**Western Balkan societies’ awareness of air pollution. Estimations using Natural Language Processing techniques**

Angela Madjar1, Ivana Gjorshoska1, Jana Prodanova2,3, Aleksandra Dedinec1,2, Ljupco Kocarev1,2  
[angela.madzhar@students.finki.ukim.mk](mailto:angela.madzhar@students.finki.ukim.mk), [ivana.gjorshoska@students.finki.ukim.mk](mailto:ivana.gjorshoska@students.finki.ukim.mk), [jprodanova@manu.edu.mk](mailto:jprodanova@manu.edu.mk), [aleksandra.kanevche@finki.ukim.mk](mailto:aleksandra.kanevche@finki.ukim.mk),  
[lkocarev@manu.edu.mk](mailto:lkocarev@manu.edu.mk)

1Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, Republic of Macedonia

2Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia

3Faculty of Economic and Business Sciences, University of Burgos, Burgos, Spain

Abstract

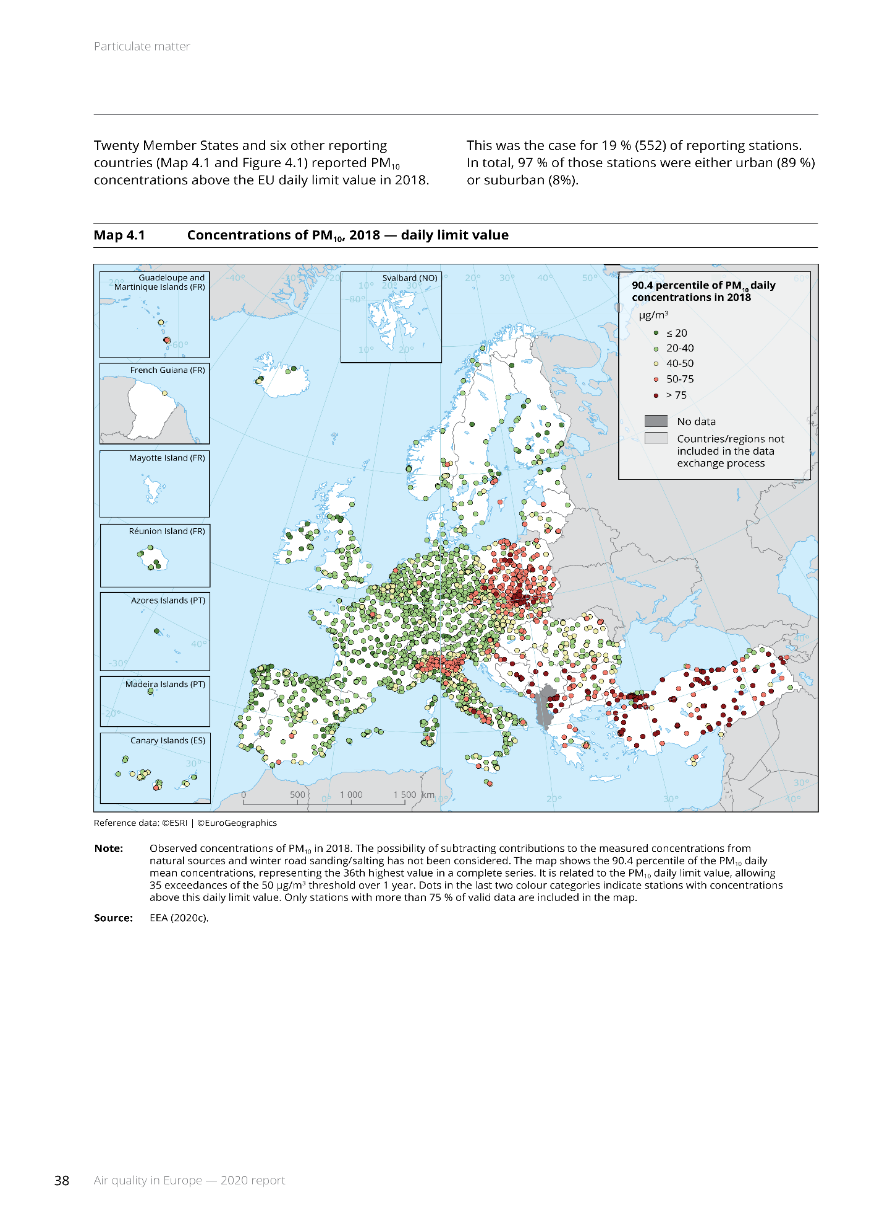
Air pollution remains a severe concern in European countries, especially in Western Balkan, where the air monitoring data point to harmful ambient pollution. The public apprehension with this issue becomes particularly critical during the fall and winter months, when the contamination is more visible, provoking a series of reactions directed principally to the government authorities as the responsible entities for regulating air pollution levels. Given that citizen-based data are valuable for forming a more objective impression of the impact of air pollution, public opinion could act as a tool for monitoring air pollution and increasing awareness and response. 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 escalation 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 websites with actual PM10 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 PM10 concentration in the air and the dangerous consequences. The content of the resources reveals several topics of concern, contributing to better identification of public opinion and possibilities for tracking the news trends. Nevertheless, attention should be paid to news interpretation, provided that sometimes they might offer a more neutral understanding of the situation, improperly impacting society’s opinions in this way. Additionally, the public might not count on sufficient or accurate information for evaluating main air polluters, emphasizing the need for more transparent communication and greater education regarding air pollution monitoring. Finally, the implications of the study provide deeper insights into the content of the data and help detect the reasons for skepticism toward pro-environmental behavior occurring in social-media discussions. Explicitly, individual frustration and disappointment with the air quality in Western Balkan countries should be taken as an inflection point by responsible parties to intervene in improving citizens’ quality of life.

Keywords: air pollution; Western Balkans; Twitter; sentiment analysis; topic modelling; cross-correlation

1. **Introduction**

Air pollution is a severe environmental issue that contributes to the annual mortality of 3.7 million people globally (Wang et al., 2021). Although developing countries are most at risk, highly developed nations are also affected by the adverse effects of ambient air pollution (Olsen et al., 2020). For example, the level of PM2.5 particles in the last week of January 2017 was more prevalent in London than in Beijing, which is a city notorious for its superior levels of air pollution (Hswen et al., 2019). Moreover, the European Environment Agency estimates that around 90% of European urban residents are exposed to harmful air pollutants, indicating the need to recognize the communities at threat and to promote initiatives and adverse exposure control approaches (European Environment Agency, 2019). Air pollution is considered the most significant health threat in Europe, with outdoor and indoor exposure to fine particulate matter resulting in respiratory diseases and premature deaths. Equally, this forms a major problem in all Western Balkan countries, as well. People living in this region are estimated to lose up to 1.3 years of life due to the prominent levels of pollutants in the air (Colovic Daul, M., Kryzanowski, M., Kujundzic, 2019). According to the Air Quality Report (2018), the crude mortality rate attributable to air pollution in Macedonia was 154.7 per 100 000 inhabitants. In comparing neighboring countries such as Serbia, it was 200.7 per 100 000; in Montenegro was 110.9 per 100 000; and in Bosnia and Herzegovina, it was 105.1 per 100 000 (Carvalho, 2019).

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 PM10 concentration in Skopje is above the World Health Organization (WHO) and European Union (EU) recommendations of 50 µg/m3 per day for around half of the year; a staggering fact given the guidelines that the recommended concentrations should not be exceeded for more than 35 days a year. As can be observed in Fig. 1, we note that 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. High percentile of PM10 daily concentrations in 2018 across the air quality monitoring stations in the EU. PM10 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.**

PM10 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 considerably brings about 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). Consequently, the daily PM10 limit is exceeded between 120 and 180 days a year, mostly during wintertime (Colovic Daul, M., Kryzanowski, M., Kujundzic, 2019). Especially during the months between November and March, every year, the air quality monitoring data, particularly PM10, 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 pollutants, and therefore an evaluation of the public level of awareness is of great importance. To measure the individuals’ interest, awareness and reaction to this situation, by 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 escalation 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, Bosnia and Herzegovina and Montenegro, is analyzed by employing 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.

1. **Related work**

Monitoring environmental changes has become increasingly important 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). However, the traditional and static monitoring air quality stations are costly and challenging with respect to discovering adequate places to position them, since the levels of pollutants in the air vary broadly over space and time (Hswen et al., 2019).

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. The study demonstrates that filtered social media messages can be used to monitor air pollution dynamics to some extent. Such messages can reveal insights into public perceptions and concerns about air pollution (Jiang et al., 2015). Another study demonstrates the feasibility of using social media to monitor PM2.5 levels as an alternative method in areas without air pollution monitoring systems, showing that citizen-led monitoring can be used to better understand the public’s interaction with air quality issues and use Twitter discussions to promote pro-environmental behavior. Previous research minimized media influence by filtering tweets containing URLs and compared the frequency of different sentiment groups against the official PM2.5 data in Greater London (Hswen et al., 2019). Moreover, public response to a particular event, such as wildfire, can be captured using a structural topic model to extract topics from tweets posted during the event, as demonstrated in previous research (Sachdeva & McCaffrey, 2018). 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 of crucial importance 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 PM10, 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 causes and diseases are mentioned in online documents (Song & Song, 2019). Even exposure to low concentrations of PM10 negatively affects subjective assessments of well-being, with an increase in PM10 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 PM10 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 & 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.

1. **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.

## **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 & 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, n.d.) to collect air quality-related tweets from November 1st, 2021, to February 28th, 2022. We used Twitter’s Standard API (Twitter, n.d.), 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: “aerozagaduvanje” (air pollution), “аерозагадудвање” (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).

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.

## **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 keywords mentioned above. 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.

## **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 PM10 particles measured by official measuring stations. The PM10 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 EEA, 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. PM10 data measured in Serbia was available from December 2nd, 2021 to February 28th, 2022.

1. **Methodology**

## **Sentiment analysis**

To analyze the data 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 & 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:



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

Before using the built-in NLTK VADER Sentiment Analyzer, we first used an automatic document translator to translate the collected data into English (DocTranslator, n.d.).

## **Time series statistical analysis**

In terms of 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 and , separated by time units (the correlation between and ), where is called a lag. The confidence interval is calculated with ±, where is the number of observations and is the lag. The correlation is significant if its absolute value is greater than (Minitab, 2022). The CCF is based on the assumption 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 & 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.

## **Topic modelling**

To understand what contributes to air pollution and who is accountable according to the 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 & 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 are obtained by calculating the sum of the weight of each topic generated by LDA multiplied by the weight of each word contained in that topic. The ranking weight of the word is computed as follows:

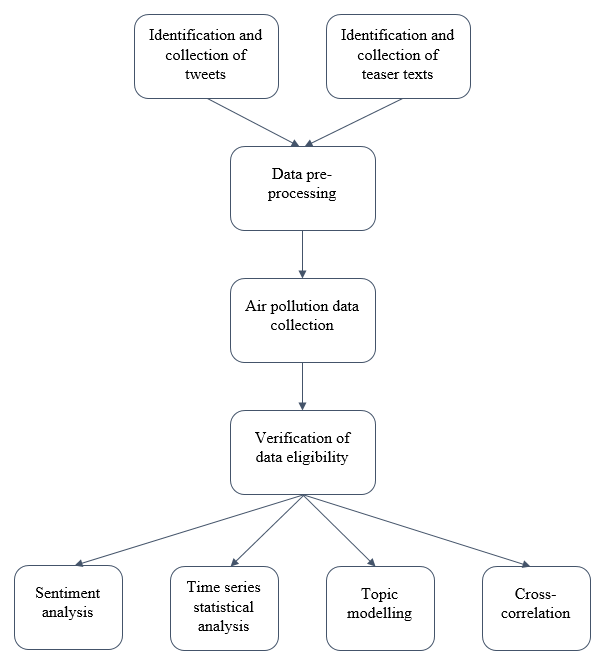


where is the number of topics and denotes the weight of the word in topic .

Contrary to the LDA’s assumption, GSDMM assumes that there is 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.

To evaluate the performance of these approaches, we employed the topic coherence score. 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 & 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.

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



**Fig. 2. Data collection and analysis process**

1. **Results**

## **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 & 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.

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

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Total number | Retweets (%) | Unique (%) | Negative (%) | Positive (%) | Neutral (%) |
| Hashtag with solid fill Macedonian Tweets | 1018 | 33.99 | 53.93 | 64.3 | 19.4 | 16.2 |
| Newspaper with solid fill Time.mk teasers | 994 |  | 48.89 | 55.2 | 39.8 | 4.9 |
| Hashtag with solid fill WB Tweets | 2664 | 47.56 | 44.52 | 54.3 | 29.09 | 16.61 |
| Newspaper with solid fill Time.rs teasers | 709 |  | 73.77 | 64.2 | 27.5 | 8.3 |

## **Cross-correlation**

The tweets and teasers obtained through the sentiment analysis were plotted weekly against the official PM10 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 PM10 data obtained from measuring stations in Macedonia. Sentiment groups of the rest of the tweets were compared against PM10 data in Serbia, Montenegro and Bosnia and Herzegovina to check for correspondence.

The maximum cross-correlation between all categories of Macedonian air pollution tweets and PM10 particles in the country was at lag 0, indicating that there is no lag or lead-time between levels of PM10 particles and tweet frequency. Cross-correlation analysis between the number of negative tweets and PM10 showed a coefficient of 0.62 (p = 0.001) and 0.54 (p < 0.001) between the number of neutral tweets and PM10 (Fig. 3). The cross-correlation between the positive tweets and PM10 particles was insignificant.

Regarding the sentiment groups of Time.mk teasers, the maximum cross-correlation between the frequency of negative teasers and PM10 data was 0.53 (p = 0.002), between the positive teasers and PM10 data was 0.52 (p < 0.001) (Fig. 4), both at lag 0; while between the neutral teasers and PM10 data it was insignificant.

|  |  |  |
| --- | --- | --- |
|  |  | |
| **Fig. 3. Weekly comparison of official average PM10 data in Macedonia and frequency of Macedonian tweets** | **Fig. 4. Weekly comparison of official average PM10 data in Macedonia and frequency of Time.mk teasers** |

There was no significant cross-correlation between the Balkan air pollution tweets and any of the PM10 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 PM10 data measured in Serbia, with a coefficient of 0.66 (p < 0.001) (Fig. 5) and at lag 3 between the neutral teasers and PM10 data measured in Montenegro with a coefficient of 0.65 (p < 0.001) (Fig. 6).

|  |  |
| --- | --- |
|  |  |
| **Fig. 5. Weekly comparison of official PM10 data in Serbia (RS) and frequency of Time.rs teasers** | **Fig. 6. Weekly comparison of official PM10 data in Montenegro (MN) and frequency of Time.rs teasers** |

Furthermore, the results of the analysis testing the correlation between tweets and media teasers reveal that 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. 7).

|  |  |
| --- | --- |
|  |  |
| **Fig. 7. Weekly comparison of the frequency of negative Time.mk teasers and the frequencies of different sentiment groups of Macedonian tweets** | **Fig. 8. Weekly comparison of the frequency of positive Time.mk teasers and the frequencies of different sentiment groups of Macedonian tweets** |

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) (Fig. 8).

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 (Fig. 9).



**Fig. 9. Weekly comparison of the frequency of neutral Time.mk teasers and the frequencies of different sentiment groups of Macedonian tweets**

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

## **Obtained topics**

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. We experimented with topic numbers ranging from 2 to 20 and different 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.

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 & 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.

**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) |

1. **Discussion**

The analyses showed that negative Macedonian tweets were most predictive of PM10 levels in the country, whereas positive tweets do not have comparable peaks and fall with the measured PM10 particles. This suggests that during times of escalated air pollution, the public expresses negative sentiments and concerns on Twitter, partly sustaining RQ1 that Twitter discussions about air pollution reflect measurements of PM10 particles. 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 PM10 data collected from Serbia, Bosnia and Herzegovina and Montenegro. However, most tweets were classified as negative, suggesting that people express negative sentiments on Twitter even when the PM10 levels are moderate or low.

As for the news media in Macedonia, the maximum cross-correlation was detected between negative teasers and PM10 data, which is consistent with RQ2, that news media information is in accordance with the communicated data on air pollution. However, the high cross-correlation of positive teasers and PM10 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 PM10 data, which, again, supports RQ2 about mass media portraying a realistic ambient condition.

The outcome of the cross-correlation analysis to detect the possible link between media teasers and the tweets reveals relatively high correlations between negative Macedonian tweets and positive or neutral news teasers, besides the expected substantial correlation between the negative tweets and negative teasers. These results confirm the suggested correlation in RQ3. However, this inconsistency among the sentiments could indicate that the public is indeed aware of the actual situation, and 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.

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, G., Dedinec, A., Taseska-Gjorgievska, V., Dedinec, 2017). The experimentation with different numbers of topics revealed that some discussions about the harmful effects of biomass are present on Twitter, suggesting awareness of this pollutant 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 a conclusion that the use of biomass in households is not among the topics obtained with the highest coherence score. At the same time, the ambient condition is often attributed to the transport. These results involve important public unawareness of the main air pollutants in the country, emphasizing the need to raise public consciousness of the dangers posed by using biomass, and pointing out the importance of promoting pro-environmental behavior on this topic.

Throughout the topic modelling, we discovered that many air pollution tweets refer to Serbia and cities in Serbia (e.g., Belgrade, Sabac). 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 as one of the central air pollutants in the Western Balkans (Todorović, 2022). However, this tendency of potential ignorance might also be partly influenced by the media, bearing in mind the noteworthy correlations stated among positive and neutral sentiments.

Additionally, the topics obtained from the news teasers from Time.mk and Time.rs expose that the most frequent concerns about air pollution in the media are associated with health issues and politics. In this way, it could be understood that great responsibility for air pollution is attributed to Western Balkan governments’ management and health implication.

1. **Conclusions**

A parallel study of media information and individual Twitter activity, reflecting public awareness of the air quality problem, 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. The regular media coverage of the subject and substantial self-perceived knowledge has evidenced people’s consciousness about air pollution and related risks, the initiative to bring the attention of individuals, companies and the government to the problem and the need for actions to preserve environmental resources and improve air quality. Bearing in mind 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 a promotion of increased motivation toward pro-environmental behavior. In this way, the media impact could be handled as a stimulus for not only increasing engagement among citizens, but also for requiring governmental implication and real action.

Previous research points out that, for example, urban areas’ residents 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 & 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.

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 toward creating environmental 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.

Future research directions for this study include using geo-localized tweets that could add additional information to the obtained data and would eliminate the possibility of capturing Croatian tweets when using keywords for collecting Western Balkan tweets. Finally, 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.

**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. *Journal of Student Research*.

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.

Banja, M., Đukanović, G., & Belis, C. A. (2020). Status of air pollutants and greenhouse gases in the Western Balkans. *Publications Office of the European Union, EUR*, *30113*, 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. *Ecological Informatics*, *52*, 35–45.

Becken, S., Stantic, B., Chen, J., Alaei, A. R., & Connolly, R. M. (2017). Monitoring the environment and human sentiment on the Great Barrier Reef: assessing the potential of collective sensing. *Journal of Environmental Management*, *203*, 87–97.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Boussalis, C., & Coan, T. G. (2016). Text-mining the signals of climate change doubt. *Global Environmental Change*, *36*, 89–100.

Carvalho, H. (2019). Air pollution-related deaths in Europe–time for action. *Journal of Global Health*, *9*(2).

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. *Behavior Research Methods*, *48*(2), 783–802.

Discomap EEA. (2022). *Download of air quality data*. Discomap EEA. https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm

DocTranslator. (n.d.). *Online Doc Translator*. DocTranslator. 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 Environment Agency. (2019). *Air pollution still harming health across Europe*. https://www.eea.europa.eu/media/newsreleases/air-pollution-still-causing-harm

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

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. *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 917–921.

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

Househ, M. (2016). Communicating Ebola through social media and electronic news media outlets: A cross-sectional study. *Health Informatics Journal*, *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. *Preventive Medicine*, *121*, 86–93.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 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. *Global Journal of Engineering Sciences*, *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* (pp. 21–40). Elsevier.

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

Meisner, C., Gjorgjev, D., & Tozija, F. (2015). Estimating health impacts and economic costs of air pollution in the Republic of Macedonia. *South Eastern European Journal of Public Health (SEEJPH)*.

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/

Olsen, Y., Nøjgaard, J. K., Olesen, H. R., Brandt, J., Sigsgaard, T., Pryor, S. C., Ancelet, T., del Mar Viana, M., Querol, X., & Hertel, O. (2020). Emissions and source allocation of carbonaceous air pollutants from wood stoves in developed countries: A review. *Atmospheric Pollution Research*, *11*(2), 234–251.

Orru, K., Orru, H., Maasikmets, M., Hendrikson, R., & Ainsaar, M. (2016). Well-being and environmental quality: Does pollution affect life satisfaction? *Quality of Life Research*, *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 Transactions on Knowledge and Data Engineering*, *34*(3), 1427–1445.

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. (n.d.). *Tweepy: Twitter for Python!* https://github.com/tweepy/tweepy

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

Sharifi, Z., & Shokouhyar, S. (2021). Promoting consumer’s attitude toward refurbished mobile phones: A social media analytics approach. *Resources, Conservation and Recycling*, *167*, 105398.

Song, J., & Song, T. M. (2019). Social big-data analysis of particulate matter, health, and society. *International Journal of Environmental Research and Public Health*, *16*(19), 3607.

Syed, S., & Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165–174.

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=The European Environmental Agency estimated,challenges around improving woodburning technology

Trajkovski, I. (2008). *How does TIME.mk work?* Time.Mk. https://time.mk/info/site

Twitter. (n.d.). *Standard v1.1*. https://developer.twitter.com/en/docs/twitter-api/v1

Venus, T. E., & Sauer, J. (2022). Certainty pays off: The public’s value of environmental monitoring. *Ecological Economics*, *191*, 107220.

Wang, B., Wang, N., & Chen, Z. (2021). Research on air quality forecast based on web text sentiment analysis. *Ecological Informatics*, *64*, 101354.

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–242.