Comparative analysis of Air Pollution-related Tweets and News Article Teasers

Teodora Perikj, Aleksandra Dedinec and Jana Prodanova
Faculty of Computer Science and Engineering Ss. Cyril and Methodius University, Skopje, Republic of Macedonia
Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia

teodora.perikj@students.finki.ukim.mk
aleksandra.kanevche@finki.ukim.mk
jprodanova@manu.edu.mk

Abstract—**Air pollution is a consequence of both natural sources and human activities in the environment, which have a negative impact on the atmosphere. Air pollution is a serious environmental problem and poses a serious threat to the health of people in Macedonia. The purpose of the study is to investigate the society's awareness of air pollution in Macedonia, compared with the findings from previous research. This study's objective is to explore whether there is a correlation between tweets related to air pollution, teasers related to air pollution, and measured values of PM_{10} particles and if there is a difference in society's awareness of air pollution compared to last year's situation. To analyze our assumptions, we used the Natural Language Processing techniques: sentiment analysis and topic modeling, along with statistical analysis, fisher's z-test and correlation analysis. The obtained results show us what feelings people express towards air pollution, what topics they talk about and help us determine if society's awareness has increased since last year.**

Keywords—air pollution, pm10, sentiment analysis, cross correlation, topic modeling, fisher's z-test, news media, tweets analysis

I. Introduction

Air pollution is a global health and environmental problem. It is estimated that, due to exceeded levels of air pollutants, people living in the countries of the Western Balkans, lose up to 1.3 years of life [1]. The primary sources of air pollution in Macedonia are: the process of electricity production using solid fuels, emissions from traffic, burning unclean fuels for heating in homes, and industrial production processes [2]. One of the air pollutants whose daily limit of 50µg/m3, established by national legislation, often exceeded in Macedonia, is PM_{10} particles. The purpose of our research is to investigate what is the awareness of society regarding air pollution compared to last year. Due to the negative effects of air pollution on human health, for this study we aim to explore if there is a correlation between air pollution-related tweets, teasers, PM_{10} particles measurements, and if there is a difference in the public awareness regarding air pollution. For such analyzes, we used the Natural Language Processing (NLP) techniques: sentiment analysis and topic modeling, along with statistical analysis and fisher's z-test.

II. Related Work

In recent years, an increasing emphasis is placed on the analysis of air pollution, the causes of its occurrence, its impact on health, and the ways to reduce the impact. NLP techniques are often used to analyze micro-blogging posts related to air pollution. Such research was conducted using comments from tourists, related to air quality for 195 tourist destinations in China, posted on the Sina Weibo website [3]. The purpose of this research was to get an idea of the perceptions of tourists regarding air quality, which will contribute to the study of the satisfaction of the tourist experience. To obtain the sentiments, the comments were subjected to content analysis and then an artificial neural network was used. Another study used tweets related to air pollution in the five most polluted states in the USA, to improve air pollution prediction [4]. The tweets were filtered using word selection and topic modeling, and then to improve the prediction, a convolutional network and over-tweet-pooling were applied. These analyzes were our guide to setting up the first three research questions (RQs), namely: RQ1: Is there a correlation between air pollution-related tweets and teasers? RQ2: Is there a correlation between air pollution-related tweets and measurements of PM_{10} particles? RQ3: Is there a correlation between air pollution-related teasers and measurements of PM_{10} particles? The fourth research question compares the first 3 RQs with data from last year's investigation [5]. RQ4: Is there a difference in society's awareness of air pollution compared to last year. Experimenting with different tools and learning methods can help us determine if there are any similarities or differences in public response to air pollution.

III. Data

A. Twitter Data

For the aims of this study, we collected Macedonian tweets related to air pollution, on a weekly basis from October 24th, 2022 to February 28th, 2023. Tweets were collected using the Python library Tweepy [6] and Twitter's Standard API [7], in a searching application. In order to collect only tweets written in Macedonian language, we used the keywords: "aerozagaduvanje" (air pollution), "аерозагадување" (air pollution), "zagaduvanje" (pollution), "загадување" (pollution), "пм10" (pm10) and "дишеме" (we breathe). Also, we concluded that twitter users refer to the terms: "pollution" and "air pollution" interchangeably.

B. News Media Data

While collecting tweets, we also collected teaser texts of news articles on a weekly basis. A teaser is a short reading suggestion for an article that is illustrative and includes elements that arouse curiosity to entice potential readers to read particular news items [8]. The teasers were gathered using the Python library Beautiful Soup [9], which pulls data out of HTML and XML files, which is useful for web scaping, and because of that we scraped the news web site Time.mk [10], which is cluster-based news aggregator, and selected only the teasers containing the keywords mentioned above.

C. Official Air Pollution Data

The PM_{10} data in Macedonia were collected from 20 monitoring sites, that are owned and funded by local authorities [11]. The daily data measured by official measuring stations, were weekly aggregated to adjust for collected tweets and teasers. This type of data were collected in order to investigate whether the tweets and teasers reflect the levels of PM_{10} particles. Regarding the PM_{10} data in the period around the New Year, some data from the measuring stations are missing due to the installation of new server equipment and the migration of systems and databases from the old to the new equipment [12].

D. Previously Collected Data

For the purposes of our study, the comparison is made with data (tweets and teasers) collected with the same keywords and data on measurements of PM_{10} particles collected in the same way in the period from November 1st, 2021 to February 28th, 2022 [5].

IV. Methodolgy

A. Sentiment Analysis

Before analyzing the collected data, we translated the data into English language using an online automatic document translator [13]. We used the built-in NLTK VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Analyzer to analyze the data [14]. VADER represents a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in micro blog texts, such as tweets. It merges together an English dictionary and five simple heuristics, to provide a sentiment or Valence scores based on an input text. For each word in the text, the sentiment scores are measured on a scale from -4 (most negative) to +4 (most positive), with 0 indicating a neutral sentiment. By summing the valence scores of each word in the lexicon, normalized to be between -1 (most extreme negative) and +1 (most extreme positive) acquire the compound score of the whole text. The applied normalization is as follows:

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

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. We used a default threshold value of - 0.05 and + 0.05 in order to separate the tweets into positive, negative and neutral sentiment groups.

B. Time Series Statistical Data

In order to assess the relationship between tweets, teasers, and the PM_{10} data, we measured the similarity between them with cross-correlation. The Cross Correlation Function (CCF) is the correlation between two time series xt and yt, separated by k time units, where k is called a lag [15]. In fact, CCF is the correlation between yt+k and xt. The CCF assumes that the mean and variance of the data, are constant and independent of time, in fact, that the data are stationary. In relation to the confidence interval of the function, it is calculated as $\pm\frac{2}{\sqrt{n-|k|}}$, where n is the number of observations and k is the lag. To decide whether a correlation is significant or not, its absolute value has to be greater than $\frac{2}{\sqrt{n-|k|}}$. We used the Mann-Kendall test to verify if the data are stationary, and the test did not indicate any monotonic trends.

C. Topic Modeling

In advance to this analysis, we did a preprocessing of the negative tweets, in such a way that, we removed the retweet symbols, special characters, URLs, emojis, extra spaces, and to bring the important information in the fore, we removed the stopwords, then tokenized the tweets and reduced the words to their lemma. In this study, we used the Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) [16], as a topic modeling approach, in an attempt to understand who, according to the public opinion, is responsible for the air pollution and what contributes to having an air pollution. This topic modeling approach is more appropriate for detecting topics in smaller documents, in our case tweets, because it assumes that there is only one topic per document [17]. It iterates through and reassign clusters based on a conditional distribution, in fact tweets are assigned to clusters based on the highest conditional probability. We used topic coherence score in order to evaluate the performance of the approach. The topic coherence is based on the distributional hypothesis that claims that words with similar meaning tend to co-occur within a similar context. When all or most of the words in the topic are related, topics are considered to be coherent. As we learn an increasing number of topics, starting from 2, we examine the differences between each model.

D. Fisher's z-test

In this study, we used the Fisher's z-test, in order to compare whether there is a significant difference between the correlation values obtained from this year's and last year's cross-correlations [18]. An online tool was used to calculate z values [19]. The null hypothesis for the test is: two correlations are not significantly different. The z value (z_{observed}), can be calculated through the following formula:

$$z_{observed} = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_a - 3} + \frac{1}{n_b - 3}}} \tag{3}$$

In (3), z_1 represents the z value of the correlation coefficient r_1, z_2 is the z value of the correlation coefficient r_2 (which can be obtained through an online calculator), n_a - number of observations from this year, which is 18, n_b - number of observations from last year, which is 17.

V. Results

A. Obtained Sentiments

The general statistics as well as the acquired sentiments for the tweets and teasers collected for this study and from previous study, are represented in Table I. The number of tweets and teasers this year is higher compared to last year, 41.47% and 33.99%, respectively. The sum of the retweets and unique tweets does not equal the total number of tweets, because sometimes, several original tweets may be identical or the original tweet being retweeted was not captured because it predates the time span of this study. Therefore, as unique tweets and teasers are considered those that have distinct preprocessed content, with removed retweet symbols and special characters. As a result, we have more unique tweets, but fewer unique teasers than last year. It should be noted that the percentage of negative tweets in both datasets is above 50%. The obtained results for positive and neutral sentiments in both studies are not notably different.

TABLE I. STATISTICS OF THE COLLECTED AIR POLLUTION TWEETS AND NEWS ARTICLE TEASERS

Name	Total number	Retweets (%)	Unique (%)	Negative (%)	Positive (%)	Neutral (%)
Tweets 2022/23	1442	41.47	63.24	51.1	24.89	23.99
Tweets 2021/22	1018	33.99	53.93	64.3	19.4	16.2
Teasers 2022/23	2485		31.34	50.1	36.65	13.23
Teasers 2021/22	994		48.89	55.2	39.8	4.9

B. *Cross-correlation*

In order to determine the similarity between the sentimental groups of tweets, teasers and air pollution data, the distinguished tweets and teasers obtained through sentiment analysis were plotted weekly against each other, and against official PM_{10} data. The maximum cross-correlation between the negative and neutral sentiment groups of tweets and PM_{10} particles was at lag 0, indicating that there is no lead time between these tweet frequencies and the levels of PM_{10} particles. The coefficient of the cross-correlation analysis between the negative tweets and PM_{10} particles was 0.69 ($p = 0.0019$) and 0.58 ($p = 0.0125$) between the neutral tweets and PM_{10} particles. Cross-correlation between the positive tweets and PM_{10} particles was 0.74 ($p = 0.0005$) at lag 2, indicating that positive tweets peak two-time units after the PM_{10} data (Fig. 1). Compared to the results obtained from the previous study, there was no lag in terms of cross-correlation. We have a significant cross-correlation in terms of positive tweets, while last year this correlation was insignificant. In terms of negative tweets, the correlation is similar. In relation to the sentiment groups of Time.mk teasers, the maximum cross-correlation between the negative and positive sentiment groups of teasers and PM_{10} particles was at lag 1, indicating that these teasers appear one time unit after the PM_{10} data. The coefficient of the cross-correlation analysis between the negative teasers and PM_{10} particles was 0.73 ($p = 0.0006$) and between the positive teasers and PM_{10} particles was 0.63 ($p = 0.0056$). Cross-correlation between the neutral teasers and PM_{10} particles was 0.54 ($p = 0.0247$), at lag 2, indicating that neutral teasers peak two-time units after the PM_{10} data (Fig. 2). Compared to the results from last year, the cross-correlation was at lag 0. We can conclude that there is a greater cross-correlation between the data of this study, unlike last year. Regarding the similarity between the negative and the positive sentiment groups of tweets against the negative teasers, the maximum cross-correlation was at lag 0, which indicates that there is no lead time. The coefficient of the cross-correlation analysis between the negative teasers and the negative tweets was 0.80 ($p = 0.0001$) and between the negative teasers and the positive tweets was 0.85 ($p = 0.0001$). Cross-correlation between the negative teasers and the neutral tweets was 0.52 ($p = 0.0315$) at lag 1, indicating that negative teasers peak one time unit after the neutral tweets. Compared to last year's results, we have a higher cross-correlation in terms of positive tweets, but the results for other tweets were similar. The coefficient of the cross-correlation analysis between the positive teasers and the negative tweets was 0.69 ($p = 0.0019$) and between 0.0498), at lag 0, indicating that there is no lead time. But the coefficient of cross-correlation between the positive teasers and the positive tweets was 0.71 ($p = 0.001$) and it was at lag 1, indicating that positive teasers peak one time unit after the positive tweets (Fig. 3). Compared to the previous research, we have a higher cross-correlation regarding positive tweets, but the cross-correlations for negative and neutral tweets were higher last year.

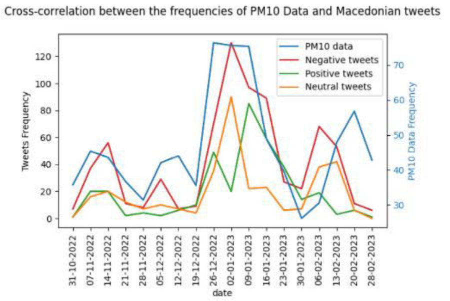

Fig. 1. Weekly comparison of official PM10 data in Macedonia and frequency of Macedonian Tweets

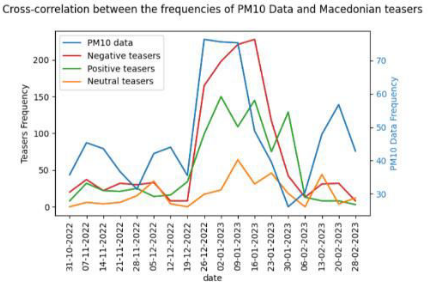

Fig. 2. Weekly comparison of official PM10 data in Macedonia and frequency of Time.mk teasers

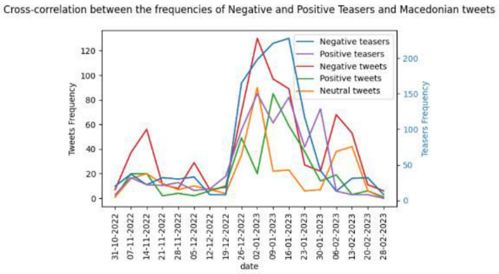

Fig. 3. Weekly comparison of negative and positive Time.mk teasers and frequency of tweets

C. *Obtained Topics*

Modeling of the topics is not a straight forward path, and deciding which model is the "best" is challenging. For that reason, to evaluate the quality of the model we used the topic coherence, along with the visual inspection of the model. To come along to the highest topic coherence score for the topic model, we experimented with different hyperparameters, and with different topic numbers ranging from 3 to 10. After the analysis, for the GSDMM model, 9 topics were selected with a coherence score of 0.47. Regarding the results of the previous research, we can conclude that the number of topics is the same, and the topic coherence scores are very similar. In relation to the negative teasers, for the GSDMM model, 8 topics were selected with a coherence score of 0.62. Regarding the number of topics and topic coherence score, we have no difference compared to last year. In Table II, we introduced some of the most important topics, together with the most important words and their frequency of occurrence.

TABLE II. TOPICS OBTAINED THROUGH GSDMM TOPIC MODELING[a]

Topic	Heating & Transport	Geography	Health & Government	Industry
Negative Tweets 2022/23	burn (19), poison (19), heat (12), oil (9), wood (9), fuel (8)	municipalities (6), bitola (5), veles (5), cities (4), countries (4)	blame (11), government (10), health (5), sick (4), danela (3)	usje (9), plant (8), criminal (7), titan (7), cement (6), enormous (6)
Topic	Electricity	Transport	Government	Health
Negative Tweets 2021/22	crisis (16), increase (16), electricity (11), cut (10), energy (9)	less (11), problem (9), car (7), need (6), tram (4)	immediately (6), mayor (6), vmro (4), municipality (4)	die (8), breathe (6), children (6), lose (6), poison (4)
Topic	Government	Health	Industry & Transport	Politics
Negative Teasers 2022/23	ministries (48), environment (42), plan (28), inform (26)	health (17), monitor (15), measure (14), extreme (10)	plant (35), usje (34), fuel (22), industries (17), oil (15)	sdsm (35), vrmodpmne (25), council (23), local (9)
Topic	Politics	Health	Landfields	Protest against industrial air pollution
Negative Teasers 2021/22	world (38), environment (33), ministry (24), Arsovska (19)	reduce (34), cause (20), health (15), death (15), fight (9)	landfield (14), illegal (9), municipality (8), waste (6), burn (5)	protest (17), citizen (14), factories (11), mills (8), prevent (4)

a. In Table III we have shown some of the most common words that appear in the corresponding topic

D. *Fisher's z-test*

Fisher's z-test was used to compare whether there is a significant difference between the cross-correlations obtained from this year's and last year's research. At significance level of alpha = 0.05, thus a confidence interval of 0.95, the number of this year's observations was $n_a = 18$ and the number of last year's observations was $n_b = 17$. In Table III, we present cross-correlations from both studies, along with the z-value and p-value (two-sided). Last year cross-correlations were insignificant between positive tweets and PM_{10} data, and between neutral teasers and PM_{10} data, and thus we have no results for the Fisher's z-test in Table III. We can conclude that we have no statistically significant difference between each comparison, which means that we can accept the null hypothesis of the test: the comparative correlations are not significantly different.

VI. Discussion and Conclusion

The main goal of our study was to understand the perception of the public in Macedonia about air pollution, through the analysis of tweets and teasers related to air pollution, compared to last year. In relation to previous similar researches, the results obtained from this research reveal noteworthy knowledge about the awareness of society about air pollution in Macedonia, which can make a significant contribution to the existing literature. RQ1 explores whether there is a relationship between air pollution-related tweets and news media teasers. Since there is a high correlation between negative and positive tweets with negative teasers we give an affirmative answer to the RQ1. In

TABLE III. FISHER'S Z-TEST RESULTS

	Cross-correlation 2022/23	Cross-correlation 2021/22	Z value	P value (two-sided)
Negative Tweets vs. PM_{10} Data	0.69	0.62	0.32	0.75
Neutral Tweets vs. PM_{10} Data	0.58	0.54	0.17	0.87
Negative Teasers vs. PM_{10} Data	0.74	0.53	0.96	0.34
Positive Teasers vs. PM_{10} Data	0.63	0.52	0.47	0.64

Negative Tweets vs. Negative Teasers	0.80	0.80	-0.04	0.97
Positive Tweets vs. Negative Teasers	0.85	0.55	1.72	0.09
Neutral Tweets vs. Negative Teasers	0.51	0.53	-0.07	0.94

addition to the expected correlation between negative tweets and negative news, which is consistent with last year's results, here we can see that there is also a high correlation with positive tweets, which may indicate that people are already starting to talk that the biggest air pollution has passed, which may imply that there is a slight delay in the negative teasers, with which it can be concluded that the news media does not really present the true picture of air pollution. RQ2 studies whether there is a relation between tweets related to air pollution and measurements of PM_{10} particles. Regarding this RQ, the results also show that there is a high correlation between negative and positive tweets with measurements of PM_{10} particles, thus we provide an affirmative answer to this RQ. There is a high correlation between negative tweets and measured values of PM_{10} particles, as in the results from last year, meaning that people express negative sentiments simultaneously with the escalation of air pollution. But there is also a high correlation with positive tweets that are two weeks late, but this only tells us that people are talking about how pleased they are that the period of high air pollution is over. RQ3 inquires whether there is a correlation between air pollution-related teasers and measurements of PM_{10} particles. We also give an affirmative answer to this RQ, as the results show that there is a high correlation between negative and positive teasers and measurements of PM_{10} particles. These teasers are delayed by a week in relation to PM_{10} particles measurements, but they also appear after people's reactions to air pollution, so it can be concluded that the media might not present the true picture of air pollution in Macedonia, compared to last year's results. We should also note that the two correlations, we have analyzed, are similar, which may indicate that even after a one-week delay, some of the internet portals are still suppressing the real picture of air pollution. RQ4 analyze whether there is a difference in society's awareness of air pollution from last year. But the results show that there is no significant difference. Topic modeling of tweets, resulted in similar topics, but the main difference compared to last year's results is that the public is increasingly talking about the negative effects of biomass used for heating, as one of the main air pollutants. Also, more and more people are talking about air pollution in other cities as well. According to the topic modeling of the teasers, the resulting topics are similar to last year's. This indicates that the media still do not talk about the negative effects of biomass. It is interesting to note that people and the news media are saying that bonfires should not be lit during Christmas Eve due to a possible increase in air pollution. As a conclusion from the analyzes made about tweets, we can say that even though the public is talking more and more about air pollution, they still need to be much more informed about air pollution through the media, in order to increase the awareness of the society. Regarding the analyzes made about the news media, we can conclude that they do not give a completely realistic picture of the main air pollutants, which could mean that to some extent they might suppress the discussions about air pollution. A recommendation regarding the media is that they should more realistically inform the public about air pollution and promote pro-environmental behavior. Future research guidelines for this study include expanding the search keywords to obtain more tweets and teasers, in order to gain a clearer picture of the topics emerging from them. In the future, a comparison of the approach used in this paper can be compared to language-agnostic or popular (cross) language models pre-trained in Macedonian to infer the sentence-level sentiment and topic from the text written in Macedonian. In conclusion, this study proves the usefulness of sentiment analysis as it helped us to analyze the sentiments that people expressed on the topic of air pollution, as well as topic modeling that helped us to find out the topics that people discussed regarding air pollution in Macedonia. The obtained results with correlation analysis, to confirm the similarity, helped us to find out what is the society's awareness about air pollution. Also, the results obtained using Fisher's z-test showed that there was no significant difference between this year's and last year's correlations.

ACKNOWLEDGMENT

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

REFERENCES

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

[2] "Information about the air pollution in cities in the Republic of Macedonia and possible risks to health," Institute of Public Health of Republic of North Macedonia, 2015.

[3] Y. Tao, F. Zhang, C. Shi, and Y. Chen, "Social Media Data-Based Sentiment Analysis of Tourists' Air Quality Perceptions," MDPI Open Access Journals, 2019.

[4] J.-Y. Jiang, X. Sun, W. Wang, and S. Young, "Enhancing Air Quality Prediction with Social Media and Natural Language Processing," 2019.

[5] A. Madjar, I. Gjorshoska, J. Prodanova, and A. Dedinec, "Investigating Public Awareness of Air Pollution in Western Balkans by analyzing Tweets and News Article Teasers," Faculty of Computer Science and Engineering: Conference papers, 2022.

[6] J. Roesslein, "Tweepy: Twitter for Python!" 2020. Available at: https://github.com/tweepy/tweepy.

[7] Twitter, "Twitter API v2." Available at: https://developer.twitter.com/en/docs/twitter-api

[8] S. K. Karn, M. Buckley, U. Waltinger, and H. Schutze, "News Article Teaser Tweets and How to Generate Them," 2019.

[9] Crummy, "Beautiful Soup," 2023. Available at: https://www.crummy.com/software/BeautifulSoup/.

[10] I. Trajkovski, "How does TIME.mk work?," Time.mk, 2008. Available at: https://time.mk/info/site.

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

[12] B. Blazhevski, "МЖССП: Податоци за аерозагадувањето не се прикажуваат поради инсталација на нова опрема," Meta.mk, 2022.

[13] Online Doc Translator, 2021. Available at: https://www.onlinedoctranslator.com/en/.

[14] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14), Ann Arbor, MI, June 2014. Available at: https://github.com/vaderSentiment/vaderSentiment.

[15] "8.2 Cross Correlation Functions and Lagged Regressions," Eberly College of Science. Available at: https://online.stat.psu.edu/stat510/lesson/8/8.2.

[16] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014.

[17] R. Pelgrim, "Short-Text Topic Modelling: LDA vs. GSDMM," 2021.

[18] S. Glen, "Fisher Z-Transformation." Available at: https://www.statisticshowto.com/fisher-z/#:~:text=Fisher's%20z'%20is%20used%20to,r2%20from%20independent%20samples.

[19] R. Lowry, "Significance of the Difference Between Two Correlation Coefficients." Available at: http://vassarstats.net/rdiff.html.