Comparative Social Media Analysis on Air Pollution Awareness in Macedonia

Marija Stojcheva[1], Aleksandra Dedinec[1,2], Jana Prodanova[2], Trifce Sandev[2], Desheng Wu[3], and Ljupco Kocarev[2]

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

[2]Macedonian Academy of Sciences and Arts, Skopje, Macedonia

[3]University of Chinese Academy of Sciences, Beijing, PR China

Abstract

Air pollution is a significant problem in cities and urban centers in Macedonia, impacting both the environment and public health. This study aims to investigate the evolution of public awareness regarding air pollution in Macedonia over the past three years, with a focus on discussions on social media platform X (Twitter). Recognizing social media platforms as influential channels for disseminating information and raising awareness, the correlation between tweets and PM_{10} particles is explored. Utilizing natural language processing techniques, specifically sentiment analysis and topic modeling, alongside statistical methods such as correlation analysis and Kruskal–Wallis H test, the study examines public sentiment and trending topics associated with air pollution. In addition to providing insight into public perceptions of air pollution, the results assist in determining if public awareness has increased from previous years.

Keywords: air pollution, pm10, natural language processing, sentiment analysis, topic modeling, correlation analysis, Kruskal–Wallis H test, social media platforms, X analysis, Twitter

1. Introduction

The increasing number of diseases and deaths caused by air pollution is becoming a global issue for public health, especially in developing countries (Pittet et al., 2008). In 2019, air pollution was linked to over 1.8 million deaths worldwide, with its impact being particularly severe in urban areas, where more than half of the global population resides (Southerland et al., 2022; Institute, 2020). Predictions indicate that by 2050, outdoor air pollution, especially particulate matter and ground-level ozone, will emerge as the

leading cause of environmentally related deaths globally (OECD, 2012). Furthermore, air pollution represents the primary source of environmental disease burden, worsening health conditions such as asthma, cancer, pulmonary illnesses, and cardiovascular diseases (NECE, 2016).

The Western Balkans, a region which primarily includes developing countries, faces serious challenges due to high levels of air pollution. The economic impact of air pollution in this region, when measured against the GDP of the EU27 during the same period, is notably higher, thereby emphasizing the severity of the problem (Belis et al., 2023). On average, people living in urban areas of the Western Balkans lose between 13 and 16 months of their lives because of air pollution, leading to almost 5,000 premature deaths annually (Daul et al., 2021). The main sources of air pollution in the Western Balkans include burning solid fuels for heating homes, vehicle exhaust, and industrial activities. Additionally, coal power plants in the region worsen the problem, emitting 16 times more particulate matter (PM) than the average European plant (CAN, 2019).

Macedonia, a part of the Western Balkans, faces significant health and economic challenges due to air pollution. With approximately 1,350 deaths annually and economic losses equivalent to around 3% of the country's GDP, Macedonia ranks fifth in Europe for $PM_{2.5}$ particle concentration (Bank, 2019). Cities like Skopje and Tetovo experience some of the highest rates of pollution-related mortality in the region, accounting for nearly one-fifth of all deaths (Daul et al., 2021). Beyond the capital Skopje, where 45% of the disease burden is concentrated, other cities also experience air quality issues due to nearby industrial activities. Industries such as energy production in Bitola, metallurgical works in Kavadarci and Tetovo, and oil refining in Miladinovci are major contributors to air pollution, as depicted in Figure 1 (Jorgensen & Shkaratan, 2020). Furthermore, In 2016, Macedonia recorded the highest number of days exceeding the limit of 35 days with PM_{10} concentration higher than 50 $\mu g/m^3$ within the Western Balkan (Banja et al., 2020).

Given the substantial health and economic impacts of air pollution, understanding public awareness and perception of this issue is crucial for effective policy implementation and community engagement. This study investigates the level of public awareness regarding air pollution in Macedonia over the past three years. Through an analysis of social media discussions and their correlation with air quality data, insights into public perceptions regarding air pollution were obtained. Employing advanced methodologies including natural language processing techniques like sentiment analysis and topic modeling, as well as statistical methods like correlation analysis and Kruskal–Wallis H test, this study explores the relationship between discussions about air pollution on social media platform X and actual air quality measurements. The aim of this study is to discover public sentiments and prevalent PM regarding air pollution, while also evaluating any improvements in public awareness compared to previous years.

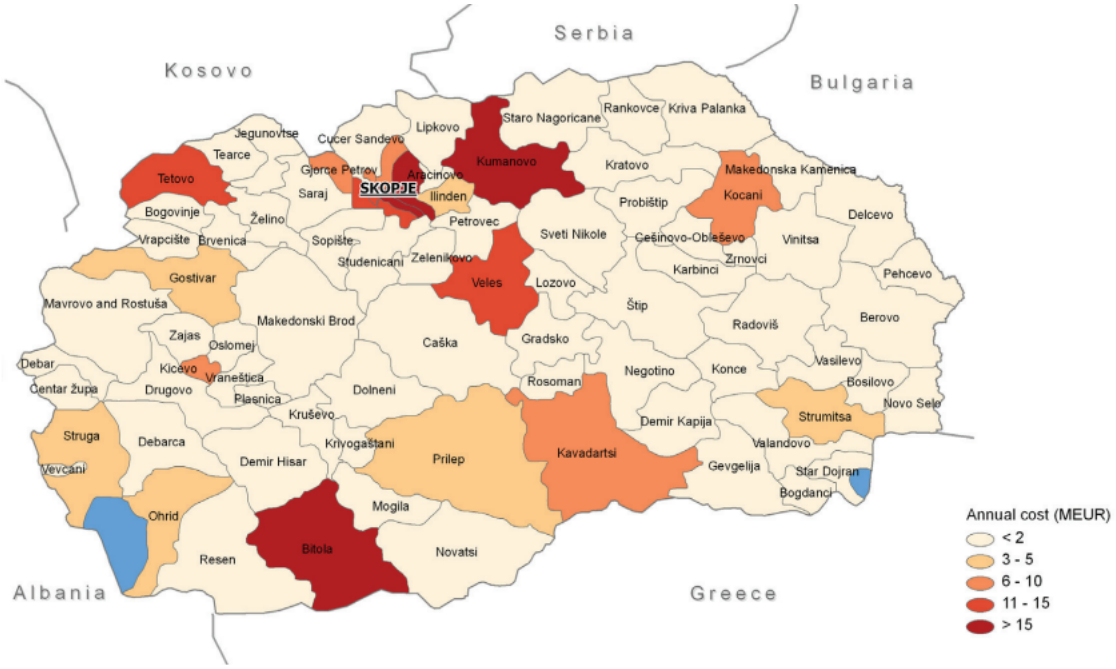

Figure 1: Air pollution in Macedonia

2. Related work

In recent years, there has been a notable increase in interest in using X (Twitter) microblogs to study air pollution dynamics. A study in China used geotagged tweets to measure the level of public happiness in relation to air quality and $PM_{2.5}$ concentrations, revealing how happiness is affected by air pollution (Zheng et al., 2019). Similarly, research in Delhi, India, used machine learning to classify tweets by sentiment and estimate $PM_{2.5}$ concentration with high accuracy (80-99%) during extreme pollution events (TS et al., 2023). Studies have expanded their scope to target specific groups, such as analyzing haze-related posts on the Chinese micro-blogging platform Sina Weibo, aiming to understand how air pollution concerns impact vulnerable groups (Wang et al., 2022). Additionally, another study compared natural language processing methods for analyzing climate change tweets, finding BERT's sentiment classification superior (Rosenberg et al., 2023). In Macedonia, ongoing studies are investigating the primary pollution sources of $PM_{2.5}$ and PM_{10} in Skopje's urban area (Mirakovski et al., 2024; Mirakovski et al., 2019). Additionally, recent research has explored semantic analysis and topic modeling, providing insights into air pollution dynamics and public perception over the past two heating seasons (Madjar et al., 2023; Perikj et al., 2023).

3. Data

For this study, three types of data have been collected: X data, official air pollution data, and air pollution data from the preceding two years relevant to this topic. These will be discussed in detail in the following chapter.

3.1 X Data

X is a popular microblogging service, widely used for sharing daily activities, seeking information, and discussing various PM (Java et al., 2007). Public opinion is formed through the exchange of views that takes place on social media platforms, where the spread of information and content can influence the perceptions and economic decisions of individuals, companies, and governments (Almaududi Ausat, 2023). During crisis events, messages on social media platforms like X can be a crucial source of information, often providing details and personal perspectives more rapidly than traditional news sources (Kruspe et al., 2021). In events like earthquakes, its real-time nature is particularly valuable, as the immediate surge of posts allows for rapid detection and response (Sakaki et al., 2010). Additionally, journalists using X are often found to be more effective in their roles, highlighting the platform's impact on news dissemination (Mcgregor & Molyneux, 2018).

Recognizing the influence and real-time capabilities of X, a search application was created using the Tweepy Python library to collect air quality-related tweets from November 1st 2023 to February 29th 2024. The X API v2 was utilized to access the Standard search API REST endpoint, enabling the retrieval of tweets from the previous week with a single search query (Roesslein, 2020; X, n.d.). This approach allows filtering tweets based on specific PM using message keywords, hashtags and URLs. Given the limited usage of hashtags in Macedonia, keywords were used to capture relevant tweets. These keywords include "aerozagaduvanje" (air pollution), "аерозагадување" (air pollution), "zagaduvanje" (pollution), "загадување" (pollution), "пм10" (pm10), "дишеме" (we breathe) and "загадено" (polluted). Although "pollution" is a broad term, in the Macedonian language users frequently use "pollution" and "air pollution" interchangeably. Consequently, tweets discussing other forms of pollution were manually excluded from the dataset.

3.2 Official Air Pollution Data

To analyze tweet activity during polluted and non-polluted periods, PM_{10} data was required. This data was collected from 20 monitoring sites operated and funded by local authorities of Macedonia (of environment & physical planning - Republic of North Macedonia, n.d.). Data from each monitoring station was gathered hourly, and the mean value for each station was computed. Subsequently, for the purpose of this research, the mean PM_{10} value for all stations across Macedonia was calculated, aiming to explore the correlation between activity on the platform X and PM_{10} particle levels throughout the country.

3.3 Previously Collected Data

To conduct a comparative analysis, tweets gathered using identical keywords were utilized, along PM_{10} particle measurement data collected similarly during the heating seasons of

the preceding two years, specifically from November 1st 2021 to February 28th 2022 and from October 24th 2022 to February 28th 2023 (Madjar et al., 2023; Perikj et al., 2023).

4. Methodology

In this section, we detail the methodology employed to analyze the sentiment of tweets, explore time series statistical data, conduct topic modeling, and perform the Kruskal-Wallis H test.

4.1 Sentiment Analysis

Sentiment analysis, also referred to as opinion mining, is a natural language processing technique used to determine whether the emotional tone conveyed within a piece of text is positive, negative, or neutral. In the context of analyzing sentiments expressed on social media platforms like X, specialized tools have been developed to handle the unique characteristics of short, informal texts like tweets. Two such tools commonly used for X sentiment analysis are VADER (Valence Aware Dictionary and Sentiment Reasoner) and Twitter-roBERTa-base.

VADER, a sentiment analysis tool tailored for micro-blog texts, operates on lexicon and rule-based principles (Hutto & Gilbert, 2014). It employs an English dictionary mapping words to sentiment scores, from -4 (most negative) to +4 (most positive), with 0 denoting neutrality. The tool computes a compound score for a text by summing up the valence scores of individual words, which is then normalized to a scale between -1 and +1. This normalization process follows the formula:

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

where x is the sum of the valence scores of constituent words, and α is a normalization constant, usually set to 15. For categorizing the tweets into positive, negative, and neutral sentiment groups, the default threshold value of -0.05 and +0.05 was applied.

On the other hand, Twitter-roBERTa-base, derived from the BERT architecture and refined by RoBERTa, is a specialized model tailored explicitly for sentiment analysis tasks within the TweetEval benchmark framework (Barbieri et al., 2020). Twitter-roBERTa-base excels in capturing nuanced contextual relationships crucial for understanding sentiment within tweets. Its multi-layer bidirectional transformer architecture generates numerical vectors that encode word meanings, enabling deeper contextual understanding necessary for sentiment analysis. Additionally, it incorporates a dense layer tailored to reduce the dimensions of RoBERTa's final layer to three, aligning with the typical sentiment classification task, where sentiments are categorized as positive, negative or neutral (Loureiro et al., 2022).

Since these models do not support the Macedonian language, the tweets were translated into English using the Deep Language translator (Translator, n.d.).

4.2 Time Series Statistical Data

Within time series analysis, evaluating similarity is important for understanding the causal relationship between two signals over time. To evaluate the correlation between tweets and PM_{10} data, cross-correlation analysis was employed. The Cross Correlation Function (CCF) measures the correlation between two time series at different lag intervals, assuming stationarity where the mean and variance of the data remain constant over time (Welsh, 1999). The confidence interval for the CCF is determined based on the number of observations and the lag. A correlation is considered significant if its absolute value surpasses a predefined threshold.

The Mann-Kendall test was utilized to verify the stationarity of the data (Sheng Yue, 2002). This non-parametric test is used to identify trends in time series data without requiring the data to follow any particular distribution.

4.3 Topic Modeling

Topic modeling is a natural language processing technique used to extract underlying themes or topics from a collection of text documents. In social media analysis, BERTopic, a modular framework, is utilized to extract valuable insights from user-generated content, such as tweets. This framework incorporates default components, but is allowing users to interchange them to their preferences or specific task requirements (Grootendorst, 2022).

At its core, BERTopic utilizes SentenceBERT, a variant of the BERT model fine-tuned for sentence-level tasks, to generate embeddings that capture the semantic context of the text. These embeddings are then reduced in dimensionality using Uniform Manifold Approximation and Projection, preserving intrinsic structure, followed by clustering with Hierarchical DBSCAN to identify clusters of varying densities without prior specification of cluster numbers.Next, the text undergoes preprocessing using CountVectorizer at a cluster level, ensuring topic-level word capture. Additionally, in BERTopic TF-IDF is adapted to work on a cluster level instead of a document level, resulting in a modified version known as c-TF-IDF. This approach enhances topic representations by considering inter-cluster differences. This involves converting each cluster into a single document and extracting word frequencies, refined by logarithmic transformations. The weight of term x in cluster c is calculated using the formula:

$$W_{x,c} = |tf_{x,c}| \times \log\left(1 + \frac{A}{f_x}\right) \tag{2}$$

where $W_{x,c}$ is the weight of term x in cluster c, $tf_{x,c}$ is the term frequency of term x in cluster c, A is the average number of words per cluster, and f_x is the document frequency

of term x across all clusters. Furthermore, BERTopic offers the flexibility for fine-tuning model parameters, allowing users to customize the analysis according to their specific requirements.

4.4 Kruskal-Wallis H test

The Kruskal-Wallis H test is employed to analyze the weekly tweet counts across three years, aiming to uncover potential trends within each year during the same time frame. The Kruskal-Wallis H test is a rank-based nonparametric test utilized to determine whether statistically significant differences exist among two or more groups of an independent variable concerning continuous or ordinal dependent variables (Kruskal & Wallis, 1952). The Kruskal-Wallis H statistic is computed as follows:

$$H = \frac{12}{N(N+1)} \left(\sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1) \right) \tag{3}$$

where N represents the total sample size, k signifies the number of groups being compared, R_j denotes the sum of ranks for group j, and n_j signifies the sample size of group j. The null hypothesis H_0 asserts that there are no significant variations among the population medians of the groups. For the study, H_0 suggests that there are no significant differences in the weekly tweet counts across the three years.

5. Results

5.1 Obtained sentiments

The prior two studies examining tweets from Macedonia relied on VADER as sentiment analysis tool. Upon the review of the results, it was evident that the previously used sentiment analysis framework struggled to effectively capture neutral tweets, often resulting in an overrepresentation of positive sentiments. This inconsistency did not match the typical sentiment distribution observed on the social platform X. Additionally, these conventional methods frequently failed to identify sarcasm in tweets, requiring the authors to manually address this issue.

Motivated by the increasing popularity of transformer models, a comprehensive reanalysis using Twitter-roBERTa-base revealed its significant outperformance compared to VADER. Unlike the aforementioned traditional methods, which rely on fixed dictionaries, the deep neural network model of Twitter-roBERTa-base provides enhanced contextual comprehension, leading to more precise detection of neutral and positive sentiments. Although VADER may provide quicker results and could be favored for initial assessments, especially with larger datasets, our findings suggest that Twitter-roBERTa-base is superior in understanding the nuances of sentiment, particularly within social media data.

Detailed comparative results are shown in Table 1.

Table 1: Comparison of Sentiment Analysis Results

Year	**Model**	**Total number**	**Retweets (%)**	**Positive (%)**	**Neutral (%)**	**Negative (%)**
2023/2024	Twitter-roBERTa-base	2075	41.64	9.5	35.4	55.1
	VADER			33.6	17.8	48.5
2022/2023	Twitter-roBERTa-base	1442	37.59	3.2	36.5	60.3
	VADER			24.8	23.2	52.0
2021/2022	Twitter-roBERTa-base	1018	34	5.9	27.5	66.6
	VADER			19.4	16.2	64.3

5.2 Cross-correlation

In the study, a comprehensive analysis was conducted to investigate the relationship between sentiment expressed in Macedonian tweets and PM_{10} levels, with a specific focus on the year 2023/2024. Weekly analyses were performed to explore correlations between different sentiment groups in tweets and PM_{10} data collected from measuring stations in Macedonia, as shown in Table 2. The Mann-Kendall test was employed to assess data trends, indicating the stationarity of the data and its suitability for correlation analysis without preprocessing.

Subsequent cross-correlation analyses were conducted using both the Twitter-roBERTa-base and VADER sentiment analysis models, revealing significant coefficients for the 2023/2024 period. Notably, the cross-correlation coefficient for All Tweets remained consistent between the two models, given the absence of inherent sentiment in this group. The observed coefficient for All Tweets was 0.88, indicating a robust correlation between the sentiment conveyed in tweets and PM_{10} levels during this timeframe. Furthermore, the Twitter-roBERTa-base model displayed coefficients of 0.81 for Positive Tweets, 0.86 for Neutral Tweets, and 0.85 for Negative Tweets, as can be seen in Figure 4. Similarly, VADER demonstrated coefficients of 0.91 for both Positive and Neutral Tweets, and 0.84 for Negative Tweets, illustrated in Figure 5. Moreover, no lag was observed for either method, suggesting an immediate association between tweet sentiment and PM_{10} levels.

Comparative analysis with data from preceding years revealed significant trends. In the 2021/2022 study, both the Twitter-roBERTa-base model and VADER exhibited identical cross-correlation coefficients: 0.43 for Positive Tweets, 0.66 for Neutral Tweets, and 0.64 for All Tweets. For Negative Tweets, the Twitter-roBERTa-base model showed a coefficient of 0.52, while VADER exhibited a coefficient of 0.64. Positive tweets lagged by -1.Similarly, in the subsequent year, 2022/2023, both the Twitter-roBERTa-base model

and VADER showed consistent values across the following metrics: 0.50 for Positive Tweets, 0.64 for Neutral Tweets, and 0.65 for All Tweets. Additionally, Negative Tweets displayed coefficients of 0.68 and 0.63, with Neutral Tweets lagging by 1 and Positive Tweets by 5 for both methods.The results from the Twitter-roBERTa-base model for the years 2021/2022 and 2022/2023 are depicted in Figure 2 and Figure 3, respectively.

Comparatively, the highest cross-correlation with no lag was observed across all groups in the 2023/2024 analysis, highlighting the evolving dynamics of public responsiveness to air pollution in Macedonia. Additionaly, the consistent cross-correlation results between the Twitter-roBERTa-base and VADER sentiment analysis models suggest their robustness in capturing overall sentiment trends and sensitivity to temporal dynamics, despite variations in sentiment distribution and underlying architectures.

Table 2: Comparison of Cross-correlation Results

Year	**Model**	**All Tweets**	**Positive Tweets**	**Neutral Tweets**	**Negative Tweets**
2023/2024	Twitter-roBERTa-base	0.88	0.81	0.86	0.85
	VADER		0.91	0.91	0.84
2022/2023	Twitter-roBERTa-base	0.65	0.50	0.64	0.68
	VADER		0.50	0.64	0.63
2021/2022	Twitter-roBERTa-base	0.64	0.43	0.66	0.52
	VADER		0.43	0.66	0.64

5.3 Obtained Topics

In addition to using BerTopic for modeling the topics in the current year, the analysis was replicated for the previous two years using the same model to enable a more comprehensive comparison of the obtained topics, as shown in Table 3. Throughout the examined periods, several recurring themes emerged. Heating systems remained a prominent topic, evident in both the previous research and the current study, emphasizing the harmful effects of burning various materials for warmth. Concerns regarding air pollution from cement factories emerged notably in 2022/2023, reflecting worries over industrial emissions. Skopje consistently surfaced as a key topic throughout all three years, highlighting ongoing efforts to address air quality challenges in the capital city. Notably, discourse expanded in 2022/2023 to include air pollution in other municipalities, indicating a broader recognition of regional environmental issues. While the Covid-19 pandemic dominated discussions in the previous two studies, the focus shifted this year to encompass broader health concerns, including polluted water's impact on human health. Topics concerning government environmental regulations and corruption remained prevalent across all years.

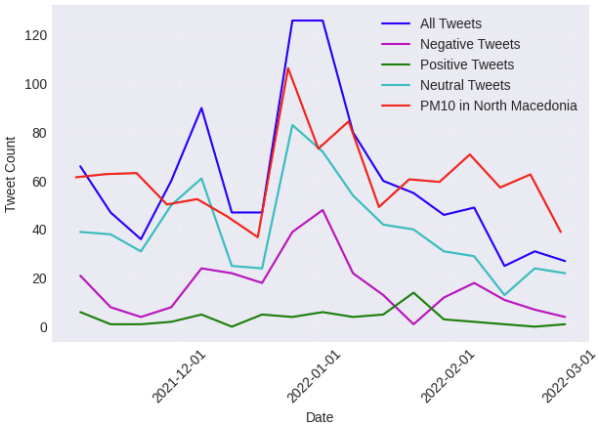

Figure 2: Weekly PM_{10} Data and Macedonian Tweet Frequency - Twitter-roBERTa-base (2021/2022).

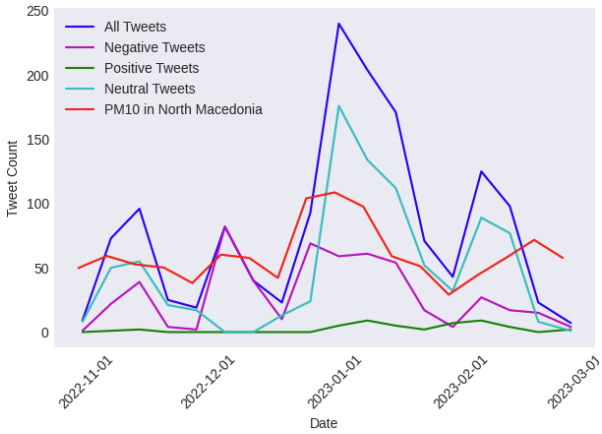

Figure 3: Weekly PM_{10} Data and Macedonian Tweet Frequency Twitter-roBERTa-base (2022/2023).

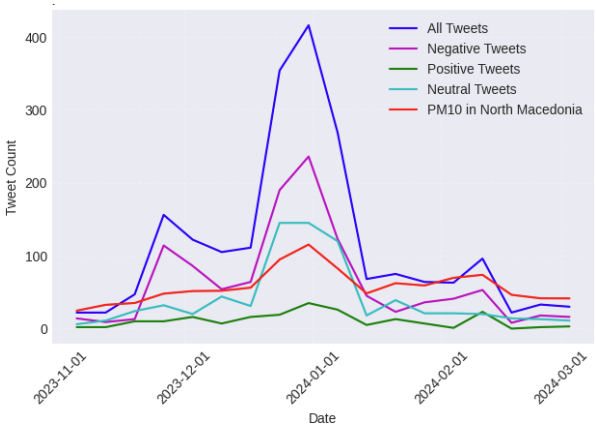

Figure 4: Weekly PM_{10} Data and Macedonian Tweet Frequency Twitter-roBERTa-base (2023/2024).

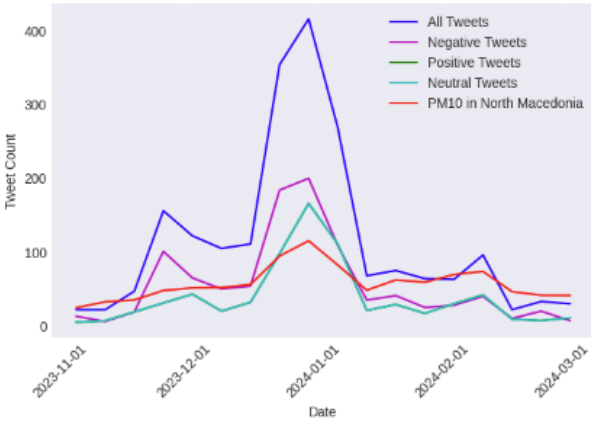

Figure 5: Weekly PM_{10} Data and Macedonian Tweet Frequency - VADER (2022/2023).

Moreover, the electricity crisis was a concern in 2021/2022 and 2022/2023, while transportation emerged as an issue in 2023/2024 and 2021/2022. Detailed topic representations for the years 2021/2022, 2022/2023 and 2023/2024 are illustrated in Figure 6, Figure 7 and Figure 8, respectively.

5.4 Kruskal-Wallis H test

With tweet frequency rising to 2075 in 2023/2024 compared to 1018 in 2021/2022 and 1442 in 2022/2023, the Kruskal-Wallis H test was utilized to examine potential matching trends within each year over the same time frame. The analysis revealed no significant differences in overall tweet counts ($p = 0.382$), as depicted in Figure 9. Similarly, no significant differences were observed in negative tweet counts ($p = 0.694$), as shown in Figure 10. While the difference in neutral tweet counts did not reach statistical significance ($p = 0.009$), as illustrated in Figure 11, there was a notable trend towards variability. However, a notable disparity was observed for positive tweets ($p = 0.003$), indicating varying levels

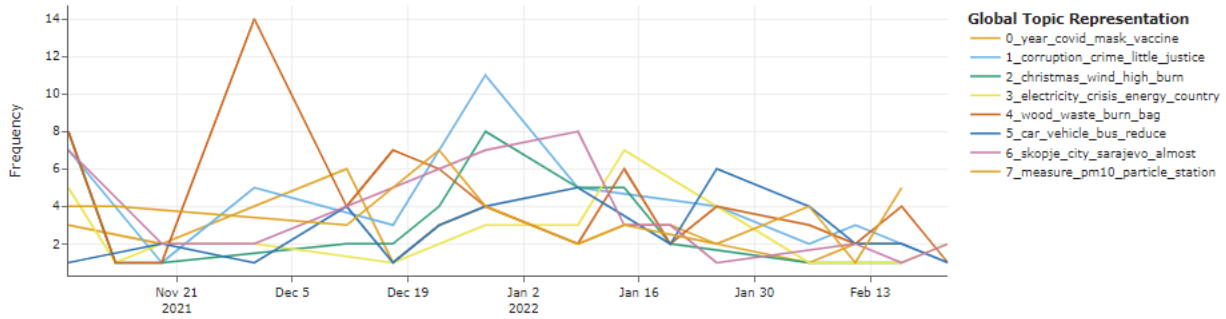

Figure 6: Topics (2022/2023).

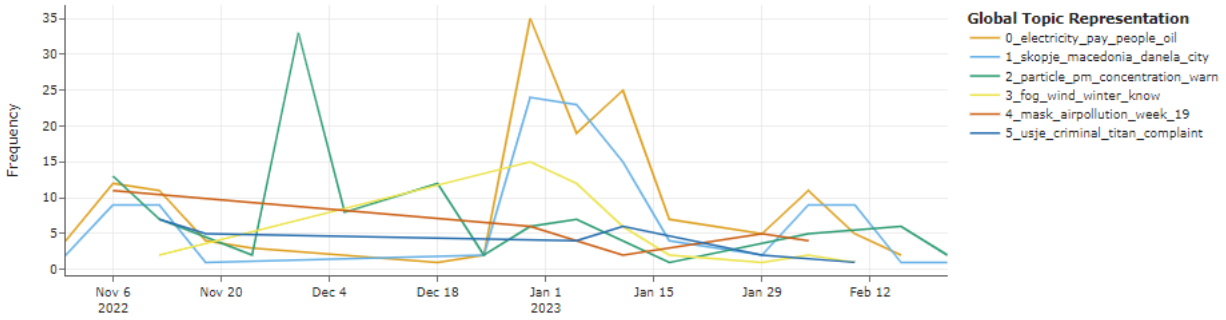

Figure 7: Topics (2022/2023).

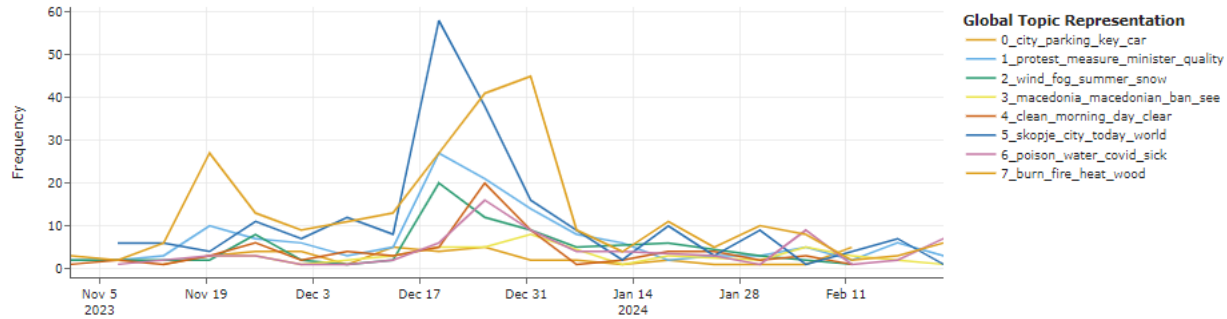

Figure 8: Topics (2023/2024).

Table 3: Comparison of Topic Modeling Results

Topic	**Year**		
	2023/2024	**2022/2023**	**2021/2022**
Domestic heating	✓	✓	✓
Industrial emissions	×	✓	×
Air pollution in urban areas	✓	✓	✓
Health concerns	✓	✓	✓
Electricity crisis	×	✓	✓
Government corruption	✓	✓	✓
Transportation	✓	×	✓

of positivity in discussions over the same time periods, as depicted in Figure 12.

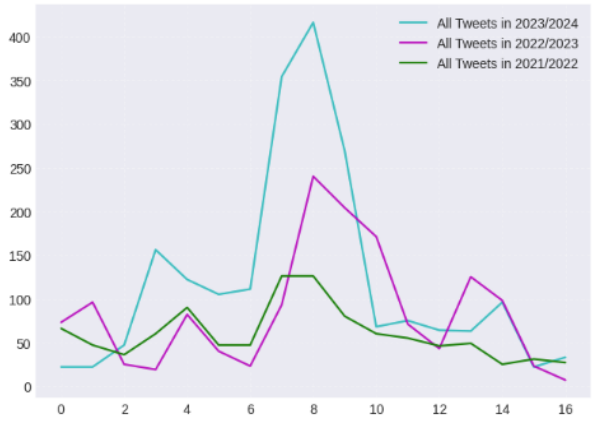

Figure 9: Comparison of the total tweet counts across the three years.

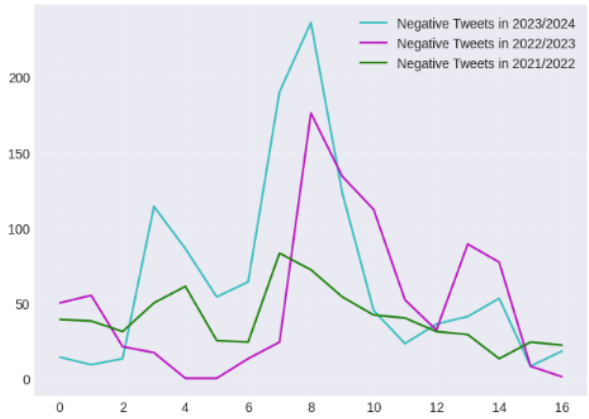

Figure 10: Comparison of the negative tweet counts across the three years.

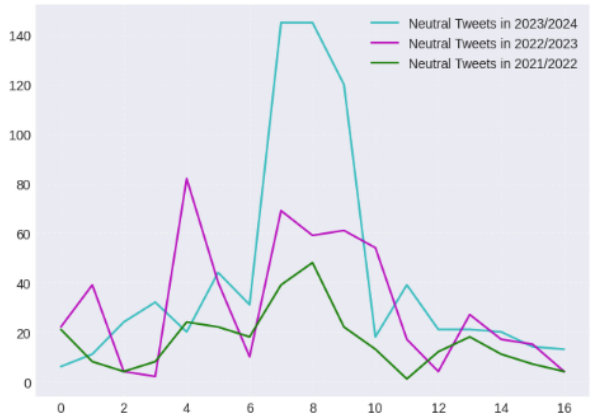

Figure 11: Comparison of the neutral tweet counts across the three years.

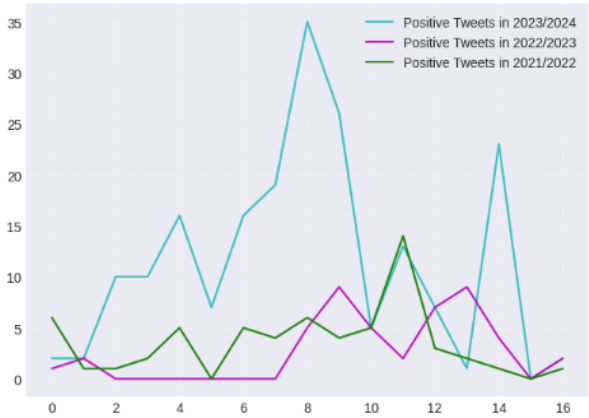

Figure 12: Comparison of the positive tweet counts across the three years.

6. Discussion

The findings from our results reveal a notable increase in awareness among people in Macedonia, as evidenced by the annual rise in the number of tweets. This trend reflects growing engagement and discourse on social media platforms. Additionally, the increase in the percentage of retweets suggests that users are more actively sharing opinions with others, indicating a stronger communal exchange of views on various topics.

Our sentiment analysis, performed using the Roberta method, shows a consistent annual decrease of nearly 5 % in the number of negative tweets. This indicates a positive shift in the general sentiment expressed on Twitter. We chose the RoBERTa method for our analysis due to its demonstrated accuracy in sentiment determination, both in our findings and when tested on manually labeled sentences, as highlighted in previous research (Jain, 2021) (Braig et al., 2023).

Although Twitter-roBERTa-base, which is already fine-tuned on X data was utilized,

additional fine-tuning specific to the dataset was not conducted due to a lack of sufficient data. Previous research suggests that further fine-tuning of the model on task-specific data can significantly enhance the results (Pei et al., 2022). Therefore, for other sentiment analysis tasks with larger datasets, additional fine-tuning should be considered to achieve even better performance and more precise sentiment classification.

Moving to topic modeling, where various approaches to topic extraction exist. However, this discussion will focus on the advantages of BerTopic over GSDMM and LDA, both of which were used in previous research on this topic in Macedonia (Madjar et al., 2023; Perikj et al., 2023). GSDMM assumes that a document belongs to a single topic, whereas LDA assumes that a document contains several topics in varying proportions (Yin & Wang, 2014); Blei et al., 2003. In both GSDMM and LDA, the number of topics is a crucial parameter for obtaining reliable results. GSDMM requires specifying the maximum number of possible topics to find the optimal number, but it faces computational challenges when this parameter is high (Alsmadi et al., 2024). On the other hand, while increasing the number of topics in LDA can enhance the model's predictive power, it tends to decrease semantic consistency (Savin, 2023). BerTopic, on the other hand, does not require predefined clusters and has demonstrated superior topic coherence scores in previous research (Lande et al., 2023; Glazkova, 2021). Additionally, this model can be fine-tuned using a corpus, further enhancing its performance (Grootendorst, 2022.

The application of BerTopic revealed several key topics of concern among residents of Macedonia. Burning practices and air pollution remain significant concerns across Macedonia, particularly in its capital, Skopje. While burning occurs throughout the country, Skopje's status as one of the most polluted cities in the world underscores the urgency of addressing this issue. Situated amidst mountains, Skopje's geographical position makes the pollution problem worse by keeping pollutants trapped in its basin, making them more harmful to people's health and well-being. Residents actively advocate for action, urging municipal authorities and government officials to implement effective measures to combat air pollution. Discussions emphasize the dual challenges faced: the necessity for individuals to resort to burning waste for warmth due to economic constraints, and the resulting environmental and health risks associated with air pollution. Proposed solutions by users also entail implementing filters and enhancing waste management practices as alternatives to burning. While the focus remains on addressing burning practices and air pollution, discussions in the following years touch upon a range of other pressing topics.

In the 2021/2022 period, discussions on various topics were nearly equally frequent, but the most prominent peaks were observed in corruption and crime-related issues during late November and early January. These discussions often demanded justice and proposed organizing campaigns. The economic and electricity crisis also received significant attention, with intensified activity in late December and January correlating with peak pollution levels, indicating heightened public concern.

In the 2022/2023 period, there was a noticeable increase in discussions concerning

air pollution warnings, reflecting a growing awareness of environmental issues. These discussions peaked at the start of the season, serving as a precautionary measure for the public. Despite this, the electricity and economic crises persisted throughout the year, accompanied by discussions on criminal activities, notably focusing on cement factories. The peaks in conversations coincided with elevated pollution levels in late December and January, reaffirming the enduring connection between environmental concerns and public discourse.

In the 2023/2024 period, there was again a sense of revolt among the people, leading to discussions emerging around protest measures and demanding accountability from government ministries. Discussions regarding poisoned water had an increase towards the end of January, highlighting emerging environmental and health risks. While COVID-19 remained a topic of discussion, attention shifted from its prominence in previous years to focus more on the health implications of polluted water. Given the shared impact on respiratory health, it's not surprising that discussions about COVID-19 persist in the context of air pollution, with people feeling its consequences. Once again, peaks in discussions aligned with periods of high pollution, emphasizing the public's awareness of the dangers posed by air pollution.

7. Conclusion

In conclusion, the study provides valuable insights into public opinion towards air pollution in Macedonia. While overall tweet counts constantly rose over the years, there has been a decrease in negative tweets, signaling a shift in sentiment despite the increasing awareness regarding this environmental issue. Moreover, the topic modeling analysis revealed that domestic heating, industrial emissions and transportation are widely perceived as major contributors to air pollution, posing significant health risks. Particularly, Skopje emerges as a focal point of concern within the country, reflecting ongoing efforts to address air quality challenges in the capital city. This study demonstrates the effectiveness of natural language processing tools like sentiment analysis and topic modeling in understanding public sentiments on critical topics like air pollution. By analyzing social media discussions, these methods provide valuable insights into prevailing attitudes and concerns. Furthermore, the correlation between online sentiments and real-world air pollution measurements serves as an indicator of public awareness. Leveraging topic modeling techniques aids in identifying underlying issues in public opinion, facilitating effective problem-solving. In future research, expanding the scope to include additional social media platforms would provide a more comprehensive understanding of public perceptions towards air pollution in Macedonia.

Acknowledgments

A previous version of this research was presented at the 21st International Conference on Informatics and Information Technologies (CIIT 2024). Following constructive feedback, the research scope was extended to encompass a broader range of findings and discussions, as detailed within this paper. This expansion notably enhances the contribution and implications of our study.

References

Almaududi Ausat, A. M. (2023). The role of social media in shaping public opinion and its in-fluence on economic decisions. *Technology and Society Perspectives*, 1(1), 35–44. https://doi.org/10.61100/tacit.v1i1.37

Alsmadi, M., Alzaqebah, M., Jawarneh, S., Almarashdeh, I., Al-Betar, M., Alwohaibi, M., Al-Mulla, N., Ahmed, E., & Al Smadi, A. (2024). Hybrid topic modeling method based on dirichlet multinomial mixture and fuzzy match algorithm for short text clustering. *Journal of Big Data*, 11(1). https://doi.org/10.1186/s40537-024-00930-9

Banja, M., Đukanović, G., & Belis, C. (2020). *Status of air pollutants and greenhouse gases in the western balkans*. Publications Office of the European Union. https://doi.org/10.2760/557210

Bank, W. (2019). *Western balkans regional aqm - western balkans report – aqm in north macedonia* (Last accessed: 11 June 2024). https://openknowledge.worldbank.org/bitstream/handle/10986/33042/Air-Quality-Management-inNorth-Macedonia.pdf?sequence=5&isAllowed=y

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *Findings of the Association for Computational Linguistics*. https://doi.org/https://doi.org/10.48550/arXiv.2010.12421

Belis, C., Matkovic, V., Ballocci, M., Jevtic, M., Millo, G., Mata, E., & Van Dingenen, R. (2023). Assessment of health impacts and costs attributable to air pollution in urban areas using two different approaches. a case study in the western balkans. *Environment International*, 182, 108347. https://doi.org/10.1016/j.envint.2023.108347

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3, 993–1022. https://doi.org/http://dx.doi.org/10.1162/jmlr.2003.3.4-5.993

Braig, N., Benz, A., Voth, S., Breitenbach, J., & Buettner, R. (2023). Machine learning techniques for sentiment analysis of covid-19-related twitter data. *IEEE Access*, PP, 1–1. https://doi.org/10.1109/ACCESS.2023.3242234

CAN. (2019). *Report: Eu action on western balkans' chronic coal pollution is a unique opportunity to improve health and productivity* [Accessed 7 March 2024]. https://caneurope.org/report-eu-action-on-western-balkans-chronic-coal-pollution-is-a-unique-opportunity-to-improve-health-and-productivity/

Daul, M. Č., Krzyzanowski, M., & Kujundžić, O. (2021). *Air pollution in the western balkans: Key messages for policymakers and the public* [Accessed 7 March 2024]. United Nations Environment Programme. https://zoinet.org/wp-content/uploads/2022/02/Pollution-Balkans-EN2.pdf

Glazkova, A. (2021). Using topic modeling to improve the quality of age-based text classification. *CEUR*, 2930. https://api.semanticscholar.org/CorpusID:248743508

Grootendorst, M. R. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, 2203.05794. https://doi.org/https://doi.org/10.48550/arXiv.2203.05794

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, 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Institute, H. E. (2020). *State of global air 2020* (Last accessed: 11 June 2024). https://www.stateofglobalair.org/sites/default/files/documents/2022-09/soga-2020-report.pdf

Jain, N. (2021). Customer sentiment analysis using weak supervision for customer-agent chat. *ArXiv*, 2111.14282v2. https://doi.org/https://arxiv.org/pdf/2111.14282

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. *of the 9th WebKDD and 1st SNA*, 43, 56–65. https://doi.org/10.1145/1348549.1348556

Jorgensen, E., & Shkaratan, M. (2020). *Former yugoslav republic of macedonia - green growth country assessment*. World Bank Group. http://documents.worldbank.org/curated/en/949621468090285546/Former-Yugoslav-Republic-of-Macedonia-Green-growth-country-assessment

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association*, 47, 583–621. https://doi.org/https://doi.org/10.1080/01621459.1952.10483441

Kruspe, A., Kersten, J., & Klan, F. (2021). Review article: Detection of actionable tweets in crisis events. *Nat. Hazards Earth Syst. Sci*, 21, 1825–1845. https://doi.org/https://doi.org/10.5194/nhess-21-1825-202

Lande, J., Pillay, A., & Chandra, R. (2023). Deep learning for covid-19 topic modelling via twitter: Alpha, delta and omicron. *PLoS One*, 18(8). https://doi.org/https://doi.org/10.1371/journal.pone.0288681

Loureiro, D., Barbieri, F., Neves, L., Espinosa Anke, L., & Camacho-collados, J. (2022). Timelms: Diachronic language models from twitter. *Proceedings of the 60th Annual*

Meeting of the Association for Computational Linguistics: System Demonstrations, 251–260. https://doi.org/https://doi.org/10.48550/arXiv.2202.03829

Madjar, A., Gjorshoska, I., Prodanova, J., Dedinec, A., & Kocarev, L. (2023). Western balkan societies' awareness of air pollution. estimations using natural language processing techniques. *Ecological Informatics*, 75, 102097. https://doi.org/10.1016/j.ecoinf.2023.102097

Mcgregor, S., & Molyneux, L. (2018). Twitter's influence on news judgment: An experiment among journalists. *Journalism*, 21(7), 146488491880297. https://doi.org/10.1177/1464884918802975

Mirakovski, D., Boev, B., Boev, I., & et al. (2019). Wintertime urban air pollution in macedonia – composition and source contribution of air particulate matter. *Proceedings of the 18th World Clean Air Congress*, 492–500.

Mirakovski, D., Zendelska, A., Boev, B., & et al. (2024). Evaluation of pm2.5 sources in skopje urban area using positive matrix factorization. *Environmental Modeling & Assessment*.

NECE, U. (2016). *Clean air for life, europe* (Last accessed: 11 June 2024). https://unece.org/sites/default/files/2021-06/Clean-air-for-life_eng.pdf

OECD, O. (2012). *Oecd environmental outlook to 2050: The consequences of inaction india*.

of environment, M., & physical planning - Republic of North Macedonia. (n.d.). *Air quality portal* [Accessed 7 March 2024]. https://air.moepp.gov.mk/?page_id=175

Pei, Y., Mbakwe, A., Gupta, A., Alamir, S., Lin, H., Liu, X., & Shah, S. (2022). TweetFinSent: A dataset of stock sentiments on Twitter. *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, 37–47. https://doi.org/10.18653/v1/2022.finnlp-1.5

Perikj, T., Dedinec, A., & Prodanova, J. (2023). Comparative analysis of air pollution-related tweets and news article teasers. *Proceedings of the 20th International Conference on Informatics and Information Technologies*. https://doi.org/http://hdl.handle.net/20.500.12188/27393

Pittet, D., Allegranzi, B., Storr, J., Nejad, S. B., Dziekan, G., Leotsakos, A., & Donaldson, L. (2008). Infection control as a major world health organization priority for developing countries. *Journal of Hospital Infection*, 68(4), 285–292. https://doi.org/https://doi.org/10.1016/j.jhin.2007.12.013

Roesslein, J. (2020). *Tweepy: Twitter for python!* [Accessed 20 May 2024]. https://github.com/%20tweepy/tweepy

Rosenberg, E., Tarazona, C., Mallor, F., Eivazi, H., Pastor-Escuredo, D., Fuso-Nerini, F., & Vinuesa, R. (2023). Sentiment analysis on twitter data towards climate action. *Results in Engineering*, 19.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. *Proceedings of the 19th International Conference on World Wide Web*, 851–860. https://doi.org/10.1145/1772690.1772777

Savin, I. (2023). Evolution and recombination of topics in technological forecasting and social change. *Technological Forecasting and Social Change*, 194. https://doi.org/https://doi.org/10.1016/j.techfore.2023.122723

Sheng Yue, C. Y. W. (2002). Applicability of prewhitening to eliminate the influence of serial correlation on the mann-kendall test. *Water Resources Research*, 38. https://doi.org/10.1029/2001WR000861

Southerland, V. A., Brauer, M., Mohegh, A., Hammer, M. S., van Donkelaar, A., Martin, R. V., Apte, J. S., & Anenberg, S. C. (2022). Global urban temporal trends in fine particulate matter (pm2·5) and attributable health burdens: Estimates from global datasets. *The Lancet. Planetary health*, 6(2). https://doi.org/https://doi.org/10.1016/S2542-5196(21)00350-8

Translator, D. (n.d.). https://deep-translator.readthedocs.io/en/latest/

TS, K., R, M., VP, R., & VP., N. (2023). Assessment of urban air quality from twitter communication using self-attention network and a multilayer classification model. *Environ Sci Pollut Res Int.*, 30(4), 10414–10425.

Wang, Z., Zhao, W., Wang, B., Liu, J., Xu, S., Zhang, B., Sun, Y., Shi, H., & Guan, D. (2022). Environmentally vulnerable or sensitive groups exhibiting varying concerns toward air pollution can drive government response to improve air quality. *iScience*, 25(6).

Welsh, W. (1999). On the reliability of cross-correlation function lag determinations in active galactic nuclei. *The Publications of the Astronomical Society of the Pacific*, 111(765), 1347–1366. https://doi.org/10.1086/316457

X. (n.d.). *X api v2* [Accessed 20 May 2024]. https://developer.twitter.com/en/support/x-api/v2

Yin, J., & Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering, 233–242. https://doi.org/10.1145/2623330.2623715

Zheng, S., Wang, J., Sun, C., Zhang, X., & Kahn, M. (2019). Air pollution lowers chinese urbanites' expressed happiness on social media. *Nat Hum Behav*, 3(3), 237–243.