A joint analysis of air pollution level and digital social media activity: A case study of Paris and its area.

PEREZ Charles(c.perez@psbedu.paris), SOKOLOVA KARINA, GUNCU HUSEYIN

Paris School of Business

Abstract

Air pollution is a critical problem which affects all human beings through the planet: according to the 2 May 2018 News Release from the World Health Organization, 9 out of 10 people in the world breathe polluted air. The World Health Organization estimates that every year there is around 7 million deaths of people due to the exposure to fine particles penetrating into the lungs and cardiovascular system, being the cause of severe diseases and health issues. The Lancet Commission on pollution and health (2017) affirms that air pollution is the largest environmental cause of disease in the world, killing mostly poor and vulnerable people: nearly 92\% of deaths related to air pollution occurring in low-and-middle-income countries, poor children being the most vulnerable. India and China being by far the most affected countries with more than 50\% of the global world's deaths due to ambient air pollution.

Studies about air pollution are very useful for governments. Because air pollution affects all of us, governments of many countries, impacted at diverse degrees, are concerned in finding solutions to reduce its importance and its harmful effects on people's health for which they are responsible. Governments mainly need efficient means to fight against air pollution in different ways, in monitoring, and also trying to forecast it, but they also have interest in knowing about how their inhabitants' feelings and perceptions regarding this issue and how the public opinion is evolving through time. People reaction to highly to peaks of pollution, their reaction follows the trend of the pollutants' concentration in the air.

Different studies have been made on ways to forecast air pollution by using recent big data and machine learning techniques, also by using digital image analysis because of the diverse impact of pollutants on the visibility. Some studies have also been made analyzing the social media activity on specific words like "haze" in highly polluted cities in China where pollution can be visible for a long duration.

In Asia where, according to the World Health Organization, the pollution level is in many places greater than in Europe, many social media activity analyses have been done about air pollution issues, but few studies have been done in European countries where the pollution is less perceivable, except at some periods of peak of pollution.

Here, in this study, we analyze the impact of air pollution on the social media activity in a European country such as France, and more particularly in the Paris area whose levels of air pollution have been many times above the World Health Organization's recommendations in the last years. France is significantly impacted by air pollution: according to the French Government, air pollution is the cause of 48 000 premature deaths, corresponding to 9\% of mortality in France. Air pollution annual cost for the French Government being estimated around 100 billion euros, a large part of which being related to health costs. Paris area representatives consider this issue very seriously and try improving their means to fight against it. The non-profit organization AirParif, created in 1979 and accredited by the French Ministry of Environment is the main association for monitoring air quality in the Paris area, measuring continuously the concentration of different pollutants (particle matters, ozone, carbon monoxide) in different locations with the use of dedicated air quality sensors. AirParif makes these pollutant concentrations data easily accessible for anybody on their website and from the Open data platform provided by the French state.

Among these different pollutants, in this study, we focus on the concentrations of the particle matters denoted PM2.5. These particles are fine particles whose diameter is lower than 2.5 micrometers and are harmful to human health: because of their little size they can be easily inhaled and penetrate bronchi, pulmonary alveoli and blood system potentially causing serious health damages. Today, thanks to their stations and pollutant sensors, AirParif is the most adapted and efficient institute to monitor and forecast the air quality in the Paris area with continuous scientific measurements of pollutant concentrations, and analysis of the data.

We analyze the air pollution in the Paris area focusing on its concentration of particle matters PM2.5 evolution through time. We also analyze the data resulting from the media social activity on the Twitter website, using Data

Mining techniques, then we compare these two dimensions and try find evidence of correlation between them.
Because of the large number of growing active users on Twitter since many years, Twitter has become a huge source of information and data.

In this study, we analyse tweets on a 5 years-period, from 1 May 2013 to 1 May 2018 by doing a search with keywords: “pollution or pollutions”. Then, we filter the retrieved data to focus on the messages posted by members located in the Paris area. The aim is to see if there is correlation between the scientific measurements done by the accredited organization AirParif and the number of tweets containing keywords about air pollution, posted by members from the Paris area.
In order to find correlation, techniques of data mining are used and statistical calculations of standard scores (Z-scores) are used for comparisons.

The hypothesis of this study are:\\
Hypothesis 1: In the Paris area, is there a relation between the activity on the social media platforms such as Twitter and the PM2.5 concentration values measured by the accredited Parisian environment organization AirParif?\\
Hypothesis 2: In the Paris area, is there a strong correlation between the number of posted tweets and the measured PM2.5 concentration values in a period including a peak of pollution?
In this study, a relation between the social media activity on Twitter and the PM2.5 concentration values measured by AirParif. The Pearson correlation coefficient between the standard scores of these two dimensions is not very strong on long duration, but for some smaller periods of time, the correlation coefficient is greater than 0.5 (which means a moderate correlation), particularly in some periods of peak of pollution.

Keywords: *Air pollution; Particle Matter 2.5; Twitter; Data Mining; Correlation*