Proceedings of the SALLD-2 Workshop @LREC2020, pages 25–30
Language Resources and Evaluation Conference (LREC 2022), Marseille, 24 June 2022

Correlating Environmental Facts and Social Media Trends Leveraging Commonsense Reasoning and Human Sentiments

Brad McNamee[1], Aparna S. Varde[2,3], Simon Razniewski[3]
1. Computational Linguistics Program, Montclair State University, Montclair, NJ, USA
2. Department of Computer Science, Montclair State University, Montclair, NJ, USA
3. Max Planck Institute for Informatics, Saarbrucken, Germany
mcnameeb1@montclair.edu, vardea@montclair.edu, srazniew@mpi-inf.mpg.de
(* A. Varde: Visiting Researcher at Max Planck Institute for Informatics)

Abstract

As climate change alters the physical world we inhabit, opinions surrounding this hot-button issue continue to fluctuate. This is apparent on social media, particularly Twitter. In this paper, we explore concrete climate change data concerning the Air Quality Index (AQI), and its relationship to tweets. We incorporate commonsense connotations for appeal to the masses. Earlier work focuses primarily on accuracy and performance of sentiment analysis tools / models, much geared towards experts. We present commonsense interpretations of results, such that they are not impervious to the masses. Moreover, our study uses real data on multiple environmental quantities comprising AQI. We address human sentiments gathered from linked data on hashtagged tweets with geolocations. Tweets are analyzed using VADER, subtly entailing commonsense reasoning. Interestingly, correlations between climate change tweets and air quality data vary not only based upon the year, but also the specific environmental quantity. We anticipate that this study will shed light on possible areas to increase awareness of climate change, and methods to address it, by the scientists as well as the common public. In line with Linked Data initiatives, we aim to make this work openly accessible on a network, published with the Creative Commons license.

Keywords: AQI, Commonsense Reasoning, Human Sentiments, Linked Data, Opinion Mining, Twitter Hashtags

1. Introduction

Human-caused climate change affects millions of lives. However, reactions are varied: from placing blame on other causes to speaking out against contributing factors. Our study focuses on a subset of USA Twitter users. This is pertinent because the USA has the second highest numbers of climate change deniers worldwide as evident from recent studies (Buchholz, 2020).

We address a significant area of climate change, namely, *AQI (Air Quality Index)*, and delve into multiple environmental quantities comprising this aggregated quantity. We compare this hard data to discussions around related topics represented by linked data via hashtags on Twitter. This is performed in order to glean insight into how people voice their opinions about climate change, and how various concerning issues can be analyzed from a commonsense knowledge standpoint. This is important rather than just appealing to experts (unlike much prior work) because the common public needs to take actions in order to deal with climate change, in addition to policy-makers and government bodies outling their decisions accordingly. Ideally, the purpose of this study is to enhance comprehension of where climate change education is potentially lacking, and thus propose steps to improve the concerned areas, by the masses as well as the classes.

In connection with this, we wish to mention the concept of linked data. Linked linguistic data is a current trend that focuses on making linguistic and Natural Language Processing (NLP) data openly available on a network, ideally accessible via a web browser. Likewise, an additional goal of our work is not only to associate sentiment analysis scoring with each tweet, but also to make this sentiment-analyzed-dataset available for use by others. It is our hope that it could be used in related work, pertinent to included model training or further climate change sentiment analysis, analogous to other literature (e.g. Iglesias et al., 2017). Pursuant to this goal, future work on this project will entail publishing this dataset under the Creative Commons (CC) license, ascribing it a URI, and ideally making the dataset accessible via a web interface. This would allow the data to become dynamic and easily accessible to others.

2. Related Work

Previous work touches upon this issue, though much of it focuses on adjacent areas. In an article on 'Tracking Climate Change Opinions from Twitter Data', the authors compare the performance of various sentiment analysis tools on climate change tweet data, and work towards accurately predicting sentiment and subjectivity in tweets with these tools (An et al., 2014). Some researchers perform sentiment analysis and topic modeling on climate data from cities worldwide, including Paris, London, and New Delhi. (Gurajala et al., 2019). This work is similar to ours, but focuses on a wider area but shorter time period, while also concentrating on topic modeling. A recent study (Puri et al, 2021) presents an overview of relevant topics pertaining to the COVID pandemic and social media trends surrounding it, touching upon some topics relevant to climate change as affected by the pandemic. While this study addresses many interesting aspects, it does not focus on AQI in particular, nor does it conduct a deeper analysis of the numerous quantities comprising the AQI quantity.

In another relevant study, researchers investigate similar tweet sentiments on China's well-known Weibo platform, and determine whether air quality predictions can be made by combining tweet sentiment with sparse air quality testing data from remote sensor locations in rural China (Wang et al., 2017). In a research article on 'Air Quality Assessment from Social Media and Structured Data', the authors present an insight into mining pollutant data and assessing air quality by focusing on fine particle pollutants PM2.5, i.e. particulate matter of diameter less than 2.5 microns, since these are the most dangerous (Du et al., 2016). Some researchers explore other aspects of climate change, e.g. water quality, via sentiment analyzed from created emotion dictionaries (Jiang et al., 2016).

Additionally, there are related works on commonsense knowledge with respect to its extraction and compilation (Razniewski et al., 2021), as well as its usefulness in various tasks involving machine intelligence in general (Tandon et al,, 2017). Since our study in this paper targets the common public, it is important to address issues from a commonsense angle, and accordingly derive interpretations of the inferences obtained from our analysis in this work. Hence, the commonsense perspectives are significant.

3. Approach and Experiments

We acquire AQI data from EPA (Environmental Protection Agency, USA). It has thirty air quality monitoring stations in NJ, for environmental quantities in AQI, including:

- Carbon Monoxide (CO)
- Sulfur Dioxide (SO2)
- Nitrogen Dioxide (NO2)
- Ozone (ground level)
- Particulate Matter 10 (PM10)
- Particulate Matter 2.5 (PM2.5)

This data is compiled for 14 years: 2007-2021.

We then shift focus to Twitter; using *snscrape* to harvest tweets on environmental quantities using the following criteria. The tweets need to range from 2007-2021, they should originate in NJ, and they must correspond to our accepted hashtags. Since hashtags typically serve well as linked data identifiers, we carefully select these based on commonsense knowledge as per the environment. Selected hashtags are: #airpollution, #airquality, #airqualityindex, #aqi, #cleanair, #ozone, #smog, #haze, #emissions, #pollution, #carbonmonoxide, #co, #nitrogendioxide, #no2, #sulfurdioxide, and #so2.

Some filtering is needed based on Named Entity Disambiguation, e.g., CO can imply Colorado. This is conducted while preprocessing. We compile hard data for different environmental quantities (SO2, ozone, etc.), and can visualize temporal changes. We utilize *Matplotlib* to plot each value, for AQI data and tweets.

3.1 VADER

After scraping tweets, we perform sentiment analysis via VADER (Valence Aware Dictionary and Sentiment Reasoner). It inadvertently entails commonsense reasoning through its "wisdom-of-the-crowd approach" and its manner of "establishing ground truth using aggregate data from multiple human raters" (Hutto and Gilbert, 2014). It is adept at evaluating and scoring human sentiments in social media text.

In sentiment analysis, we use the *compound* score, i.e. the normalized weighted composite of all scores, normalized between (-1, +1), thus enhancing analysis from a commonsense standpoint. If it is >=0.05, we assign the tweets a positive sentiment; if it is > -0.05 and <0.05, tweets are neutral; if it is <=-0.05, tweets are negative. (Hutto and Gilbert, 2014).

3.2 Experimental Process and Algorithm

The diagram in Figure 1 below summarizes the high-level process adapted in this study, and detailed next.

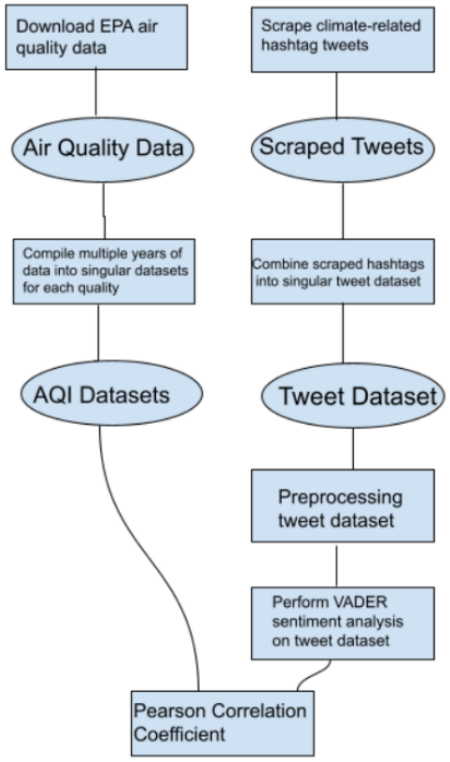


Figure 1: Overview of Experimental Process

3.3 Algorithm 1 on Compilation of AQI

We now present two succint algorithms proposed in our work. Algorithm 1 summarizes the process for finding and compiling AQI records using the EPA source. This is outlined below.

Algorithm 1 : Compile AQI records from EPA

```
FUNCTION compileAQI (EPA_record):
#Each separate AQI factor is compiled separately
AQI_Dataset_$quality = []
FOREACH year in EPA_record :
        DOWNLOAD AQI data
AQI_Dataset += EPA_record
return AQI_Dataset
```

3.4 Algorithm 2 on Acquisition of Tweets

The next algorithm, i.e. Algorithm 2, describes the process for scraping the tweets, and combining them into a singular tweet dataset. This is presented below.

Algorithm 2 : Acquire tweets and construct dataset

```
FUNCTION getTweets(list_of_hashtags) :
Tweet_Datset = []
FOREACH hashtag in list_of_hashtags :
#separate dataset for each hashtag
        Dataset_$hashtag = []
#scrape Twitter w/ snscrape for that hashtag, as well as other
#parameters
        Dataset_$hashtag += snscrape(hashtag)

FOREACH Dataset_$hashtag :
        Tweet_Datset += Dataset_$hashtag
FOREACH tweet in Tweet_Dataset :
#remove tweets that are nonlegible, nonsensical, or completely
#unrelated
        PREPROCESS tweet
FOREACH tweet in Tweet_Dataset :
        CONDUCT sentiment analysis
        ADD results of analysis to column in Tweet_Dataset
```

---

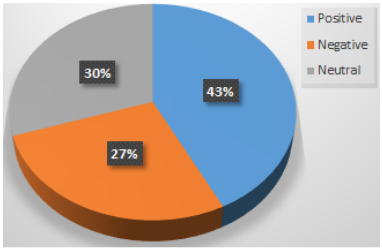
4. Results and Discussion

4.1 AQI Values and Tweet Sentiments

The results of our experiments are summarized in Figures 2-7. The tweets emanate from 2972 unique users, the most frequent ones (with 421 and 228 tweets respectively) being an industrial cooling cleaning company and a private user.

Text	Date
Everyone deserves clean air ðŸŒŽ #cleanair #AirPollution #uconn https://t.co/KHYZDmGgwB	2021-12-04
Study suggests #children w specific genetic allele may be more susceptible to neurobehavioral problems when exposed to traffic #airpollution https://t.co/q1lo1OkM2r	2019-02-16
Exposure to particulate air pollutants associated with numerous types of cancer https://t.co/BNRmYWoBfM #cancer #airpollution #cleanair #USA	2017-10-15
Come check out our brand new website! https://t.co/lLz48hwAGf ðŸŒ. The Pandora Project is a global ground based atmospheric trace gas network. #NO2 #O3 #HCHO #SO2 #Pandora #instrument #NASA ðŸŒŽ	2020-01-28
Proud of the students in and around #District36 who have taken a stand to protect our environment. #cleanenergy #cleanair #cleanwater #NewJersey https://t.co/jH8TaV4m0c	2019-06-27

Figure 2: Sample Tweets from Dataset



Sentiment Classification	Number of Tweets
Positive	2490
Negative	1603
Neutral	1755
Total	*5848*

Figure 3: Sentiment Distribution of Tweets

Figure 4: Word Cloud Visualization of All Terms

Figure 5 : Word Cloud Visualization of Positive Tweets

Figure 6 : Word Cloud Visualization of Negative Tweets

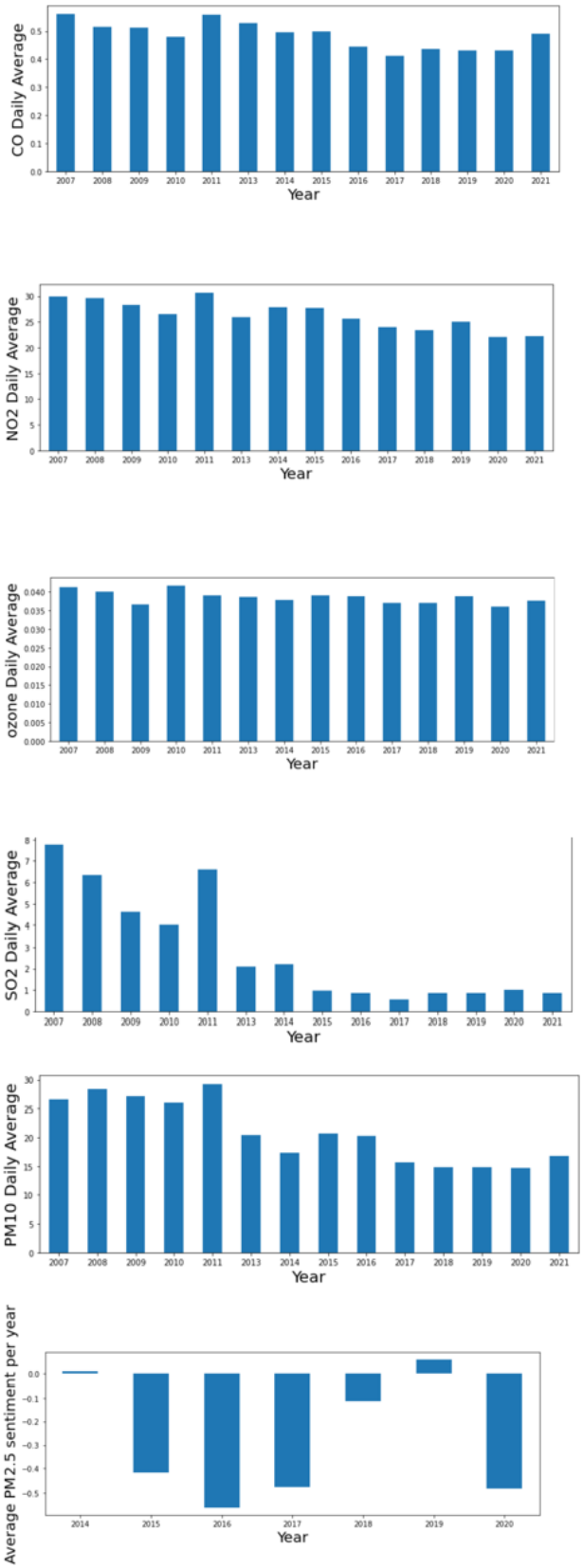


Figure 7: Average Daily Values for Quantities

Figure 2 in this paper depicts a snapshot of sample tweets from our dataset, subjected to analysis. Figure 3 illustrates the sentiment distribution of all the tweets after analysis. Figures 4, 5, and 6 present the Word Cloud Visualization of terms in all the tweets, the positive tweets, and the negative tweets, respectively. Figure 7 includes bar charts portraying the average daily values for all the quantities analyzed in the overall AQI quantity, i.e. CO, PM2.5 etc. Figure 8 comprises bar charts for the average sentiment values on the same quantities, synopsizing the analysis.

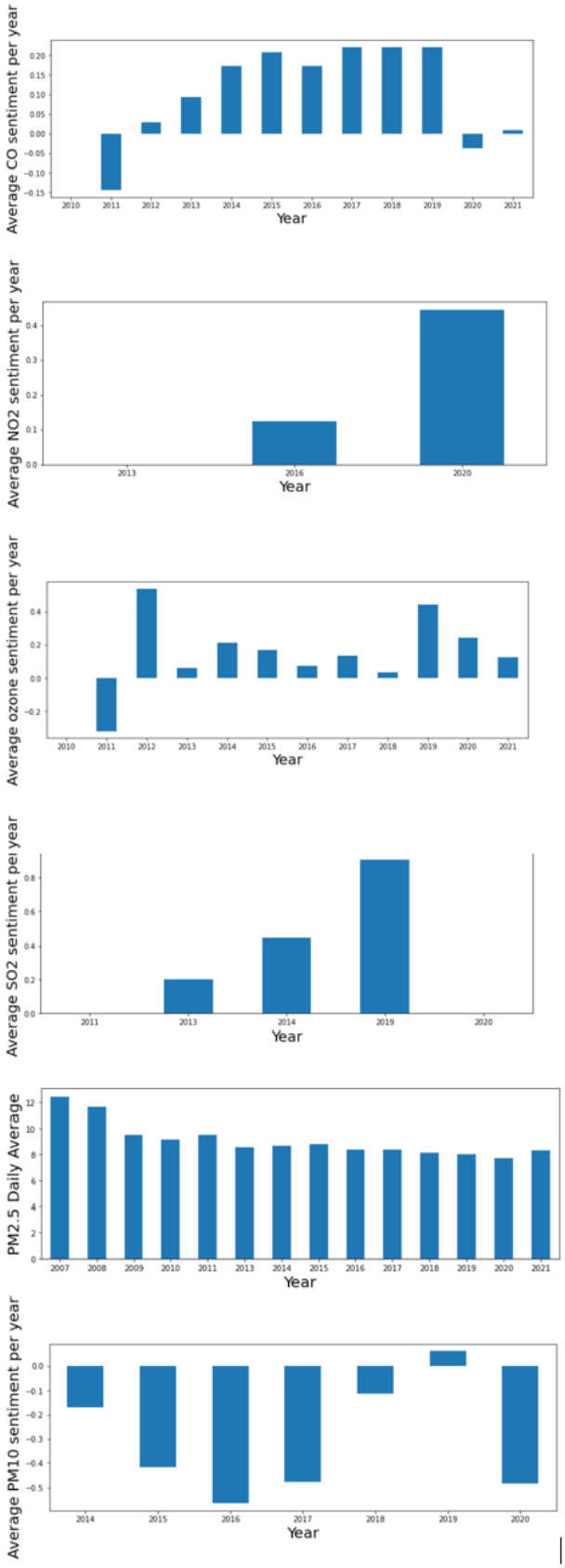


Figure 8: Average Sentiment for Quantities

4.2 Pearson's Correlation Coefficient

In order to better understand the relationship between quantity values and tweet sentiments, we utilize the Pearson's correlation coefficient (PCC). This measurement details the strength and direction of the linear association between two variables with no assumption of causality (Nickolas, 2021). The table below, i.e. Table 1, provides the names of each quantity within AQI (analyzed in our work) and the associated Pearson's Correlation Coefficient.

Quantity	Pearson's Correlation Coefficient
Ozone (ground)	0.0095
Carbon Monoxide	-0.0616
Sulfur Dioxide	-0.3489
Nitrous Dioxide	-0.3530
Particulate Matter 2.5	0.3398
Particulate Matter 10	0.1866

Table 1: Pearson's Correlation Coefficient for AQI quantities showing relationships between the actual quantity values and their respective tweet sentiments

In order to interpret the results as shown in this table, it is important to understand that a correlation coefficient >0 indicates a positive relationship between two values, while a coefficient <0 indicates a negative relationship. Additionally, if two values have a correlation coefficient >0.1 and <0.1, they are said to have no/very weak linear relationship. Finally, while this coefficient does provide unique insights, it is important to note that it is a measurement of correlation, not causality.

5. Conclusions and Roadmap

Surprisingly, most tweets have positive sentiments because people celebrate the success of climate initiatives and their own participation therein. Common climate terms (CO / ozone) have more positive sentiments than uncommon terms (PM10 / PM2.5).

Overall, we can deduce some commonsense interpretations based on human sentiments, listed as follows.

- CO, ozone, NO2, PM2.5, and PM10 depict no fluctuations in data, hence sentiment shifts in tweets must be due to other influences.
- NJ residents notice improvements in SO2 levels.
- People often tweet positively when they recognize improvements in climate change.
- The more specific / uncommon an environmental quantity is, the more negative its tweet sentiment is likely to be.

Such interpretations can enhance strategies to educate people about climate change. As future work, this can entail further questions. If people are willing to voice positive climate change work, how do we best address this through the lens of success stories? If we see more frequent usage of commonsense related climate terms (pollution, ozone), how do we harness that to strengthen climate awareness? Conversely, how can we raise awareness of less common but important aspects of AQI? Much work remains, and natural language expressions of social media can provide valuable insights into how it can be accomplished. Further investigations from commonsense standpoints can occur, leveraging the plethora of work on commonsense reasoning from sources in the literature.

6. Acknowledgements

Aparna Varde acknowledges the NSF grants 2018575 (MRI: Acquisition of a High-Performance GPU Cluster for Research & Education); and 2117308 (MRI: Acquisition of a Multimodal Collaborative Robot System (MCROS) to Support Cross- Disciplinary Human-Centered Research & Education). She is a visiting researcher at Max Planck Institute for Informatics, Saarbrucken, Germany.

7. References

An, X., Ganguly, A. R., Fang, Y., Scyphers, S. B., Hunter, A. M., and Dy, J. G. (2014). Tracking climate change opinions from twitter data, Workshop on Data Science for Social Good, pp. 1-6.

Buchholz, K. (2020). Infographic: Where Climate Change Deniers Live, https://www.statista.com/chart/19449/countries-with-biggest-share-of-climate-change-deniers/.

Du, X., Emebo, O., Varde, A., Tandon, N., Chowdhury, S. N., and Weikum, G. (2016). Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning, IEEE ICDE (International Conference on Data Engineering) workshops, pp. 54-59.

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

Hutto, C., and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text, AAAI conf. on web and social media, 8(1):216-225.

Iglesias, C. A., Sanchez-Rada, J. F., Vulcu, G., & Buitelaar, P. (2017). Linked Data Models for Sentiment and Emotion Analysis in Social Networks, Sentiment Analysis in Social Networks (pp. 49-69). Morgan Kaufmann.

Jiang, H., Lin, P., and Qiang, M. (2016). Public-opinion sentiment analysis for large hydro projects, Journal of Construction Engineering and Management, 142(2), 05015013.

McNamee B, Varde A. (2022), https://github.com/brad-mcnamee/climate-sentiment

Nickolas, S. (2021). What Do Correlation Coefficients Positive, Negative, and Zero Mean?. pp. 1-3. https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

Puri, M., Dau, Z., Varde, A. (2021) COVID and social media: analysis of COVID-19 and social media trends for

smart living and healthcare. ACM SIGWEB (Autumn): 5:1-5:20.

Razniewski, S., Tandon, N., and Varde, A. S. (2021), Information to wisdom: commonsense knowledge extraction and compilation, ACM International Conference on Web Search and Data Mining, WSDM, pp. 1143-1146.

Tandon, N., Varde, A. S., and de Melo, G. (2018), Commonsense knowledge in machine intelligence, ACM SIGMOD Record, 46(4): 49-52.

Wang, Y. D., Fu, X. K., Jiang, W., Wang, T., Tsou, M. H., and Ye, X. Y. (2017). Inferring urban air quality based on social media, Computers, Environment and Urban Systems, 66, 110-116.