Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning

Conference Paper · May 2016

DOI: 10.1109/ICDEW.2016.7495616

6 authors, including:

Xu Du

Montclair State University

7 PUBLICATIONS **20** CITATIONS

SEE PROFILE

Onyeka Emebo

Covenant University Ota Ogun State, Nigeria

12 PUBLICATIONS **38** CITATIONS

SEE PROFILE

Aparna S. Varde

Montclair State University

92 PUBLICATIONS **351** CITATIONS

SEE PROFILE

Air Quality Assessment from Social Media and Structured Data

Pollutants and Health Impacts in Urban Planning

Xu Du[2], Onyeka Emebo[1], Aparna Varde[1,2], Niket Tandon[3], Sreyasi Nag Chowdhury[3] and Gerhard Weikum[3]
[1] Department of Computer Science, Montclair State University, Montclair, NJ, USA
[2] Environmental Management Program, Montclair State University, Montclair, NJ, USA
[3] Databases and Information Systems, Max Planck Institute for Informatics, Saarbruecken, Germany
(dux3 | emeboo | vardea)@montclair.edu; (ntandon | sreyasi | weikum)@mpi-inf.mpg.de

***Abstract*— This paper describes our work on mining pollutant data to assess air quality in urban areas. Notable aspects of this work are that we mine social media and structured data in a domain-specific context, incorporate commonsense knowledge in mining media opinions and focus on the urban planning domain in a multicity environment. The results of mining are useful for predictive analysis in urbanization. A significant contribution is that we provide useful information on urban health impacts.**

Keywords—Air Pollution, Commonsense Knowledge, Health Impacts, Opinion Mining, Predictive Analysis, Urban Planning

I. INTRODUCTION

The quality of air in urban regions is important with respect to health impacts. A significant aspect of air quality is the presence of pollutants and their effects on human health [1]. Given this, an important sub-problem in our work is to mine real data on pollutants from *structured repositories* to assess air quality. We propose an approach entailing the classical data mining paradigms of association rules, clustering and classification for this purpose.

Another important aspect today is public reaction typically expressed through social media. Opinions entered by urban residents on sites such as Twitter give an idea of user satisfaction. This brings us to another interesting sub-problem, i.e., mining *social media* data on pollutants to assess air quality. One of the biggest challenges here is to review relevant information intuitively as a human would. We thus incorporate commonsense knowledge [2] in this process and develop domain-specific knowledge bases in order to guide the social media mining. We also incorporate lexical databases [3] of words with sentiments to mine public opinions.

The results of these mining processes can be used to help urban residents plan lifestyles, assist government bodies in urban policies and give inputs to environmental scientists for research. Accordingly, we conduct predictive analysis based on the results of mining. The broader impact of this work includes developing *smart cities* catering to the *smart environment* characteristic [4] by monitoring air quality, enhancing greenness and improving health. Domain KBs developed here can be useful in *smart governance* [4] by promoting automation and providing at-a-glance information for decision support. To the best of our knowledge, this is one of the first works to incorporate structured data mining and public opinion mining for urban planning.

II. MINING STRUCTURED DATA ON POLLUTANTS

A. Background and Goals

In the first sub-problem, we focus on mining pollutant data. More specifically, we consider fine particle pollutants PM2.5 (Particulate Matter, diameter < 2.5 μm). Finer pollutants are worse as the human respiratory system cannot easily filter them [1]. High PM2.5 concentration could cause severe health problems; long term exposure to it could lead to cardiovascular and respiratory diseases, genotoxicity, mutagenicity and cancer. Since PM2.5 has highly negative effects, it is desirable to avoid it, thus it is *smart* to live in a *city* with negligible PM2.5 concentration [1]. A major source of PM2.5 is traffic in urban areas. Hence, we collect real data on traffic conditions from structured sources and mine it with the following goals:

- Analyze the causes of PM2.5 occurrence in air based on multicity traffic conditions
- Predict the impact of PM2.5 presence on air quality with respect to health standards

B. Data and Standards

We propose to use the AQI (Air Quality Index) by EPA (Environmental Protection Agency, USA) [5] as *ground truth*. This is because it is a widely accepted global standard and is recommended by experts in Environmental Management for health impacts. This is shown in TABLE I. For example, an index of 401-500 implies that PM2.5 concentration is between 350.5 and 500 μg/m^3. This is "Hazardous" for health. Note that color coding is significant (e.g., green: good, red: unhealthy).

The structured data sources for PM2.5 used here are from WHO (World Health Organization) [6] and World Bank [7]. The time frame of this data is mainly the last ten years and the geographic scope is worldwide. Attributes analyzed are: *Region, Income Group, Diesel Consumption, Gasoline Consumption, Road Density, Cars per k people, Vehicles per k*

people, Vehicles per km and PM2.5 Range (µg/m^3). *Region* is the area analyzed, e.g., East Asia, Middle East etc. *Income Group* is categorical: it considers OECD (Organization for Economic Cooperation & Development) countries and others.

TABLE I. AQI STANDARDS FOR HEALTH BASED ON PM2.5

AQI Category	Index Value	Breakpoints (mcg/m^3, 24-hour average)
Good	0-50	0-12
Moderate	51-100	12.1-35.4
Unhealthy for sensitive groups	101-150	35.5-55.4
Unhealthy	151-200	55.5-150.4
Very Unhealthy	201-300	150.5-250.4
Hazardous	301-400	250.5-350.4
	401-500	350.5-500

C. Approach and Experiments

We propose an approach of combined analysis with classical mining paradigms. We deploy Apriori for association rules, k-means for clustering and decision trees for classification.

We mine association rules with Apriori, as we need to study potential impact of parameters on each other. For this, we discretize numeric data with equal frequency binning. After discretizing continuous data into ranges, we assign categorical values to a few variables, e.g., "high", "low" etc. for gas consumption using domain-specific mapping [5]. After running experiments with Apriori, we get useful inferences. There are rules showing that income groups could influence other traffic conditions. This is reasonable as economic conditions affect traffic facility construction. It is also found that high diesel consumption is not directly related to high concentration of PM2.5 in air. Examples of interesting rules are shown below.

Region=Europe & Central Asia Vehicles_Per_KM=VERY LOW => PM25_Class=GOOD conf:(1)

Gasoline_Consumption=VERYLOW Road_Density=VERY LOW Cars_Per_K_People=LOW => PM25_CLASS=MODERATE conf: (0.91)

The terms GOOD and MODERATE, pertain to the PM2.5 ranges with respect to their impact on air quality index (see TABLE I). For example, PM2.5 class = GOOD implies that the resulting AQI category is good since its index value is in the safe range of 0-50, which occurs with PM2.5 concentration of 0.0 to 12.0 µg/m^3. Likewise, we can interpret the other ranges.

Clustering is performed with k-means, an algorithm well-suited to numerical attributes, as found in this data set. We disregard the *Region* attribute here to avoid obvious clusters. An example of experimental results with clustering is shown in TABLE II. The numbers in brackets are the number of items in each cluster. We note a few interesting observations as listed next.

- Cluster 0 has relatively low traffic indicators, yet its PM2.5 range is not within safe standards
- Income of Cluster 0 is the lowest
- Cluster 2 has the highest PM2.5 concentration, yet it is not the highest traffic indicator
- Countries in Cluster 2 may have other significant PM2.5 sources or poor regulation of car emission
- Cluster 1 and cluster 3 both have the PM2.5 within safe standards and are OECD countries

TABLE II PARTIAL SNAPSHOT OF CLUSTERING

Attribute	Cluster0 (58)	Cluster1 (36)	Cluster2 (26)	Cluster3 (22)
Income Group	Upper Middle	High: OECD	High: non-OECD	High OECD
Diesel Consumption	108.63	416.61	208.7	266.42
Gasoline Consumption	96.9	341.07	286.03	186.47
Road Density	39.33	140.83	149.42	97.51
Cars per k people	120.54	493.28	234.14	290.31
Vehicles per k people	151.99	588.38	288.69	345.04
Vehicles per km	37.9	50.23	86.21	27.88
PM2.5 Range	15.12 - 18.43	-inf - 5.85	21.76 - inf	5.85 - 11.98

In these observations, it is signficant that high gas consumption does not associate with high PM2.5 concentration. In fact, *medium gas consumption is associated with higher PM2.5 concentration*. With further analysis, this can be reasoned as:

- High gas consumption usually associates with better economic conditions and better pollutant regulations
- The *Income* attribute is also significant
- High income groups & high gas consumption groups have better regulatory facilities, so PM2.5 concentration does not rise much

Decision tree classification is conducted with J4.8, the Java version of the classical C4.5 algorithm, to inductively learn a decision tree from categorical attributes. This is useful because we aim to learn potential causes of the *PM2.5 range*, which thus forms the classification target. Mapping from numeric to categorical attributes is done in a manner similar to that for association rules. A partial snapshot of results is shown below. It is found that the *Region* attribute has the strongest influence here. It is also discovered that PM2.5 pollution is highly associated with local conditions.

Region = East Asia & Pacific
| Gasoline_Consumption <= 427.7
| | IncomeGroup = High income: nonOECD: '(18.43-21.755]' (2.0)
| | IncomeGroup = High income: OECD: '(21.755-inf)' (2.0)
| | IncomeGroup = Low income: '(18.43-21.755]' (2.0/1.0)
| | IncomeGroup = Lower middle income: '(11.98-15.12]' (2.0)
| | IncomeGroup = Upper middle income
| | | Diesel_Consumption <= 114.38: '(21.755-inf)' (2.0)
| | | Diesel_Consumption > 114.38: '(11.98-15.12]' (2.0)
| Gasoline_Consumption > 427.7: '(-inf-5.845]' (5.0)

Thus, we have analyzed the causes of PM2.5 occurrence in air based on traffic conditions, which caters to the first goal of this sub-problem. The results of this are used for predictive analysis to address the second goal, i.e., predicting the health impact of PM2.5 on air quality, as elaborated in Section IV.

III. OPINION MINING ON POLLUTION FROM SOCIAL MEDIA

A. Motivation and Problem Defintion

Opinion mining or sentiment analysis deals with automated discovery of knowledge about public reactions from sites such as weblogs, review pages etc. This is important to assess user satisfaction. It motivates us to mine social media based on entries relevant to our issue, i.e., pollution and air quality. We focus on Twitter here, since it is a micro-blogging site with concise information. Thus, the goals of this sub-problem are:

- Analyze tweets on pollutants and related terms to discover knowledge useful in air quality assessment
- Use the discovered knowledge to predict potential health impacts in the context of urban planning

B. Proposed Methodology

We propose a two-phase approach for opinion mining. Phase 1 involves developing domain-specific knowledge bases (domain KBs) bootstrapped from Commonsense Knowledge (CSK). These provide the *background knowledge* to classify domain specific information. This background knowledge comprises the concepts and instances (named entities) within our domain. Phase 2 involves a domain-specific tweet crawler using the background knowledge of phase 1 (e.g., spotting concepts and instances in a tweet), and analyzing sentiments in the crawled tweets, followed by data visualization.

1) Developing Domain-Specific Knowledge Bases: We propose using domain KBs to employ background knowledge like an expert. The KB creation is outlined in the steps below.

a) Harnessing Commonsense Knowledge: Humans possess the ability to tell apart relevant content (in our case, relevant tweets) due to CSK. On the other hand, machines do not possess such knowledge. We propose to provide this background commonsense knowledge through a large, automatically mined commonsense knowledge repository, WebChild [2], which contains commonsense facts about concepts. WebChild provides a mapping from a domain to concepts and commonsense properties of these concepts.

b) Slicing WebChild: WebChild comprises a large list of domains (illustrated in TABLE III) however, we require a subset of these domains. We thus manually specify a smaller list of WebChild domains that are relevant to our context (*urban planning*). This is depicted in TABLE IV. It is conceivable to automate this process via a *probabilistic domain classifier* [1, 8] to derive a subset of domains, but would be an overkill for our usecase herewith. Thus, are now left with a slice of WebChild that contains concepts relevant to *urban planning*.

c) Curating the sliced WebChild: The selected domains provide us with a list of concepts for the given domain (e.g., *pollutant* for the domain *environment*). The sliced WebChild can be incomplete or noisy for certain concepts. We curate this slice of WebChild by designing a smart GUI (see Fig. 1) that assists the curator by automatically proposing relevant attribute values. For example, using the WebChild knowledge, the GUI knows that *small* is a size and that a pollutant is comparable to a toxin. Fig. 1 shows an example of curation for the concept *pollutant* in the domain *environment*. As discussed in [8], this curated knowledge about our *urban planning* domain is used to propose relevant Wikipedia categories. These Wikipedia categories lead to the Wikipedia entries where the categories appear, enabling the compilation of encyclopedic entries for concepts, e.g., PM2.5.

d) From domain KB to tweets: We propose a mapping from *Commonsense Concept Classes* → *Wiki Categories* → *Wiki entries* → *Hashtags* to set the stage for mining social media [8]. In essence, we spot the presence of a domain relevant encylopedic entry (e.g., PM2.5) or a domain relevant commonsense concept (e.g., pollutant) in a tweet's hashtag which highlights the main topic or subject of a tweet. If there is an overlap of the tweet's hashtag in our domain vocabulary, we consider that the tweet is relevant to our domain. As explained in [8], it is conceivable to make a more sophisticated model (e.g., a language model over our domain) that estimates whether a given tweet can be generated by the language model representing the big context (urban planning).

TABLE III. POTENTIAL LIST OF DOMAINS (PARTIAL SNAPSHOT)

acoustics	administration	agriculture	anatomy	animals
archery	architecture	Art	astrology	aeronautics
biology	banking	buildings	chemistry	cinema …

TABLE IV. CURATED LIST OF RELEVANT DOMAINS FOR KB SLICING

environment	**transport**	**buildings**	**vehicles**	**town_planning**

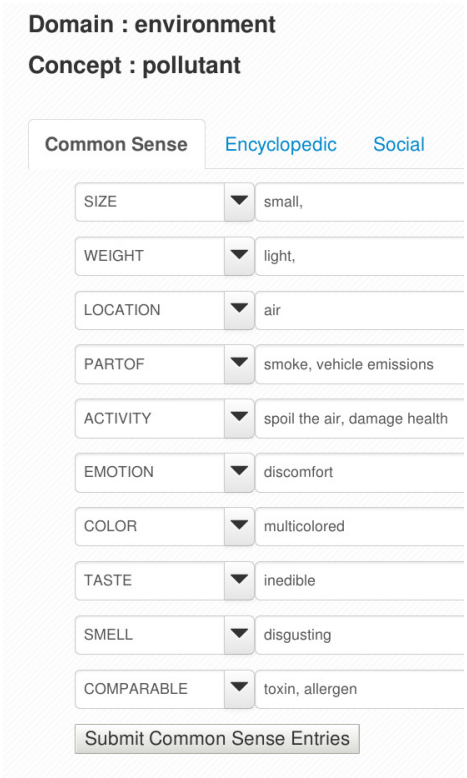

Fig. 1. Example of populating domain specific KB

Note that besides being useful in opinion mining from social media, domain KBs can be helpful in broader settings. This includes giving inputs to smart cities for smart environment and smart governance; utility in machine learning to automate various learning processes; and providing domain knowledge to mine visual commonsense from multimodal content [4, 8].

2) Building a Sentiment Analyzer: Using the domain KBs, NLP and other resources, the analyzer is built as follows.

a) Tweet Collection with Hashtags: To collect tweets, we use a Twitter API and a script written in Python. The Twitter API gives us access to user tweets using the OAuth, while the Python script collects tweets with keyword combinations and hashtags. These hashtags are derived from domain KBs, currently using the domain-specific commonsense concepts and encyclopedic entities as a dictionary. We have a tunable support threshold, the higher the support *sup* (at least *sup* number of dictionary entries are expected in the tweet), the higher the accuracy and lower the coverage. As described in [8], an alternative approach is to construct language models over domain-specific data to estimate the likelihood of the language model to generate the tweet. This step is crucial in filtering tweets and collecting only pertinent ones. For example, from 750 million tweets, we got 2.5 million *urban* domain-specific tweets, with *sup* being set to 1.

b) Storage and Cleaning: Once pertinent raw tweets are collected, the file is downloaded, converted into CSV and imported to a MySQL database for further computation. The data on tweets is then cleaned before further processing. Unnecessary characters, hashtags, usernames are removed. Any duplicate posts such as retweets and identical tweets are removed as well. It is important to clean the tweets to enhance classification accuracy by removing unwanted details that do not contribute to sentiment analysis. Consequently, any URLs in tweets are also removed. It is possible to design a more complex system that deeply analyzes the content of URLs. However, our design decision was simplicity and efficiency, as this is a pre-processing step. Also, we do not want URL content to affect polarity classification through sentiwords.

c) Text Processing of Tweets: We use a sentence level model for processing (not document level) because Twitter is a microblogging site where a tweet is at most 140 characters, therefore a sentence level model is preferable over a document level model. Text processing of tweets is conducted with TextBlob, a Python library that provides a consistent API for common NLP tasks including part-of-speech tagging, noun phrase extraction, classification, translation and more.

d) Polarity Classification with Sentiwords: The S*entiwords* lexicon is used to analyze sentiments expressed in tweets. *Sentiwords* are words pertaining to emotions. We use SentiWordNet 3.0 from LREC [3] for this purpose. The sentiwords are mapped to the content of the tweet to determine whether it is closest to expressing a positive or negative or neutral sentiment. Thus, the polarity of tweets is classsified into one of the three categories and used for further analysis.

e) Analysis and Visualization: The information about the polarity of each tweet is computed and stored in a json file. A Python script is written to aggregate this information, thus as an output, we get a set of positive, negative and neutral tweets. Using this polarity information, we plot graphs with the given data (discussed in the next section). Graph plotting is done using IPython Notebook. The plotted results displayed in graphical form allow users to see public reaction at-a-glance.

C. Experiments and Observations

We summarize our experiments pertaining to tweets in South East Asia on pollution caused by peatland fires [9]. We briefly explain the background for our experiments here. Peatlands have vast organic matter due to low decomposition of plant residue. Indonesia has the most peatlands in South East Asia. Pollutants due to these fires also affect neighboring countries, e.g., Malaysia and Singapore. Thus, Indonesian Peatland Fires (IPFs) are considered to be an international problem in Environmental Management. Pollution caused by airborne particulates is of primary concern. Studies show that rhinitis, asthma, and respiratory infections increase when particulate concentration is of hazardous level [10]. Singapore has built an air quality system called Pollutant Standards Index (PSI), which incorporates six pollutants: sulphur dioxide (SO_2), particulate matter (PM10), fine particulate matter (PM2.5), nitrogen dioxide (NO_2), carbon monoxide (CO) and ozone (O_3). The Singapore national environment agency publicly publishes the PSI level hourly through websites (e.g., haze.gov.sg). Twitter is one of the most visited social media sites. People get the information about PSI levels through this, and more importantly, express their reaction to the daily PSI level and air quality. We thus use this Twitter data in the experiments shown here. Note that it is important to conduct this analysis, since it also has the broader impact of catering to *smart cities*. Public opinion expressed through social media is useful for the *smart governance* characteristic. Also, counter-balancing the effect of hazards to maintain public health and safety is important in the *smart environment* characteristic.

In the experiments shown here, we collect pertinent tweets using hashtags and store them in a MySQL database. Based on KB knowledge, some hashtags used in collection of these tweets are: *CO_2, clean air, air pollution, Singapore, climate change*, etc. The tweet collection parameters are as follows:

i) q=air+pollution+singapore+%22air+pollution%22+%23 singapore; This shows the query used
ii) lang= en; This is the language, which in our case is English
iii) count =100; The number of tweets to return per page, up to a max of 100
iv) until =2015-10-01; Returns tweets generated before the given date.
v) since_id= ? Returns results with an ID greater than (i.e., more recent than) the specified ID.

These tweets are limited by geographical range, in our case, Singapore (though we consider a multicity context, tweets are collected from Singapore for experiments here; yet they reflect reactions of people in other cities also, constituting multicity

analysis). The date range for these tweets is the end of October to the first week of December, 2015. The tweets are then fetched from the table one by one for cleaning. Fig. 2 shows a code snippet of the functions used for cleaning the tweets.

Once the cleaning is completed, the clean tweets undergo classification either as positive, negative or neutral tweets. These results of sentiment analysis are then visualized by graphical plotting as the last step of the analyzer. Fig. 3 shows an example of visualization. This provides an at-a-glance view of mining public opinion in the area.

```
def replaceDuplictes(s);
        #replace repetitions
        pattern = re.compile(r"(.)\1{1,}", re.DOTALL)
        return pattern.sub(r"\1\1",s}

def processTweet(tweet);
        #clean the tweets
        #Convert to lower case
        tweet = tweet.lower()
        #Remove www.* or https?//*
        tweet = re.sub('((www\.[\s]+)|(https?://[^\s]+))','',tweet)
        #Remove @username
        tweet = re.sub('@[^\s]+','',tweet)
        #Remove additional white spaces
        tweet = re.sub('[\s]+', ' ', tweet)
        #Replace hashtags with word
        tweet = re.sub(r'#([^\s]+)', r'\1', tweet)
        #trim
        tweet = tweet.strip('\'"')
        return tweet
#end
```

Fig. 2. Code snippet of functions for cleaning tweets

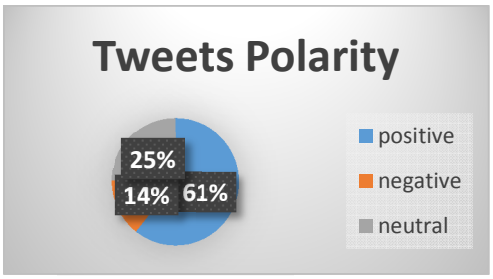

Fig. 3. Example of visualizing opinion mining results

From this figure, it appears that policies to counter-balance the effect of pollution seem fairly satisfactory since 61% of users have expressed positive sentiments. However, there is scope for improvement due to 25% of the users being neutral and 14% being negative. This opinion mining thus provides useful inputs to government bodies in urban planning and also to prospective residents and environmental scientists.

IV. PREDICTIVE ANALYSIS AND DISCUSSION

Results from the mining can be used for predictive analysis in Environmental Management, more specifically urban planning. To demonstrate this, we develop a prototype prediction tool. Programming for this tool is done in Java. We summarize the evaluation herewith. Sample executions are shown in Figs. 4 and 5. Users enter input conditions and the tool estimates the range of PM2.5 based on health impacts. We use terms "very good", "moderate" etc. to describe PM2.5 safety range as per the chance of affecting public health based on AQI (see TABLE I). For example in Fig. 4, if a user enters East Asia & Pacific with gas consumption: 582, vehicles per k people: 700, high income OECD group, road density: 11, vehicles per km: 20, diesel consumption: 467 and cars per k people: 550, the tool predicts that PM2.5 range is "very good". It means that, as learned by mining over existing data, the PM2.5 range for the given user entry is predicted as 0 - 12.0 $\mu g/m^3$, which is within safe limits for good health. Similarly, we can interpret Fig. 5.

Many experiments are conducted with the prototype tool and useful predictions are obtained. This tool is evaluated by scientists in Environmental Management who consider it to be helpful in urban planning. For example, government bodies can get an idea of how PM2.5 concentration is affected by change in traffic conditions with respect to health impacts. This can help them plan policies. Residents can estimate air quality based on various inputs to plan their current lifestyles and prospective future moves.

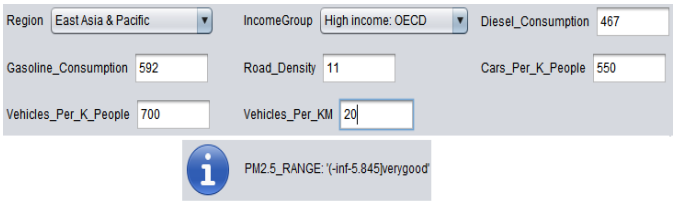

Fig. 4. Evaluation example with good PM2.5 range

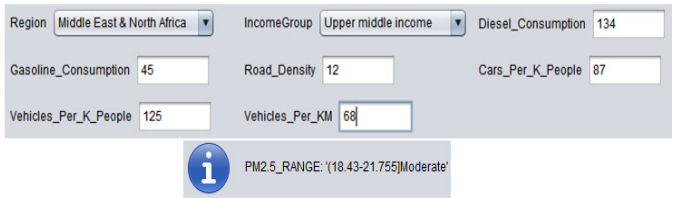

Fig. 5. Evaluation example with moderate PM2.5 range

Likewise, the polarity classification of tweets on air quality is also very useful in predictive analysis. As an output of the social media mining, the tweets are stored in a database along with their polarities. Visualization of opinion mining results is also stored. This serves as the basis to perform predictive analysis. For example, in the specific scenario here, it enables studying the correlations between users' sentiments and the actual PSI (Pollutants Standards Index) level.

Furthermore, this helps predict potential concern of users given certain PSI levels (based on opinion mining of existing data and correlation). In other words, if a particular PSI level is maintained, it helps estimate whether user sentiments would be positive, negative or neutral. This predictive analysis is useful in urban planning by allowing government bodies to estimate public opinion in advance while making regulations. It helps in catering to the satisfaction of current and future residents. It also provides inputs to environmental scientists for research, e.g., factors leading to PSI and potential measures for improvements from a health standpoint.

V. RELATED WORK

Applied data mining research appears in many fields today as the amount of available data increases and there is also a need to automate analysis from a domain perspective, e.g., in Environmental Management [11]. In urban planning, mining is applied in calibration of cellular automata transition rules that potentially relate to theories on relocation [12]. In this paper, we address issues that are the not the focus of earlier works. We consider fine particle pollutants as these are especially harmful due to not being easily filtered by the respiratory system. Also, prior research focuses mostly on single cities while we consider a multicity global context.

An overview of sentiment analysis appears in [13]. They describe approaches for opinion-oriented IR. In SentiWordNet 3.0, a lexical resource to support sentiment classification is developed [3]. It is the result of annotating WordNet synsets by degrees of positivity, negativity and neutrality. In [14] they use an approach to extract sentiments with polarities for specific subjects from a document. They have a syntactic parser and sentiment lexicon for finding sentiments in Web pages and news. Our work fits in this category, orthogonal to the existing literature. We do opinion mining in a domain-specific context, incorporating commonsense knowledge to extract concepts from social media as a human expert would. We build domain KBs useful for other tasks as well.

Studies have been conducted on pollutants. Zhou et al. analyze relationships of indoor and outdoor pollutant concentration, finding that they depend on individuals' situations [10]. Forsyth analyzes articles from representative newspapers in affected nations to help provide public opinions to pollutant problems [15]. This shows that public reaction is significant to develop urban regulations. Our research takes a step ahead and mines public reaction from online social media. Since this reaction is crucial in the urban planning area, our paper makes an important contribution here through public opinion mining.

VI. CONCLUSIONS

In this paper, we conduct mining on pollutant data from social media and structured sources to discover knowledge on air quality from a health standpoint. We use association rules, clustering and classification to mine structured data from global sources on urban air pollution. In social media mining we use Twitter, incorporate CSK and build domain KBs to guide extraction as a human expert would. We use this domain knowledge, lexical databases and text processing for polarity classification of tweets and visualize the results. Knowledge discovered by mining is useful in predictive analysis. To demonstrate this, we build a prototype tool to estimate air quality with respect to health standards. This is evaluated by domain experts and found useful in urban planning. Estimation from predictive analysis can be helpful to government bodies for urban polices, residents for lifestyle decisions and environmental scientists for further research.

Notable contributions of this work include: *mining social media and structured data in a domain-specific context; using CSK for mining tweets; addressing a multicity environment in urban planning; and conducting predictive analysis on air quality for human health*. Ongoing work includes enhancing domain KBs to provide inputs to smart cities, using CSK in the automation of learning processes and potentially deploying CSK with domain KBs for mining from photo-blogs. Another ongoing task is the use of CSK and social media mining to automate identification of IMRs (Implicit Requirements) in Software Requirement Specifications for inputs to AI tools.

ACKNOWLEDGMENT

Xu Du is funded by a Doctoral Assistantship from Montclair State University. Onyeka Emebo is supported by a Fullbright Scholarship for International PhD students. Aparna Varde has been a Visiting Senior Researcher at Max Planck Institute of Informatics during a part of this research.

REFERENCES

[1] S. Zauli Sajani, I. Ricciardelli, A. Trentini, D. Bacco, C. Maccone, S. Castellazzi, P. Lauriola, V. Poluzzi and R. Harrison, "Spatial and indoor/outdoor gradients in urban concentrations of ultrafine particles and PM2.5 mass and chemical components", Atmospheric Environment, 2015, Vol. 103, 307-320.

[2] N. Tandon, G de Melo, F. Suchanek and G. Weikum, "WebChild: harvesting and organizing commonsense knowledge from the Web", ACM WSDM, Feb 2014, pp. 523-532.

[3] S. Baccianella, A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining", May 2010, LREC, Vol. 10, pp. 2200-2204.

[4] Vienna University of Technology (TU Wien), "European Smart Cities", Tech Rep, Vienna, Austria, 2015.

[5] US EPA, "Policy Assessment for Review of the Particulate Matter National Ambient Air Quality Standards (NAAQS)", epa.gov, 2015.

[6] http://www.who.int/gho/en/, World Health Organization, Data Repository, 2015.

[7] http://data.worldbank.gov, The World Bank, Data By Country, 2015.

[8] A Varde, N. Tandon, S. Nag Chowdhury and G. Weikum, "Commonsense knowledge in domain-specific knowledge bases", Technical Report, Max Planck Institute for Informatics, Saarbruecken, Germany, Aug 2015.

[9] Y. Fujii, S. Tohno, N. Amil, M. T. Latif, M. Oda, J. Matsumoto, and A. Mizohata. "Annual Variations of Carbonaceous PM2.5 in Malaysia: Influence by Indonesian Peatland Fires." Atmospheric Chemistry and Physics Discussions, 2015, Vol. 15, pp. 22419-22449.

[10] J. Zhou, A. Chen, Q. Cao, B. Yang, W. Victor, C. Chang, and W. Nazaroff. "Particle Exposure during the 2013 Haze in Singapore: importance of the built environment." Bldg and Env., 2015, pp. 14-23.

[11] A. Pampoore-Thampi, A. Varde and D. Yu, "Mining GIS data to predict urban sprawl", ACM KDD (Bloomberg Track), 2014, pp. 118-125.

[12] X. Li, Y. Gar-On and A. Yeh, "Data mining of cellular automata's transition rules". International Journal Of Geographical Information Science, 2004, Vol. 18, No. 8, pp. 723-744.

[13] B. Pang and L. Lee, "Opinion mining and sentiment analysis", Foundations & Trends in Information Retrieval, 2008, Vol. 2, pp. 1-135.

[14] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing", ACM K-Cap, New York City, NY, Oct 2003, pp. 70-77.

[15] T. Forsyth, "Public concerns about transboundary haze: a comparison of Indonesia, Singapore and Malaysia.", Global Environmental Change, 2014, Vol. 25, pp. 76-86.