simple walk to the destination. Bergstra et al. (2018) found the causality of PM_{2.5} and NO_x pollutants on the respiratory system of the children. According to their results, it is evident that the working of the lungs of children in the near vicinity of large industrial areas is severely affected. Not only children but adults also get affected by these harmful pollutants, which increases mortality rates and health issues (Schultz et al., 2017; Yao, 2019; Vahedi, 2020). Our proposed model aims to mitigate these health risks by analysing real-time sentiment data from Twitter to identify areas where pollution or pollution-related health issues are a concern. We propose a novel approach to identify environment-related distress (specifically air pollution complaints) on Twitter using a BERT-based transformer model. Our approach extends beyond existing BERT sentiment classifiers by focusing on pollution-specific content and integrating geographic information to pinpoint affected locations. Unlike generic sentiment analysis models, we fine-tune BERT to distinguish pollution-related tweets and link them to specific areas, enabling targeted interventions (e.g. dynamic traffic re-routing in pollution hotspots). If a surge in pollution is detected in a specific area, the model provides insights for authorities to take targeted actions to curb pollution. This approach emphasizes identifying problematic locations based on community-reported concerns rather than directly managing traffic flow. Results show that the BERT-based classifier effectively captures public sentiment about pollution and successfully identifies critical locations suffering from poor air quality. This work demonstrates the novel use of transformer-based NLP in urban environmental sensing, bridging the gap between social media sentiment and actionable urban planning strategies. While the BERT transformer model itself is well established, our novel contribution lies in its application to detecting pollution-related distress and integrating these outputs with geolocation data to identify actionable urban hotspots. This contextual focus and interdisciplinary linkage between NLP and urban environmental management distinguish our work from prior sentiment analysis studies.

Rerouting based on an increase in pollution measure is suggested in the model. If the sentiments collected from the people show a surge in pollution or pollution-related health problems in a particular area, then based on the location identified the traffic from that street can be

diverted onto different paths if possible. It can generate a messaging-based alarm system to alert the riders that the air quality of the path on which they are forwarding is far more polluted so they will be shown some alternate better paths to follow. Collecting the sentiments from Twitter leads to a better analysis of the problems suffered by the people, suppose the AQI of a particular area is found to be a very critical level then the authorities will take appropriate steps to curb the pollution at that area. But there might be some areas of the city in which the air pollution is not at that critical level, but the residents of that area might be suffering from some problems, there might be some medical facilities at that location which might be at high risk due to this pollution-related hazards. The work suggested in this paper will identify such areas in which the people of that particular area facing serious issues related to air pollution.

Related Work

Traffic monitoring or traffic scheduling is not a new dimension of the study; it originates way back in the early '80s when traffic-related problems were started to arise very often. Initially, the solutions were mostly related to curbing the traffic from one part of the city to reduce the load on that part, which is termed as traffic rationing. A similar case study is presented by Ogunsanya (1984) of the city Lagos in Nigeria, where they applied a scheme to allow the vehicles with registration number ending with even digits to pass through the path on Monday, Tuesday, and Wednesday, And similarly for the other half of vehicles. But as the technology advanced and the vehicles improved in their efficiency along with the high emission factors the traffic scheduling was not possible by the same old methods of traffic rationing. Now we are living in the era of megacities and smart cities and for this scenario, an intelligent traffic system is an apt solution.

We surveyed traffic scheduling related work which has done in the past and presented the findings in the form of a comparison Table 1. The method for traffic modelling is categorized into three models which are Macroscopic, Mesoscopic, and microscopic. The macroscopic model considers the aggregate details of the vehicles in the traffic whereas the mesoscopic model considers every individual vehicle and their effect on the traffic. The microscopic model is the combination of both the macro and meso approach. Several simulation software is available for simulating the traffic environment and apply the proposed policies. The simulation software is selected as the desired level of study (Macro, Meso, or Micro). Along with traffic simulators Osorio and Nanduri (2015) have used emission simulators which simulate the emission from a particular vehicular organization.

Recent research on sentiment analysis of pollution-related tweets using BERT has explored various methodologies to effectively leverage the model's capabilities for this specific task, the comparative analysis

is presented in Table 2. While prior studies (Wan, 2024; Pota, 2021; Rodríguez-Ibáñez, 2023; Behera, 2021; Islam, 2021) apply BERT to general sentiment tasks or specific health-crisis domains, none specifically focus on distinguishing pollution-related environmental distress as a signal for urban traffic management. Our work is distinct in combining domain-specific fine-tuning with geographic hotspot detection.

Recent scholarship in environmental humanities and urban studies highlights the intersection of technology and public environmental engagement. Citizen sensing research (Gabrys, 2016) examines how digital tools empower citizens to monitor and report environmental changes.

Studies of environmental justice movements in India emphasize how pollution disproportionately affects marginalized groups, linking our work to broader socio-political struggles. Similarly, digital media theorists have analysed how affective publics emerge around environmental crises, shaping public discourse and policy debates. By integrating these perspectives, our study connects computational modelling with the lived realities and civic implications of pollution.

Recent advancements in Natural Language Processing (NLP) and transformer-based architectures have significantly improved the capability of analysing social media data for urban sensing and environmental monitoring applications. Earlier studies on tweet classification and sentiment analysis primarily relied on traditional machine learning approaches such as Support Vector Machines (SVM), Logistic Regression, Naïve Bayes, and Decision Trees combined with handcrafted features or TF-IDF representations. While these methods achieved moderate success in identifying traffic incidents and public opinions, they often failed to capture contextual semantics, implicit sentiment, sarcasm, and multilingual expressions present in social media content.

The emergence of transformer-based architectures, particularly Bidirectional Encoder Representations from Transformers (BERT), introduced substantial improvements in contextual understanding and semantic representation of textual data. Devlin et al. (2019) proposed BERT as a deep bidirectional transformer model capable of learning contextual relationships between words using attention mechanisms. Since then, BERT-based models have demonstrated superior performance across several NLP tasks including sentiment analysis, question answering, disaster response classification, and social media analytics.

Study Area

Delhi or the national capital territory of Delhi (NCT) is a union territory containing New Delhi the capital of India with an estimated population density of 11,297 person/km² according to census of India 2011 and also having the largest number of populations of youth 356 million (Government of India, 2011).

Table 1: Comparative study of existing traffic scheduling techniques

Author	Target area	Technique	Key findings	Experimental results
Diaz et al. (2020)	Madrid, Spain	Complex event processing + Petri nets-based simulation model	A model able to simulate the real behaviour of traffic and find the best path between source and destination along with keeping the pollution levels steady	Yes
Rajé et al. (2018)	Nairobi, Kenya	Holistic planning-based change in transport scheme	A study is presented which justifies the need for changes in policies both at the infrastructural and political levels. The import of used cars having a large impact on pollution levels in Kenya.	No
Hilpert et al. (2019)	New York, USA	Use of real-time Radar calibrated crowdsource traffic data with pollution meters	Real-time traffic data is collected via radars on three different sites for heavy and medium vehicles. A correlation and a regression model is developed to show the relation between Black carbon level and the traffic conditions	Yes
Xie et al. (2020)	Tianjin, China	Inter Spatio-temporal Apriori algorithm + K-Means++ with FCM Clustering Algorithm	Association between congestion in grids or cross-sections is identified along with the calculation of speed and flux as supportive parameters. Results have shown that congestion in one location can cause congestion in the connecting lanes.	Yes
Choudhary and Gokhale (2019)	Guwahati, India	Upgradation of emission standards from Euro-III to Euro-VI & change of fuel from gasoline to CNG for auto rickshaws and by regulating the frequency of buses	A significant reduction in the emission of harmful gases was observed by changing the emission standards and by upgrading the fuel options. Also, by monitoring and properly scheduling the frequency of bus ply can reduce traffic congestion	Yes
Osorio, and Nanduri (2015)	Lausanne, Switzerland	Combination of High-resolution microscopic Model with Low-resolution Macroscopic Model	The study suggests that the combined model of macroscopic data and microscopic data can reduce the emission levels of CO, NO, PM, and HC. Also, traffic conditions will be improved as per the simulations.	Yes
Li et al. (2019)	Chicago, USA	Multi-agent based hierarchical framework using communication between drivers and controlling stations	The Model uses the communication between vehicle to vehicle and vehicle to infrastructure as the backbone of the MAS framework, which proves to result in a shorter queuing length of the vehicles at the intersections.	Yes
Hamidi and Kamankesh (2018)	N/A	Swarm Optimization based multi agent model for traffic flow management	A Multi agent-based model based on Ant Colony optimization algorithm is built to improve traffic flow and to minimize the load on a particular lane	Yes
Percoco (2015)	Milan, Italy	Spatio temporal traffic modelling	The study proves that the spatiotemporal model is free from the correlation from the temporal aggregation if it is separable in space and time	Yes
Jain et al. (2020)	Delhi, India	Inference rule based automatic dynamic traffic light control system	Based on six different parameters, inference rules of IF-Then kind are developed to control the timing of traffic signals.	No
Farrag et al. 2020	Muscat, Oman	Traffic incident management technique using CAR2X communication	Simulation results show that car2x communication-based model results in less emission of pollutants and smoother traffic with fewer blockages.	Yes
Tian et al. (2019)	Wuhan, China	Quantum genetic algorithm-based route selection	The proposed algorithm uses a genetic algorithm in a quantum computing environment and by using both dynamic and static features of road traffic management it proves to reduce traffic congestion when the traffic increases above a threshold value.	Yes

Author	Target area	Technique	Key findings	Experimental results
Dabiri and Heaslip (2019)	N/A	Using Deep learning architecture and tweets related to traffic incidents	The model first identifies the tweets based on their content into two classes, traffic incident, and non-traffic incident. Then by identifying the location specified in the tweet proper actions can be made.	Yes
D'Andrea et al. (2015)	Rome, Italy	Using machine learning methods to identify real time traffic related tweets.	A machine learning model is presented which classifies the tweets in to three classes with traffic related and traffic and location related tweets and then by text processing identifies the location in real time.	Yes

And due to heavy rise in traffic related pollution this young and healthy adult population is also at a higher risk of respiratory malfunctioning (Baig et al., 2020). This increase in pollution is due to increase of migrants from nearby villages and from other states due to job and other purposes, and increase in headcounts also results in increase of the vehicles on road. The total number of vehicles registered in Delhi in 2017 was 2.5 million (approx.) and in just one year it rose to 10 million (approx.) in 2018 (Government of Delhi, 2018). Due to this heavy rise in vehicle demand the air pollution also increases at a great rate. The atmospheric concentration of NH₃ also got affected or positively correlated by the emission of gases from automobiles (Kotnala et al., 2020). To reduce this effect of emission from vehicles the Delhi Govt. applied Road Space Rationing Programme (RSRP) which was called the even-odd rule. In which vehicles with registration number ending in even digit and odd digit were allowed to ply on the road in an alternate fashion. According to a survey 80% of the people found this approach as effective to solve the problem of air

pollution (Rao et al., 2017). Still, there is a great space for improvement in this strategy of traffic rationing.

We also examine the sentiments of people all over the country to get a view of the bigger picture of the problem. For that task, apart from our geotagged data collected for Delhi location we collected tweets from every city of the country. This data is collected on some mostly used hashtags for air pollution such as #airpollution, #cleanair, #pollution, #airquality. From the collected data we analysed the tweets to identify the location for which that tweet is meant for. For accomplishing this task, we applied tokenization method and Tf-Idf method of vectorization (A detailed description of Tf-idf vectorization method used for text classification is given in section 4). The resultant is the names of the cities along with the names of some state and some gibberish words. The meaningless words and the name of the states were filtered and a final list is maintained containing the ratio of tweets per city. The Fig. 1 represents the major cities and their contribution in the text corpora.



Fig. 1: Comparative participation of several cities of India to tweet on the hashtags (#airquality, #airpollution, #pollution, #cleanair) related to air pollution

This analysis also helped us to identify the most discussed city on Twitter with air pollution problem i.e. Delhi. The ratio of tweets pertaining to Delhi is high in case of each and every hashtag considered. A total of 12000 tweets, 3000 for each hashtag have been collected over a period of one month from December 2018 to January 2019. After this identification only, we further proceeded with our final dataset collection.

Materials and Methods

The method we used to model our ITS is based on the principle of traffic rationing but with dynamic selection of routes to be selected for rationing. From the studies it is obvious that not a single ITS model is able to completely solve the problems related to traffic congestion and vehicular emission. To guarantee a feasible solution, we formulated three basic necessities to be satisfied by any ITS scheme to fulfil the demand of society. The rules can be described as follows:

- i) It should be easy to adapt. As it should not increase the complexity and overall cost of the existing transportation system
- ii) It must be responsive and consistent enough to be used in a real time environment consisting of some critical situations (e.g. Traffic accidents)
- iii) It must capture all the aspects related to transportation, such as pollution control, risk assessment, throughput, stress of drivers, etc

As suggested by the method of demand spreading by Mori et al. (Mori et al., 2012) and currently used even-odd policy of Delhi Govt., we propose a load balancing scheme for traffic termed as left-right scheme. In the proposed method, first a busy street or area is identified where the AQI is significantly poor and a high traffic congestion is observed during the peak hours. This identification is basically done on the basis of user sentiments on social media about pollution level at a particular street or area. Our model does not perform traditional sentiment analysis (positive-negative-neutral). Instead, we classify tweets into specific categories related explicitly to environmental distress (pollution-specific and location-specific content). Previous studies used the level of harmful gases or pollutants to identify a locality with pollution related issues but it is seen from studies that even if the pollution is not up to the critical limit prescribe by the authority, people may suffer from harmful effects (Levy et al., 2020; Almeida, 2020).

For finding user sentiments about pollution, we used the standard search API of Twitter which is free of cost for standard usage. The API retrieves the tweets based on three parameters:

- II) Language, the language of the tweets to be searched
- III) Geocode, the longitude, and latitude of the location for which the tweets are to be searched along with a search radius

Data collection was performed daily using automated scripts, retrieving up to 100 tweets per API call within the rate limits of the Standard Search API. Duplicate tweets identified by exact text match were removed, resulting in a final dataset of 21,549 unique tweets.

The complete procedure of our model is shown in Fig. 2, the very first step is to extract the tweets by using the standard Twitter API. After extraction, some pre-processing is required to make the tweets in the proper format for classification. This pre-processing consists of the tokenization of tweets into distinct tokens followed by filtering of stop words consisting of English language punctuation marks and other stop words.

After which the task of feature extraction is performed on the tweets by using the tf-IDF method of vectorization. Out of more than 21,000 tweets, 68 different features were extracted by this method. After feature extraction, six machine learning algorithms are applied to the data for classification and their results have been compared based on accuracy measure. Along with accuracy precision, recall, and f1-score for all the algorithms have been calculated and compared. The complete process is described in detail in the following sections.

Collection and Labelling of Tweets

For our work, tweets are extracted with pollution-related content based on the hashtags given in Table 2.

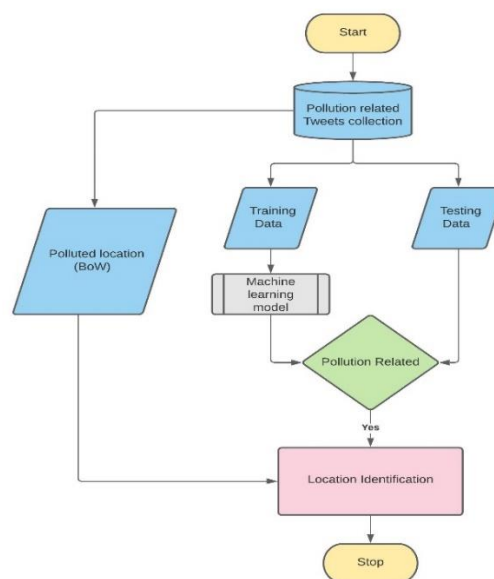


Fig. 2: Overall model architecture for identification of polluted location via user sentiments

- I) Search query, which is the hashtags for which the tweets are to be searched

Table 2: Performance comparison of BERT-based models in recent studies

Study	Model Architecture	Dataset Used	Key Performance Metrics	Limitations
Wan (2024)	Enhanced BERT	Emergency-related tweets	Not specified	Lacks detail on model evaluation and performance metrics
Pota (2021)	BERT	Twitter (English & Italian)	Insights on preprocessing strategies	Does not quantify performance improvements; focus is preprocessing
Rodríguez-Ibáñez (2023)	BERT, BiLSTM, Attention-based	IMDB, Social Networks	Superior performance demonstrated	General performance claims without specific quantitative benchmarks
Behera (2021)	LSTM	Twitter	BERT achieved high accuracy	No ablation studies to validate individual component impact
Das and Pedersen (2024)	BERT (BASE)	SemEval-2017 Task 4A	Accuracy: 0.6337, F1-score: Not specified	Moderate performance; lacks comparison with newer transformer models
Islam (2021)	BERT & BiLSTM	COVID-19 Vaccine Tweets	BERT: 89.99% accuracy, 85.08% MCC	Results dataset-specific; generalization to other domains not tested

Table 3: Hashtags to find tweets related to air pollution

#AIRPOLLUTION	#CLIMATECHANGE	#SMOG	#PM10
#CLEANAIR	#AIRQUALITY	#HAZE	#PARTICULATES
#EMISSIONS	#POLLUTION	#PM2.5	

We collected the tweets over a period of one year from January 2019 to January 2020. As we extracted the tweets restricted by the geolocation of the tweets, the number of tweets gathered were very few even we collected the tweets daily. The daily collection and filtering of the tweets resulted in a dataset of 21,549 tweets for analysis. Out of the total 21,549 collected tweets, the class-wise distribution was as follows:

Not Critical (NC): 9,120 tweets (~42%)
Pollution Specific (PS): 8,254 tweets (~38%)
Location Specific (LS): 4,175 tweets (~20%)

This distribution indicates a mild imbalance, which was addressed using balanced sampling techniques during model training.

The tweets were manually labelled into three classes based on their contextual content. Table 3 presents the hashtags used for collecting pollution-related tweets. These hashtags were carefully selected to capture a wide range of discussions related to air pollution and environmental distress on social media platforms, ensuring comprehensive data collection. The language of the tweets we searched for as of now is English.

The geocode values for Delhi are used as latitude=28.644800 and longitude=77.216721. An example of the extracted tweets is shown in Fig. 3. We can see from the tweets that some of the issues discussed in there might not be captured by the pollution trackers or meters installed at different locations. Such as some tweets are for a particular vehicle spreading a huge smoke out and causing breathing problems for the people nearby.

Such incidents can be quickly captured by sentiments on social media and proper action can be taken in due time.

Tweets Pre-Processing

Before the labelling of the tweets, we pre-process the tweets to make them suitable for the classification process. The following are the few steps that we followed to modify the tweet dataset:

- Remove duplicate tweets, tweets containing the same text are considered to be duplicate. So, we have to drop them
- Short tweets containing only 5 letters or a few are discarded
- All abbreviations related to the street names or area names are unfolded to their full names

For BERT-based classification, raw tweet text was additionally cleaned by removing URLs, @mentions, and the hashtag symbol (#) while retaining the keyword text. Tweets exceeding 128 WordPiece tokens were truncated to the maximum sequence length supported by BERT-base.

Labelling of Tweets

The next step is to classify the tweets into different categories, for which initially we made three classes. The present model is perhaps considering only a 2-class model in which one class of tweets will be for the tweets which contain information regarding the pollution-based tweet for a particular street or area and rest other tweets lies in the second class which is not related to any location in the context of pollution.

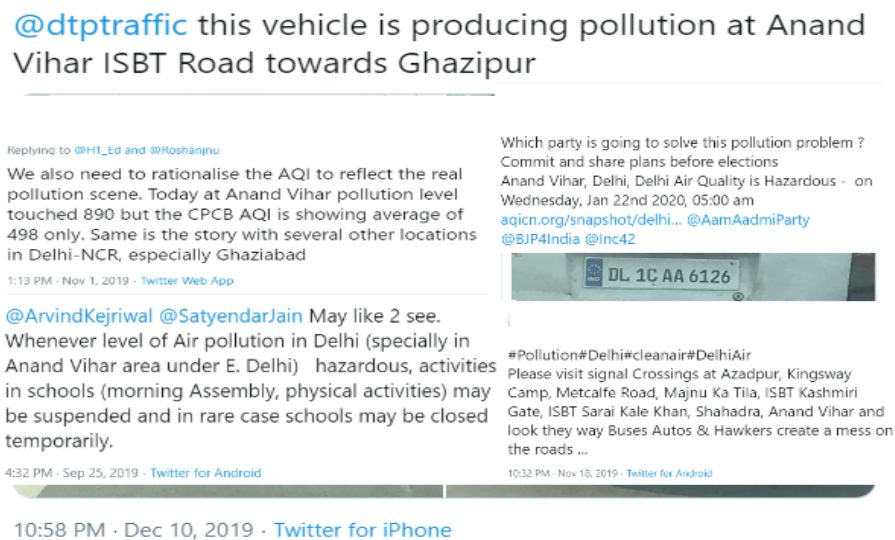


Fig. 3: Tweets extracted showing the environment-related distress in the locality of Anand Vihar Delhi

But we created our dataset to extend the research further into a 3-class identification problem, for that we categorized the tweets into three classes:

- i) NC (Not critical): These are the tweets in which there is no discussion of the pollution-related problem, there are several tweets that we came across in which there is a general discussion or suggestion related to pollution is present
- ii) These are not useful for our scheduling and location identification purposes
- iii) PS (Pollution specific): The tweets in which pollution-related content is there but there is no discussion about the location. These types of tweets are also not useful for scheduling purposes at present
- iv) LS (Location specific): This is the category of tweets which is used for the identification of pollution-related distress for a particular location. The tweets which contain pollution-related keywords and a discussion of the location lies in this category

Table 4 presents representative examples of tweets along with their corresponding class labels. These examples illustrate how tweets are categorized into Pollution Specific (PS) and Location Specific (LS) classes, forming the basis for subsequent analysis and location identification.

Tokenization and Stop Word Removal

The first step for processing a text document for any analysis through machine learning-based algorithms is to convert the sentence into a collection of tokens. Tokenization is important for NLP (natural language processing) projects because it breaks a continuous stream of words into separate tokens so that further processing can be done. In NLP tokenization can be performed at various levels such as character level, word level, or sub-word level. Several algorithms are present for each level of the tokenization method, for our processing we used the word level tokenization in this project.

Table 4: Samples of tweets and their related classes for the 2-class dataset

S. NO.	Text	Class
1	#AirQuality to fall to moderate levels later today. Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.	PS
2	Scientists tested people while exercising and measured what percentage of particles stuck in their lungs. Here's what they found. #Health #AirPollution	PS
3	Air Quality Index (AQI) at 431 in 'severe' category in <u>Anand Vihar</u> and at 372 in 'very poor' category in RK Puram, according to Delhi Pollution Control Committee (DPCC) data	LS
4	#Delhi: Air quality in 'Very Poor' category in <u>Anand Vihar</u> , Lodhi Road and India Gate circle areas, as per Central Pollution Control Board. #DelhiAirQuality #AirPollution #AirQuality	LS
5	#Delhi: Air Quality Index (AQI) at 463 in ITO and 467 around <u>Anand Vihar</u> - both in 'Severe' category. ANI #AirQuality #AirPollution #DelhiAirEmergency #DelhiAirQuality #DelhiPollution	LS

In word-level tokenization, we break the sentence into different words to create our vocabulary. For this task, there are various methods available such as white space tokenization, rule-based tokenization, dictionary-based tokenization, etc. We applied the word tokenize method available in the nltk library of python which is a dictionary-based tokenization method.

Stop words are those words which do not have enough importance in the meaning of the sentence such as “an”, “is”, “of” etc. In any word processing model, the step before feature extraction is the removal of these stop words which reduce the number of unnecessary words from the document. We can use various NLP libraries to stop word removals like NLTK, Gensim, and SpaCy, we used the NLTK package of python for the stop word removal.

Feature Extraction

After pre-processing the next step is to extract the features from the tweets for making them appropriate for the classification algorithms. For feature extraction, we used the *tf-IDF* approach which provides a measure statistically about how a word is important to a particular document when several documents are present. The *tf-idf* approach has been widely used in the domain of text classification (Almeida, 2020; Zhang et al., 2005; White, 2018). Where *tf* stands for term frequency and *IDF* stands for inverse document frequency. The following formulae represent the calculation of *tf-idf*:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

$$idf_i = \log \frac{D}{\{d_j: t_j \in d_j\}} \quad (2)$$

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

In Eq. 1. The term $tf_{i,j}$ represents the frequency of the term i in the document j . $n_{i,j}$ is the number of times the word i appears in the document j and $\sum_k n_{k,j}$ is the length of the document j or the number of words present in document j . D is the total number of documents which is divided by the times the word i appear in different documents. Finally, *tf* and *idf* values are multiplied to yield the final metric of *tf-idf*, this value represents the importance of any word in the document. The higher the frequency of a word lower will be the *tf-idf* value of that word or in other words higher the *tf-idf* value of any word than that word is considered as important to the document.

Classification

After pre-processing of tweets, we applied a suite of algorithms to classify tweets into the two categories,

with a particular focus on the performance of BERT vs. other approaches. In total, we evaluated seven classifiers: Decision Tree, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naive Bayes, Linear Discriminant Analysis, and a BERT-based transformer model. The first five are traditional machine learning models often used in text classification, and they were trained on the TF-IDF features of the tweets. The BERT model, in contrast, operates on the raw text and leverages deep contextual embeddings. We fine-tuned the pre-trained BERT-base model (uncased) on our labelled tweet dataset for binary classification. Fine-tuning involved adding a dense output layer on top of BERT's [CLS] token representation and training the model to predict LS vs. PS class. We used a 75/25 train-test split on the data for model development and performed 10-fold cross-validation to ensure robust evaluation for each classifier. Recognizing the potential for overfitting concerns with transformer-based models like BERT, we additionally implemented an independent hold-out validation (80% training, 10% validation, 10% testing) specifically for the BERT fine-tuning procedure. Early stopping and regularization techniques (dropout) were strictly employed during training to mitigate overfitting risks.

The performance results reported reflect both the cross-validation outcomes (for baseline models) and the dedicated hold-out test set evaluation for BERT, ensuring robust validation of the model. The BERT model was fine-tuned for several epochs with a learning rate in the order of 10^{-5} (typical for BERT fine-tuning) represented in Fig 4., using the Adam optimizer. The model was fine-tuned for 4 epochs with a batch size of 32, a dropout rate of 0.3 on the final classification layer, and a weight decay of 0.01. Training was conducted on a single NVIDIA GPU (Tesla T4) using the Hugging Face Transformers library (v4.x) in a Python 3.8 environment. Early stopping was employed to prevent overfitting, given the relatively modest dataset size for a deep model. For the classical models, we used scikit-learn implementations with default parameters, tuning them minimally (for instance, the decision tree's depth was not explicitly constrained since the dataset was manageable). All model training and evaluation was conducted in a Python Jupyter environment; the transformer model was implemented via the Hugging Face Transformers library, while the others used scikit-learn. Unlike previous BERT applications that focused on general sentiment classification, our model is trained specifically to distinguish pollution-related tweets and extract actionable geographic references, which has not been demonstrated in prior urban sensing works.

The performance of algorithms was measured using the accuracy measure, but other features like recall, precision, and f1-score are also calculated by class. The calculation method for these factors is described below:

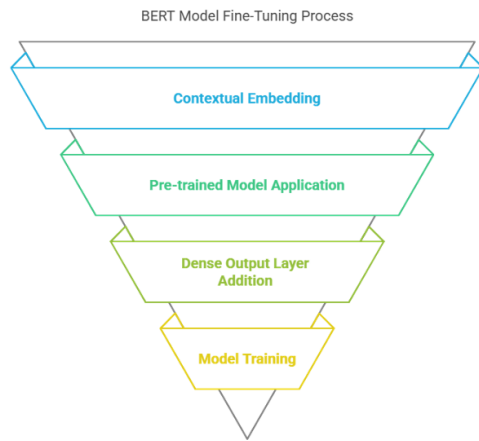


Fig. 4: Proposed BERT model fine-tuning process

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$Prec = \frac{TP}{TP + FP} \quad (5)$$

$$Rec = \frac{TP}{TP + FN} \quad (6)$$

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (7)$$

For the explanation of the metrics shown in equations let's consider an example of binary classification. The validity of any classification can be measured by four parameters:

- i) TP (True Positive): The number of samples that are correctly classified as the positive class
- ii) TN (True negative): The number of samples which are correctly classified as a negative class
- iii) FP (False positive): The number of samples which are negative but classified as positive
- iv) FN (False negative): The number of samples that are positive but are classified as negative

Based on these four parameters we calculate our accuracy (Eq. 4) which is the measure of the overall performance of the classification measured as the number of samples classified correctly divided by the total number of samples. Precision (Eq. 5) is the measure of the number of samples classified correctly of a class divided by the total number of samples classified into that class. Recall (Eq. 6) is the measure of the number of samples classified correctly of a class divided by the actual number of samples of that class. The harmonic mean of precision and recall is termed as the f1 score (Eq. 7) and it is also the measure of overall correctness of the method. It is calculated by the values of precision and recall calculated per class.

Location Identification

The next phase after classification of the “Location-specific” tweets from the dataset is to identify the locations from the tweets. As for now, we have collected the tweets with geolocation of New Delhi, we will search for the major polluted areas if any discussed in the tweets. For identification of location, we created a bag of words of the major pollution affected the location of New Delhi. The Bag of a word is created in accordance with the data provided by the CPCB (Centre for Pollution Control Board) of Government of India. There is a total of 31 locations in Delhi which are regularly monitored by the CPCB. We count the occurrences of these different locations in the classified tweets and count their frequencies day wise. Then the location with the highest number of tweets regarding pollution is selected as the current location for traffic scheduling due to pollution, by traffic rationing method.

Our method directly addresses the research question through a two-stage process:

- First, we fine-tune a BERT model to detect environment-related distress in tweets (what is being said)
- Second, we extract and aggregate location keywords from classified tweets (where it is being said). This integration of tweet content and location context allows the system to detect pollution hotspots as inferred from public sentiment

Results

Classification Comparison

The comparison of the performance of all 6 classification algorithms is shown in Fig. 5. The results are obtained after applying 10-fold cross-validation on every classification algorithm to improve the result. Despite the Support Vector Machine (SVM) model achieving perfect scores in certain metrics, such as recall or precision for specific classes, its performance tends to fluctuate across different evaluation criteria. In contrast, the Proposed Model demonstrates consistently high performance across all key metrics, including accuracy, precision, recall, and F1-score, for both location-specific and pollution-specific classification tasks. This consistent behaviour indicates a better generalization capability, reducing the risk of overfitting often associated with models showing perfect scores. Furthermore, the Proposed Model balances performance trade-offs more effectively, achieving robustness without sacrificing precision or recall, thus making it a more reliable choice for real-world applications.

Error analysis revealed key challenges. For instance, sarcastic tweets such as ‘Love breathing smog every morning!’ were misclassified as positive due to literal interpretation. Tweets using local slang or code-switching

between Hindi and English also posed difficulties. These examples highlight the importance of cultural and linguistic nuance in distress detection.

The comparison of proposed model with the existing BERT based models have been done and the proposed outperformed the existing models in both the parameters of Accuracy and F1 score (Fig. 6). Table 5 shows the comparison of the models based on Accuracy and F1-Score.

The BERT-based classification model achieved an overall accuracy of 96.5% with a weighted F1-score of 96% using 10-fold cross-validation and a 75/25 train-test split. The class-wise performance metrics demonstrate the robustness of the proposed model across all distress

categories. Precision, recall, and F1-score values indicate (Table 6) balanced model performance with minimal class bias, particularly for minority classes such as Location Specific (LS).

Location Identification

After the classification of the tweets into PS and LS Class, the next task is to identify the current location of the city where the pollution-related sentiments created a lot of distress to the people around. For this task, the tweets classified as Location-specific tweets are parsed and checked for the occurrences of the location titles we stored in our bag of the word. For every occurrence per tweet of a location, a counter value is incremented.

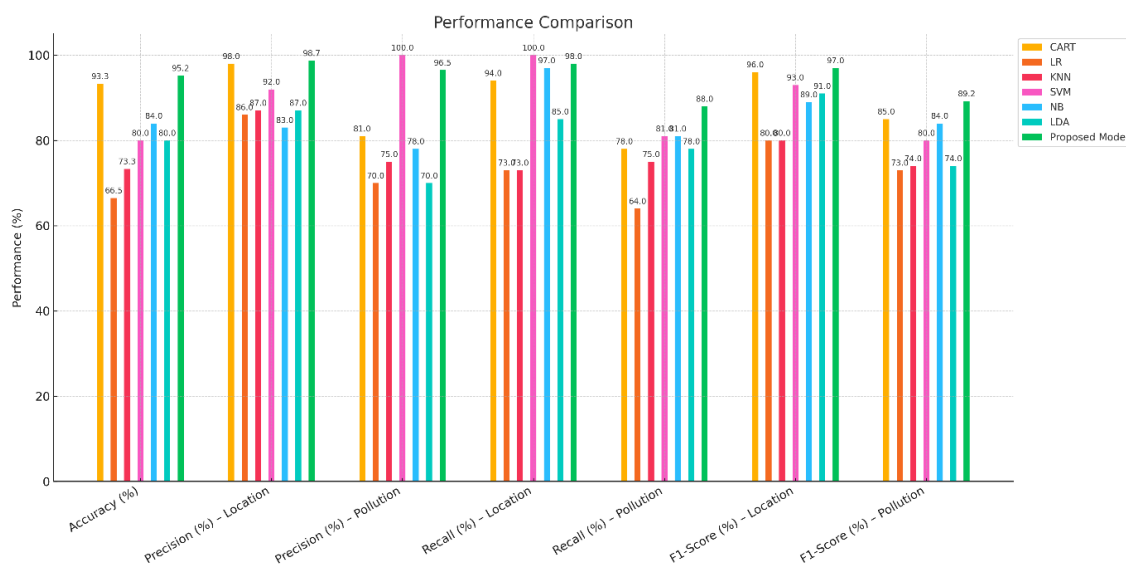


Fig. 5: Comparison of accuracy measures for the classification algorithms

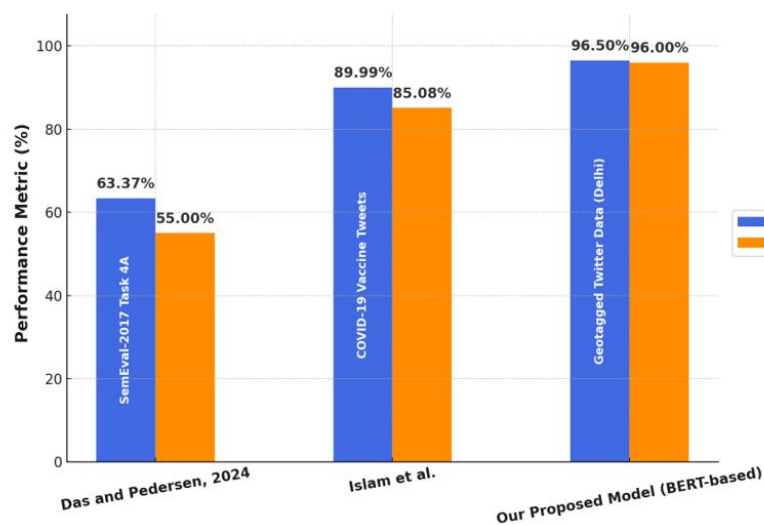


Fig. 6: Model comparison with BERT based models

Table 5: Model comparison with BERT based models

Study (Author, Year)	Accuracy (%)	F1-score (%)
Das and Pedersen (2024)	63.37	55
Islam (2021)	89.99	85.08
Our Proposed Model (BERT-based)	96.5	96

Table 6: Class-wise classification performance of the proposed BERT model

Class	Precision	Recall	F1-Score	Support
Not Critical (NC)	0.97	0.96	0.96	9,120
Pollution Specific (PS)	0.96	0.97	0.96	8,254
Location Specific (LS)	0.95	0.94	0.95	4,175
Weighted Avg	0.96	0.965	0.96	21,549

At the end of the parsing of the entire tweet set, the locations and their counter values are obtained and the location with the highest value of the counter is selected as the location where traffic scheduling measures have to be performed. Fig. 7 represents the percentage-based sentiment score for our current dataset, the location which is identified with the largest number of pollution-related tweets is “Anand Vihar”.

The reason behind the Anand Vihar area of New Delhi for getting the highest number of problem-related tweets is, that it is one of the most crowded areas of the city (Nath, 2016). Several other researchers also considered this area as their target for pollution study because of its highly pollution affected nature. Anand Vihar is reported to be more vulnerable to the harmful effects of traffic congestion and vehicular emission (Mishra, 2019). A satellite view of the Anand Vihar area is shown in Fig. 8 with the following commodities:

- Transportation
- Hotels
- Hospitals
- Residences

These maps are taken with the help of the ISRO Bhuvan platform which is a national Geoportal containing the 2-D and 3-D images regarding various natural resources. These map images can also show us the high number of organizations and offices running in a very short area of 5 km. Also, there is a large number of residential properties lies in this area in which thousands of people living in it, and they will include children, old age people, and a person suffering from other ailments. In peak hours the concentration of pollutants touches the harmful threshold level almost every day, due to which people sometimes feel less visibility also. A fog-like situation arises for the people of this area and almost every part of the city. Also, the satellite view of Bawana area is shown in Fig. 9 which is as per the AQI index is more polluted to the Anand Vihar area but as we can see that the population in the Bawana area is low and due to that the negative sentiment generated for that area is low. This point proves the effectivity of our presented approach. The frequency of classified LS tweets mentioning Anand Vihar indicates a high density of pollution-related complaints for that location.

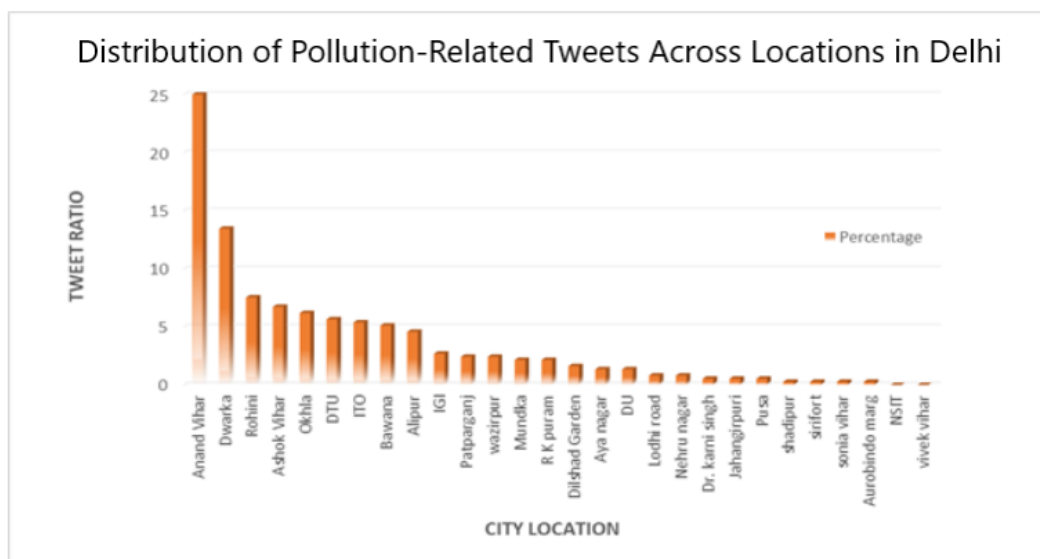


Fig. 7: The Percentage of location-specific pollution-related tweets for different locations of Delhi

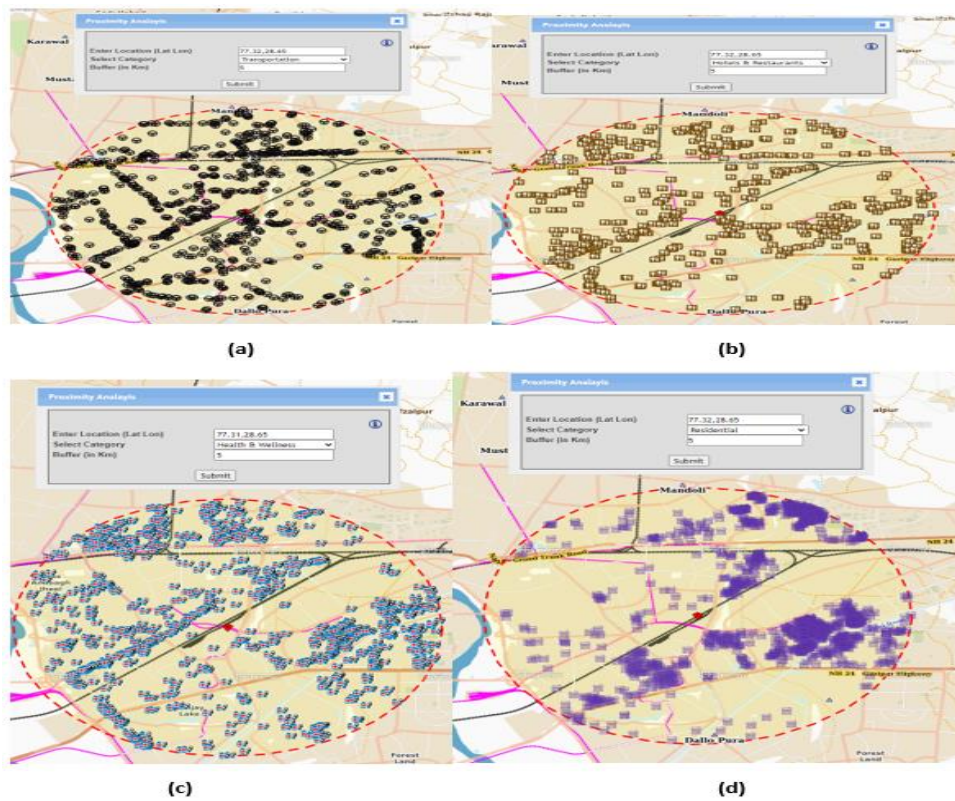


Fig. 8: Satellite images collected from ISRO Bhuvan data showing the density of a) Transport, b) Hotels, c) Hospitals and d) Residential space in Anand Vihar locality of Delhi

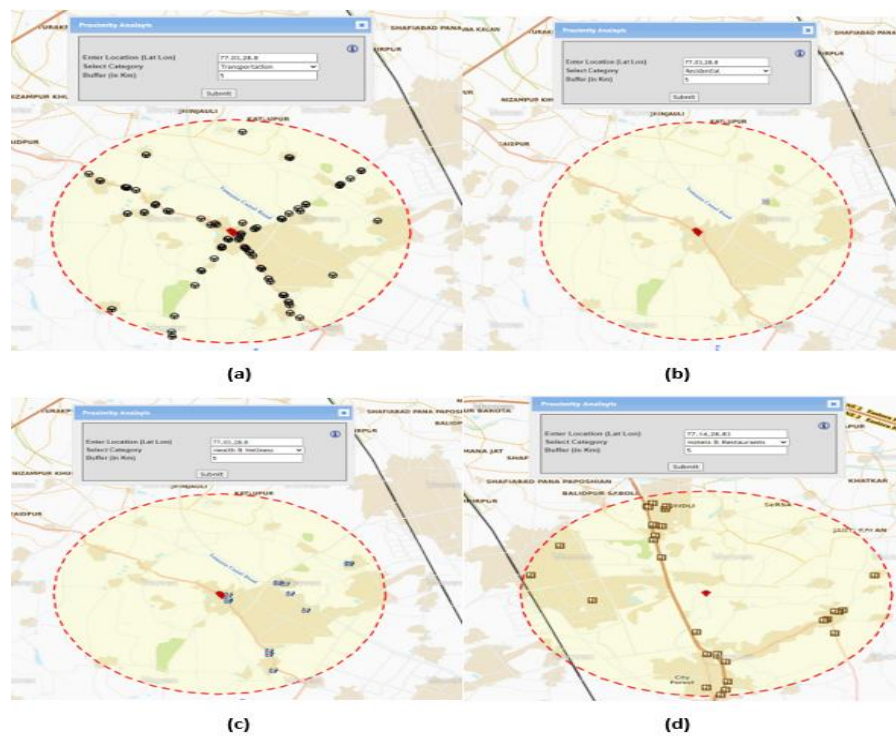


Fig. 9: Satellite images collected from ISRO Bhuvan data showing the density of a) Transport, b) Residentials, c) Hospitals and d) Hotel space in Bawana locality of Delhi

This demonstrates how the model answers the RQ: It uses social media as a sensor, where public complaints (detected through NLP) signal environmental distress in specific areas.

Discussion

This study suggests a traffic scheduling policy based on the sentiments collected of people on online social media regarding air pollution in an overpopulated city of India i.e. Delhi. It was found out that there is a significant impact of the population of an area on the sentiments shown by people on online social media related to pollution as it is discussed that social media is an exact measure of human sentiment (De Carolis et al., 2020). Previous studies also suggest that there can be some places where even if the pollution is not marked at its critical level but still the residents of that area can suffer from respiratory problems due to pollution such as places with a greater number of hospitals and schools. The comparison of two such places based on satellite images is also shown for one of which is a densely populated area and another one is a low populated area. From the results, it is obvious that the people in densely populated areas containing a high number of residential buildings, hospitals, transport facilities, and hotels suffer more from the problems of air pollution and need immediate solutions. A traffic scheduling policy is also suggested to curb the traffic of that area because an increase in traffic is the main cause of an increase in air pollution.

Future research should incorporate multilingual data to better reflect different perspectives, incorporate social media data along with sensor-based AQI data to validate those sources of information, and to understand how expressions of distress differ based on class, caste or gender. Qualitative methods (e.g., interviews, ethnography) will also provide valuable insight into the actual experiences of those affected by pollution in addition to what is found through computational methods. Policymakers could use this model to prioritize high-risk zones particularly near schools, hospitals, and densely populated residential areas for immediate traffic management interventions, AQI monitoring upgrades, or public health advisories. While the model achieves high classification accuracy, several limitations should temper interpretation. The dataset is restricted to English-language, geotagged tweets from Delhi, which represents only a small fraction of total pollution-related discourse. The bag-of-words location extraction method may miss misspelled location names or informal references (e.g., 'Anand Vihar station' vs. 'AV'). Furthermore, the high accuracy of 96.5% should be interpreted cautiously given that the dataset was collected during a specific season (winter 2019–2020), when Delhi's pollution is historically at its peak, which may inflate performance for that pollution regime.

Biases and Limitations

Despite providing valuable insights into pollution-related concerns, the dataset used in this study is subject to several biases and limitations that may affect the generalizability and comprehensiveness of the findings. These biases include.

Geotagging Bias

Only tweets with a geocode for Delhi, latitude 28.644800 and longitude 77.216721, could be included in the dataset for analysis, creating a relatively small number of geotagged tweets out of the total number of tweets that were posted to social media during one year of daily collection. Due to this geospatial restriction, some tweets concerning pollution that were relevant in the surrounding area or in less populated areas may have been excluded.

Language Bias

Limiting the data collection to tweets in only one language. In this case, English. Will limit possible responses from people who do not speak English. It might also skew the representation of how people feel about pollution if the majority of those responding to the question or issue are only English speakers. Delhi and the surrounding regions are multilingual. Consequently, the bias of limiting data to English-speaking people will affect the representativeness of the data.

Temporal Bias

The tweets were collected over a one-year period (January 2019 to January 2020), which may not reflect long-term or seasonal variations in pollution levels and public concerns. Events such as festivals, industrial shutdowns, or sudden weather changes could disproportionately impact the data collected during this period.

Selection Bias

Hashtags of the sample in Table 2 could potentially remove any tweets related to pollution that do not have those specific hashtags present; therefore, tweets where citizens are discussing pollution but are not using trending tags would be removed from this analysis and would result in fewer insights from citizens impacted by pollution who do not currently use any trending hashtags to engage with others who talk about pollution. Because citizens who have made a comment about pollution may not have been included due to not including specific hashtags related to pollution (e.g., #pollution), this may introduce skews to the study population's data and thus lead to skewed findings to urban areas/digitally adept citizens.

Sentiment Interpretation Bias

Textual clues are primarily used to determine public sentiment in the area of sentiment analysis. Misclassifying

or misrepresenting opinions surrounding pollution could result from subtleties within the text such as sarcasm, humour, or language that is specific to a certain cultural context/region. Manual labelling also creates opportunities for human error.

Small Dataset

The dataset size (~21,500 tweets) raises potential overfitting concerns for transformer models. Cross-validation was used to mitigate this, but generalization remains limited.

Triangulation Limitation

Findings were not cross validated with external data such as AQI readings or health records. Future work should triangulate results with independent sources for validation.

Mitigation and Future Directions

To address these biases in future work, several strategies can be explored:

- **Incorporating Multilingual Data:** Expanding the dataset to include non-English tweets can help capture a broader spectrum of public sentiment
- **Combining Data Sources:** Integrating pollution data from sensors, government reports, and social media can improve the reliability of the findings
- **Reducing Geotagging Dependency:** Exploring methods to identify location information from tweet content or user metadata may help mitigate the geographic bias
- **Extending the Time Frame:** Collecting data over a longer period or during specific pollution-prone seasons can provide more comprehensive insights

By acknowledging these biases and proposing mitigation strategies, this study aims to pave the way for more inclusive and representative Environmental Distress Classification of pollution-related concerns.

Conclusion and Future Scope

The proposed traffic scheduling framework is conceptual. Actual implementation would require real-time traffic data and optimization modelling, which are beyond the scope of this paper and planned for future research. The presented model leverages transformer-based classification to identify environment-related distress (pollution-specific and location-specific content) rather than generic sentiment polarity. The identified locations are thus based on the frequency of location-specific complaints, offering actionable insights for targeted interventions. This paper proposed a method for the identification of problematic areas suffering from the

problem of pollution and its related consequences. The task was divided mainly into 3 modules, Tweet collection, Tweet processing, and location identification. For the task of identifying pollution-related distress in tweets, we evaluated multiple machine learning algorithms. While the algorithms like Decision Tree and SVM achieved high raw accuracy in one setting, the fine-tuned BERT model demonstrated more consistent and generalizable performance, making it the foundation of our proposed framework. We selected BERT as the core model due to its contextual understanding, adaptability to nuanced language, and future extensibility (e.g., to handle multilingual data or extract entities). This makes it better suited for real-time urban deployment despite slight differences in baseline accuracy. After the classification of tweets, the location with the highest number of tweets is identified and the proposed traffic scheduling on the streets will be performed. It is the need of the hour for emerging smart cities to incorporate smart transport systems as well and the proposed model is an efficient model for traffic scheduling based on human sentiments. The implementation of the traffic scheduling algorithm is not discussed in this paper which can be the prospective future work, to collect the real-time data and to find out the time difference to be applied to rationalise the traffic on the streets.

In answering our research question, we show that public sentiment on social media, when processed with fine-tuned NLP models and mapped to locations, can reliably identify urban pollution hotspots in need of intervention. This bridges public concern and actionable city responses. Future research should pursue the following concrete directions:

- (1) **Multilingual extension:** Fine tune a multilingual BERT variant (mBERT or XLM-R) on Hindi-English code-switched tweets to capture a broader population
- (2) **Sensor fusion:** Integrate real-time AQI sensor readings with tweet frequency scores to triangulate pollution hotspots and validate the social media signal against ground truth
- (3) **NER-based location extraction:** Replace the bag-of-words location matching with a fine-tuned Named Entity Recognition model to handle informal, abbreviated, or misspelled location names
- (4) **Temporal modeling:** Apply time-series analysis on tweet volumes to predict pollution surges before they peak, enabling proactive rather than reactive traffic management
- (5) **Ethical deployment:** Conduct a bias audit of the classification model across demographic groups and explore federated learning approaches to address data privacy concerns

Furthermore, the integration of our model with urban planning platforms could provide policymakers with valuable insights to prioritize pollution mitigation efforts. This could include identifying pollution hotspots near schools, hospitals, or densely populated residential areas and informing traffic policies or urban development plans. We envision deploying this model on smart city dashboards or integrating it with Intelligent Transportation Systems (ITS) to provide real-time alerts and route suggestions to commuters. This can help reduce pollution exposure for vulnerable populations and improve overall urban air quality. By incorporating real-time feedback loops and dynamic data inputs, future iterations of this work will aim to bridge the gap between sentiment analysis and actionable urban planning strategies. While promising, these results are preliminary. The model should be seen as a complementary tool to traditional environmental monitoring rather than a definitive system. Its outputs require validation and contextual interpretation before informing policy or real-time interventions.

Acknowledgment

The authors would like to thank all individuals and organizations that contributed to this work. We acknowledge JUT for providing computational resources and support during the experiments. We also thank the Twitter Developer community for facilitating data access through the API.

Funding Information

The authors declare that no external funding was received and the research was conducted with internal resources.

Authors' Contributions

Utkarsh Sharma: Conceptualization and Methodology: Conceived the research idea and designed the methodology; Writing (Original Draft), wrote the initial draft of the manuscript, especially the introduction and methods sections; Project Administration coordinated the research activities and managed the project timeline. Data Curation, Collected the Twitter data and performed data preprocessing and labeling; Software, implemented the data collection scripts and contributed to the code for model training; Validation: Helped design validation experiments and reviewed the results for consistency.

Prateek Pandey: Model Development Developed and fine-tuned the BERT-based model; Formal Analysis, analyzed the experimental results and carried out error analysis; Writing (Review and Edited), contributed to

writing the results section and revising the manuscript for important intellectual content.

Shishir Kuma: Supervision, Supervised the research project and provided guidance on methodology; Visualization, created figures and charts to illustrate data distributions and model performance; Writing (Review and Edited), critically reviewed the manuscript draft, provided substantial edits, and approved the final version for submission.

Ethics

This study involved the analysis of publicly available social media data and did not involve any intervention or interaction with human subjects. Ethical Approval: In accordance with prevailing research guidelines, analysis of public tweets is generally exempt from Institutional Review Board (IRB) review because the data are publicly available and anonymized. Nonetheless, the authors took care to maintain ethical standards in data handling. No attempts were made to identify or contact the individuals behind the accounts. We removed or masked any user-identifying information (such as Twitter handles or real names) during data processing to protect user privacy. The study was conducted in compliance with Twitter's Terms of Service and developer policies, which allow collection and analysis of public content for research purposes.

Informed Consent: Because the data consist of public tweets, obtaining individual consent was not feasible; however, we only used data that users had made public. No private or direct messages were accessed. We acknowledge the ethical considerations of using social media content and have referenced guidelines for internet research ethics to ensure respect for user privacy and intent.

Data Availability: The Twitter data used in this study are subject to Twitter's data sharing policies. We provide tweet IDs and processed feature data upon reasonable request, which will allow other researchers to retrieve the original tweets directly from Twitter (ensuring compliance with Twitter's terms). Any additional derived data or analysis code is available in the project's online repository or from the corresponding author on request.

Conflict of Interest

The authors have no competing financial or personal interests to disclose.

References

- Almeida, L. O. (2020). Influence of urban forest on traffic air pollution and children respiratory health. *Environmental Monitoring and Assessment*, 192, 175. <https://doi.org/10.1007/s10661-020-8142-4>

- Ashmore, M. R., Batty, K., Machin, F., Gulliver, J., Grossinho, A., Elliott, P., Tate, J., Bell, M., Livesley, E., & Briggs, D. (2000). Effects of Traffic Management and Transport Mode on the Exposure of Schoolchildren to Carbon Monoxide. *Environmental Monitoring and Assessment*, 65(1–2), 49–57. <https://doi.org/10.1023/a:1006431301620>
- Baig, N. A., Yawar, M., Jain, K., Singh, G., Singh, S., Siddharthan, D., & Habib, G. (2020). Association between traffic emissions mixed with resuspended dust and heart rate variability among healthy adults in Delhi. *Air Quality, Atmosphere & Health*, 13(3), 371–378. <https://doi.org/10.1007/s11869-020-00800-2>
- Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., & Stanaway, S. S. (2019). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India. *The Lancet Planetary Health*, 3(1), e26–e39. [https://doi.org/10.1016/S2542-5196\(18\)30261-4](https://doi.org/10.1016/S2542-5196(18)30261-4)
- Behera, R. K. (2021). Cognitive computing in healthcare. *Information Processing & Management*, 58(1), 102435.
- Bergstra, A. D., Brunekreef, B., & Burdorf, A. (2018). The effect of industry-related air pollution on lung function and respiratory symptoms in school children. *Environmental Health*, 17(1), 30. <https://doi.org/10.1186/s12940-018-0373-2>
- Choudhary, A., & Gokhale, S. (2019). Evaluation of emission reduction benefits of traffic flow management and technology upgrade in a congested urban traffic corridor. *Clean Technologies and Environmental Policy*, 21(2), 257–273. <https://doi.org/10.1007/s10098-018-1634-z>
- D’Andrea, E., Ducange, P., Lazzarini, B., & Marcelloni, F. (2015). Real-time traffic detection from Twitter. *IEEE Transactions on Intelligent Transportation Systems*, 16(4), 2269–2283. <https://doi.org/10.1109/TITS.2015.240443>
- Dabiri, S., & Heaslip, K. (2019). Developing a Twitter-based traffic event detection model using deep learning architectures. *Expert Systems with Applications*, 118, 425–439. <https://doi.org/10.1016/j.eswa.2018.10.017>
- Das, R. K., & Pedersen, T. (2024). *SemEval-2017 Task 4: Sentiment Analysis in Twitter using BERT*. arXiv. <https://doi.org/10.48550/arXiv.2401.07944>
- De Carolis, B., D’Errico, F., & Rossano, V. (2020). Socio-affective technologies [SI 1156 T]. *Multimedia Tools and Applications*, 79(47–48), 35779–35783. <https://doi.org/10.1007/s11042-020-10015-3>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). <https://doi.org/10.18653/v1/N19-1423>
- Diaz, G., Macià, H., Valero, V., Boubeta-Puig, J., & Cuartero, F. (2020). An Intelligent Transportation System to control air pollution and road traffic in cities integrating CEP and Colored Petri Nets. *Neural Computing and Applications*, 32(2), 405–426. <https://doi.org/10.1007/s00521-018-3850-1>
- Farrag, S. G., Outay, F., Yasar, A. U.-H., Janssens, D., Kochan, B., & Jabeur, N. (2020). Toward the improvement of traffic incident management systems using Car2X technologies. *Personal and Ubiquitous Computing*, 25(1), 163–176. <https://doi.org/10.1007/s00779-020-01368-5>
- Gabrys, J. (2016). *Program Earth: Environmental sensing technology and the making of a computational planet*. 49.
- Government of Delhi. (2018). Vehicle registration data. *Transport Department Portal, Delhi*.
- Government of India. (2011). Census report. *Census India Web Portal*.
- Hamidi, H., & Kamankesh, A. (2018). An Approach to Intelligent Traffic Management System Using a Multi-agent System. *International Journal of Intelligent Transportation Systems Research*, 16(2), 112–124. <https://doi.org/10.1007/s13177-017-0142-6>
- Heydon, J., & Chakraborty, R. (2020). Can portable air quality monitors protect children from air pollution on the school run? An exploratory study. *Environmental Monitoring and Assessment*, 192(3), 195. <https://doi.org/10.1007/s10661-020-8153-1>
- Hilpert, M., Johnson, M., Kioumourtoglou, M.-A., Adria-Mora, B., Peters, A., Ross, J., & Chillrud, S. (2019). Traffic-related air pollution inference. *Environment International*, 127, 142–159.
- INRIX. (2018). Global traffic scorecard report. *INRIX Research*. https://static.poder360.com.br/2019/02/INRIX_2018_Global_Traffic_Scorecard_Report_final.pdf
- Islam, M. R. (2021). COVID-19 vaccine tweet analysis. *Axioms*, 10(4), 270.
- Jain, A., Yadav, S., Vij, S., Kumar, Y., & Tayal, D. K. (2020). A Novel Self-Organizing Approach to Automatic Traffic Light Management System for Road Traffic Network. *Wireless Personal Communications*, 110(3), 1303–1321. <https://doi.org/10.1007/s11277-019-06787-z>
- Kim, S.-W., & Gil, J.-M. (2019). Research paper classification. *Human-Centric Computing and Information Sciences*, 9, 30. <https://doi.org/10.1186/s13673-019-0192-7>
- Kotnala, G., Sharma, S. K., & Mandal, T. K. (2020). Influence of Vehicular Emissions (NO, NO₂, CO and NMHCs) on the Mixing Ratio of Atmospheric Ammonia (NH₃) in Delhi, India. *Archives of Environmental Contamination and Toxicology*, 78, 79–85. <https://doi.org/10.1007/s00244-019-00689-8>

- Levy, I. (2020). A hybrid model for evaluating exposure of the general population in Israel to air pollutants. *Environmental Monitoring and Assessment*, 192, 4. <https://doi.org/10.1007/s10661-019-7960-8>
- Li, Z., Al-Hassan, R., Shahidehpour, M., Bahramirad, S., & Khodaei, A. (2019). A Hierarchical Framework for Intelligent Traffic Management in Smart Cities. *IEEE Transactions on Smart Grid*, 10(1), 691–701. <https://doi.org/10.1109/tsg.2017.2750542>
- Mishra, S. (2019). Comparative emission study by real-time congestion monitoring for stable pollution policy on temporal and meso-spatial regions in Delhi. *Journal of Cleaner Production*, 224, 465–478. <https://doi.org/10.1016/j.jclepro.2019.03.122>
- Mori, H., Kitaoka, H., & Asakura, Y. (2012). Simulation-based Analysis for Reducing Traffic Congestion in Real Traffic Networks by Demand Spreading over Time. *International Journal of Intelligent Transportation Systems Research*, 10(2), 93–100. <https://doi.org/10.1007/s13177-012-0045-5>
- Myllyvirta, L., & Sivalingam, S. D. (2016). Out of sight: How coal burning advances India's air pollution crisis. *Greenpeace Report*. <https://wayback.archive-it.org/9650/20200505115958/http://p3-raw.greenpeace.org/india/Global/india/cleanairnation/Reports/Out%20of%20Sight.pdf>
- Nath, D. (2016). Anand Vihar pollution report. *The Hindu*.
- Ogunsanya, A. A. (1984). Improving urban traffic flow by restraint of traffic: The case of Lagos, Nigeria. *Transportation*, 12(2), 183–194. <https://doi.org/10.1007/bf00167375>
- Osorio, C., & Nanduri, K. (2015). Urban transportation emissions mitigation: Coupling high-resolution vehicular emissions and traffic models for traffic signal optimization. *Transportation Research Part B: Methodological*, 81, 520–538. <https://doi.org/10.1016/j.trb.2014.12.007>
- Papacharissi, Z. (2015). We Have Always Been Social. *Social Media + Society*, 1(1), 1–2. <https://doi.org/10.1177/2056305115581185>
- Percoco, M. (2015). Temporal aggregation and spatio-temporal traffic modeling. *Journal of Transport Geography*, 46, 244–247. <https://doi.org/10.1016/j.jtrangeo.2015.07.001>
- Picone, M., Busanelli, S., Amoretti, M., Zanichelli, F., & Ferrari, G. (2015). *Advanced Technologies for Intelligent Transportation Systems*. 139. <https://doi.org/10.1007/978-3-319-10668-7>
- Pota, M. (2021). Multilingual sentiment analysis. *Expert Systems with Applications*, 168, 114481.
- Rajé, F., Tight, M., & Pope, F. D. (2018). Traffic pollution: A search for solutions for a city like Nairobi. *Cities*, 82, 100–107. <https://doi.org/10.1016/j.cities.2018.05.008>
- Rao, A. M., Madhu, E., & Gupta, K. (2017). Impact of Odd Even Scheme on Transportation Systems in Delhi. *Transportation in Developing Economies*, 3(1), 4. <https://doi.org/10.1007/s40890-017-0035-z>
- Rodríguez-Ibáñez, M. (2023). Deep learning in sentiment analysis. *Expert Systems with Applications*, 213, 119862.
- Schultz, E. S., Litonjua, A. A., & Melén, E. (2017). Effects of Long-Term Exposure to Traffic-Related Air Pollution on Lung Function in Children. *Current Allergy and Asthma Reports*, 17(6), 41. <https://doi.org/10.1007/s11882-017-0709-y>
- Tian, Y., Hu, W., Du, B., Hu, S., Nie, C., & Zhang, C. (2019). IQGA: A route selection method based on quantum genetic algorithm- toward urban traffic management under big data environment. *World Wide Web*, 22(5), 2129–2151. <https://doi.org/10.1007/s11280-018-0594-x>
- Vahedi, S. (2020). Long Term Impact of Childhood Exposure to Pollution on Children's Test Scores. *Journal of Family and Economic Issues*, 41(4), 729–748. <https://doi.org/10.1007/s10834-020-09675-y>
- Wan, X. (2024). Sentiment classification of emergency tweets using BERT. *Information Processing & Management*, 61(1), 103609.
- White, H. D. (2018). Bag of works retrieval: TF*IDF weighting of works co-cited with a seed. *International Journal on Digital Libraries*, 19, 139–149. <https://doi.org/10.1007/s00799-017-0217-7>
- Xie, D., Wang, M., & Zhao, X. (2020). A Spatiotemporal Apriori Approach to Capture Dynamic Associations of Regional Traffic Congestion. *IEEE Access*, 8, 3695–3709. <https://doi.org/10.1109/access.2019.2962619>
- Yao, L. (2019). Spatio-Temporal Differences in Health Effect of Ambient PM2.5 Pollution on Acute Respiratory Infection Between Children and Adults. *IEEE Access*, 7, 25718–25726. <https://doi.org/10.1109/ACCESS.2019.2900539>
- Zhang, Y., Gong, L., & Wang, Y. (2005). An improved TF-IDF approach for text classification. *Journal of Zhejiang University SCIENCE*, 6A(1), 49–55. <https://doi.org/10.1007/BF02842477>