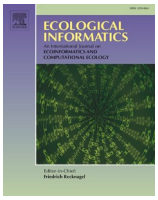
Ecological Informatics 75 (2023) 102097

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Western Balkan societies' awareness of air pollution. Estimations using natural language processing techniques

Angela Madjar [a], Ivana Gjorshoska [a], Jana Prodanova [b,c], Aleksandra Dedinec [a,b,*], Ljupco Kocarev [a,b]

[a] Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia
[b] Macedonian Academy of Sciences and Arts, Skopje, Macedonia
[c] Faculty of Economic and Business Sciences, University of Burgos, Burgos, Spain

ARTICLE INFO

Keywords:
Air pollution
Western Balkans
Twitter
Sentiment analysis
Topic modelling
Cross-correlation

ABSTRACT

Air pollution remains a severe concern in European countries, especially in Western Balkan, where the air monitoring data point to harmful ambient pollution. The public concern with this issue becomes particularly critical during the fall and winter months, when the contamination is more visible, provoking a series of reactions directed principally to the government authorities as the responsible entities for regulating air pollution levels. Since citizen-contributed data are generally considered valuable additional information for assessing the impacts of air pollution, the public contribution could act as a tool for increasing awareness and response about air pollution. Consequently, this study's objective focuses on researching public awareness of air pollution in Western Balkan. The study assumes that citizens' reactions will grow more intensely during the months with an increase in air pollution levels, principally due to winter heating. Therefore, Twitter activity and news articles related to air pollution have been investigated for the case of Macedonia, Serbia, Bosnia and Herzegovina and Montenegro, from November 2021 to March 2022. Natural Language Processing techniques such as sentiment analysis, topic modelling, and cross-correlations statistical analysis were employed to determine the relationship between Twitter discussions and news with actual PM_{10} levels measured by official air monitoring stations. The aim was to observe whether tweets and news teasers reflect the realistic air pollution situation. The results affirm that social media discussions, mainly with a negative connotation, can serve as a measure of public awareness of temporal changes in the PM_{10} concentration in the air and the negative consequences. The content of the resources reveals several topics of concern, contributing to better identification of public opinion and possibilities for tracking news trends. Nevertheless, attention should be paid to news interpretation, provided that sometimes they might offer a more neutral understanding of the situation, failing, in this way, to present the actual air conditions and possibly impacting society in forming an unrealistic opinion. Additionally, the public might not be able to obtain sufficient or accurate information about the primary sources of air pollution, emphasizing the need for more transparent communication and greater education regarding air pollution monitoring. Finally, the study provides deeper insights into the content of the data and helps detect the reasons for skepticism towards pro-environmental behavior occurring in social media discussions. Explicitly, personal disappointment with the air quality should be taken as an inflection point by responsible parties to intervene in improving citizens' quality of life.

1. Introduction

Air pollution is a severe environmental issue and a leading risk factor for health problems and mortality globally (The Lancet, 2020). Developing countries are most at risk, but highly developed nations are also affected by the adverse effects of ambient air pollution (European Environment Agency, 2022a). In the previous two years, 2020 and 2021, 96% of the urban population of the European Union (EU) was exposed to harmful air pollutants with levels surpassing the limits set by the World Health Organization (WHO), where Italy and Central-eastern European

* Corresponding author at: Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Macedonia.
E-mail addresses: angela.madzhar@students.finki.ukim.mk (A. Madjar), ivana.gjorshoska@students.finki.ukim.mk (I. Gjorshoska), jprodanova@manu.edu.mk (J. Prodanova), aleksandra.kanevche@finki.ukim.mk (A. Dedinec), lkocarev@manu.edu.mk (L. Kocarev).

https://doi.org/10.1016/j.ecoinf.2023.102097
Received 27 September 2022; Received in revised form 20 February 2023; Accepted 2 April 2023
Available online 6 April 2023

countries reported the highest concentrations of particulate matter (PM) levels, indicating the need to recognize that communities are still at threat and the urge to promote initiatives and adverse exposure control approaches (European Environment Agency, 2022b).

Recent reports indicate that the Republic of Macedonia is classified among the most polluted areas in the region and Europe (European Environment Agency, 2020). The average daily PM_{10} concentration in Skopje is above the WHO recommendations and EU standards of 50 μg/m3 per day for around half of the year, a staggering fact given the limit values of the concentrations, recommended not to be exceeded for >35 days a year. The air quality problem is not restricted only to urban centers in the Western Balkans, as the air pollution in several urban centers in the EU often exceeds WHO and EU guidelines (Fig. 1.).

PM_{10} is among the air pollutants whose levels are most frequently above the legislation limits in this region and are mainly emitted by human activities such as industry, household heating and transport (Banja et al., 2020; Meisner et al., 2015). Uncontrolled urbanization, illegal construction, and poor planning and design also endanger the environment by contributing to such ambient conditions (Jovanovic, 2019). Additionally, transboundary pollution from within and outside the region contributes significantly to the observed concentrations (Banja et al., 2020). However, the Health and Environment Alliance (HEAL) has reported that the primary sources of air pollution in the Western Balkans are the production of electricity from coal and using wood for heating households (Todorović, 2022). It is in line with the indication of the daily PM_{10} exceeding the limit between 120 and 180 days a year, mostly during wintertime (Colovic Daul et al., 2019). Especially during the months between November and March every year, the air quality monitoring data, particularly PM_{10}, show a peak in air pollution in the Western Balkan, causing increased concern and disappointment in its inhabitants.

The activities of individuals significantly contribute to the air pollution situation, considering, for example, that household heating is one of the main sources of pollution. Therefore, an evaluation of the general level of awareness is of great importance. To measure the individuals' interest, awareness and reaction to this situation, through quantitative and qualitative analyses, we aim to explore social media activity and media news related to air pollution, in association with the actual data obtained from the continuous monitoring of the air quality. Accordingly, we suppose that the air pollution increase in Western Balkans during winter will provoke more intense activity on social media, especially on Twitter, where people share their opinions and feelings. This study analyses a comprehensive dataset of Twitter entries and media news as responses to the monitoring air quality data between November 2021 and March 2022, aiming to detect the most reliable sentiments and environmental awareness regarding the air pollution problem in the Western Balkans. The public perception of air pollution in several Western Balkan countries, including Macedonia, Serbia,

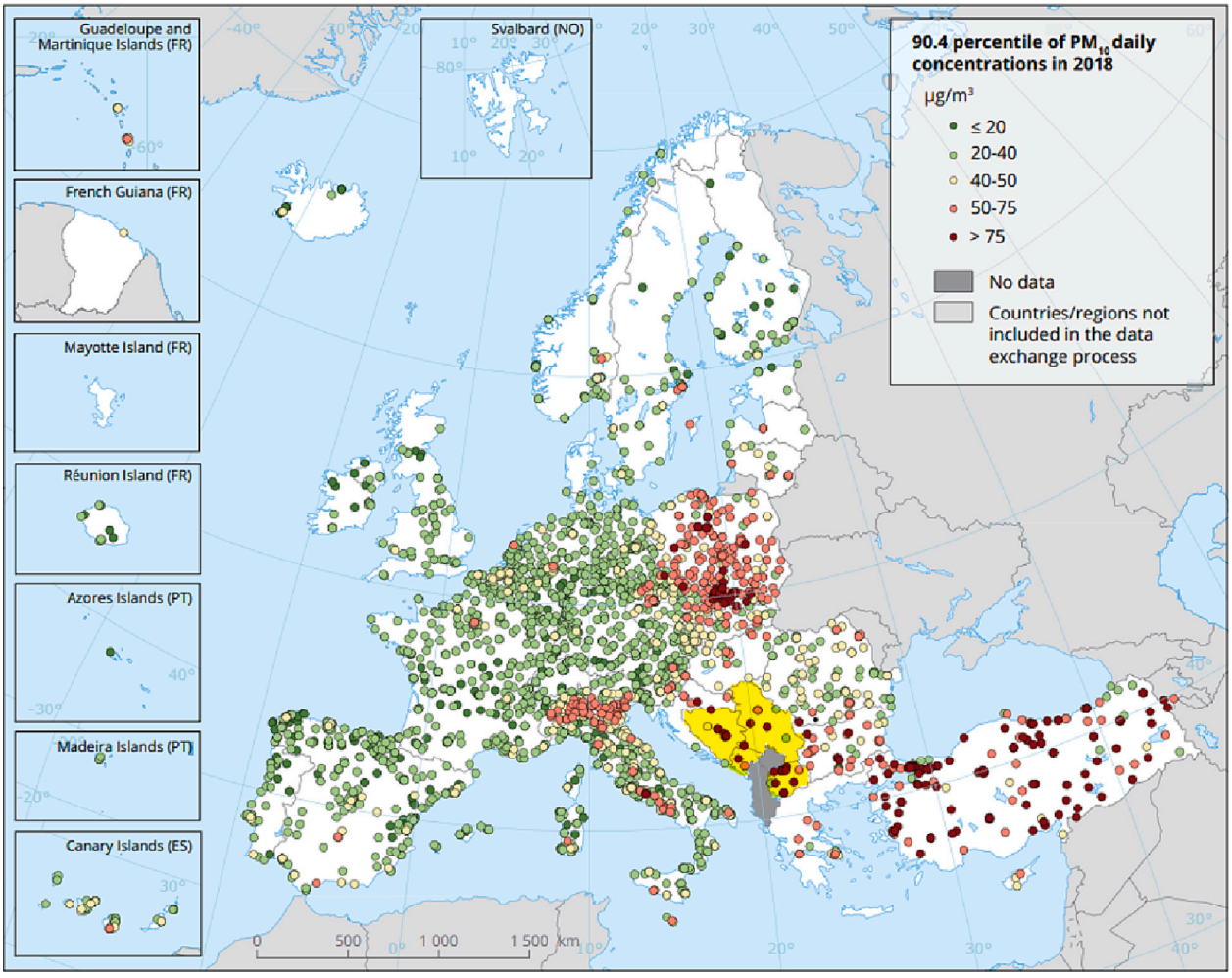


Fig. 1. High percentile of PM_{10} daily concentrations in 2018 across the air quality monitoring stations in the EU. PM_{10} levels in several areas exceed the limit of 50 μg/m3 per day in several countries and areas of the EU, but the Western Balkans are behind in terms of air quality and monitoring infrastructure. Source: EEA Air quality in Europe–2020 report. Note: the area of study is marked with a yellow color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Bosnia and Herzegovina and Montenegro, is analyzed using Natural Language Processing (NLP) techniques such as sentiment analysis and topic modelling, as well as statistical analyses. The positive vs negative classification of the tweets and news, their distribution and correlation in time, as well as the distinctive analysis of topics related to air quality in Western Balkan, offered interesting insights into the landscape of environmental attention. Generally, the data showed a tendency for environmentally aligned practices concerning air pollution mitigation.

2. Related work

Monitoring environmental changes has become crucial at times of accumulating and accelerating stress on ecosystems (Becken et al., 2019). The creation of a more profound consideration of air quality variation over a period by regional zone is a crucial element in enhancing public reaction to unhealthy air impurities (Wang et al., 2021). The air quality monitoring stations require costly equipment and complex positioning plans so as to offer high-quality measures and comparable data on the air quality in precise areas. Given the restricted monitoring within a specific radius, for certain areas, different methods for assessment become necessary, such as air pollution modelling, which can reveal the spatial distribution of air pollution (Shen et al., 2022). Another source of data that is increasingly being analyzed in this regard is social media, where the users act as sensors and the content they share as sensory information (Sakaki et al., 2010). Analyses of micro-blogging posts referring to air pollution have been an appealing research field for many researchers over recent years. Such analyses were primarily done using Weibo (Chinese Twitter) content by separating the positive and negative sentiments with a manual qualitative classification method and using their frequency to improve correlation with the daily Air Quality Index (Jiang et al. (2015)). The study demonstrates that filtered social media messages can be used to monitor air pollution dynamics to some extent and reveal insights into public perceptions and concerns about air pollution. Another study demonstrates the feasibility of using social media to monitor air quality, especially for high-pollution episodes such as wildfires or other specific events, showing that citizen-led monitoring can be used to understand better the public's interaction with air quality issues and Twitter discussions to promote pro-environmental behavior (Sachdeva and McCaffrey, 2018). Previous research showed minimized media influence by filtering tweets containing URLs and compared the frequency of different sentiment groups against the official $PM_{2.5}$ data in Greater London (Hswen et al., 2019). As indicated in Gurajala et al. (2019), experimenting with a wide range of supervised and unsupervised learning methods can provide information about the evolution of topics over time and determine similarities and differences in the public response to air quality information.

In line with this, it is crucial to identify how people feel and what they think about a particular topic, which can often be influenced by sentiments communicated in news articles. Bearing in mind the harmful exposure to PM_{10}, our purpose is to study citizens' sentiments towards air pollution in their country and explore how news media reflects the air pollution ambient in an attempt to promote pro-environmental behavior. Previous findings suggest that negative emotions towards particulate matter increase when PM-related sources and diseases are mentioned in online documents (Song and Song, 2019). Even exposure to low concentrations of PM_{10} negatively affects subjective well-being assessments, with an increase in PM_{10} annual concentrations by 1 μg/m3 contributing to a significant reduction in life satisfaction of 0.017 points on the ESS 10-point life satisfaction scale (Orru et al., 2016).

With that being said, our RQ1 explores if the sentiments expressed in Twitter discussions about air pollution are related to the actual measurements of PM_{10} particles. Our RQ2 examines if the news media, aiming to promote pro-environmental behavior, provides truthful information in accordance with the actual air pollution ambient. Moreover, in previous research, news media reports are shown to influence the number of social media discussions that occur, and it has been suggested that some news values can determine the intensity of Twitter activity (Araujo and van der Meer, 2020; Househ, 2016). Therefore, our RQ3 proposes testing whether news media correlates to the public sentiments towards air pollution in their country, expressed on Twitter.

3. Data

For the purpose of this study, we have collected three types of data: Twitter data, news media data and official air pollution data, which are explained in this chapter.

3.1. Twitter data

Often, unusual or extreme events attract larger volumes of social media activity, especially tweets, than everyday conditions (Becken et al., 2019). In this line, Twitter has attracted considerable research activity, mainly with a focus on event detection, due to the advantages arising from its speed and coverage as well as the spatial and temporal information associated with the tweets (Girish et al., 2022; Li et al., 2012), and especially relevant for situations of crisis or emergencies (Wang et al., 2021). As an example, the literature (Sachdeva and McCaffrey, 2018) demonstrates that crowdsourced data is a viable low-cost source of information about where and when the air quality is affected by a wildfire event.

For the purpose of this study, we coded an application using the Python library Tweepy (Roesslein, 2023) to collect air quality-related tweets from November 1st, 2021, to February 28th, 2022. We used Twitter's Standard API (Twitter, 2023), which allows searching for tweets posted within the last 7 days prior to the search time. Thus, tweets were collected on a weekly basis using the keywords: "aero-zagaduvanje" (air pollution), "аерозагадување" (air pollution), "zagaduvanje" (pollution), "загадување" (pollution), "пМ10" (pm10), "дишеМе" (we breathe) to capture tweets written in the Macedonian language. To collect Western Balkan tweets (excluding tweets written in the Albanian language), we used the terms: "zagadjenje" (pollution), "загађење ваздуха" (air pollution) and "zagadjenje vazduha" (air pollution).

Before analyzing the data, we first used an automatic document translator to translate the collected data into English (DocTranslator, 2023) so that we could use state-of-the-art tools for Natural language processing. To verify the data eligibility, the contents of the weekly collected data for each keyword were carefully inspected by the authors. Although "pollution" is a broad concept, we empirically concluded that Twitter users most frequently refer to the terms "pollution" and "air pollution" interchangeably. The very few tweets discussing different kinds of pollution were excluded manually.

3.2. News media data

Parallel to the collection of tweets, the powerful web-crawler tool Octoparse (Almaqbali et al., 2019) was employed to weekly assemble teaser texts of news articles containing the abovementioned keywords. A teaser is an illustrative short reading suggestion for an article that entices potential readers to read particular news items (Karn et al., 2018). We crawled the news websites Time.mk and Time.rs (Trajkovski, 2008), which are cluster-based news aggregators that analyze 15.000 news daily, collected from 120 distinct sources. Following the translation to English and verification of Twitter data eligibility, the news media data was manually inspected on a weekly basis to exclude potential data not referring to air pollution.

3.3. Official air pollution data

Air pollution data were acquired in order to investigate the frequency of tweets and news articles during the peaks and falls of PM_{10} particles measured by official measuring stations. The PM_{10} data in

Macedonia (Ministry of environment and physical planning - Republic of North Macedonia, 2022), Serbia (Republic of Serbia - Open Data Portal, 2022), Montenegro (Environmental Protection Agency of Montenegro, 2022) and Bosnia and Herzegovina (Discomap, 2022) was collected from 21, 37, 6 and 16 monitoring sites respectively, owned and funded by local authorities. The hourly data measured from November 1st, 2021, to February 28th, 2022, was aggregated by week so that the air pollution data is adjusted to the frequency of collection of Tweets and teasers. For each country, data from all of its monitoring stations was aggregated to encompass the entire country area. PM_{10} data measured in Serbia was available from December 2nd, 2021 to February 28th, 2022.

4. Methodology

4.1. Sentiment analysis

To analyze the data[1] in this study, we used VADER (Valence Aware Dictionary and sEntiment Reasoner) - a lexicon and rule-based sentiment analysis instrument optimized to find semantics in micro-blog texts, such as tweets (Hutto and Gilbert, 2014). VADER relies on an English dictionary that maps lexical features to emotion intensities called sentiment or Valence scores. Valence scores of each word are measured on a scale from - 4 (most negative) to +4 (most positive), with 0 indicating a neutral sentiment. The compound score of the whole text is obtained by summing the valence scores of each word in the lexicon, normalized to be between −1 (most extreme negative) and + 1 (most extreme positive) by using the following normalization:

$$x = \frac{x}{\sqrt{x^2 + \alpha}} \tag{1}$$

In (1), x is the sum of the Valence scores of constituent words, and α is a normalization constant with a default value equal to 15. For sveralizing the tweets into positive, negative, and neutral sentiment groups, the default threshold value of - 0.05 and + 0.05 was used.

4.2. Time series statistical analysis

Regarding time series analysis, measuring similarity is essential to assess the relationship between two signals in time. We used cross-correlation to compare the tweets and teasers against the PM10 data. The Cross Correlation Function (CCF) is the correlation between the observations of two-time series x_t and y_t, separated by k time units (the correlation between y_{t+k} and x_t), where k is called a lag. The confidence interval is calculated with $\pm\frac{2}{\sqrt{n-|k|}}$, where n is the number of observations and k is the lag. The correlation is significant if its absolute value is greater than $\frac{2}{\sqrt{n-|k|}}$ (Minitab, 2022). The CCF assumes that the data is stationary, meaning that the mean and variance are constant and independent of time. If a time series has an upward or downward trend, it is commonly made stationary by differencing (Dean and Dunsmuir, 2016).

We used the nonparametric Mann-Kendall test to assess whether the obtained sentiment groups of tweets and teasers and the PM10 data are stationary. The test indicated a decreasing trend in the teasers obtained from Time.rs. Before performing the CCF test, we conducted first-order differencing to make this data stationary. The Mann-Kendall test did not indicate any trends for the rest of the data.

4.3. Topic modelling

To understand what contributes to air pollution and who is accountable according to public opinion and the news media, we experimented with two topic modelling approaches: Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) (Yin and Wang, 2014).

LDA assumes that a single document (tweet in our study) covers a small set of concise topics and calculates the contribution of each topic to the document. Topics are identified based on the likelihood of co-occurrences of words contained in them (Korzycki et al., 2017). Ranked lists of words associated with a given word w_n are obtained by calculating the sum of the weight of each topic generated by LDA multiplied by the weight of each word w_n contained in that topic. The ranking weight of the word i is computed as follows:

$$w_i = \sum_{j=1..N} w_{ij} * w_{nj} \tag{2}$$

where N is the number of topics and w_{ij} denotes the weight of the word i in topic j.

Contrary to the LDA's assumption, GSDMM assumes only one topic per document, making it suitable for detecting topics in smaller documents, such as tweets. Gibbs sampling describes the method of iterating through and reassigning clusters based on a conditional distribution. In the same manner that the Naïve Bayes Classifier works, documents (tweets) are assigned to clusters based on the highest conditional probability.

Topic modelling is a challenging NLP task, and choosing the "best model" is not an exact science. Despite visual inspection of the topic models, we also depended on the topic coherence to assess the quality of each model. The coherence of a topic, used as a proxy for topic quality, is based on the distributional premise that words with similar meanings tend to co-occur within a similar context (Syed and Spruit, 2017). Topics are considered to be coherent when all or most of the words in the topic are related. We consider the differences between each model as we learn an increasing number of topics, starting from 2. Prior to this analysis, we pre-processed the negative tweets to remove retweet symbols, special characters, URLs, emojis and extra spaces. To give more focus to the critical information, we also removed stop-words, tokenized the tweets and reduced the words to their lemma. In order to obtain the best model, we experimented with topic numbers ranging from 2 to 20 and various hyper-parameters. With a small number of topics, we only extracted broad topics, achieving a low topic coherence score. On the other hand, with a large number of topics, it was difficult to distinguish between them, and the topic coherence score was low again, indicating that the chosen number of topics was probably wrong.

The steps employed in the process of data collection, selection and analysis have been presented in the following diagram (Fig. 2).

5. Results

5.1. Obtained sentiments

The statistics about the tweets and the news article teasers collected during the course of this 4-months study, as well as the sentiments obtained with sentiment analysis for all of the datasets, are displayed in Table 1. It is noticeable that the percentage of negative sentiments prevails in every dataset, while the percentage of neutral sentiments is the lowest. The relatively high number of retweets indicates a sense of agreement and approval among the users (Sharifi and Shokouhyar, 2021). An important point is that the total number of tweets does not equal the sum of retweets and unique tweets (including replies). Sometimes, the original tweet that is being retweeted is not captured since it dates long before the time scope of this study. Moreover, it can happen for several original tweets to be identical. Thus, tweets with distinct pre-processed content (without retweet symbols and special characters) are considered as unique. Similarly, media teasers with distinct content count as unique.

[1] In the absence of geo-localized data, as stated in the previous chapter, air-pollution data are aggregated at country level and at weekly level. That data is then compared to the weekly number of Tweets and teasers.

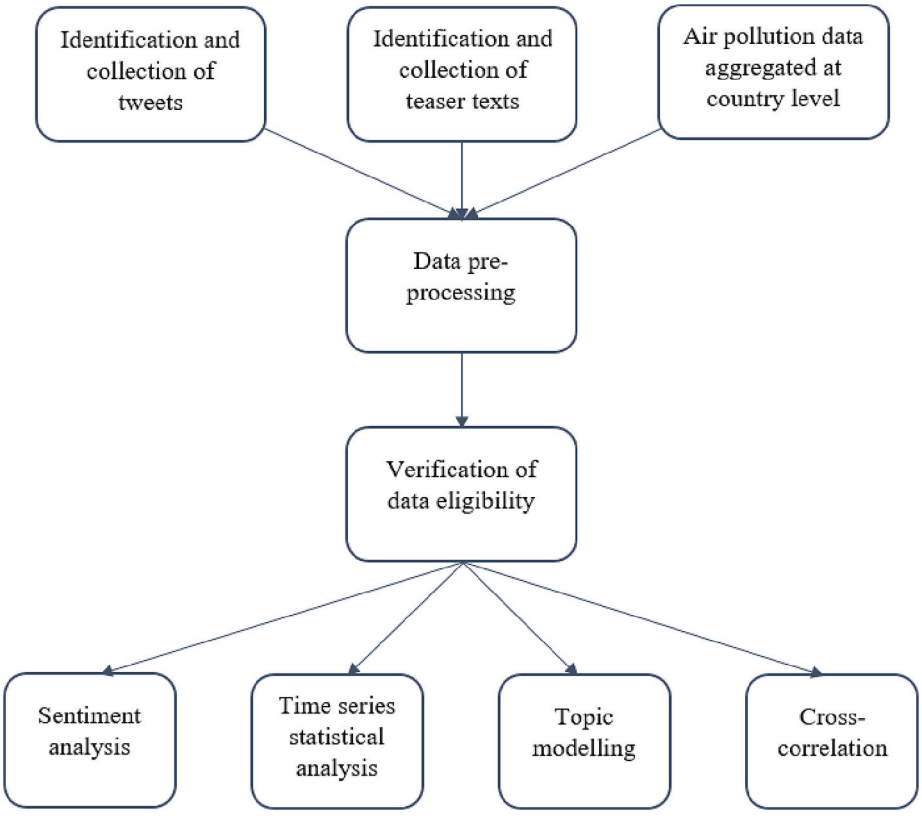


Fig. 2. Data collection and analysis process.

Table 1
Collected tweets and news article teasers about air pollution.

Name	Total number	Retweets (%)	Unique (%)	Negative (%)	Positive (%)	Neutral (%)
#Macedonian Tweets	1018	34	54	64	20	16
Time.mk teasers	994		49	55	40	5
#WB Tweets	2664	48	45	54	29	17
Time.rs teasers	709		74	64	28	8

5.2. Temporal variations and cross-correlation

The tweets and teasers obtained through the sentiment analysis were plotted weekly against the official PM_{10} data to determine the resemblance between the sentiment groups and the actual air pollution data. The different sentiment groups of the Macedonian tweets were compared against the PM_{10} data obtained from measuring stations in Macedonia. Sentiment groups of the rest of the tweets were compared against PM_{10} data in Serbia, Montenegro and Bosnia and Herzegovina to check for correspondence.

The weekly comparison of average PM10 data and the frequency of the negative Macedonian tweets show an exact match of the peaks in both series (Fig. 3). The statistical analysis also shows that the maximum cross-correlation between all categories of Macedonian air pollution tweets and PM_{10} particles in the country was at lag 0, indicating that there is no lag or lead time between levels of PM_{10} particles and tweet frequency. Cross-correlation analysis between the number of negative tweets and PM_{10} showed a coefficient of 0.62 ($p = 0.001$) and 0.54 ($p < 0.001$) between the number of neutral tweets and PM_{10}. The cross-correlation between the positive tweets and PM_{10} particles was insignificant.

Regarding the sentiment groups of Time.mk teasers, the maximum cross-correlation between the frequency of negative teasers and PM_{10}

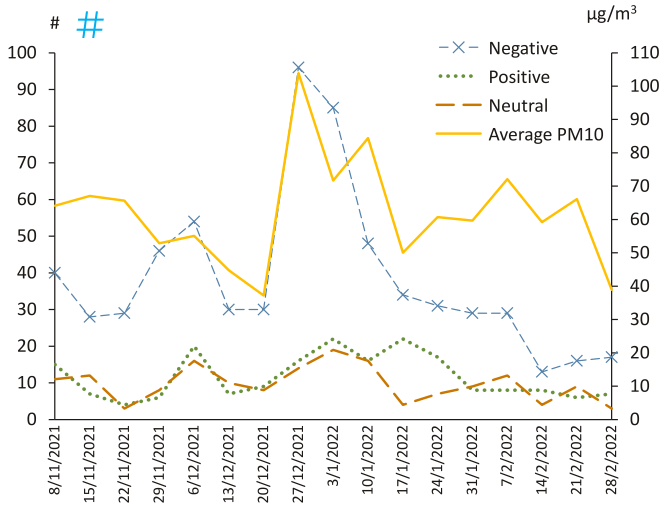


Fig. 3. Weekly comparison of official average PM_{10} data in Macedonia and frequency of Macedonian tweets.

data was 0.53 ($p = 0.002$), between the positive teasers and PM_{10} data was 0.52 ($p < 0.001$) (Fig. 4), both at lag 0; while between the neutral teasers and PM_{10} data, it was insignificant. Although the maximum cross-correlation is at lag 0, we can notice a lag in the peaks of the teasers in relation to the official average PM_{10} data.

There was no significant cross-correlation between the Balkan air pollution tweets and any of the PM_{10} data measured in Serbia, Bosnia and Herzegovina and Montenegro. As for the Time.rs teasers, there was significant cross-correlation at lag 0 between the negative teasers and PM_{10} data measured in Serbia, with a coefficient of 0.66 ($p < 0.001$) (Fig. 5). It can be clearly noted that there is a visual overlap of the peaks in the two time-series. Additionally, there is significant cross-correlation at lag 3 between the neutral teasers and PM_{10} data measured in Montenegro with a coefficient of 0.65 ($p < 0.001$) (Fig. 5).

Also, the time variance of the tweets demonstrates that their maximum peak precedes the highest peak of the number of media teasers (negative, positive and neutral). However, in the case of the second-highest peak, the peak of the negative teasers, for instance, comes before the peak of the negative tweets. Therefore, the results of the correlation analyses between tweets and media teasers reveal the maximum cross-correlation between negative Macedonian tweets and negative Time.mk teasers is 0.8 at lag 0 ($p = 0.0001$), between positive Macedonian tweets and negative Time.mk teasers is 0.55 at lag 1 ($p = 0.02077$) and between neutral Macedonian tweets and negative Time.mk teasers is 0.53 at lag 0 ($p = 0.02854$) (Fig. 6).

Furthermore, the maximum cross-correlation between negative Macedonian tweets and positive Time.mk teasers is 0.77 at lag 0 ($p = 0.00028$), between positive Macedonian tweets and positive Time.mk teasers is 0.59 at lag 2 ($p = 0.01307$) and between neutral Macedonian tweets and positive Time.mk teasers is 0.59 at lag 0 ($p = 0.01315$).

Additionally, the cross-correlation results between negative Macedonian tweets and neutral Time.mk teasers show the maximum value of 0.58 at lag 0 ($p = 0.01419$); between positive Macedonian tweets and neutral Time.mk teasers, it is 0.55 at lag 2 ($p = 0.02319$), while between neutral Macedonian tweets and neutral Time.mk, there is no significant cross-correlation.

The statistical tests affirmed no significant cross-correlation between any sentiment group of the Balkan tweets and any sentiment group of the Time.rs teasers.

5.3. Obtained topics

According to the analysis, for GSDMM on the Macedonian negative tweets, 9 topics were selected with a coherence score of 0.51. In contrast, for LDA, the choice of 15 topics presented a coherence score of 0.41. As for the Time.mk news article teasers, 5 topics (coherence score = 0.43) and 10 topics (coherence score = 0.41) were selected for GSDMM and LDA, respectively.

In addition, for GSDMM on the Western Balkan negative tweets, 7 topics were chosen (coherence score = 0.57), while for LDA, 15 topics (coherence score = 0.38). For the Time.rs news article teasers, 5 topics were selected (coherence score = 0.48) for GSDMM and 16 topics (coherence score = 0.40) for LDA.

In other studies, it has been shown that traditional topic models, such as LDA, experience large performance degradation over short texts (Qiang et al., 2020; Syed and Spruit, 2017), which is in accordance with our results. Therefore, we focus on the topics obtained with GSDMM and present some of the most important ones in Table 2 and Table 3, along with the most important words and their frequency of occurrence.

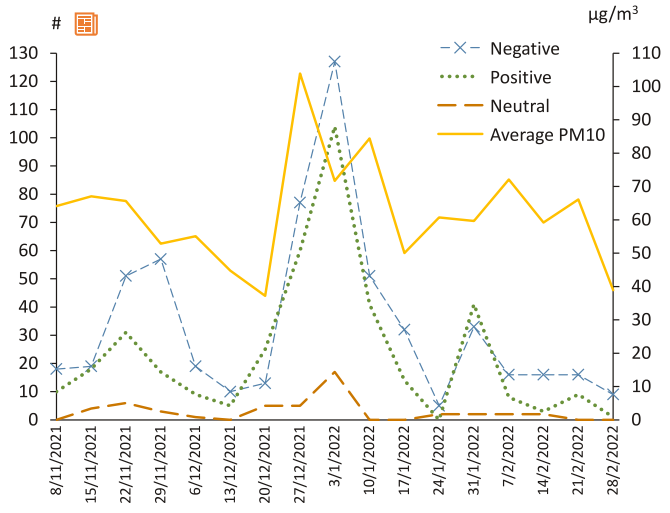


Fig. 4. Weekly comparison of official average PM_{10} data in Macedonia and frequency of Time.mk teasers.

6. Discussion

The analyses showed that during times of increased air pollution, the public expresses negative sentiments and concerns on Twitter, partly sustaining RQ1 that Twitter discussions about air pollution reflect measurements of PM_{10} particles, which is in line with the findings in (Jiang et al. (2015)). This is especially apparent at the maximum pollution peak, broadly consistent with the conclusion in (Sachdeva and McCaffrey, 2018), which initially contends that social media could be valuable for estimating emissions in severe circumstances. Nevertheless, these results could only be confirmed for the Macedonian case, given the non-significant cross-correlation of other countries' data. Namely, opposing the RQ1, there was no resemblance between any of the sentiment groups of Western Balkan Tweets and the actual PM_{10} data collected from Serbia, Bosnia and Herzegovina and Montenegro.

As for the news media in Macedonia, the maximum cross-correlation was detected between negative teasers and PM_{10} data, which is consistent with RQ2, that news media information is in accordance with the communicated data on air pollution, which is also noted in some studies in the literature (Murukutla et al., 2019). However, the high cross-correlation of positive teasers and PM_{10} data should not be neglected, since it might imply that news media try to suppress public discussions about air pollution hazards. Additionally, although the number of articles regarding air pollution decreases over time, a high cross-correlation was detected between negative Time.rs teasers and Serbian PM_{10} data, which, again, supports RQ2 about mass media portraying a realistic ambient condition.

The outcome of the analysis confirms the suggested correlation in RQ3. However, the inconsistency among the sentiments, revealing relatively high correlations between negative Macedonian tweets and positive or neutral news teasers, could indicate that the public is indeed aware of the actual situation. The media news might not be the only source of information they rely on, suggesting, in this way, the presence of critical thinking and search for truthful information by citizens, which are becoming more widely available now. Moreover, the positive and neutral sentiments correlations might speak about air pollution news that likewise contain information about pro-environmental actions needed to be undertaken by citizens and institutions as proof of increasing awareness and action on the subject of air pollution responsibility.

The distinction between the cross-correlation results obtained for the regions covered in this study could be due to the noteworthy difference in the number of collected tweets and teasers in Macedonia compared to the rest of the countries. Since the contents of the Balkan tweets and Time.rs news article teasers on air pollution refer to three countries (Serbia, Bosnia and Herzegovina and Montenegro) collectively, it can be plausibly expected that the ratio of collected tweets and teasers for these regions and Macedonia should be at least 3:1. However, the number of Time.mk teasers is significantly greater than the number of Time.rs teasers. The high cross-correlation between various sentiment groups

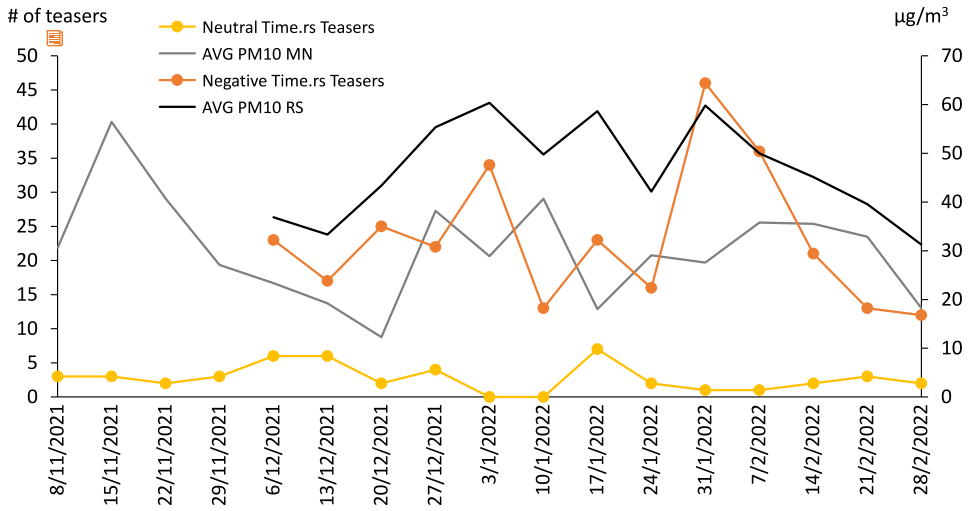


Fig. 5. Weekly comparison of official PM10 data in Montenegro (MN) and Serbia (RS) and frequency of Time.rs teasers.

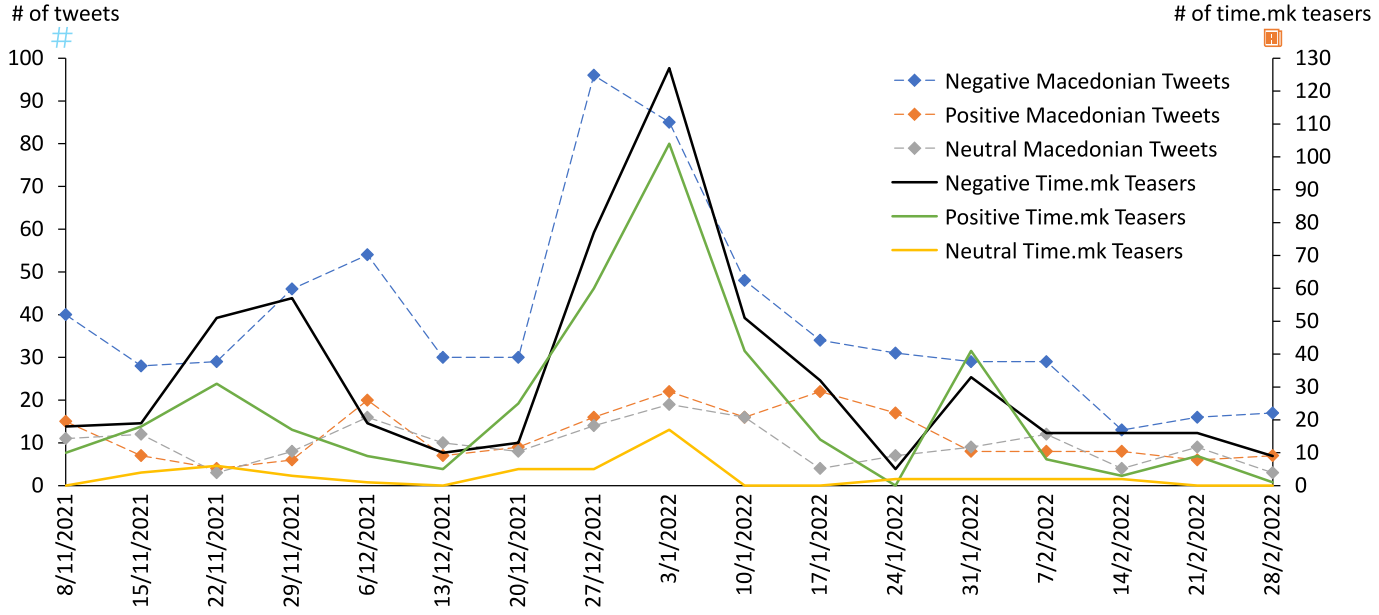


Fig. 6. Weekly comparison of the frequency of different sentiment groups of Time.mk teasers and the frequencies of different sentiment groups of Macedonian tweets.

detected in Macedonian tweets and teasers, in contrast to the non-significant cross-correlation between such data in the remaining countries, suggests that the frequency of sharing a public opinion on social media platforms, such as Twitter, is possibly influenced by the news media coverage of air pollution. Considering the high cross-correlation of negative Time.rs teasers measured solely with the Serbian PM_{10} data, there might be a possible lack of news articles on air pollution in Bosnia and Herzegovina and Montenegro. Despite the media coverage influence, the non-significant cross-correlation between the Balkan tweets and PM_{10} data in Serbia, Bosnia and Herzegovina and Montenegro could be impacted by the tweets' aggregation over these regions. Additionally, the possible mix of Croatian twits with those from Serbia, Bosnia and Herzegovina and Montenegro, because of language similarities, might have restricted the correct attribution of the air quality twits to the data from the corresponding measuring stations.

Furthermore, the topic modelling showed that the Macedonian public thinks that electricity production and transport lead to great air pollution and, thus, health issues. They expect the government to take action and address the problem. However, studies conducted in Macedonia show that the air pollution ratio caused by household heating and transport in the country is about 95% to 5% (Kanevce et al., 2017). The experimentation with different numbers of topics revealed that some discussions about the harmful effects of biomass burning are present on Twitter, suggesting awareness of this pollution source to some extent. Even so, the size of such clusters was significantly smaller compared to the size of clusters regarding transportation and a lower coherence score was achieved. This leads to the conclusion that the use of biomass in households is not among the topics obtained with the highest coherence score, meaning that the awareness of household biomass burning as a pollution source is rather low. At the same time, the ambient condition is often attributed to the transport. These results involve important public unawareness of the main sources of air pollution in the country, emphasizing the need to raise public awareness of the dangers posed by using biomass, and pointing out the importance of promoting pro-environmental behavior on this topic.

Previous research points out that, for example, urban areas' residents

Table 2
Topics obtained through GSDMM topic modelling on Tweets.

Topic (GSDMM)	Negative Macedonian Tweets	Topic (GSDMM)	Negative Balkan Tweets
Electricity	pollution (52), crisis (16), increase (16), price (14), air (11), electricity (11), cut (10), energy (9), terrible (9)	Power plant	pollution (407), air (188), cities (62), Belgrade (53), Serbia (48), terrible (45), plant (34), power (31), problem (24)
Transport	pollution (38), air (12), less (11), problem (9), car (7), brt (7), cities (6), need (6), tram (4)	Climate change	pollution (293), air (97), Serbia (83), problem (45), climate (30), change (30), increase (21), death (18), environment (17)
Government	pollution (12), people (11), anything (9), pm10 (7), immediately (6), mayor (6), don't (6), vmro (4), municipality (4)	Industry	pollution (49), Sabac (25), miner (19), poison (17), plant (8), industry (7), factories (7), burn (7), suffer (7)
Health	die (8), pollution (6), breathe (6), people (6), lose (6), children (6), poison (4), burn (4), poor (4)	Health	pollution (7), Lazarevac (6), children (5), problem (3), nausea (3), dizziness (3), cough (3), heart (3), symptom (3)

Table 3
Topics obtained through GSDMM topic modelling on news article teasers.

Topic (GSDMM)	Negative Time.mk Teasers	Topic (GSDMM)	Negative Time.rs Teasers
Politics	pollution (183), air (176), skopje (88), measure (87), pm10 (41), environment (33), world (38), ministry (24), Arsovska (19)	Health	air (277), pollution (249), Serbia (53), Belgrade (51), Sarajevo (41), environment (38), unhealthy (29), protect (25), health (23)
Health	air (105), reduce (34), cause (20), health (15), death (15), government (13), fight (9), research (9), project (8)	Protest against air pollution	air (101), pollution (92), protest (51), citizen (35), Serbia (34), gather (20), health (20), group (17), change (12)
Landfills	pollution (41), air (32), landfills (14), skopje (10), illegal (9), municipality (8), waste (6), mayor (5), burn (5)	Heating	pollution (64), air (61), heat (16), mask (8), high (7), season (7), smog (7), increase (6), winter (6)
Protest against industrial air pollution	pollution (35), air (32), protest (17), citizen (14), factories (11), mills (8), winter (7), chimney (6), prevent (4)	Industry	air (18), pollution (15), citizen (5), warn (5), politics (5), cause (4), factories (4), reduce (4), need (3)

are more likely to be actively involved with pro-environmental behaviors in comparison to rural populations, leading to a higher likelihood to engage in pro-environmental responses (Anderson and Krettenauer, 2021). It might explain the public perception of specific polluters, such as transport or industrial air pollution, as more important. Therefore, it becomes critical to promote immediate attention and maximize citizens' social responsibility and participation in favor of an optimal sustainable transition towards a healthier environment (He and Shi, 2023).

Throughout the topic modelling, we deduced an indication of unparalleled media coverage of the air pollution in Bosnia and Herzegovina and Montenegro, compared to Serbia. However, when tweeting about air pollution, the public generally expresses worry about pollution caused by power plants. They also relate air pollution to climate change and have concerns about industrial air pollution and health. Again, this indicates Balkan societies' unawareness of biomass burning as one of the central sources of air pollution in the Western Balkans (Todorović, 2022). Bearing in mind the noteworthy correlations stated among positive and neutral sentiments, the oversight of the actual sources of air pollution may be partly influenced by the media, as well.

Additionally, the topics obtained from the news teasers from Time.mk and Time.rs expose that great responsibility for air pollution is attributed to Western Balkan governments' management and health implication. It is noteworthy to underline that media news discussed air pollution implications to health, primarily recognized as being of utmost importance (Ramondt and Ramírez, 2020). The accountability associated with the governments' management is in a way similar to that in the EU countries, which consider that public authorities have not taken enough action and that air pollution should be addressed at the national or EU level, but also on an international and global level (European Commission, 2022). It is important to emphasize that, while a few years ago, the primary responsibility for repairing air pollution damage pointed to the big polluters (European Commission, 2017), in the later years, citizens increasingly recognized personal actions as essential for reducing harmful emissions (European Commission, 2022). This current should serve as an example for Western Balkan media to focus not only on communicating the consequences of air pollution and looking for those responsible, but likewise on encouraging additional citizens' individual activities for tackling the air quality problem.

Considering the power of influence that media have and the relation of the presented information with the actual air pollution data, mass media coverage should be followed by promoting increased motivation towards pro-environmental behavior. In this way, the media impact could be employed to stimulate not only an increasing engagement among citizens, but also to incite greater contribution by governmental institutions and their actual involvement in actions for mitigating air quality problems (Malik et al., 2022).

Citizens seem to be conscious of air pollution and concerned for their health and well-being. Nevertheless, the degree of knowledge about the leading cause of pollution, together with its precise environmental impact, emphasizes the necessity to increase the level of sensitive education, attitude and behavior towards creating environmental quality (Aytun and Akin, 2022).

Intending to overcome the limitations of capturing Croatian tweets when using keywords for collecting Western Balkan tweets, future research directions for this study include using geo-localized tweets that could add additional information to the obtained data, such as in the case of (Hswen et al., 2019). Moreover, a separate analysis of the data obtained from each measuring station could provide insights into air pollution in smaller regions in every country, giving a more detailed representation of the results. Additional direction for research is the possibility of complementing the VADER lexicon used for sentiment analysis by including more air-pollution-related terms and exploring how this influences the obtained results. Finally, observing a broader perspective of individuals could add to the relevant body of knowledge, by exploring personal variables describing socioeconomic and demographic traits as a predictor of citizens' involvement in pro-environmental actions (Yang and Arhonditsis, 2022).

7. Conclusions

A study of associations between media information, individual Twitter activity and air quality reflects the public awareness of the air quality problem and gives a general image of the consistency of monitoring data, presented news and citizens' reactions to air pollution in Western Balkan. The results of the tweets and media news analysis highlighted that only a small portion of these included messages about positive behavior or environmental responsiveness. The general sentiment that prevails in the public reaction is negative and calls for an immediate reflection on improving Western Balkan's air quality.

A unique implication of this study is the comparative correlation assessment among sources of air quality-related public opinion and

actual information from air pollution measurements, contrasting the findings of the case of Macedonia with the broader perception in the Western Balkan region. This study confirms the convenience of NLP techniques such as sentiment analysis and topic modelling for analyzing public thoughts and feelings on essential topics such as air pollution, as well as the reliability of the information transmitted by news media. The cross-correlation between sentiments detected in social media discussions and real air pollution measurements in a country can serve as a measure of public awareness of air pollution. The correlation between media news and public concern evidences the media's crucial role and impact on the public. Therefore, the media should use its power of communication and persuasion, as a tool for enhancing public awareness and action, in addition to the institutional one. In turn, topic modelling techniques can reveal issues in public opinion and, thus, contribute to tackling such problems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

A previous version of this research was presented at the 19th International Conference on Informatics and Information Technologies (CIIT 2022). After the initial feedback, we broadened the scope of our research, including a more varied set of findings and discussions as reported in the current paper, thus extending the contribution and implications of the study.

References

Almaqbali, I.S.H., Al Khufairi, F.M.A., Khan, M.S., Bhat, A.Z., Ahmed, I., 2019. Web scrapping: data extraction from websites. J. Stud. Res. 2019: Fourth Middle East College Student Research Conference, Muscat, Sultanate of Oman.

Anderson, D.J., Krettenauer, T., 2021. Connectedness to nature and pro-environmental behaviour from early adolescence to adulthood: a comparison of urban and rural Canada. Sustainability 13 (7), 3655.

Araujo, T., van der Meer, T.G.L.A., 2020. News values on social media: exploring what drives peaks in user activity about organizations on twitter. Journalism 21 (5), 633–651.

Aytun, C., Akin, C.S., 2022. Can education lower the environmental degradation? Bootstrap panel granger causality analysis for emerging countries. Environ. Dev. Sustain. 24 (9), 10666–10694.

Banja, M., Đukanović, G., Belis, C.A., 2020. Status of Air Pollutants and Greenhouse Gases in the Western Balkans, 30113. Publications Office of the European Union, EUR, pp. 1–53.

Becken, S., Connolly, R.M., Chen, J., Stantic, B., 2019. A hybrid is born: integrating collective sensing, citizen science and professional monitoring of the environment. Ecol. Inform. 52, 35–45.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (Jan), 993–1022.

Colovic Daul, M., Kryzanowski, M., Kujundzic, O., 2019. Air Pollution and Human Health: The Case of the Western Balkans. https://www.developmentaid.org/api/frontend/cms/file/2019/06/Air-Quality-and-Human-Health-Report_Case-of-Western-Balkans_preliminary_results.pdf.

Dean, R.T., Dunsmuir, W., 2016. Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: the importance of constructing transfer function autoregressive models. Behav. Res. Methods 48 (2), 783–802.

Discomap, E.E.A., 2022. Download of air quality data. In: Discomap EEA. https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm.

DocTranslator, 2023. Online Doc Translator. In: DocTranslator. Retrieved November 1, 2021, from. https://www.onlinedoctranslator.com/en/.

Environmental Protection Agency of Montenegro, 2022. Measurement Data Archive. Environmental Protection Agency of Montenegro. http://www.epa.org.me/vazduh/arhiv/7.

European Commission, 2017. Special Eurobarometer 468 - "Attitudes of European Citizens Towards the Environment". https://europa.eu/eurobarometer/surveys/detail/2156.

European Commission, 2022. Special Eurobarometer 524 - "Attitudes of Europeans Towards Air Quality". https://europa.eu/eurobarometer/surveys/detail/2660.

European Environment Agency, 2020. Air Quality in Europe 2020. https://www.eea.europa.eu/publications/air-quality-in-europe-2020-report.

European Environment Agency, 2022a. Air Qual. Europe 2022. https://www.eea.europa.eu/publications/air-quality-in-europe-2022/air-quality-in-europe-2022.

European Environment Agency, 2022b. Europe's Air Quality Status 2022. https://www.eea.europa.eu/publications/status-of-air-quality-in-Europe-2022/europes-air-quality-status-2022.

Girish, K.K., Moni, J., Roy, J.G., Afreed, C.P., Harikrishnan, S., Kumar, G.G., 2022. Extreme event detection and management using twitter data analysis. In: 2022 International Conference on Decision Aid Sciences and Applications (DASA), pp. 917–921.

Gurajala, S., Dhaniyala, S., Matthews, J.N., 2019. Understanding public response to air quality using tweet analysis. Soc. Media Soc. 5 (3), 2056305119867656.

He, X., Shi, J., 2023. The effect of air pollution on Chinese green bond market: the mediation role of public concern. J. Environ. Manag. 325, 116522.

Househ, M., 2016. Communicating Ebola through social media and electronic news media outlets: a cross-sectional study. Health Inform. J. 22 (3), 470–478.

Hswen, Y., Qin, Q., Brownstein, J.S., Hawkins, J.B., 2019. Feasibility of using social media to monitor outdoor air pollution in London, England. Prev. Med. 121, 86–93.

Hutto, C., Gilbert, E., 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media, 8(1), pp. 216–225.

Jiang, W., Wang, Y., Tsou, M.-H., Fu, X., 2015. Using social media to detect outdoor air pollution and monitor air quality index (AQI): a geo-targeted spatiotemporal analysis framework with Sina Weibo (Chinese twitter). PLoS One 10 (10), e0141185.

Jovanovic, M., 2019. Environmental impact of illegal construction, poor planning and Design in Western Balkans: a review. Glob. J. Eng. Sci. 3 (5), 1–4. https://doi.org/10.33552/GJES.2019.03.000572.

Kanevce, G., Dedinec, A., Taseska-Gjorgievska, V., Dedinec, A., 2017. Transport in Skopje – Realities and Challenges, Path to Green Transport. https://api.klimatskipromeni.mk/data/rest/file/download/71dea57f28b%0A54b8c5f35e41a364a586d58c97edef47aacf36268b5d4296667ec.pdf.

Karn, S.K., Buckley, M., Waltinger, U., Schütze, H., 2018. News Article Teaser Tweets and How to Generate Them. ArXiv Preprint (ArXiv:1807.11535).

Korzycki, M., Gatkowska, I., Lubaszewski, W., 2017. Can the human association norm evaluate machine-made association lists?. In: Cognitive Approach to Natural Language Processing. Elsevier, pp. 21–40.

Li, R., Lei, K.H., Khadiwala, R., Chang, K.C.-C., 2012. Tedas: A twitter-based event detection and analysis system. In: 2012 IEEE 28Th International Conference on Data Engineering, pp. 1273–1276.

Malik, S., Arshad, M.Z., Amjad, Z., Bokhari, A., 2022. An empirical estimation of determining factors influencing public willingness to pay for better air quality. J. Clean. Prod. 372, 133574.

Meisner, C., Gjorgjev, D., Tozija, F., 2015. Estimating health impacts and economic costs of air pollution in the republic of Macedonia. SE Eur. J. Public Health (SEEJPH), IV 22–29.

Ministry of environment and physical planning - Republic of North Macedonia, 2022. Air Quality Portal. Ministry of Environment and Physical Planning - Republic of North Macedonia. https://air.moepp.gov.mk/?page_id=175.

Minitab, 2022. Interpret all Statistics and Graphs for Cross Correlation. Minitab. https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/cross-correlation/interpret-the-results/all-statistics-and-graphs/.

Murukutla, N., Kumar, N., Mullin, S., 2019. A review of media effects: implications for media coverage of air pollution and cancer. Ann. Cancer Epidemiol. 2, 7–10.

Orru, K., Orru, H., Maasikmets, M., Hendrikson, R., Ainsaar, M., 2016. Well-being and environmental quality: does pollution affect life satisfaction? Qual. Life Res. 25 (3), 699–705.

Qiang, J., Qian, Z., Li, Y., Yuan, Y., Wu, X., 2020. Short text topic modeling techniques, applications, and performance: a survey. IEEE Trans. Knowl. Data Eng. 34 (3), 1427–1445.

Ramondt, S., Ramírez, A.S., 2020. Media reporting on air pollution: health risk and precautionary measures in national and regional newspapers. Int. J. Environ. Res. Public Health 17 (18), 6516.

Republic of Serbia - Open Data Portal, 2022. Air Quality - Unverified Real-Time Clock Data. Republic of Serbia - Open Data Portal. https://data.gov.rs/sr/datasets/kvalitet-vazduha/.

Roesslein, J., 2018. Tweepy: Twitter for Python!. Retrieved October 26, 2021, from. https://github.com/tweepy/tweepy.

Sachdeva, S., McCaffrey, S., 2018. Using social media to predict air pollution during California wildfires. In: Proceedings of the 9th International Conference on Social Media and Society, pp. 365–369.

Sakaki, T., Okazaki, M., Matsuo, Y., 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860.

Sharifi, Z., Shokouhyar, S., 2021. Promoting consumer's attitude toward refurbished mobile phones: a social media analytics approach. Resour. Conserv. Recycl. 167, 105398.

Shen, Y., de Hoogh, K., Schmitz, O., Clinton, N., Tuxen-Bettman, K., Brandt, J., Christensen, J.H., Frohn, L.M., Geels, C., Karssenberg, D., 2022. Europe-wide air pollution modeling from 2000 to 2019 using geographically weighted regression. Environ. Int. 168, 107485.

Song, J., Song, T.M., 2019. Social big-data analysis of particulate matter, health, and society. Int. J. Environ. Res. Public Health 16 (19), 3607.

Syed, S., Spruit, M., 2017. Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 165–174.
The Lancet, 2020. Global Burden of Disease chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.thelancet.com/pb-assets/Lancet/gbd/summaries/risks/air-pollution.pdf.
Todorović, I., 2022. HEAL: Biomass is one of Main Sources of Air Pollution in Western Balkans. Balkan Green Energy Group. https://balkangreenenergynews.com/heal-biomass-is-one-of-main-sources-of-air-pollution-in-western-balkans/#:~:text=TheEuropeanEnvironmentalAgencyestimated,challengesaroundimprovingwoodburningtechnology.
Trajkovski, I., 2008. How does TIME.mk work? Time.Mk. https://time.mk/info/site.
Twitter, 2023. Standard v1.1. Retrieved October 10, 2021, from. https://developer.twitter.com/en/docs/twitter-api/v1.
Wang, B., Wang, N., Chen, Z., 2021. Research on air quality forecast based on web text sentiment analysis. Ecol. Inform. 64, 101354.
Yang, C., Arhonditsis, G.B., 2022. What are the primary covariates of environmental attitudes and behaviours in Canada? A national-scale analysis of socioeconomic, political, and demographic factors. Ecol. Inform. 69, 101661.
Yin, J., Wang, J., 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 233–242.