Measuring the health effects of severe air pollution incidents using spatiotemporally tagged tweets

Zach Calhoun (zdc6@duke.edu)[1],
Michael Bergin, PhD[1]
David Carlson, PhD[1,2]

[1]Department of Civil and Environmental Engineering
[2]Department of Biostatistics and Bioinformatics

Abstract

Motivation

- Tracking air pollution's impact on human health is **expensive and time-consuming:** hospital data can be difficult to collect, and linking health outcomes to unhealthy air pollution may be difficult or impossible.
- Prior research has demonstrated that Twitter can be used to monitor local air quality conditions, and that negative sentiment among twitter users can be correlated with air quality.

Problem Statement

- Previous methods used targeted search terms.
- Less than 0.1% of tweets are geotagged.
- Need largest number of meaningful search terms to capture signal of air quality effect.

Our Proposal

- Look at all geotagged tweets in a location.
- Create a model that evaluates topics that correlate most with poor air quality (high AQI).
- Efficiently learned using neural networks.

Model Setup

Given a count vector, x_i, representing the counts of words used on a certain day in a certain city, find a latent topic that predicts the AQI, $\hat{y}$. In this model:

$$x_{pi} \sim \text{Pois}\left(\sum_{k=1}^{K} \phi_{pk}\theta_{ki}\right)$$

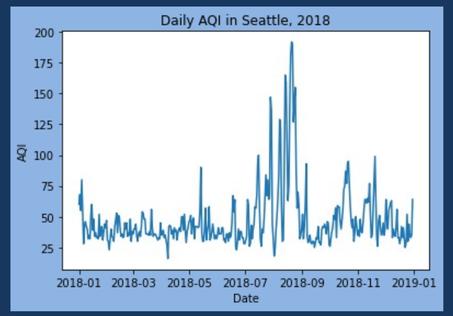
$$\hat{y_i} = \sum_{k=1}^{K} \beta_k \theta_{ik}$$

θ_{ki} represents the weight of sample x_i belonging to topic k, and ϕ_{pk} represents the weight of word p in topic k.

Methodology

1. Collect, clean, and vectorize data

Criteria: urban areas (i.e., high tweet density) with high AQI variability.



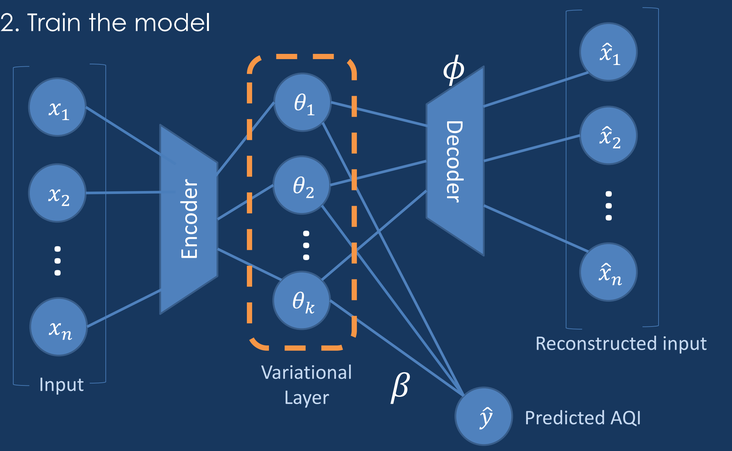
Final Dataset

City	Date Range	# of Tweets	AQI Range	Train/Test
Los Angeles, CA	7/2018 - 9/2018	2,273,134	53-201	Train
San Francisco, CA	5/2018 - 12/2018	1,770,446	3-228	Train
Phoenix, AZ	7/2018 - 9/2018	569,193	47-240	Train
Portland, OR	8/2020 - 9/2020	226,157	21-477	Train
Seattle, WA	7/2018 - 11/2018	751,181	18-192	Test
Orange County, CA	7/2020 – 9/2020	626,365	35-218	Test

Cleaning the data:

1. Remove stop words.
2. Lemmatize tweets (e.g., turn "coughing" into its lemma "cough")
3. Count vectorize tweets. Word must appear >200 times, but in less than 1% of tweets.
4. Create dataset by sampling 1000 count vectorized tweets from each day/city combo, multiple times for each day/city. This gives a set of tweets that have information about the distribution of words used on a specific day in a specific location.

2. Train the model



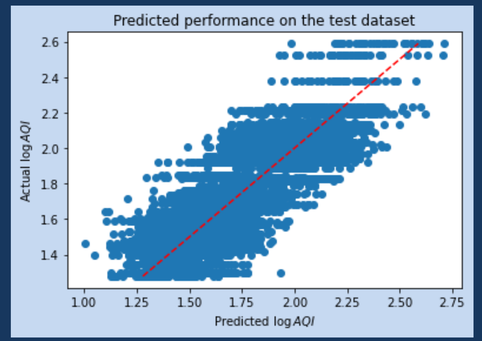
Training Loss = KLD + PNLL + MSE

KLD (Kullback-Leibler Divergence): Regularizes the variational layer **(S)**

PNLL (Poisson Negative Log Likelihood): Ensures that the input count vector is likely, thus ensuring topics found are accurate.

MSE (Mean squared error): Ensures that the topics predict $\hat{y}$, the AQI.

3. Evaluate model performance



Metric	Value
R^2	0.689
MSE (log)	0.02
PNLL	0.4
KLD	0.9 x 10^4

Assessing Topics and Words

Topic importance:

$$\text{Information(k)} = \text{Var}(\theta_k)\beta_k^2$$

Within-topic word importance:

$$\zeta_{pik} = \frac{\phi_{pk}\theta_{ki}}{\sum_{k=1}^{K}\phi_{pk}\theta_{ki}}$$

Using the equations above, we can get a sorted list of topics, along with the most unique words defining that topic.

Initial results

Topic	Words
Locations	WeHo, Haight, Sepulveda, etc.
Weather	Rain, humidity, fog, wind, etc.
Air Quality	Smoke, hazardous, smell, odor, etc.
Wildfires	Tree, firefighter, burning, panic, evacuate, etc.
Health	Eyes, itch, cough, diabetes, migraine, etc.

Next Steps:

- Fine-tune training loss to balance KLD, PNLL, MSE.
- Pull more data from Twitter.
- Interpret topics, focusing on health related words; start looking for lagged effects of AQI in tweets.
- Start using learned search terms.

Acknowledgements

Thanks to Marilyn Black, Underwriters Laboratory; and Nick Silva, Duke University.