Exploring Public Reaction to Air Quality on Social Media: The 2022 Saharan Dust Episode in Lisbon

Antonino Candeias[1], Rogério Dionísio[1,2] and Fernando Ribeiro[1,2]
[1]Polytechnic Institute of Castelo Branco
[2]CISeD—Research Center in Digital Services
E-mail: fribeiro@ipcb.pt

Keywords: Air quality, social media data, sentiment analysis, environmental dataset, geolocated tweets, saharan dust

Received: February 6, 2026

This study analyses the relationship between air quality and sentiments expressed on social media, focusing on the Saharan dust episode that affected Lisbon in March 2022. Environmental data were collected alongside geolocated Twitter posts (now rebranded as X) from the Lisbon area. The content of these posts was analysed using the Valence Aware Dictionary and sEntiment Reasoner algorithm to identify emotional patterns and assess correlations with fluctuations in atmospheric pollutants. A predictive analysis was also conducted using Decision Tree and Random Forest machine learning algorithms. The findings suggest that despite high levels of inhalable particulate matter (PM10) during the event, there was no significant change in online sentiments, which remained mostly neutral or slightly positive. This implies that the publics environ-mental perception may not be clearly reflected on social media, possibly due to cultural factors or the limited influence of Twitter in Portugal. The study concludes that integrating environmental and social data is feasible and valuable, but further research is needed to explore longer periods, other cities, and more advanced techniques, thus contributing to the understanding of air pollution and public perception in Portugal.

Povzetek: Študija preučuje povezavo med kakovostjo zraka in razpoloženjem na družbenih omrežjih ter ugotavlja, da močno onesnaženje ni bistveno vplivalo na izražena čustva uporabnikov.

1 Introduction

In recent decades, the growing concern about the impacts of atmospheric pollution on public health and the well-being of urban populations has driven the development of new approaches to environmental monitoring [1]. At the same time, the exponential rise in the use of social media, as a means of public expression provides a unique opportunity for the indirect analysis of environmental phenomena through the observation of emotions expressed online [2].

Despite these advances, the intersection between air quality and social media–based sentiment analysis remains relatively underexplored, particularly in the European context [3]. Most previous studies have focused either on traditional air quality monitoring using IoT sensors [4] and satellite data or on social media analyses that do not integrate environmental parameters [5]. Moreover, many works concentrate on regions of high population density mainly in China and where pollution events are more frequent and data availability is higher [6]. For example, research conducted in Beijing and California has shown that peaks in PM2.5 are often associated with increased expressions of negative emotions such as fear, anger, or frustration [7]. However, such studies are rarely extended to European urban areas, where social, cultural, and linguistic factors may influence how individuals perceive and express environmental conditions online.

This gap highlights the need for studies that combine environmental and social data, exploring different regions and new methodologies capable of offering a more integrated perspective on the phenomenon. In this context, the present research seeks to explore the correlation between air quality and the emotions expressed on social media, with a particular focus on the analysis of data collected during March 2022, a period marked by an extreme atmospheric event associated with the passage of dust from the Sahara Desert over the Iberian Peninsula. By integrating environmental and meteorological data with geolocated Twitter posts, this study aims to investigate not only the variation in pollution levels but also how these changes are reflected in the emotions expressed by the population. The study is based on the application of Natural Language Processing (NLP) techniques and sentiment analysis, aiming to identify dominant emotional patterns and assess to what extent these are related to changes in air quality. For this, environmental data from official sources were collected and cross-referenced with geolocated posts from the city of Lisbon. The work also aims to identify limitations and methodological challenges associated with using social media as an environmental inference tool, proposing approaches that contribute to a more representative and reliable analysis. The study of the analytical aspect based on Machine Learning (ML) algorithms could help identify behaviour patterns [8], assess the relative importance of environmental variables, and estimate potential relationships between atmospheric parameters and the sentiments expressed on social media. The integration of this quantitative approach can contribute to a more comprehensive and rigorous understanding of the analysed phenomenon, a topic still underexplored in the Portuguese context.

2 Methodology

The development of this research followed a structured sequence of stages designed to analyse the potential relationship between air quality and emotions expressed on social media. The methodological framework combines environmental monitoring, social media data extraction, and computational text analysis to establish a correlation between environmental parameters and public sentiment.

In the initial stage, a comprehensive literature review was conducted to identify the main methodological approaches used in recent studies related to air quality assessment, sentiment analysis, and environmental perception. This review provided the conceptual foundation for the definition of the research scope and supported the selection of analytical techniques suitable for the study objectives.

Subsequently, data was collected from both official sources and digital platforms. Environmental datasets included atmospheric pollutant concentrations, while meteorological data were obtained from weather stations geographically close to the study area. In parallel, geolocated posts from Twitter were collected for the city of Lisbon, covering the month of March 2022, which coincided with a significant atmospheric event — the passage of Sahara dust over the Iberian Peninsula.

After pre-processing, exploratory analysis and graphical visualization techniques were applied to identify general trends, patterns, and possible anomalies in the data. The sentiment analysis of social media posts was then performed using Natural Language Processing (NLP) tools, allowing the classification of tweets into positive, negative, or neutral categories and the detection of dominant emotional patterns across time. In the final stage, predictive and inferential models were applied to further explore relationships between environmental and emotional variables.

The results obtained were statistically analysed and visually represented through correlation graphs, time-series plots, and performance metrics, enabling a comprehensive interpretation of the observed relationships.

3 State of art

The relationship between social media and air quality has been explored from multiple perspectives in scientific literature, including sentiment analysis, pollution forecasting, and public perception of environmental and health impacts. To position the present study within the existing body of research and to understand how air quality and environmental conditions are reflected in emotions expressed online, this section provides a literature review of the most relevant studies published over the past five years. The aim of this review is to better characterize the context in which previous research was conducted and to examine how variations in air quality correlate with emotions expressed on social media, as well as to identify the predominant emotional responses associated with changes in pollution levels.

The search for articles was conducted in the Scopus and Web of Science (WoS) databases, using a search string constructed from a combination of queries relevant to this study. These queries specifically focused on topics such as: air quality, influence or impact, health, emotions, well-being, and social networks.

("air quality" OR aqi OR pollution) AND (perception OR impact OR influenc) AND (health OR wellbeing OR happiness OR consumer OR sentiment*) AND ("social networks" OR twitter OR facebook OR weibo OR "social media" OR "X Social Media" OR "X Social network")*

The initial search was performed based on the titles, abstracts, and keywords of the articles. When necessary, the structure, introduction, methodology, results, and conclusions were analysed to determine the relevance of the study. A total of 65 articles were found in WoS and 94 in Scopus. After the analysis, 47 articles from WoS and 64 from Scopus were excluded. After combining the results from both searches, 12 duplicate records were removed, resulting in 36 articles to be included in the analysis, Table 1.

Table 1: Summary of the analysed articles.

Article	Input Parameters	Algorithm	Location	Air Quality-Emotion Correlation	Main Limitation
[9]	AQI, PM2.5, Twitter Data	ROST EA (Emotion Analysis), ANN (Artificial Neural Network)	China	✓	Study only on tourists.
[7]	PM2.5, Twitter Data, Weibo, Facebook	LSTM (Long Short-Term Memory), SVM (Support Vector Machine)	Beijing, China	✓	Study conducted in a single location.

[10]	PM2.5, Fire Data, Twitter	LIWC (Linguistic Inquiry and Word Count)	California, USA	✓	Study focused on a specific event.
[1]	AQI, Weibo	Regression, NLP	China	✓	Online trends may influence data.
[11]	PM2.5, Weibo	Louvain Algorithm	China	✓	No sensor data validation.
[12]	AQI, Social Media	LSTM (Deep Learning)	China	✓	Complex models are difficult to interpret.
[13]	PM2.5, Online Opinions, Weibo	Factor Analysis	China	✓	Inconsistencies in social media data.
[14]	AQI, Weibo	Predictive Models	China	✓	Insufficient data for direct causality.
[6]	AQI, Weibo	Sentiment Analysis	China	✓	Use of a single emotional indicator.
[15]	Fire Posts on Twitter	Thematic Analysis	USA	✓	Qualitative analysis can be subjective.
[16]	AQI, Online Searches	Statistical Models	China	✓	Cultural differences may affect results.
[2]	Environmental Data, Twitter, Reddit, YouTube	Pointwise Mutual Information (PMI)	Global	✓	Data does not represent the entire population.
[17]	Twitter	Content Analysis	USA	✓	Lack of longitudinal data to assess long-term changes.
[4]	IoT Data, Air Quality	ML,	Portugal	X	Focus on air quality, not emotions.
[18]	PM2.5, Interviews, Social Media	Qualitative Analysis	Australia	✓	Correlation between interviews exists.
[19]	AQI, Online Searches	Social Network Models	China	✓	Social factors may influence more than air quality.
[20]	AQI, Weibo	Statistical Analysis	China	✓	Only urban data analyzed.
[21]	AQI, Migration Data	Probit Regression Analysis	China	✓	Other economic factors not considered.
[22]	AQI, Social Media	Temporal Analysis Models	Hong Kong	✓	Specific event.
[23]	Social Media	Predictive Models	USA	✓	Difficult to predict long-term effects.
[24]	AQI, Traffic, Twitter	Combined Models	Florida, USA	✓	Focus on interaction between environmental factors.
[25]	Weibo	Inference Models	Shandong Province, China	✓	Data may be influenced by media.
[26]	Social Media Data, Internal Data	Deep Learning, NLP	Global	✓	Focus on indoor health.

[5]	Twitter	NLP	Japan and Italy	✓	External factors may influence the study.
[27]	Social Media, AQI	NLP	Balkans	✓	Regional differences not fully explored.
[28]	AQI, Social Media	Spatial Modeling	Wuhan, China	✓	Data may not capture rural areas.
[29]	AQI, Twitter	Sentiment Analysis	Paris, London, New Delhi	✓	Data representativeness is limited.
[30]	AQI, Mobile Device Data	ML	Singapore	✓	Difficult to differentiate pollution effects from other factors.
[31]	Twitter Data, AQI	NLP	Minneapolis, USA	✓	Only public posts analyzed.
[32]	AQI, Social Media Reviews	Thematic Analysis	China	✓	Errors in review selection may occur.
[33]	AQI, Social Media	Factor Analysis	Muscat, Oman	✓	Demographic differences may influence responses.
[34]	Health Data, Environment, AQI	Static Models	Hamburg, Germany	✕	Complex relationship with other urban factors.
[35]	AQI, Vulnerable Groups Data, Environmental Concerns	Predictive Models	China	✓	Socioeconomic factors may be more influential.
[3]	AQI, Twitter	NLP	Delhi, Kolkata, Mumbai, Hyderabad, India	✓	Regional differences may affect results.
[36]	Fire Data, AQI, Twitter	Sentiment Analysis	Wildfires in California, USA	✓	Specific event.
[37]	Spatial Data and PM2.5	BERT Model for data extraction from Weibo, Geographically Weighted Regression (GWR)	Beijing, China	✕	Difficult to capture individual impacts.

The reviewed literature reveals several important intersections with questions related to air quality and its implications on emotions expressed on social media. The Studies indicate a significant correlation between atmospheric pollution levels and the emotions shared by users on platforms such as Twitter and Weibo [1],[29],[36]. During pollution peaks, there is a prevalence of negative emotions, such as frustration and fear reflecting concerns about environmental impacts and public health. On the other hand, during periods of better air quality, emotions tend to be more neutral or positive [3],[9],[13],[24]. These patterns vary according to cultural and regional factors, with some communities expressing indignation and demanding government action [33],[31], while others seem more resigned to adverse environmental conditions [2]. Emotion analysis also highlights the important role of social mobilization during extreme pollution events, such as wildfires [10], which can lead to greater pressure for more effective public policies [18]. China stands out as the country with the most interest in combining environmental issues with sentiment analysis (18 in 36 studies), reflecting the growing concern over the effects of pollution on public well-being [14], [28].

Nonetheless, using emotions expressed on social media to measure air quality faces significant challenges, such as the influence of external factors (e.g., political crises, sports…) [22] that may distort environmental perception, and the lack of representativeness in data collected from digital platforms [2]. Recent studies have proposed a more sophisticated approaches, such as combining emotional analysis with contextual models that consider variables such as geographic location and weather conditions, have been proposed to improve the accuracy of conclusions [11], [12]. Although existing studies provide valuable insights, many focus on sociocultural contexts distant from Portugal, making direct comparisons difficult. Therefore, the study in this article aims to fill this geographical gap by exploring locally the correlations between air quality and public opinion, offering a more accurate view of the impact of pollution on citizens' emotions.

4 Dataset description

This chapter details the data sources and processing steps that support this study. Two main datasets were created, one with environmental data and the other with social media data. From the integration of these two, a third dataset was generated. The environmental dataset includes air quality and meteorological parameters collected from official monitoring platforms. The social media dataset consists of geolocated Twitter posts published throughout March 2022, used for sentiment analysis. Finally, the merging stage aligns both sources temporally and spatially, allowing for a joint exploration of correlations between atmospheric conditions and the emotions expressed online.

4.1 Environmental dataset

Air quality data were obtained from the official platform of the Portuguese Environment Agency (Agência Portuguesa do Ambiente) through the 2022 annual report for the Entrecampos monitoring station, located in central Lisbon [38]. This station was selected due to its representative urban location. Key parameters included the overall air quality index, PM10, PM2.5, NO_2, among others referred above. The dataset was filtered to retain only data from March 2022, with hourly granularity preserved.

Complementary weather data were sourced from the MeteoManz platform [39], selecting the station geographically closest to Entrecampos to ensure spatial consistency between datasets. The same period (March 2022) and hourly resolution were used. Variables included temperature, relative humidity, wind speed, and others.

The original environmental data files (from the QualAR system) included multiple stations and long-term records (full year of 2022 and earlier) [38]. Therefore, data were filtered to retain only hourly records from March 2022 and specifically from the Entrecampos station. Irrelevant columns were removed, and numeric fields were cleaned and validated to ensure no missing or corrupted entries.

The meteorological data was processed by restructuring the time format and ensuring that each hourly record was correctly assigned. Missing values were identified and treated to guarantee that, for each hour in the dataset, a one-to-one match existed between air quality, meteorological, and future sentiment-related data.

This data cleaning and normalization process ensured that all datasets were temporally aligned and ready for integration in subsequent analysis.

The structure of the environmental dataset is presented in Table 2.

Table 2: Environmental dataset structure.

Variable/Parameter	Description
Date	Date and time of the measurement in the format YYYY-MM-DD HH:MM, with hourly values (24 per day).
Sulfur Dioxide (µg/m³)	Concentration of sulfur dioxide (SO_2) in the air.
Particles < 10 µm (µg/m³)	Concentration of inhalable particles with a diameter smaller than 10 micrometers (PM10). Particularly relevant due to variability caused by Saharan dust.
Particles < 2.5 µm (µg/m³)	Concentration of fine particles (PM2.5), smaller and more harmful to health.
Ozone (µg/m³)	Concentration of tropospheric ozone (O_3).
Nitrogen Dioxide (µg/m³)	Presence of nitrogen dioxide (NO_2), a typical urban traffic pollutant.
Carbon Monoxide (mg/m³)	Concentration of carbon monoxide (CO).
Benzene (µg/m³)	Concentration of benzene, a volatile organic compound with toxic effects.
Temp. (ºC)	Air temperature in degrees Celsius.
Rel. Humidity (%)	Relative air humidity, expressed as a percentage.
Wind Speed (Km/h)	Wind speed in kilometers per hour.
Precip (mm)	Accumulated precipitation in the previous hour, in millimeter’s.
Period	Hourly division used in the study, distinguishing between measurements taken before 18:00 and after 18:00.

4.2 Social media dataset

The sentiment dataset was built using data collected from X (formerly Twitter), focusing on posts published within a 200 km radius of Lisbon throughout March 2022. Data extraction was automated using Microsoft Power Automate [40], which enabled structured and consistent retrieval of tweets directly from the platform's interface.

A geolocated search query was manually opened for each day, formatted to include only public posts (excluding replies), and restricted to the desired location and date range:

geocode:38.7169,-9.1399,200km since:2022-03-31 until:2022-04-01 -filter:replies

The resulting data comprising username, post content, and timestamp were captured from the web page and written into an Excel file using a Power Automate flow. The flow scrolled the page gradually and extracted each visible set of tweets in a loop, with short delays added to ensure proper page rendering, summarizing in pseudocode the data extraction in this web scraping work using Power Automate:

1. Open excel with a blank or existing workbook
 → Store the instance as "ExcelInstance"

2. Open google chrome and navigate to the dynamically generated URL:
 Format:
 'geocode:38.7169,-9.1399,200km since:2022-03-XX until:2022-03-XX - filter:replies'
 Note: Replace "XX" with the desired day of March 2022

3. Set variable NLine = 2 // starting row in excel

4. For loopindex = 1 to 400 (step 1):
 Note: 400 or the times the user thinks it's necessary
 a. Set variable VerticalScroll = 1080
 b. Set variable Time = 3 // seconds
 c. Extract data from the current view of the web page based on the previously defined pattern
 Note: The browser opens and allows the user to manually select the data fields they want to extract, the program then automatically identifies patterns in the selected data to build a data extraction template.
 → Save result as DataFromWebPage
 d. Write datafromwebpage to excel at:
 → Column A, Row = NLine
 e. Increment NLine by 1
 f. Execute JavaScript on the webpage to scroll:
 → window.scrollBy(0, VerticalScroll);
 g. Wait for Time seconds before the next loop
 → End Loop

During this process, common issues included duplicate posts, blank entries (mostly image-only tweets), and reverse chronological ordering (posts listed from 23:59 to 00:00). These issues were addressed in Excel: duplicates and empty entries were removed, and the ordering was corrected to ensure temporal alignment with the air quality and weather datasets.

To prepare the data for sentiment analysis, all tweet texts originally in Portuguese were translated into English. The translated file was saved in CSV format to facilitate integration with Python scripts.

Sentiment classification was performed using Valence Aware Dictionary and sEntiment Reasoner (VADER), part of the NLTK library in Python [41]. VADER is optimized for short, informal text such as social media posts, offering high accuracy without the need for prior training. Its sensitivity to elements like capitalization, punctuation, and emojis made it a suitable choice for this dataset.

The algorithm assigned a sentiment score to each post, which was then labelled as positive, neutral, or negative. These labels were aggregated by day to generate a daily distribution of sentiment types. The results, including the sentiment labels and scores, were stored in a separate xlsx file and appended to the original dataset structure, Table 3.

To cross-check the robustness of the results, a comparison was made with a second model, TextBlob, which uses a different polarity-based scoring method [42]. This helped verify consistency in sentiment classification.

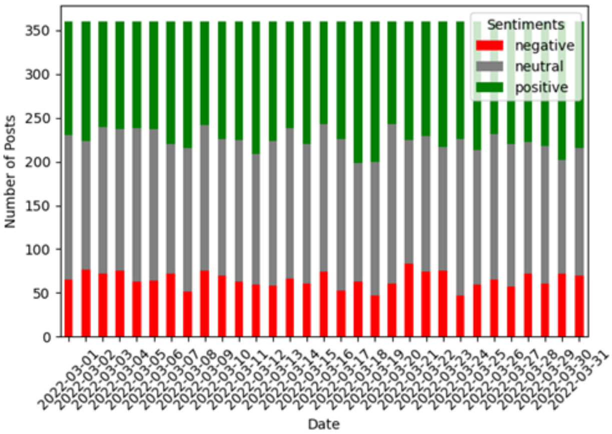

Figure 1: Vader results.

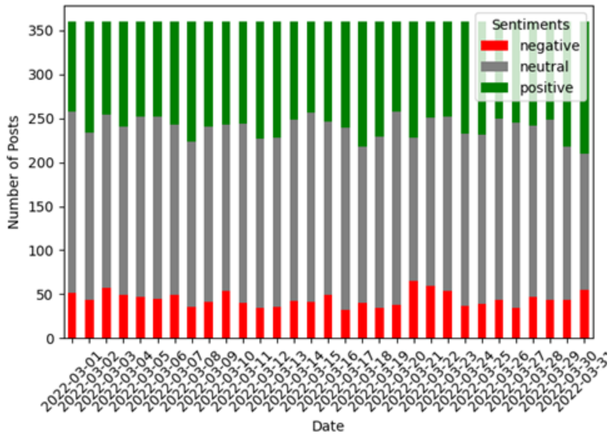

Figure 2: Textblob results.

By comparing the results of both algorithms (Figure 1 and Figure 2), it can be observed that the values are generally similar, although the second algorithm (TextBlob) exhibits a greater tendency toward neutral scores. This consistency supports the reliability of the data. Therefore, the analysis proceeded with the results obtained from the first algorithm, VADER.

Table 3: Social media dataset structure.

Variable/Parameter	Description
Posts	Total number of tweets collected, including the full text of the tweets used for analysis.
Date	Timestamp indicating what day each tweet was posted.
Score	Numerical sentiment score assigned to each tweet, representing its emotional tone.
Sentiment	Qualitative classification of the tweet's polarity: positive, negative, or neutral, based on the assigned score.
Period	Hourly division used in the study, distinguishing between measurements taken before 18:00 and after 18:00.

4.3 Dataset merging

During the planning process for the sentiment analysis, the relevance of conducting a more detailed temporal analysis was also discussed. As a result of this discussion, it was decided to divide the data into two distinct periods: before and after 18:00 each day. This separation was chosen because it roughly corresponds to the end of the typical workday, a time when many people leave work and are more likely to access social media, interact with content, and express their opinions. Additionally, it is common for significant news, event updates, or government measures to be released during the afternoon, which can trigger more pronounced reactions in the post-work period. Therefore, this segmentation allows for the identification of potential variations in the sentiment of posts depending on the time of day.

Since it was not possible to automatically obtain the exact time of all posts, it was necessary to manually identify the post published at 18:00 each day, which served as a reference point. A column was then created, called "period," with two categories: Before 18:00 and After 18:00.

After cross-referencing the environmental data with the sentiment analysis results, a consolidated version of the dataset was built, reflecting all the relevant variables for the study. This final version integrates hourly measurements of air quality and meteorological conditions with the sentiment values extracted from social media posts, as shown in Table 4.

Table 4: Merged dataset structure.

Variable/Parameter	Description
Date	Date in the format YYYY-MM-DD.
Average Sulfur Dioxide (μg/m³)	Average concentration before/after 18:00 of sulfur dioxide (SO2) in the air.
Average Particles < 10 μm (μg/m³)	Average concentration before/after 18:00 of inhaling particles (PM10). A variable of greater relevance due to its variability during the analyzed month.
Average Particles < 2.5 μm (μg/m³)	Average concentration before/after 18:00 of fine particles (PM2.5), which are more harmful to health.
Average Ozone (μg/m³)	Average concentration before/after 18:00 of tropospheric ozone (O3).
Average Nitrogen Dioxide (μg/m³)	Average concentration before/after 18:00 of nitrogen dioxide (NO2), associated with urban traffic.
Average Carbon Monoxide (mg/m³)	Average concentration before/after 18:00 of carbon monoxide (CO).
Average Benzene (μg/m³)	Average concentration before/after 18:00 of C6H6, a toxic volatile compound.
Average Temperature (ºC)	Average air temperature in degrees Celsius before/after 18:00.
Average Relative Humidity (%)	Average Relative air humidity (%) before/after 18:00.
Average Wind Speed (Km/h)	Average wind speed in km/h before/after 18:00.
Average Precipitation (mm)	Average precipitation accumulated in the previous hour (mm) before/after 18:00.
Period	Hourly division used in the study, distinguishing between measurements taken before 18:00 and after 18:00.
Average Score	Sentiment analysis value (VADER) applied to the textual content. Ranges from -1 (negative) to 1 (positive). In

the final dataset, it is the only variable from Twitter.

5 Results

This section presents a joint analysis of sentiment data expressed in online posts and atmospheric pollution levels, with a daily division between before and after 18:00, focusing specifically on the concentration of inhalable particles (PM10), which was most affected and where the effect of dust storms is observed. Additionally, a correlation analysis was conducted involving all environmental variables and the sentiment score, with the goal of evaluating possible associations between these parameters and the sentiment scores.

5.1 Analysis of sentimental score and PM10

The chart, shown in Figure 3, was created to visualize the levels of inhalable particles (PM10) over time, based on the daily average values, first and third quartiles, segmented by periods of the day (before and after 18:00 PM).

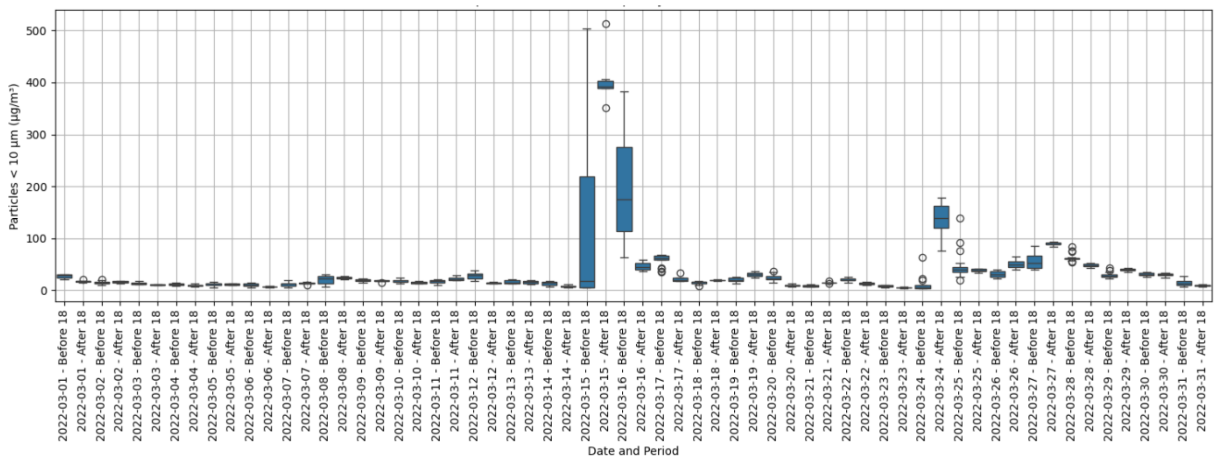

Figure 3: PM10 x date/period.

As shown on Figure 3 on days with high PM10 peaks such as March 15, 16, 24, and 27 the boxplots revealed wider boxes and clear outliers, indicating high dispersion and anomalous pollution episodes. These days often showed the mean diverging from the median, suggesting skewed distributions driven by extremely high values. For example, on March 15 after 18:00, the mean exceeded 400 µg/m³, on that same day and period, the interquartile range was narrow, and all values were well above 350 µg/m³, marking the peak intensity of the event.

In the following chart, Figure 4 the sentiment analysis (sentiment score) of the posts for the same time of year can be seen, divided by date and period of the day (before and after 18:00 PM).

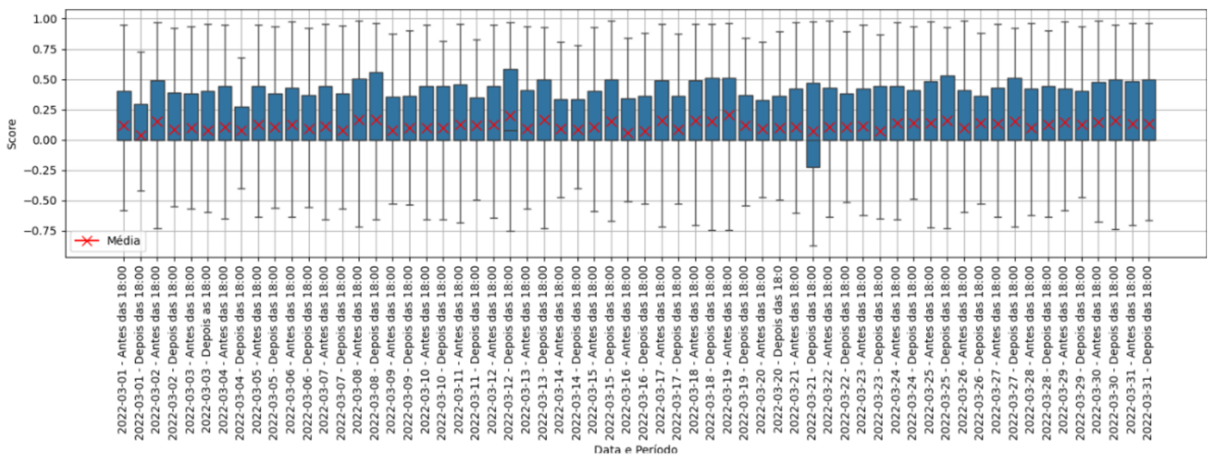

Figure 4: Sentiment score x date.

Throughout March 2022, it was observed that the median sentiment scores remained stable at 0.0 in almost all periods. As mentioned earlier, these values are expected since there are many scores exactly at 0, representing neutral sentiment score, both before and after 18:00 PM. This indicates that, overall, the posts maintained a neutral tone, with no predominance of positive or negative sentiments. However, there is some variation in the higher values, especially in the third quartile, which reached values between 0.4 and 0.5 on many days, suggesting that, although most content was neutral, there was a significant portion of posts with a slightly positive sentiment. The average follows this trend, mainly fluctuating between 0.08 and 0.16, with occasional peaks, such as on March 12th after 18:00, when it reached 0.20, reflecting brief moments of heightened positivity. Another point of interest is March 21st after 18:00, as it is the only day where Q1 is negative, meaning 25% of the values were below 0, which could indicate an event that influenced the sentiment of the population. Despite this, overall, the data suggests emotional consistency in the posts, with slight fluctuations toward positivity on certain days and times.

To specifically investigate the relationship between the sentiment score and PM10 particle levels, the Pearson correlation coefficient [43] was calculated and a correlation graph was built, Figure 5.

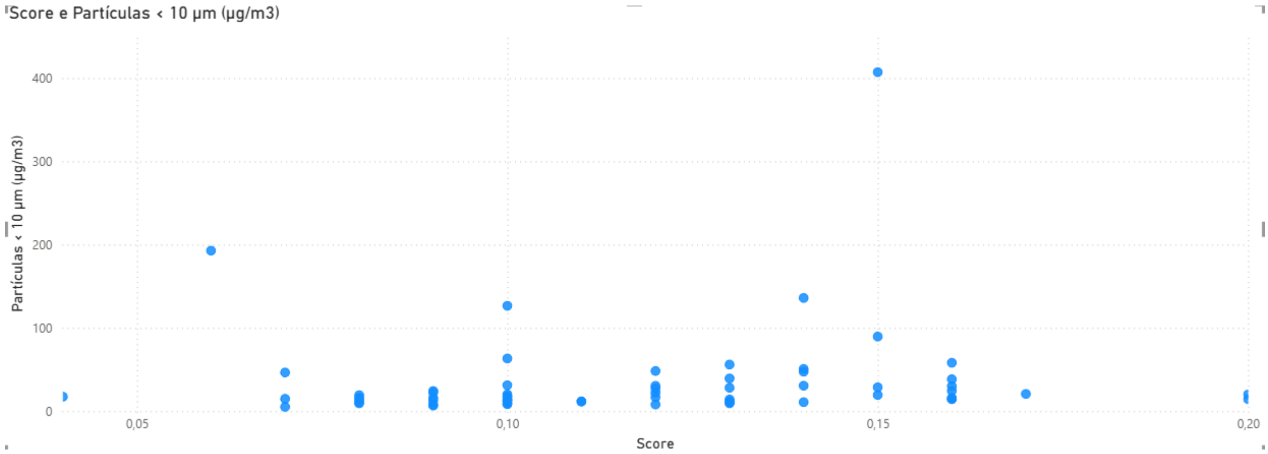

Figure 5: Correlation PM10 x sentiment score.

The calculated correlation coefficient was approximately 0.098, indicating a very weak linear relationship between the variables. It's important to note that the correlation coefficient ranges from -1 to 1, with values close to 0 suggesting a very weak or negligible linear relationship.

Moreover, the obtained p-value was 0.45, much higher than the usual significance level of 0.05, implying that there is insufficient statistical evidence to reject the null hypothesis of no linear correlation. Therefore, based on this analysis, it can be concluded that there is no significant linear relationship between the sentiment score and PM10 levels during the analysed period.

5.2 Machine learning

Two algorithms were selected: Decision Trees (DT) and Random Forest (RF) [8]. These models were chosen due to their suitability for both exploratory and predictive tasks in environmental data. DTs are particularly useful for their direct interpretability, allowing quick identification of the most influential variables affecting air quality. On the other hand, RF, which combines multiple DT, offers greater robustness and accuracy, reducing the risk of overfitting and providing a more stable estimate of variable importance.

By utilizing these models, we can complement previous statistical analyses, identifying not only linear relationships but also uncovering nonlinear patterns and interactions between pollutants and meteorological variables, thus offering a richer understanding of the monitored parameters' behaviour.

5.2.1 Decision tree

Initially, a DT was built using all available variables, employing the DecisionTreeRegressor algorithm from the Scikit-learn library (Python). The model was trained with default parameters: the splitting criterion used was Mean Squared Error (MSE), the tree depth was unlimited (max_depth=None), and the minimum number of samples required to split a node was set to 2 (min_samples_split=2). Additionally, random_state=42 was used to ensure the reproducibility of the results. The resulting model was found to be large and complex, making visual interpretation challenging. Therefore, an analysis of variable importance was conducted, identifying the following (Figure 6) as the most influential: carbon monoxide (CO), temperature, fine particles (PM2.5), and relative humidity.

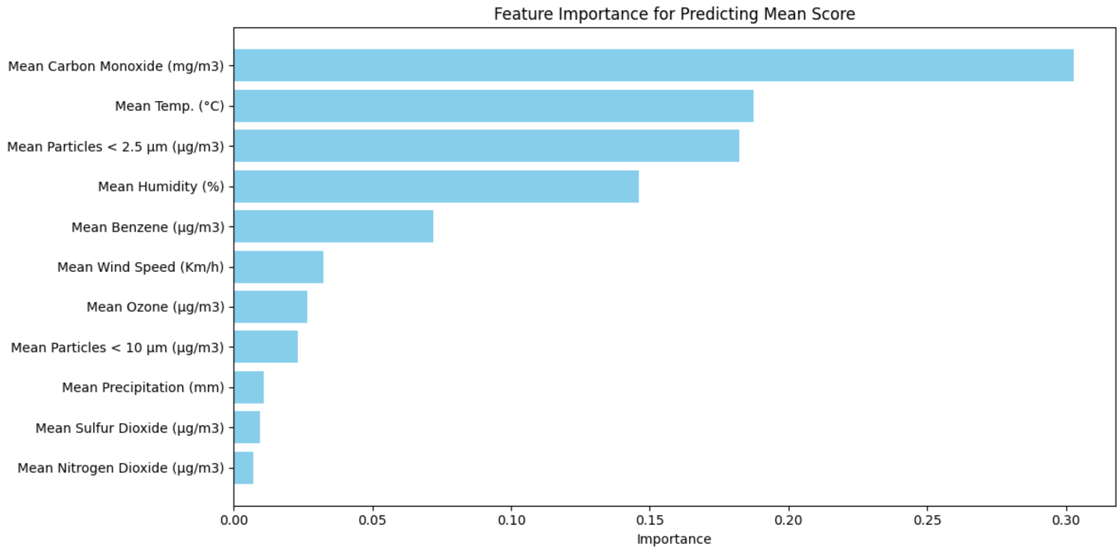

Figure 6: Order of variable importance in the decision tree.

Based on this selection, Figure 7 shows a simplified DT constructed using only these four most relevant variables. The reduced tree has a more compact structure and is easier to interpret, facilitating the understanding of the impact of the main parameters on the prediction.

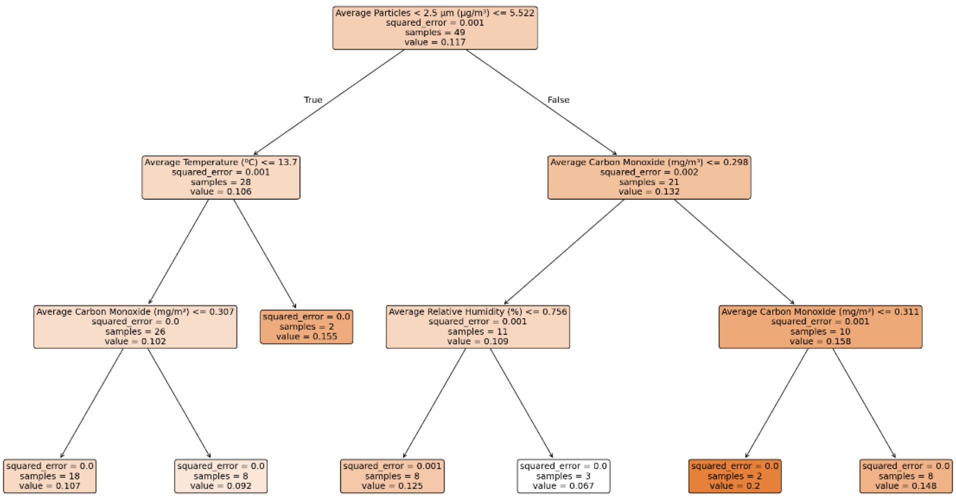

Figure 7: Decision tree.

The DT, presented in Figure 7, performs a regression to predict the target variable related to people's emotions based on four key environmental variables: Average Fine Particles (PM2.5 < 2.5 µm), Average Temperature (°C), Average Carbon Monoxide (CO, mg/m³), and Relative Humidity (%).

In the root node, the tree splits the data using a PM2.5 concentration threshold of 5.522 µg/m³. This suggests that fine particle pollution plays a central role in the initial separation of samples, aligning with scientific evidence of pollution's impact on both physical and emotional health. For cases with PM2.5 ≤ 5.522 µg/m³, the tree evaluates average temperature with a cutoff at 13.7°C. This division implies that lower thermal conditions may influence pollutant dispersion, thus affecting well-being. Next, the model uses CO levels to refine the prediction, showing its direct importance in modelling sentiment. For groups with lower pollution and temperature, the predicted target value is 0.102, which corresponds to a slightly positive or neutral sentiment on a scale from -1 (very negative) to 1 (very positive).

On the other side of the tree, where PM2.5 > 5.522 µg/m³, the model continues to use CO as the main splitting criterion (threshold 0.298 mg/m³). Even with a relatively low CO value (0.2), it suggests a tendency toward slightly more positive or neutral sentiments, but not extreme positivity. Relative humidity is then used to further split the data, highlighting the relevance of climate variables in highly polluted conditions.

Despite the logical structure of the tree, the global model metrics indicate poor predictive performance: the MSE is 0.00295, indicating a small average error, but the R^2 is negative (-2.10), suggesting the model explains less data variability than a simple mean. This means the tree is good at distinguishing groups with distinct and consistent behaviors but struggles to generalize well across the full dataset. With only 62 data points for training, the tree likely overfits the specific details of these samples, making it less reliable for new, unseen data.

Although the model has limitations for precise predictions, it serves as an effective exploratory tool, helping to identify and understand the potential relationships between air pollution and people's emotional states. This study also paves the way for future development and testing of new algorithms that could enhance predictive accuracy.

5.2.2 Random forest

To further deepen the analysis conducted with the DT, the RF model was used. This technique combines multiple DT to enhance the robustness and accuracy of predictions, as well as to provide a more consistent view of the variable importance in the problem under study.

The RF model was applied to predict people's sentiments based on the environmental variables studied. Figure 8 presents the variables' importance in the RF model, listed in order.

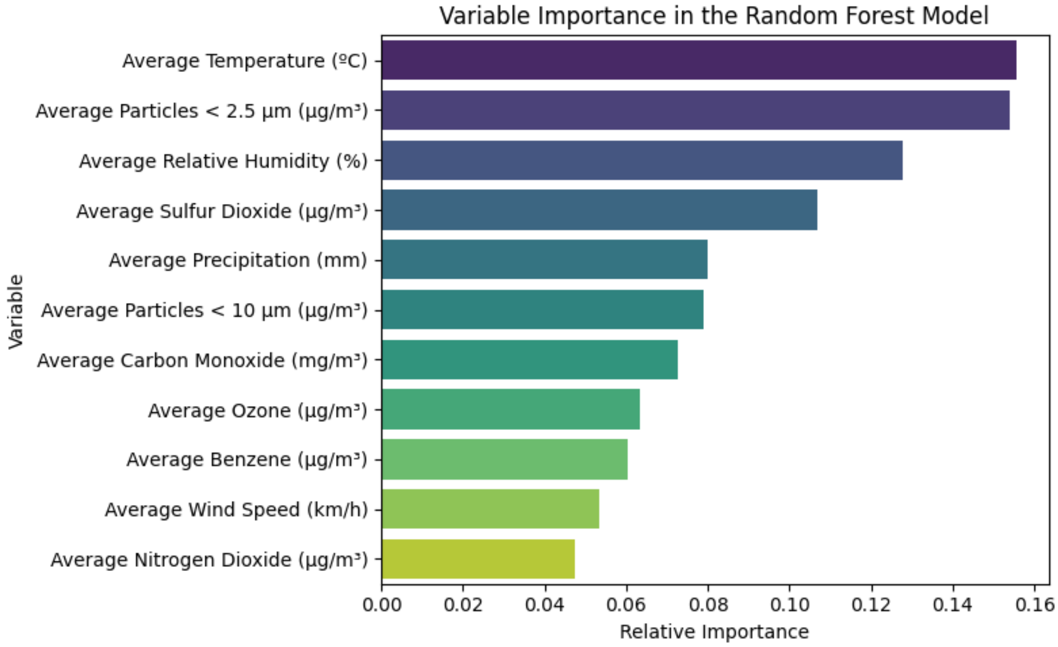

Figure 8: Order of feature importance in the random forest model.

When compared to the DT, is shown that the four most important variables for the model are now Average Temperature (°C), Average Fine Particles (PM2.5 < 2.5 µm), Average Relative Humidity (%), and Average Sulfur Dioxide (SO_2, µg/m³). Although the order is different, the importance rankings are fairly similar, apart from Carbon Monoxide (CO), which was the most important variable in the DT but is ranked only seventh in the RF model. This difference can be attributed to the distinct functioning of the two algorithms: the DT is based on a single tree and is highly sensitive to the initial splits in the data, which may overly emphasize a variable that happens to create a good separation in the training set. In contrast, RF builds multiple trees with random subsets of data and features, calculating the average importance of each feature across all trees. Thus, variables that are consistent across multiple trees, such as temperature or PM2.5, tend to gain more importance. If CO is highly correlated with other variables or is only useful in very specific contexts, its average importance in RF will be reduced, explaining its lower ranking in this more robust model.

Regarding performance, the RF model achieved a Mean Squared Error (MSE) of approximately 0.00109, lower than the DT's MSE, indicating a better fit to the data. However, the Coefficient of Determination (R^2) remained negative at around -0.14, an improvement compared to the DT but still indicating that the model does not explain the variability in sentiment better than the simple mean.

One important advantage of RF is its ability to generate individual predictions for each sample, facilitating a direct comparison between real and predicted values. This granularity helps to better understand the model's behaviour and identify its strengths and limitations in its predictions.

Finally, as indicated by the MSE and R^2, despite the improvement in fit, the RF model still suggests that the environmental variables analysed explain only a limited portion of the variation in sentiment. Nevertheless, the model fulfills its exploratory role well by identifying and validating the relative importance of environmental factors in modelling sentiment and serves as a basis for future research aiming to deepen this relationship and develop higher-performing algorithms.

5.2.3 Classification algorithm

In addition to the regression analysis, a classification approach was also explored using a DT to classify sentiments into “Low,” “Medium,” and “High” categories. For this purpose, the continuous sentiment variable, originally ranging from -1 to 1, was divided into three categories, considering that the observed values were concentrated approximately between 0.04 and 0.2. Thus, scores lower than 0.09 were defined as belonging to the “Low” class (0), scores between 0.09 and 0.14 to the “Medium” class (1), and scores equal to or higher than 0.14 to the “High” class (2). This categorization allows for a clearer interpretation of the results and enables the use of classification methods to analyze the relationship between air quality and emotional state.

The classification report, Table 5 indicates that the model shows moderate performance, with higher precision, recall, and F1-score for the “High” class (F1-score 0.67), while struggling to correctly predict the “Low” class (F1-score 0.00). The overall accuracy reached 46%, showing that, despite its limitations, the model captures some differences among the classes.

Table 5: Classification report results.

Class	Precision	Recall	F1-score
Low	0.00	0.00	0.00
Medium	0.33	0.40	0.36
High	0.57	0.80	0.67
Accuracy			0.46

The data were also represented in a confusion matrix, Figure 9.

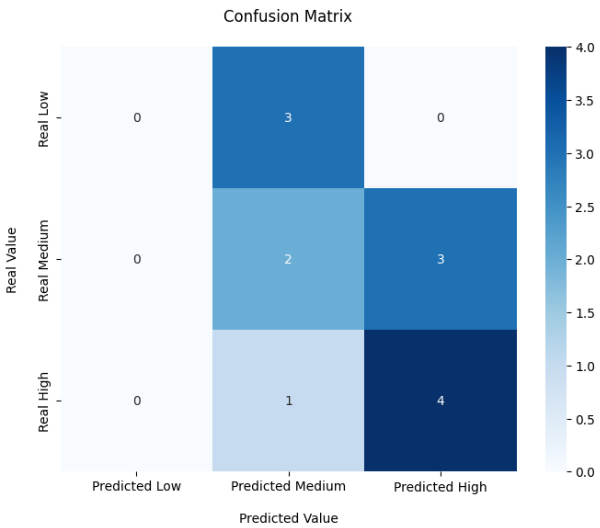

Figure 9: Confusion matrix.

Each row of the confusion matrix corresponds to the actual classes, while each column represents the predicted classes by the model. For the "Low" class (row 1), no samples were correctly classified, with 3 misclassified as “Medium” and none as “High”. For the "Medium" class (row 2), 2 samples were correctly identified as “Medium”, while 3 were incorrectly predicted as “High”. In the "High" class (row 3), 4 out of 5 samples were correctly classified, with 1 misclassified as “Medium”.

These results indicate that the model struggled to correctly classify the "Low" class, grouping all examples as "Medium". The "Medium" class was partially identified, but some were confused with the "High" class. The "High" class performed best, with a higher accuracy in classification. Overall, the model has difficulty distinguishing between adjacent classes, especially the lower ones.

The classification analysis using the DT method revealed limitations, particularly with the "Low" and "Medium" classes. Given this, we chose not to apply the classification approach to the RF model. The sentiment data is continuous, concentrated in a narrow range (0.04 to 0.2), making discrete classification challenging. Additionally, RF typically requires more data to generalize effectively, and using fewer classes could result in limited performance. Therefore, regression remains the more suitable method, as it better captures the continuous nature of sentiment values.

6 Discussion

The analysis conducted in this study provided insights into the relationship between a specific air quality degradation event, during the passage of Saharan dust over Portugal in March 2022, and the public perception expressed on social media. The applied methodology integrated environmental and meteorological data, collected from the Qualar and Meteomanz platforms, with social data from Twitter, processed through Power Automate. This integration allowed for cross-referencing measurable physical phenomena (such as particulate and gas concentrations) with social manifestations (sentiments expressed in text).

The results confirm that the event had a significant environmental impact. PM10 levels reached very high values (exceeding 400 µg/m³ on March 15, after 18:00), far above usual limits and clearly identified as episodes of poor air quality. These values align with what other studies on Saharan dust in Portugal and the Iberian Peninsula have documented, reinforcing that the phenomenon has direct and measurable atmospheric consequences.

However, contrary to what some previous studies have observed in prolonged urban pollution contexts, no strong correlation was found between the worsening air quality and the negativity expressed on social media in this case. The Pearson correlation coefficient between the two variables (sentiment score and PM10) was approximately 0.1, indicating a weak and statistically insignificant relationship. Nonetheless, some interesting specific cases were observed, such as on March 15 after 18:00 PM10 levels were extremely high, but the sentiments expressed online remained virtually unchanged. This contrast suggests that the physical effects of the event were much more evident in the environmental data than in the social data, raising the question of how much social media truly reflects public perception during extraordinary environmental events.

When compared with related studies, relevant differences emerge. International studies that explored the relationship between air pollution and social media found, in some cases, moderate to strong correlations, especially in countries where the population is more environmentally aware, such as China. In these contexts, increases in pollution were often associated with a greater number of negative posts, frequently accompanied by hashtags related to health or the environment. In contrast, in the Portuguese case analyzed, the online response was more muted, suggesting that cultural, social factors, or even the relatively low Twitter usage in Portugal may have influenced the results. This contribution shows that similar methodologies applied in different contexts can yield very different patterns, requiring adaptations.

Finally, it is important to highlight some limitations of the present study. First, the dataset of posts was relatively restricted, focusing solely on March 2022 and a geographically defined area. Studies that used longer time series were able to capture seasonal variations and broader trends. Second, the way Portuguese Twitter users express themselves may not directly reflect environmental concerns: while the phenomenon was visible in everyday life (orange skies, accumulated dust), it was not necessarily associated with health or pollution impacts in online posts. Additionally, more advanced tools, such as dedicated APIs or large-scale text mining models, could allow for broader and more diverse data collection.

7 Conclusions and future work

This study achieved its primary objective: to integrate environmental and social data to analyze air quality in an urban context, focusing on the specific episode of Saharan dust arrival that affected Lisbon in March 2022. The research clearly confirmed the atmospheric impact of the phenomenon, with PM10 levels far exceeding reference values and visible effects in the daily life of the city. By cross-referencing these data with Twitter posts, it was observed that the sentiments expressed online remained largely neutral or slightly positive, even during the times of highest particulate concentration. This lack of significant correlation shows that intense environmental phenomena do not necessarily translate into reactions on social media, particularly in the Portuguese context, where cultural and social factors may have conditioned public expression. Furthermore, the results suggest that other themes of a political, social, or even sports-related nature may exert greater influence over the sentiments expressed online than environmental issues, relegating them to a secondary position in digital attention.

One of the key lessons from this study is the evidence of a disconnect between physical reality and digital reality in our culture. While sensors recorded severe environmental impact, the online community showed limited responsiveness. This finding reinforces the need to improve the adopted methodology but also highlights its potential and how it can be expanded in broader contexts of time, space, and platform diversity. Overall, this work demonstrates that understanding the relationship between the environment and public perception is a complex challenge, but also an opportunity to build more comprehensive perspectives by crossing approaches from various academic fields on environmental issues, thereby strengthening the utility of the methodology and its potential for expanding analysis.

7.1 Future work

Although the integration of environmental and social media data enabled an exploration of air quality perception during the Saharan dust event, several opportunities for further research have been identified.

First, future work should consider expanding the temporal scope beyond a single-month event. Longer and multi-season analyses would facilitate the study of behavioral trends over time, including the potential relation between pollution exposure and public reaction.

A second direction for future research is to enhance data collection methods. Using official social media APIs or large-scale text mining models would enable a more diverse and comprehensive data collection. This would allow the inclusion of additional features such as hashtags, emojis, and images, which could provide a deeper understanding of how environmental awareness is expressed in social media.

Third, from a methodological perspective, future studies should explore more NLP and ML techniques that can capture nuances beyond simple sentiment polarity, offering a more detailed and accurate interpretation of public reactions.

7.2 Limitations and threats to validity

The dataset was relatively restricted, focusing only on March 2022 and a specific geographic area, which limited the ability to capture broader seasonal trends. Additionally, the way Portuguese Twitter users express themselves may not directly reflect environmental concerns, as posts often did not explicitly link visible phenomena, such as orange skies or dust accumulation, to health or pollution. Regarding the ML models, both the DT and RF showed limitations in predictive performance, struggling with generalization due to overfitting and limited training data. While they identified key environmental factors, their ability to explain sentiment variation remained limited, suggesting that other factors beyond those considered in this study may also play a significant role.

References

[1] B. Ye, P. Krishnan, and S. Jia, “Public Concern about Air Pollution and Related Health Outcomes on Social Media in China: An Analysis of Data from Sina Weibo (Chinese Twitter) and Air Monitoring Stations,” *Int. J. Environ. Res. Public Health*, vol. 19, no. 23, Dec. 2022, doi: 10.3390/IJERPH192316115.

[2] D. Amangeldi, A. Usmanova, and P. Shamoi, “Understanding Environmental Posts: Sentiment and Emotion Analysis of Social Media Data,” *IEEE Access*, vol. 12, pp. 33504–33523, Jan. 2024, doi: 10.1109/ACCESS.2024.3371585.

[3] K. Wadhwa, D. Mehra, H. Gosain, and A. K. Haritash, “Sentiment Analysis of Air Quality Perception in Major Metro Cities of India,” *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, p. Delhi, doi: 10.1109/ICCCNT56998.2023.10307448.

[4] G. Marques, I. M. Pires, N. Miranda, and R. Pitarma, “Air quality monitoring using assistive robots for ambient assisted living and enhanced living environments through internet of things,” *Electronics (Switzerland)*, vol. 8, no. 12, p. 1375, Dec. 2019, doi: 10.3390/electronics8121375.

[5] T. Carpi, A. Hino, S. M. Iacus, and G. Porro, “The Impact of COVID-19 on Subjective Well-Being: Evidence from Twitter Data,” *Journal of Data Science*, vol. 21, no. 4, pp. 761–780, Oct. 2023, doi: 10.6339/22-JDS1066.

[6] S. Zheng, J. Wang, C. Sun, X. Zhang, and M. E. Kahn, “Air pollution lowers Chinese urbanites’ expressed happiness on social media,” *Nat. Hum. Behav.*, vol. 3, no. 3, pp. 237–243, Mar. 2019, doi: 10.1038/s41562-018-0521-2.

[7] S. Shan, X. Ju, Y. Wei, and Z. Wang, “Effects of pm2.5 on people’s emotion: A case study of weibo (chinese twitter) in beijing,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 10, May 2021, doi: 10.3390/IJERPH18105422.

[8] A. Liaw and M. Wiener, “Classification and Regression by randomForest,” vol. 2, no. 3, 2002, Accessed: Nov. 05, 2025. [Online]. Available: http://www.stat.berkeley.edu/

[9] Y. Tao, F. Zhang, C. Shi, and Y. Chen, “Social media data-based sentiment analysis of tourists’ air quality perceptions,” *Sustainability (Switzerland)*, vol. 11, no. 18, p. 5070, Sep. 2019, doi: 10.3390/su11185070.

[10] Y. E. García *et al.*, “Wildfires and social media discourse: exploring mental health and emotional wellbeing through Twitter,” *Front. Public Health*, vol. 12, 2024, doi: 10.3389/FPUBH.2024.1349609.

[11] B. Wang, N. Wang, and Z. Chen, “Research on air quality forecast based on web text sentiment analysis,” *Ecol. Inform.*, vol. 64, Sep. 2021, doi: 10.1016/J.ECOINF.2021.101354.

[12] W. Zhai and C. Cheng, “A long short-term memory approach to predicting air quality based on social media data,” *Atmos. Environ.*, vol. 237, Sep. 2020, doi: 10.1016/J.ATMOSENV.2020.117411.

[13] J. Min, H. Yunxiu, S. Yong, J. Fengxiang, and L. Ting, “Research on Analysis of Evaluation Influence Factors of Air Quality Opinions Sentiment Value and Quantification Method,” *Water Air Soil Pollut.*, vol. 232, no. 7, p. 256, Jul. 2021, doi: 10.1007/s11270-021-05197-x.

[14] H. Ji *et al.*, “Research on adaption to air pollution in Chinese cities: Evidence from social media-based health sensing,” *Environ. Res.*, vol. 210, p. 112762, Jul. 2022, doi: 10.1016/j.envres.2022.112762.

[15] J. M. Postma *et al.*, “Assessing community response to wildfire smoke: A multimethod study using social media,” *Public Health Nurs.*, vol. 40, no. 1, pp. 153–162, Feb. 2023, doi: 10.1111/phn.13140.

[16] G. Yang, Y. Ju, and W. Ni, “Does the air pollution level information matter in public perception? Insights from China,” *J. Environ. Manage.*, vol. 349, p. 119582, Jan. 2024, doi: 10.1016/J.JENVMAN.2023.119582.

[17] C. E. Slavik, D. A. Chapman, A. S. Cohen, N. Bendefaa, and E. Peters, “Clearing the air: evaluating institutions’ social media health messaging on wildfire and smoke risks in the US Pacific Northwest,” *BMC Public Health*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12889-024-17907-1.

[18] M. T. Marfori *et al.*, “Public Health Messaging During Extreme Smoke Events: Are We Hitting the Mark?,” *Front. Public Health*, vol. 8, Sep. 2020, doi: 10.3389/FPUBH.2020.00465.

[19] W. Hong, Y. Wei, and S. Wang, “Left behind in perception of air pollution? A hidden form of spatial injustice in China,” *Environment and Planning C: Politics and Space*, vol. 40, no. 3, pp. 666–684, May 2022, doi: 10.1177/23996544211036145/ASSET/FDB83BA2-7C31-4935-88EF-99653086DD9C/ASSETS/IMAGES/LARGE/10.1177_23996544211036145-FIG1.JPG.

[20] L. Yan, F. Duarte, De Wang, S. Zheng, and C. Ratti, “Exploring the effect of air pollution on social activity in China using geotagged social media check-in data,” *Cities*, vol. 91, pp. 116–125, Aug. 2019, doi: 10.1016/J.CITIES.2018.11.011.

[21] C. Zhang and G. Zhang, “How Does Air Pollution Impact Residence Intention of Rural Migrants? Empirical Evidence from the CMDS,” *Sustainability 2024, Vol. 16, Page 5784*, vol. 16, no. 13, p. 5784, Jul. 2024, doi: 10.3390/SU16135784.

[22] P. Brimblecombe and Y. Lai, “Effect of fireworks, Chinese new year and the COVID-19 lockdown on air pollution and public attitudes,” *Aerosol Air Qual. Res.*, vol. 20, no. 11, pp. 2318–2331, Nov. 2020, doi: 10.4209/AAQR.2020.06.0299.

[23] M. Burke *et al.*, “Exposures and behavioural responses to wildfire smoke,” *Nat. Hum. Behav.*, vol. 6, no. 10, pp. 1351–1361, Oct. 2022, doi: 10.1038/S41562-022-01396-6.

[24] H. O’Leary, S. Parr, and M. M. H. El-Sayed, “The breathing human infrastructure: Integrating air quality, traffic, and social media indicators,”

Science of the Total Environment, vol. 827, Jun. 2022, doi: 10.1016/J.SCITOTENV.2022.154209.

[25] Y. Sun, F. Jin, Y. Zheng, M. Ji, and H. Wang, "A new indicator to assess public perception of air pollution based on complaint data," *Applied Sciences (Switzerland)*, vol. 11, no. 4, pp. 1–18, Feb. 2021, doi: 10.3390/APP11041894.

[26] M. Ashayeri, "Decoding Global Indoor Health Perception on Social Media Through NLP and Transformer Deep Learning," in *Artificial Intelligence in Performance-Driven Design: Theories, Methods, and Tools*, wiley, 2024, pp. 159–185. doi: 10.1002/9781394172092.ch8.

[27] A. Madjar, I. Gjorshoska, J. Prodanova, A. Dedinec, and L. Kocarev, "Western Balkan societies' awareness of air pollution. Estimations using natural language processing techniques," *Ecol. Inform.*, vol. 75, p. 102097, Jul. 2023, doi: 10.1016/j.ecoinf.2023.102097.

[28] R. Yu, C. Zeng, M. Chang, C. Bao, M. Tang, and F. Xiong, "Effects of Urban Vibrancy on an Urban Eco-Environment: Case Study on Wuhan City," *Int. J. Environ. Res. Public Health*, vol. 19, no. 6, p. 3200, Mar. 2022, doi: 10.3390/ijerph19063200.

[29] S. Gurajala, S. Dhaniyala, and J. N. Matthews, "Understanding Public Response to Air Quality Using Tweet Analysis," *Social Media and Society*, vol. 5, no. 3, Jul. 2019, doi: 10.1177/2056305119867656.

[30] Y. Yang, K. Y. Goh, H. H. Teo, and S. S. L. Tan, "The Impact of Air Pollution Information on Individuals' Exercise Behavior: Empirical Study Using Wearable and Mobile Devices Data," *JMIR Mhealth Uhealth*, vol. 12, p. e55207, Jan. 2024, doi: 10.2196/55207.

[31] H. O'Leary, D. Smiles, S. Parr, and M. M. H. El-Sayed, "'I Can't Breathe:' The Invisible Slow Violence of Breathing Politics in Minneapolis," *Soc. Nat. Resour.*, vol. 36, no. 9, pp. 1098–1118, 2023, doi: 10.1080/08941920.2023.2194068.

[32] Y. Tao, W. Liu, Z. Huang, and C. Shi, "Thematic analysis of reviews on the air quality of tourist destinations from a sentiment analysis perspective," *Tour. Manag. Perspect.*, vol. 42, Apr. 2022, doi: 10.1016/J.TMP.2022.100969.

[33] H. K. Al-Shidi, A. K. Ambusaidi, and H. Sulaiman, "Public awareness, perceptions and attitudes on air pollution and its health effects in Muscat, Oman," *J. Air Waste Manage. Assoc.*, vol. 71, no. 9, pp. 1159–1174, Jan. 2021, doi: 10.1080/10962247.2021.1930287.

[34] M. von Szombathely, B. Bechtel, B. Lemke, J. Oßenbrügge, T. Pohl, and M. Pott, "Empirical evidences for urban influences on public health in Hamburg," *Applied Sciences (Switzerland)*, vol. 9, no. 11, p. 2303, Jun. 2019, doi: 10.3390/app9112303.

[35] Z. H. Wang *et al.*, "Environmentally vulnerable or sensitive groups exhibiting varying concerns toward air pollution can drive government response to improve air quality," *iScience*, vol. 25, no. 6, p. 104460, Jun. 2022, doi: 10.1016/j.isci.2022.104460.

[36] M. L. Loureiro, M. Alló, and P. Coello, "Hot in Twitter: Assessing the emotional impacts of wildfires with sentiment analysis," *Ecological Economics*, vol. 200, p. 107502, Oct. 2022, doi: 10.1016/j.ecolecon.2022.107502.

[37] Y. Zhu *et al.*, "Quantifying Spatiotemporal Heterogeneities in PM2.5-Related Health and Associated Determinants Using Geospatial Big Data: A Case Study in Beijing," *Remote Sens. (Basel).*, vol. 14, no. 16, p. 4012, Aug. 2022, doi: 10.3390/rs14164012.

[38] "QualAr." Accessed: Jun. 18, 2025. [Online]. Available: https://qualar.apambiente.pt/

[39] "Datos meteorológicos de SYNOPS/BUFR - Predicciones GFS/ECMWF - Meteomanz.com." Accessed: Jun. 18, 2025. [Online]. Available: http://www.meteomanz.com/

[40] "Official Microsoft Power Automate documentation - Power Automate | Microsoft Learn." Accessed: Jul. 02, 2025. [Online]. Available: https://learn.microsoft.com/en-us/power-automate/

[41] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225, May 2014, doi: 10.1609/ICWSM.V8I1.14550.

[42] "TextBlob: Simplified Text Processing — TextBlob 0.19.0 documentation." Accessed: Aug. 12, 2025. [Online]. Available: https://textblob.readthedocs.io/en/dev/

[43] J. J. Berman, "Understanding Your Data," *Data Simplification*, pp. 135–187, 2016, doi: 10.1016/B978-0-12-803781-2.00004-7.