Investigating Public Awareness of Air Pollution in Western Balkans by analyzing Tweets and News Article Teasers

Angela Madjar
Macedonian Academy of Sciences and Arts
Skopje, Republic of Macedonia
angela.madzhar@students.finki.ukim.mk

Ivana Gjorshoska
Macedonian Academy of Sciences and Arts
Skopje, Republic of Macedonia
ivana.gjorshoska@students.finki.ukim.mk

Jana Prodanova
Macedonian Academy of Sciences and Arts
Skopje, Republic of Macedonia
jprodanova@manu.edu.mk

Aleksandra Dedinec
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius University
Skopje, Republic of Macedonia
aleksandra.kanevche@finki.ukim.mk

***Abstract*—Air pollution is a serious threat to the health of people living in Western Balkans, with the biomass being one of the main pollutants. This study is based on the assumption that air pollution escalation in Western Balkan countries during winter will provoke a more intense activity on Twitter, while acknowledging that people's opinions and feelings can often be influenced by content presented in news articles. The objective of this study is to investigate public awareness of air pollution in Macedonia, Serbia, Bosnia and Herzegovina and Montenegro. Natural Language Processing techniques such as Sentiment Analysis and Topic Modeling, as well as Statistical Analysis are employed to determine whether Twitter discussions regarding air pollution reflect the PM_{10} levels measured by official air monitoring stations in these countries. Such analyses are also performed on news article teasers, attempting to determine whether mass media portrays a realistic ambient condition and aims to promote pro-environmental behavior. The results of this study suggest that using Correlation Analysis to check for resemblance between the frequency of sentiments detected in social media discussions and the temporal changes in the PM_{10} concentration in the air, can serve as a measure of public awareness of air pollution. Analyzing the content of the tweets can reveal issues in the public opinion and thus, contribute to tackling them down.**

Keywords—air pollution, pm10, western balkans, social media, tweets analysis, news media, sentiment analysis, topic modelling, cross correlation

I. Introduction

Air pollution is a major problem in all Western Balkan countries. 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 [1]. 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 comparable 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 [2]. PM_{10} is among the air pollutants whose levels are most frequently above the legislation limits in this region and are mainly emitted by human activities such as industry, household heating and transport [3][4]. Uncontrolled urbanization, illegal construction, poor planning and design also endanger the environment by contributing to such ambient condition [5]. Additionally, transboundary pollution from within and outside the region considerably bring about the observed concentrations [3]. However, the Health and Environment Alliance (HEAL) has reported that the main sources of air pollution in the Western Balkans are the production of electricity from coal and using wood for heating households [6]. Consequently, the daily PM_{10} limit of 40μg/m3 set out under national legislation is exceeded between 120 and 180 days a year, mostly during wintertime [1]. Exposure to such particles can affect the heart and lungs and cause serious health effects [7].

For this study, we supposed that the air pollution escalation in Western Balkans during winter will provoke a more intense activity on social media, especially on Twitter, where people share their opinions and feelings. However, the way people feel and think about a certain topic can often be influenced by sentiments expressed in news articles. Knowing that exposure to PM_{10} particles harms the human health, our initial hypotheses were that the public has justified negative sentiments towards air pollution in their country and that news media reflects the real air pollution ambient, aiming to promote pro-environmental behavior.

To test our hypotheses and to investigate public perception of air pollution in several Western Balkan countries including Macedonia, Serbia, Bosnia and Herzegovina and Montenegro, we employ Natural Language Processing techniques such as Sentiment Analysis and Topic Modeling, as well as Statistical Analyses.

II. Related work

Analyses of micro-blogging posts referring air pollution has been an appealing research field for many researchers over the 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 dynamics of air pollution to some extent. Such messages can reveal insights about public perceptions and concerns of air pollution [8]. Another study demonstrates the feasibility of using social media to monitor $PM_{2.5}$ levels as an alternative method in areas without air pollution monitoring systems, showing that citizen-led monitoring can be used to better understand public's interaction with air quality issues and use Twitter discussions to promote pro-environmental behavior. The authors minimized media influence by filtering

tweets containing URLs and compared the frequency of different sentiment groups against the official $PM_{2.5}$ data in Greater London [9]. Moreover, public response to a particular event, such as wildfire, can be captured by using a structural topic model to extract topics from tweets posted during the event, as demonstrated in [10]. As indicated in [11], 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 public response to air quality information.

III. DATA

A. Twitter Data

For the purpose of this study, we coded a searching application using the Python library Tweepy [12] to collect air quality-related tweets from November 1st, 2021 to February 28th, 2022. We used Twitter's Standard API [13], which allows searching for tweets posted within the last 7 days prior to the search time. Thus, tweets were collected on weekly basis using the keywords: "aerozagaduvanje" (air pollution), "аерозагадување" (air pollution), "zagaduvanje" (pollution), "загадување" (pollution), "пм10" (pm10), "дишеме" (we breathe) to capture tweets written in Macedonian language. To collect Western Balkan tweets (excluding tweets written in 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.

B. News Media data

Parallel to the collection of tweets, the powerful web-crawler tool Octoparse [14] was employed to weekly assemble teaser texts of news articles containing the above-mentioned keywords. A teaser is an illustrative short reading suggestion for an article that entices potential readers to read particular news items [15]. We crawled the news web sites Time.mk [16] and Time.rs [17], which are cluster-based news aggregators that analyze 15 000 news daily, collected from 120 distinct sources.

C. Official Air pollution data

Air pollution data was acquired in order to investigate the frequency of tweets and news articles during the peaks and falls of PM_{10} particles measured by official measuring stations. The PM_{10} data in Macedonia [18], Serbia [19], Montenegro [20] and Bosnia and Herzegovina [21] was collected from 21, 37, 6 and 16 monitoring sites respectively, owned and funded by local authorities. The hourly data measured from November 1st, 2021 to February 28th, 2022 was aggregated by week, so that the air pollution data is adjusted to the frequency of collection of Tweets and teasers. For each country, data from all of its monitoring stations was aggregated to encompass the entire country area. PM_{10} data measured in Serbia was available from December 2nd, 2021 to February 28th, 2022.

IV. METHODOLOGY

A. 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 [22]. VADER relies on an English dictionary that maps lexical features to emotion intensities called sentiment or Valence scores. Valence scores of each word are measured on a scale from - 4 (most negative) to + 4 (most positive), with 0 indicating a neutral sentiment. The compound score of the whole text is obtained by summing the valence scores of each word in the lexicon, normalized to be between -1 (most extreme negative) and +1 (most extreme positive) by using the following normalization:

$$x = \frac{x}{\sqrt{x^2+\alpha}} \quad (1)$$

In (1), x is the sum of Valence scores of constituent words and α 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 [23] to translate the collected data into English language.

B. Time Series Statistical Analysis

In terms of time series analysis, measuring similarity is important to assess the casual relationship between two signals in time. To compare the tweets and teasers against the PM_{10} data, we used cross-correlation. The Cross Correlation Function (CCF) is the correlation between the observations of two time series x_t and y_t, separated by k time units (the correlation between y_{t+k} and x_t), where k is called a lag. The confidence interval is calculated with $\pm\frac{2}{\sqrt{n-|k|}}$, where n is the number of observations and k is the lag. The correlation is significant if its absolute value is greater than $\frac{2}{\sqrt{n-|k|}}$ [24]. 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 [25].

To test whether the obtained sentiment groups of tweets and teasers and the PM_{10} data are stationary, we used the nonparametric Mann-Kendall test. The null hypothesis is that no monotonic trend is present in the data. The test indicated a decreasing trend in the teasers obtained from Time.rs. Prior to performing the CCF test, we conducted first order differencing to make this data stationary. For the rest of the data, the Mann-Kendall test did not indicate any trends and thus, the null hypothesis was accepted.

C. Topic Modelling

To understand what contributes to the air pollution and who is accountable according to the public opinion, we experimented with two topic modelling approaches: Latent Dirichlet Allocation (LDA) [26] and Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) [27].

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 [28]. Ranked lists of words associated with a given word w_n are obtained by calculating the sum of the weight of each topic generated by LDA multiplied by the weight of each word w_n contained in that topic. The ranking weight of the word i is computed as follows:

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

In (2), N is the number of topics and w_{ij} denotes the weight of the word i in topic j.

Contrary to the LDA's assumption, GSDMM assumes 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 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 hypothesis that states that words with similar meaning tend to co-occur within a similar context [29]. 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 preprocessed the negative tweets to remove retweet symbols, special characters, URLs, emojis and extra spaces. To give more focus to the important information, we also removed stopwords, tokenized the tweets and reduced the words to their lemma.

V. Results

A. Obtained sentiments

The statistics about the tweets and the news article teasers collected during the course of this 17-week study, as well as the sentiments obtained with the Sentiment Analysis for all of the datasets, are displayed in Table I. 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 [30]. 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 preprocessed content (without retweet symbols and special characters) are considered as unique. Similarly, media teasers with distinct content count as unique.

TABLE I. Statistics of the collected air pollution tweets and news article teasers

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

B. Cross-correlation

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

The maximum cross-correlation between all categories of Macedonian air pollution tweets and PM_{10} particles in the country was at lag 0, indicating that there is no lag or lead-time between levels of PM_{10} particles and tweet frequency. A cross-correlation analysis between the number of negative tweets and PM_{10} particles showed a coefficient of 0.62 ($p = 0.001$) and a coefficient of 0.54 ($p < 0.001$) between the number of neutral tweets and PM_{10} particles (Fig. 1). The cross-correlation between the positive tweets and PM_{10} particles was insignificant. Regarding the sentiment groups of Time.mk teasers, the maximum cross-correlation between the frequency of negative teasers and PM_{10} data was 0.53 ($p = 0.002$), between the positive teasers and PM_{10} data was 0.52 ($p < 0.001$) (Fig. 2) both at lag 0; while between the neutral teasers and PM_{10} data it was insignificant.

There was no significant cross-correlation between the Balkan air pollution tweets and any of the PM_{10} data measured in Serbia, Bosnia and Herzegovina and Montenegro. As for the Time.rs teasers, there was significant cross-correlation at lag 0 between the negative teasers and PM_{10} data measured in Serbia, with a coefficient of 0.66 ($p < 0.001$) (Fig. 3) and at lag 3 between the neutral teasers and PM_{10} data measured in Montenegro with a coefficient of 0.65 ($p < 0.001$) (Fig. 4).

C. Obtained topics

Topic Modelling is a challenging NLP task and choosing the "best model" is not an exact science. Context, and translation in our study, make for a wide range of possible meanings that can be best interpreted by human eye.

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 with different hyperparameters. With a small number of topics, we only extracted broad topics, achieving 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 is probably wrong.

According to the analysis, for GSDMM on the Macedonian negative tweets, 9 topics were selected with a coherence score of 0.51, while for LDA 15 topics (coherence score = 0.41).

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).

In other studies, it has been shown that traditional topic models such as LDA, experience large performance degradation over short texts [29] [31], 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 II, along with the most important words and their frequency of occurrence.

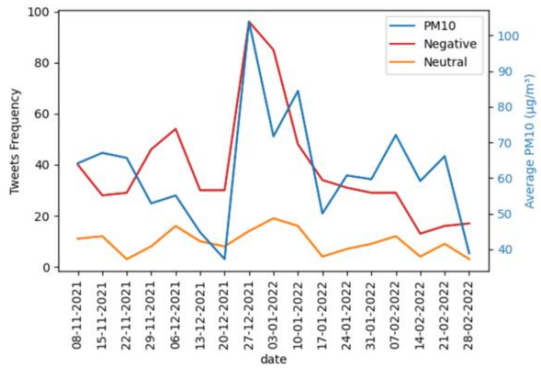

Fig. 1. Weekly comparison of official PM_{10} data in Macedonia and frequency of Macedonian Tweets

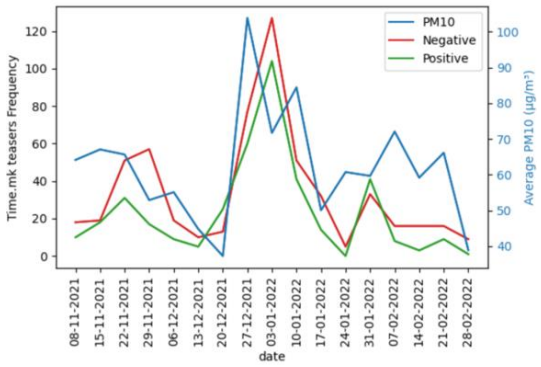

Fig. 2. Weekly comparison of official PM_{10} data in Macedonia and frequency of Time.mk teasers

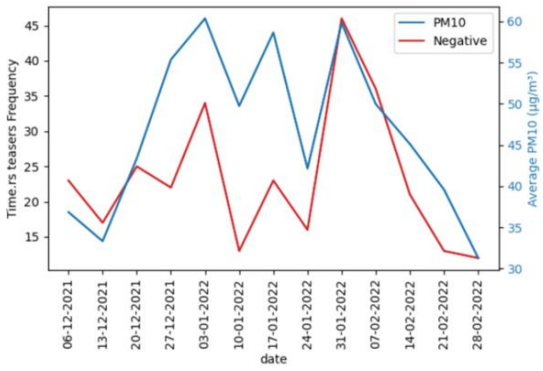

Fig. 3. Weekly comparison of official PM_{10} data in Serbia and frequency of Time.rs teasers

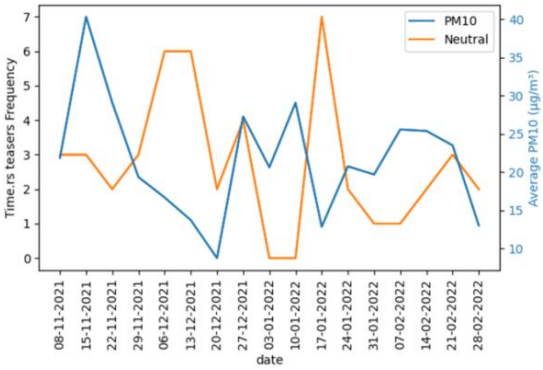

Fig. 4. Weekly comparison of official PM_{10} data in Montenegro and frequency of Time.rs teasers

VI. Discussion and Conclusion

The analyses showed that negative Macedonian tweets were most predictive of PM_{10} levels in the country, whereas positive tweets do not have comparable peaks and falls with the measured PM_{10} particles. This suggests that during times of escalated air pollution, the public expresses negative sentiments and concerns on Twitter, which is in accordance with our first hypothesis that Twitter discussions about air pollution reflect measurements of PM_{10} particles.

As for the News Media in Macedonia, the maximum cross-correlation was detected between negative teasers and PM_{10} data, which again, is consistent with our second hypothesis that News Media provides truthful information on air pollution. However, the high cross correlation of positive teasers and PM_{10} data should not be neglected, since it implies the possibility that news media tries to suppress public discussions about air pollution hazards.

The topic modelling indicated that the Macedonian public thinks that electricity production and transport lead to great air pollution, and thus, to health issues. They expect the government to take action and address the problem. However, studies conducted in Macedonia show that the ratio of air pollution caused by households and transport in the country is about 95% to 5% [32]. The experimenting with different numbers of topics revealed that some discussions about the negative 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 a highest coherence score, while the ambient condition is often attributed to the transport. These results suggest high 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 to promote pro-environmental behavior on this topic.

Contrary to the first hypothesis about Twitter discussions reflecting PM_{10} concentration in the air, there was no resemblance between any of the sentiment groups of Western Balkan Tweets and the actual PM_{10} data collected from Serbia, Bosnia and Herzegovina and Montenegro. However, most of the tweets were classified as negative, which suggests that people express negative sentiments on Twitter even when the PM_{10} levels are moderate or low.

Although the number of articles regarding air pollution decreases over time, a high cross-correlation was detected between negative Time.rs teasers and Serbian PM_{10} data, which again, confirms our second hypothesis about mass media portraying a realistic ambient condition.

Throughout the topic modeling, we discovered that many air pollution tweets refer to Serbia and cities in Serbia (Belgrade, Sabac, etc.). When tweeting on air pollution, the public generally expresses worries about health, climate change, industrial air pollution and pollution caused by power plants. Again, this indicates public unawareness of the biomass being a major air pollutant in the Western Balkans [6].

TABLE II. TOPICS OBTAINED THROUGH GSDMM TOPIC MODELLING

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)	Industrial air pollution	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), caugh (3), heart (3), symptom (3)

Future research directions for this study include a more supervised translation of the collected data to improve VADER resistance to slang and sarcasm. Additionally, using geo-localized tweets in the future would eliminate the possibility of capturing Croatian tweets when using keywords for collecting Western Balkan tweets. Finally, separate analysis of the data obtained from each measuring station could provide insights on air pollution in smaller regions in every country, since levels of PM_{10} vary largely over space and time.

Collectively, this study confirms the convenience of NLP techniques such as Sentiment Analysis and Topic Modelling for analyzing public thoughts and feelings on important topics such as air pollution, as well as the truthfulness 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. Topic Modelling techniques can reveal issues in the public opinion and thus, contribute to tackling down such problems.

ACKNOWLEDGMENT

The research presented in this article was conducted during an internship in the Macedonian Academy of Sciences and Arts.

REFERENCES

[1] Colovic Daul, M., M. Kryzanowski, and O. Kujundzic. "Air Pollution and Human Health: The Case of the Western Balkans." UN Environ (2019).

[2] Helotonio, Carvalho. "Air pollution-related deaths in Europe – time for action." Journal of Global Health 9.2 (2019).

[3] Banja, M., G. Đukanović, and C. A. Belis. "Status of air pollutants and greenhouse gases in the Western Balkans." Publications Office of the European Union, EUR 30113 (2020): 1-53.

[4] Meisner, Craig, Dragan Gjorgjev, and Fimka Tozija. "Estimating health impacts and economic costs of air pollution in the Republic of Macedonia." South Eastern European Journal of Public Health (SEEJPH) (2015).

[5] Jovanovic, Mica. "Environmental Impact of Illegal Construction, Poor Planning and Design in Western Balkans: A Review."

[6] I. Todorović, "HEAL: Biomass is one of main sources of air pollution in Western Balkans", Balkan Green Energy Group, 27-Jan-2022. Available at: https://balkangreenenergynews.com/heal-biomass-is-one-of-main-sources-of-air-pollution-in-western-balkans/#:~:text=The%20European%20Environmental%20Agency%20estimated,challenges%20around%20improving%20woodburning%20technology. [Accessed: 13-Mar-2022]

[7] Environmental Health, "Particulate matter (PM10 and PM2.5)", Environmental Health, 25-Nov-2020. Available at: https://www.health.nsw.gov.au/environment/air/Pages/particulate-matter.aspx. [Accessed: 13-Mar-2022]

[8] Jiang, Wei, et al. "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 (2015): e0141185.

[9] Hswen, Yulin, et al. "Feasibility of using social media to monitor outdoor air pollution in London, England." Preventive Medicine 121 (2019): 86-93.

[10] Sachdeva, Sonya, and Sarah McCaffrey. "Using social media to predict air pollution during California wildfires." Proceedings of the 9th International Conference on Social Media and Society. 2018.

[11] Gurajala, Supraja, Suresh Dhaniyala, and Jeanna N. Matthews. "Understanding public response to air quality using tweet analysis." Social Media+ Society 5.3 (2019): 2056305119867656.

[12] J. Roesslein, "Tweepy: Twitter for Python!", 2020. Available at: https://github.com/tweepy/tweepy. [Accessed: 10-Oct-2021]

[13] Twitter, "Standard v1.1", Available at: https://developer.twitter.com/en/docs/twitter-api/v1. [Accessed: 10-Oct-2021]

[14] Almaqbali, Iqtibas Salim Hilal, et al. "Web Scrapping: Data Extraction from Websites." Journal of Student Research (2019).

[15] Karn, Sanjeev Kumar, et al. "News Article Teaser Tweets and How to Generate Them." arXiv preprint arXiv:1807.11535 (2018).

[16] I. Trajkovski, "How does TIME.mk work?", Time.mk, 2008. Available at: https://time.mk/info/site, [Accessed: 01-Nov-2021]

[17] I. Trajkovski, Time.rs, 2008. Available at: https://time.rs/, [Accessed: 01-Nov-2021]

[18] Ministry of environment and physical planning – Republic of North Macedonia, "Air Quality Portal". Available at: https://air.moepp.gov.mk/?page_id=175, [Accessed: 01-Mar-2022]

[19] Republic of Serbia, Open Data Portal, "Air quality - unverified real-time clock data". Available at: https://data.gov.rs/sr/datasets/kvalitet-vazduha/, [Accessed: 01-Mar-2022]

[20] Environmental Protection Agency of Montenegro, "Measurement data archive". Available at: http://www.epa.org.me/vazduh/arhiv/7, [Accessed: 01-Mar-2022]

[21] Discomap EEA, "Download of air quality data". Available at: https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm, [Accessed: 01-Mar-2022]

[22] Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the international AAAI conference on web and social media. Vol. 8. No. 1. 2014.

[23] Online Doc Translator, 2021. Available at: https://www.onlinedoctranslator.com/en/, [Accessed: 01-Nov-2021].

[24] Minitab, “Interpret all statistics and graphs for Cross Correlation”, 2022. Available at: 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/, [Accessed: 15-Mar-2022]

[25] Dean, Roger T., and William Dunsmuir. "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 (2016): 783-802.

[26] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3 Jan (2003): 993-1022.

[27] Yin, Jianhua, and Jianyong Wang. "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. 2014.

[28] Korzycki, Michał, Izabela Gatkowska, and Wiesław Lubaszewski. "Can the human association norm evaluate machine-made association lists?." Cognitive Approach to Natural Language Processing. Elsevier, 2017. 21-40.

[29] Syed, Shaheen, and Marco Spruit. "Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation." 2017 IEEE International conference on data science and advanced analytics (DSAA). IEEE, 2017.

[30] Sharifi, Zahra, and Sajjad Shokouhyar. "Promoting consumer's attitude toward refurbished mobile phones: A social media analytics approach." Resources, Conservation and Recycling 167 (2021): 105398.

[31] Qiang, Jipeng, et al. "Short text topic modeling techniques, applications, and performance: a survey." IEEE Transactions on Knowledge and Data Engineering (2020).

[32] G. Kanevce, A. Dedinec, V. Taseska-Gjorgievska, A.Dedinec. “Transport in Skopje – realities and challenges, Path to green transport”, Third Biennial Update Report on Climate Change, 2017. Available at: https://api.klimatskipromeni.mk/data/rest/file/download/71dea57f28b54b8c5f35e41a364a586d58c97edef47aacf36268b5d4296667ec.pdf, [Accessed: 28-Mar-2022]