Public Opinion Matters: Mining Social Media Text for Environmental Management

Xu Du[1], Matthew Kowalski[2], Aparna S. Varde[2], Gerard de Melo[3], and Robert W. Taylor[1]

1. Department of Earth and Environmental Science, Montclair State University, NJ, USA

2. Department of Computer Science, Montclair State University, NJ, USA

3. Department of Computer Science, Rutgers University, NJ, USA

Social media mining has proven useful in multiple research fields as a tool for public opinion extraction and analysis. Such mining can discover knowledge from unstructured data in booming social media sources that provide instant public responses and also capture long-term data. Environmental scientists have realized its potential and conducted various studies where *public opinion matters*. We focus our discussion in this article on mining social media text on environmental issues, with particular emphasis on sentiment analysis, fitting the theme of *Data Science and Sustainability*. The data science community today is interested in topics that overlap with environmental issues and their broader impacts on sustainability. Such work appeals to scientists focusing on areas such as smart cities, climate change and geo-informatics. Future issues emerging from this research include domain-specific multilingual mining, and advanced geo-location tagging with demographically focused sentiment analysis.

DOI: 10.1145/3352683.3352688 http://doi.acm.org/10.1145/3352683.3352688

1. INTRODUCTION

The rapid growth of the social media industry has attracted vast numbers of users, offering valuable insights to researchers. The total number of social media users is over 2 billion worldwide. The data therein have proven pivotal due to the vast amounts of timely information as well as of long-term data for longitudinal studies. Social media mining provides opportunities to extract opinions on topics such as political issues, consumer products, emergency incidents and environmental concerns. In light of this, many environmental scientists emphasize the power of opinions expressed on social media. These include researchers on urban policy, energy conservation, transportation aspects, health issues, sustainable living, and smart cities [Zou et al. 2018; Taylor 2012; Gandhe et al. 2018].

This survey article aims to provide a review on social media text mining applications from an environmental perspective. This encompasses aspects such as climate change, smart cities, traffic management, urban policy and energy conservation, thereby fitting *Data Sci-*

ence and Sustainability, a prevalent theme today. For instance, ACM KDD 2014 had *Data Science for Social Good* as its the theme, while ACM CIKM 2017 had the theme *Smart Cities, Smart Nations*, which is closely related.

2. ENVIRONMENTAL APPLICATIONS

2.1 Climate Change and Global Warming

The public often expresses concern via social media posts related to *climate change* and its adverse impacts on sustainable living. Recent NYC TV News during climate week in September 2019 showed numerous students from local schools joining peaceful protest marches on climate change related issues. They carried banners with slogans such as "There is no Planet B" voicing their views about having nowhere to go if planet Earth deteriorated drastically in the future due to climate change and global warming. Likewise, many users post climate change-related opinions on social media platforms such as Twitter. These hence facilitate public opinion mining across time and space, since geo-tagged postings contain timestamps and geographic coordinates of latitude and longitude.

In a recent study [Dahal et al. 2019], geo-tagged tweets with keywords on climate change are mined using topic modeling and sentiment analysis. LDA is deployed for topic modeling to draw inferences from various discussion issues, while a Valence Aware Dictionary and Sentiment Reasoner are applied to conduct sentiment analysis for gauging the feelings and attitudes in the tweets.

LDA (Latent Dirichlet Allocation) is a well-known technique for topic modeling, widely surveyed in many studies [Rozeva and Zerkova 2017]. It is a generative statistical model used for sets of observations to be explained by unobserved groups that describe why some portions of data are similar to others. For example, if observations are words gathered within documents, LDA postulates that each document is a combination of a few topics and that the presence of each word is caused by one of the topics in the document. LDA can thus be used for topic modeling in environmental studies to find the prevalence of subjects in public posts.

Sentiment analysis often takes the form of polarity classification, i.e., judging whether the concerned sentiment in the text is *positive*, *negative* or *neutral*, and often the extent to which it heads in that direction, e.g., *strongly positive* etc. For instance, with reference to environmental management, users may state that they are "dissatisfied" with a climatic occurrence or legislative policy, which is a negative emotion, versus that they are "infuriated", which is a much stronger negative emotion.

The authors perform a comparison between climate change discussions across several countries over different time periods. Not surprisingly, the overall sentiment in the tweets is *negative*. This negative sentiment is even more emphatic when users express opinions on "political situations" affecting climate change or on events related to "extreme weather conditions". Interestingly, the study reveals that the climate change discussion is diverse, yet some topics are more prevalent, e.g., climate change posts in the USA are less focused on policy-related topics than corresponding posts in other countries. A broader impact of this study could entail further investigation on policy-related matters in climate change. It is important to fathom why the public currently expresses fewer opinions on policy-related

issues in the USA in comparison with other parts of the world.

In another interesting study [Wang et al. 2015], a supervised classification method is designed to process 76 million Weibo posts on climate change, aiming to discover connections between public responses and air pollution levels. The authors collect 93 million messages from 74 cities, finding that the volume of relevant posts is connected to pollution levels. Potentially relevant words on "pollution" from a probabilistic topic model are shown in Figure 1. These words are utilized to filter the Weibo posts. The study builds a 2-level classifier with randomly selected messages as the training data: the 1st level is designed to distinguish between related and unrelated messages; the 2nd level is meant to classify the related messages into "request-for-action" category versus a "pollution-experience" type. The authors suggest that combining a supervised method with an unsupervised method using LDA for topic modeling can yield higher correlations. As a broader impact, their research indicates that social media mining can be effective for air pollution monitoring even with a light-weight method.

Fig. 1: Topics about pollution learned from a probabilistic topic model [Wang et al. 2015]

There is interesting work on air quality assessment by mining over structured data and unstructured social media text [Du et al. 2016]. In this work, the authors apply association rules, clustering, and decision tree classification over structured data sources on fine particle air pollutants. They also conduct opinion mining on tweets about Indonesian Peatland Fires (IPF) and their impact on the nearby country of Singapore, since these may affect the climate. The results are used for air quality analysis from a health standpoint by using worldwide AQI (Air Quality Index) standards. A commonsense knowledge repository called *WebChild* [Tandon et al. 2014] is consulted to build domain-specific knowledge bases that capture useful domain knowledge and enable subtle human reasoning in opinion mining. A sentiment polarity classification of tweets is conducted to analyze public responses using SentiWordNet 3.0 [Baccianella et al. 2010], a lexical resource. Methods in this work can be applied to social media text mining on related topics, e.g., water quality (analogous to air quality) gauging crucial sentiments of the public, with demographics.

A related study [Sachdeva et al. 2016] conducts research on social media activity in response to the 2014 King fire in northern California. The authors induce topic models with unsupervised feature selection methods to scrutinize users' behavior on Twitter. They compare spatial and temporal variations of the most frequent topics in tweets. The results show that there are significant differences between tweets of users from regions closer to vs. further from the fire. Also, discussions about arson and threatened houses are not as

persistent as air quality concerns and potential health impacts. The authors conclude that a deeper sentiment analysis of tweets with wider data coverage could yield better results. As broader impacts, they suggest that combining social media text mining and spatio-temporal analysis can support inferences on related issues about the environment.

2.2 Urban Policy and Local Laws

Social media serves as a powerful tool for urban residents to express views on policies passed by their legislature. Likewise, the public also expresses opinions on various urban trends, including population growth/decline, and associated policies.

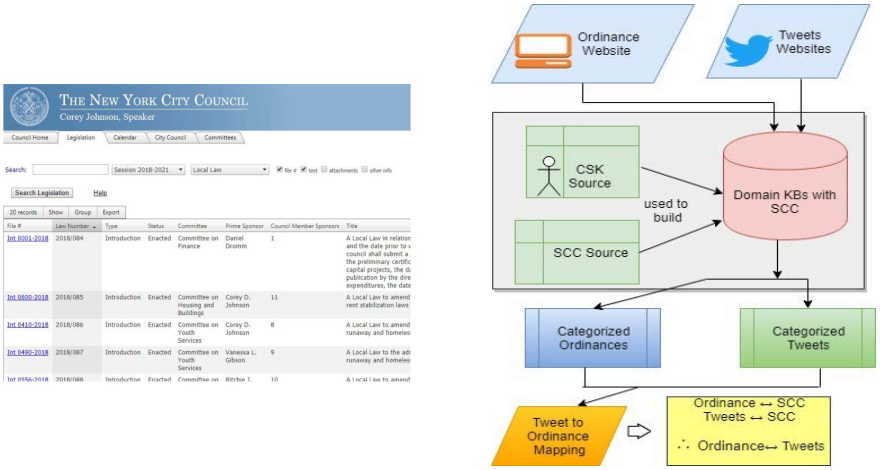

Fig. 2: NYC Council ordinance website (left) and approach for ordinance–tweet mapping (right) [Puri et al. 2018]

In urban policy mining, a process for analyzing ordinances (local laws) and their public reactions expressed via tweets has been proposed in recent work [Puri et al. 2018]. Ordinances in this study stem from the publicly available NYC Council ordinance website. The authors aim to analyze how closely a given urban region heads towards developing into a smart city by mapping groups of ordinances and tweets to smart city characteristics (SCCs) and conducting sentiment analysis of tweets on the respective SCCs to assess public satisfaction. They consider a set of six SCCs: *Smart Environment*, *Smart Governance*, *Smart Living*, *Smart Mobility*, *Smart People*, and *Smart Economy*, as defined in the literature [TU-Wien 2015]. The mapping process is depicted in Figure 2. It entails adopting commonsense knowledge from sources such as WebChild [Tandon et al. 2014] and WordNet [Fellbaum 1998] to harness human judgment involved in mapping, in line with the concept of deploying humanoid common sense within the realm of machine intelligence [Tandon et al. 2017]. The authors map groups of ordinances with tweets using the transitive property: "If ordinances map to SCCs and if the tweets map to the same SCCs, the ordinances are likely to be broadly related to the respective tweets" . This substantially reduces the sample space for ordinance–tweet mapping, since ordinances and tweets are of the order of thousands and millions, respectively. Further mapping is conducted with the

word2vec approach (widely used by many researchers, see Rozeva and Zerkova [2017]), for finding contextual similarity in the reduced mapping space.

This mapping sets the stage for sentiment analysis on the tweets by polarity classification encompassing commonsense knowledge [Tandon et al. 2017]. Results of the ordinance–tweet mining reveal the overall public satisfaction on ordinances related to various SCCs. The results suggest a positive public sentiment towards New York City as a smart city. The authors also analyze avenues for potential improvement based on public feedback. This information can be useful to urban agencies to adjust policies accordingly. This concept addresses *smart governance*, a smart city characteristic, that leverages transparency in urban decision-making through public involvement. More generally, the proposed approach [Puri et al. 2018] can be useful to map other data for opinion mining, e.g. *News and tweets*, since it is desirable to assess public reactions to various news articles, current and historical. The approach herein for mapping formal legalese in ordinances to informal acronym-ridden tweets, is potentially helpful for mapping news and tweets since these also feature formal and informal text, respectively. This mining of public opinion on news leverages *smart governance* to a considerable extent, since it entails news scrutiny in order to assess public feedback.

In a relatively recent working paper [Hollander and Renski 2015], the authors conduct exploratory research on attitudes of people in urban areas. They focus on a study in the Urban Attitudes Lab, where micro-blogging data from Twitter are assessed with quantitative and qualitative methods, such as content analysis and advanced multivariate statistics. These methods are used for a detailed study on urban experience and its implications for public policy. The authors apply a propensity scoring mechanism to create matched pairs of mid-sized cities in the Northeast and Midwest United States, where the most significant difference between each pair is that of *population decline*. The outcome is a group of 50 declining cities paired with 50 growing/stable cities. More than 300,000 tweets over a 2-month time span are analyzed, for *positive* or *negative* sentiment. The authors conduct difference of means tests, concluding that the sentiment in declining cities does not vary much in a statistically significant manner compared to that in stable and growing cities. These findings, though rather surprising, present the scope for further research. They indicate that opportunities are available to enhance the comprehension of urban attitudes based on sentiment analysis of tweets from the respective areas. Reasons for a lack of significant differences among attitudes of growing vs. declining cities would potentially be interesting to urban planning agencies, environmentalists and data scientists. Hence, this exploratory research presents promising avenues for future work on studies related to urban population growth/decline and urban policy.

The proposition of a sentiment analysis approach to extract emotions in tweets based on polarity and subjectivity, using partially labeled data, is described in recent work [Gandhe et al. 2018]. The authors put forth a hybrid approach combining supervised and unsupervised learning to take advantage of labeled training data if available, while also classifying tweets that lack specific labels. They build a classifier for sentiment analysis based on the Naive Bayes machine learning algorithm. They analyze tweets on issues such as political elections, stock markets and urban policy. The urban policy tweets are on general legislation as a whole, and on specific actions related to significant events, e.g., disasters. They implement this using TextBlob, a software library for text data processing that builds on

NLTK (the Natural Language Toolkit) to better handle human language data. This research contributes to the idea of sustainable urban development being made *smarter* by mining social media data. This is achieved using hybrid approaches that produce useful results even when fully labeled training data are not easily available.

2.3 Traffic and Mobility Issues

Social media users often post reactions about incidents that occur on the road. Issues on traffic and mobility also pertain to population relocation. Thus, sentiment analysis of the text in these posts can produce valuable results to support sustainable development and traffic optimization.

A classification based method is proposed by Gu et al. [2016] for mining tweets to extract incident information for highways and smaller roads. This method offers cost-effective data collection (see Figure 3) based on the Twitter API to obtain tweets and related information, especially location data. Using these data and metadata, the authors compare the tweets to an existing dataset to observe whether any discussed traffic incident matches with regard to the details and specifications, i.e., to authenticate its validity based on facts vs. opinions. They are also able to use this process to find additional incidents absent in the dataset but commonly discussed in the tweets. This sets the stage for further investigation on such incidents. This research is a leap in the direction of sustainable development and optimization of traffic management, by offering cost-effective social media text collection and propelling investigatory studies based on the workflow.

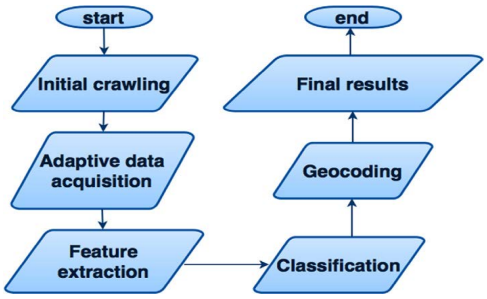

Fig. 3: Workflow of Twitter data acquisition, processing and analysis for traffic problems [Gu et al. 2016]

The idea of utilizing large-scale social media data and valuable information therein (geo-locations, times, dates and places) to infer land use within a given area has been the basis of appealing research [Zhan et al. 2014]. The researchers focus on collecting tweets having geo-locations in NYC and deploying a 3rd-party location-based service *Foursquare* to get more accurate location information from the tweets. If a user on Foursquare performs a "check-in" at a specific location, the user can share that information on Twitter. Then, by referencing the geo-location of the tweet alongside the Foursquare data within the tweet, the proposed approach can draw inferences about the content of a tweet and its neighborhood. As these tweets are continuously collected, they are categorized with respect to one of the following categories: *home*, *work*, *eating*, *entertainment*, *recreation*, *shopping*, *social service*, *education*, and *travel*. After this sorting, the approach obtains specific details

from each category, and thus performs intricate land use inference. The authors suggest that similar approaches would allow cities of varying sizes to analyze land use and interaction in a given area, thus providing greater insights into the precise activities therein. This fits the theme of *smart living* in smart cities [TU-Wien 2015].

[Wang et al. 2017] present a new method to report traffic conditions, addressing shortcomings of prior approaches. The authors base their research on the idea that while GPS (Global Positioning System) probe data are extremely useful in our everyday lives, they prove rather inadequate at fully estimating traffic conditions due to the low sampling frequency. To overcome this problem, the authors propose using social media to collect further information on traffic events not common within the geological area. To correlate the GPS probe data with social media, they focus on a deep analysis of the incoming social media data. This entails dissecting the social media posting text such that significant traffic-related phrases and locations can be separated and stored. Using these processed texts, they take the GPS probe data and fill in the missing data. From here, interesting patterns can be discovered about traffic conditions, which can be used to gain a deeper understanding of traffic commonalities in a given area. This research highlights that the process of collecting texts from social media in conjunction with GPS probes, and utilizing them to analyze specific issues, bears substantial potential in our modern world. Although the study focuses on GPS data, it exemplifies the fact that social media data can be used to augment other data sources for enhanced mining. This work caters to the *smart mobility* aspect of smart cities [TU-Wien 2015] due to the relevance of smart monitoring of traffic.

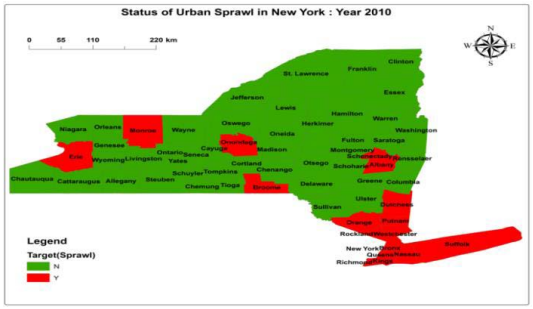

Fig. 4: Interactive map of New York State with display of sprawl affected regions generated from GIS data [Pampoore-Thampi et al. 2014]

The works of [2016], Zhan et al. [2014], and Wang et al. [2017] additionally highlight the potential for a deeper analysis on related issues such as urban sprawl. The term *urban sprawl* mainly implies the unrestricted growth of housing, transportation and commercial development over vast expanses of urban land. Researchers aim to study sprawl-causing parameters to mitigate its effects. The study by Pampoore-Thampi et al. [2014] investigates sprawl using association rules and decision tree classifiers with GIS (Geographic Information System) sources. The authors generate interactive maps (using the classical ArcGIS software) that superimpose sprawl data on the respective geographic areas to provide at-a-glance views of sprawl-affected regions (see Figure 4 for New York State). Based on these data, they analyze the impact of various sprawl-related parameters on each other and on the sprawl itself. These parameters are various spatio-temporal features related to

real GIS data, e.g., population growth, travel time to work, number of vehicles etc. These are mined to discover knowledge for a spatial decision support system (SDSS). This SDSS provides a predicted output on whether urban sprawl is likely to occur, given input parameters. It also estimates values of pertinent sprawl-related parameters to help understand their mutual impacts.

Research such as this can potentially benefit from sentiment analysis of social media text relevant to sprawl. For example, in addition to mining sprawl-related parameters, the mining of pertinent social media posts may yield further information to augment knowledge discovery. The reactions of the common public, legislators and scientists on sprawl, its causes and effects can be beneficial in understanding the gravity of certain aspects and assessing the relative importance of sprawl-related parameters. Such information can potentially be used to enhance systems such as the SDSS therein, to provide more well-informed decision support based on opinion mining. These studies thus relate to sustainable living.

2.4 Energy and Resource Conservation

Fossil fuel combustion is a major source of ambient carbon dioxide (CO_2) concentrations. The increasing public awareness of greenhouse gas emissions leads to concerns about energy types and conservation of natural resources. Social media mining can provide valuable information about public opinions on energy usage and natural resources.

Understanding public reactions on energy and resources can be extremely powerful. In the thought-provoking study by Nuortimo [2018], the authors argue that collecting data from various social media platforms is beneficial when insight on a given topic is needed, especially if that topic is rather complex. They propose a system called Case Carbon Capture and Storage to reduce harmful CO_2 emissions. Such emissions can cause widespread environmental problems. The stages in their proposed system focus on: capture and compression of CO_2 from power stations; transportation of CO_2; and storage of this captured CO_2 in a manner that keeps it out of the atmosphere. While the introduction and development of the system in this work is not being actively investigated by regulators due to low incentives, the research makes use of public reactions to emphasize the need for such a system, hoping that it will allow the project to gain more traction. In order to obtain these public reactions, text mining of social media is performed. As the social media texts are entered, the processing is conducted such that only information on the concerned system (Case Carbon Capture and Storage) is retained. From here, an analysis is conducted to observe public opinions on this environmental system.

Figure 5 provides a snapshot of opinion mining results on this system, based on posts from Social Media (SoMe). As seen here, the majority of the posts convey positive sentiments, however this majority is less than 50% of the total. While the number of negative posts are somewhat less than positive ones, they outnumber the mixed and neutral posts. This could potentially yield the inference that considerable further work might be needed for a much more widespread acceptance of the proposed technology by the public. Since concerns of many people expressed via social media are heavily opinionated, critical and often specific, important insights are gained into specific aspects needed to increase the public acceptance of the given system. Similar arguments can be applied to other such systems. Hence, in this work, sentiment analysis of social media text serves as a tool to increase the *public*

awareness and *potential acceptance* of a new technology in environmental management, geared towards solving critical energy related problems.

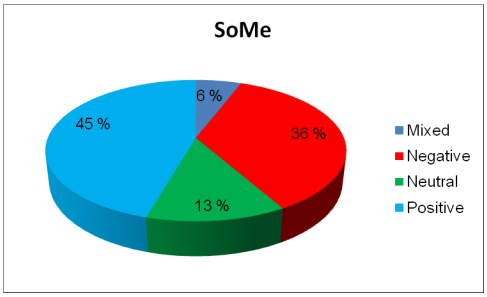

Fig. 5: Summary of sentiment analysis over social media posts about the Case Carbon Capture and Storage system [Nuortimo 2018]

Researchers often focus on opinion mining to better understand the psychological determinants of social acceptance of environment-based technologies. In useful energy-related research, Nuortimo and Harkonen [2018] focus on an analysis of *failed technologies* in which *social acceptance has been a primary factor* in the failure. This allows for better predictability during the introduction of new technology. This work emphasizes that early acceptance of a technology by the public is extremely important as new work is developed. Public opinions are obtained via machine-based social media data mining and analysis. This research furthers the idea that social media mining offers valuable insights in assessing eco-friendly technologies. This can be applied to *sustainable computing*, where the public often has mixed views about "greenness and energy conservation" versus "productivity and efficiency". Social media mining can reveal highly useful results here on the universal acceptance of eco-friendly technologies and policies.

In order to gain awareness for wildlife conservation, environmental scientists in China [Wu et al. 2018] applied social media to their research. They considered WeChat, one of China's largest social networking platforms, studying online news and relevant public comments in media posts about "Sousa chinesis" (Indo-Pacific humpback dolphin), a flagship species in China. They analyze media releases on dolphins straying into the Dongping, Beijiang and Baisha rivers of China. They deploy Content Analysis (CA) to discover knowledge from wildlife conservation information found in articles and public opinions. Their results suggest that the public feels highly doubtful about conservation efforts proposed by government bodies and experts. This is a useful and rather unfortunate observation. An interesting finding of this study is that greater efforts are needed to promote awareness on wildlife conservation, e.g., rescue operations, so as to reduce public misunderstanding. This seems quite a debatable issue on whether the public is right in expressing views on governmental lack of concern for conservation efforts, or whether the government is right in issuing appropriate conservation measures that simply need better dissemination. This work has broader impacts on *sustainable living*. Findings from this research prove that social media posts are valuable in analyzing wildlife conservation, where the public is highly opinionated, and consequently various debates continue.

2.5 Disaster and Resilience

Researchers of environmental management pay attention to the issue of disasters, since it influences natural resources and human society. Disaster-related events often propel further social media activities. By analyzing these, knowledge about efficient disaster responses can be discovered to support environmental management. The concerned studies can also help in enhancing resilience, which enables a faster recovery and may mitigate the damage brought by the disaster.

Wang et al. [2018] remark that the scarcity of hyper-resolution data for urban flooding prevents a detailed flood risk analysis. To address this issue, the authors introduce social media and crowdsourcing data into the mix. They apply NLP (Natural Language Processing) and computer vision techniques to the data they collect from Twitter and from a crowdsourcing app called MyCoast. From there, they utilize the processed data to complement existing data. In particular, they validate the extracted information against precipitation data and road closure reports to examine the quality of the data, and then utilize the results as required. The introduction of this approach for the procurement of fresh and easy-to-collect data is extremely beneficial to current environmental management techniques, in terms of its broader impacts. The application of social media in conjunction with crowdsourcing to augment data collection of otherwise rare datasets, is a useful contribution.

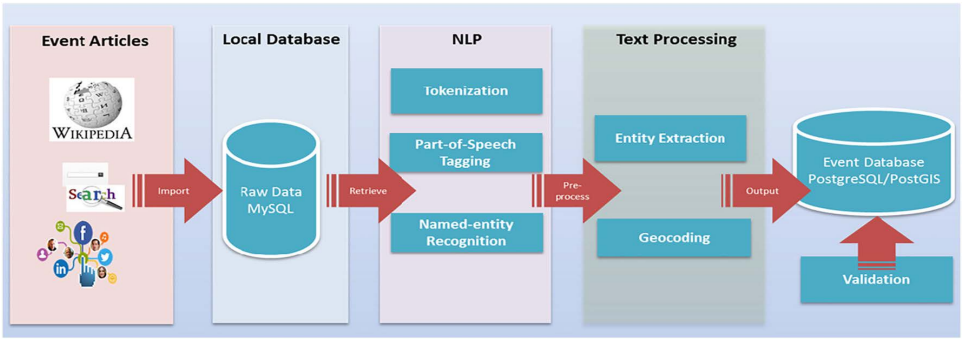

Fig. 6: Workflow for a disaster event database [Wang et al. 2018]

The framework of Huang et al. [2017] synthesizes multi-sourced data (including social media postings, remote sensing data and Wikipedia) with spatial data mining and text mining for a solution that supports disaster analysis of historical and future events. This is illustrated in Figure 6. While Wikipedia is a primary source in their work, data from Twitter and other social media platforms are also utilized to obtain more information on disasters. Using all the collected data enables the discovery of patterns in disasters through various pattern mining methods. This allows us to obtain further information that may be missing from historical reports. This framework offers advantages for disaster analysis, since data sources are added through social media and other platforms in addition to historical reports on disasters. The authors claim that a more intricate analysis and processing can facilitate real-time event tracking, highly useful for enhanced performance in disaster management and recovery.

In disaster resilience on hurricane activity, there is research [Zou et al. 2018] that focuses on mining public reactions via Twitter. The authors focus on spatio-temporal patterns of Twitter activities during Hurricane Sandy that impacted the Northeastern USA in October 2012. The study leverages 126 counties impacted by Sandy. An important finding is that social and geographic disparities are prevalent in Twitter usage. Public communities with higher socioeconomic status are found to post more hurricane-related tweets. This study also derives common indexes from Twitter data including normalized ratios to facilitate comparison across regions, and to aid in emergency management and resilience analysis. Adding Twitter indexes to a damage estimation model is found to enhance the performance. The authors thus conclude that social media data can benefit post-disaster damage estimation, provided other pertinent environmental and socioeconomic parameters are also included. Although their research addresses one particular hurricane, their results and the knowledge gained from the study can yield extremely valuable insights into strategies for using social media to increase disaster resilience. This ability is imperative in understanding how a disaster truly affects the public and how social media can be used for a deep analysis of the concerned reactions. Disaster control and resilience are by far the most critical aspects of environmental management. The works surveyed here demonstrate that social media mining plays a vital role in providing additional data for analysis beyond other recorded sources. Moreover, it helps in monitoring public reactions on disaster repercussions and the availability of recovery mechanisms, which constitute the true tests of a good disaster management system.

3. DISCUSSION ON OPEN ISSUES

Based on the above survey of the literature, we outline several thought-provoking ideas on social media text mining that offer the scope for future work from a generic web and text mining standpoint, as well as a domain-specific angle.

Demographics of posts: In the works on urban policy and local laws, it is useful to address the demographics of location-based social media posts in order to analyze public reactions to urban policies based on the backgrounds of people that post online. This could pertain to their educational, social and cultural background, as well as age and gender.

Historical diagnosis: Urban policy research can potentially entail the diagnosis of information pertaining to the historical analysis of various ordinances and their respective media posts, i.e., gauging public opinion before and after ordinance passing to assess the opinions of the public. Similar issues apply to the analysis of news and social media, i.e., these posts can be analyzed before and after a given news item is published.

Levels of granularity: While addressing the relevance of various social media posts to specific aspects of interest (e.g. tweets with regard to smart city characteristics [Puri et al. 2018]), it would be beneficial to consider the posts at a finer level of granularity. For instance, one might consider posts relating to the notion of *smart environment* in response to a given news item or a local law. It would be interesting to focus on a specific aspect within smart environment, such as green energy, and thereby assess its impacts. The same reasoning can apply to analyzing media posts in response to news etc. considering other aspects, such as climate change or disaster recovery.

Automated geo-tagging: Third-party services for collecting data on social media mining

(e.g. [Zhan et al. 2014]) may not be as effective as utilizing the social media platform itself. Not all media posts have geo-locations attached to them. Thus, methodologies can be formulated for better approaches in automatically geo-tagging the posts, which would further help in more precisely mapping a mined post to a given location. This would also propel advanced demographic analysis.

Crowdsourcing and monitoring: If the utilization of crowdsourcing apps and social media is required in the enhancement of hyper-resolution monitoring in some applications (e.g. [Wang et al. 2018]), issues would arise if a certain geographical area does not have a multitude of people providing constant updates via these apps. This motivates further research in methods used for crowdsourcing, with the goal of promoting better monitoring and analysis.

Veracity of posts: While investigating matters such as disaster recovery and resilience, data being collected via social media must be filtered and verified to avoid data manipulation. False data on such highly sensitive topics can result in misunderstandings. Hence, more research is needed on addressing the veracity of each social media post, especially pertaining to sensitive issues such as disaster repercussions and resilience. While there is much research on veracity in general, and it constitutes one of the *Vs* of big data, some of this research needs to target a more domain-specific angle, especially for sensitive subjects.

Multilingual and multicultural issues: In current studies on social media text mining, the influence of language and culture has not been an item of significant focus. There is a lack of research providing comparisons of social media posts on a given topic among people speaking different languages and emerging from various cultures. This could be a potential topic of ongoing and future research, that forms multilingual and multicultural domain-specific social media analysis, driven by recent advances in cross-lingual natural language processing [de Melo 2017; Dong and de Melo 2019]. This is particularly relevant in our current era of increasing globalization.

Irony and sarcasm: Sentiment identification in social media text mining does not particularly emphasize linguistic subtleties such as irony and sarcasm. These aspects are rather difficult to measure in media posts. Also, expressions of irony as well as sarcasm may vary across different languages and cultures. These present avenues for further research, where the state-of-the-art in idiomatic expressions and emotion detection from formal written texts can play a significant role. Such analysis in informal texts, especially on domain-specific aspects such as environmental issues, can be quite challenging. This calls for further research.

Error correction tools: Social media text is usually noisy, with both spelling errors and erroneous or non-standard grammar. Hence, progress on techniques to cope with these issues has the potential to benefit social media analytics in numerous different domains.

Abbreviations and acronyms: Excessive usage of abbreviations and acronyms in social media text often presents difficulties in mining. This challenge is even more pronounced when multiple domains are involved, each with different meanings of acronyms, thus leading to greater degrees of ambiguity and adding to the confusion caused by colloquial terms in public posts. Existing techniques in Named Entity Extraction (NEE), Named Entity Disambiguation (NED) and related areas need further research to be applicable to such informal language in social media posts, particularly with reference to context.

Big versus small data: Much of social media mining utilizes only small parts of the big data on social media. In many published studies, there is a lack of discussion about whether the small sample of data used is sufficiently robust and whether the exclusion of the bulk of remaining data can lead to adverse impacts and incorrect inferences with respect to the result interpretation. This calls for further research and discussion. For example, big data is useful to understand the big picture in a given context along with its hidden correlations. Small data may be too specific here. Hence, focusing on analyzing big data in social media with respect to the several *Vs* such as volume, velocity, variety etc. could present more interesting insights into social media mining. Some of these could be useful in domain-specific applications, where obtaining the big data itself could pose considerable challenges.

Multiple media sources: Very few studies use data from multiple sources of social media to address a single common topic. The comparison between posts on different social media sources (e.g., Twitter and Facebook) on the same topic could potentially be addressed in greater depth. This may yield even more meaningful and interesting results than analyzing each source individually, since the sources could provide a broader perspective on the opinions expressed. Such in-depth text mining over multiple media across a common thread of topics could be an aspect of future work.

4. CONCLUSION

This survey paper disseminates an overview of social media text mining applications in the environmental management area. It covers a number of facets, including climate change and global warming, urban policy and local laws, traffic and mobility issues, energy and resource conservation, as well as disaster and resilience. The topics discussed herein have significant broader impacts, as outlined in the respective subsections. These encompass news scrutiny, pollution monitoring, healthcare related decision-making, legislative transparency, traffic safety, climate change investigation, disaster repercussions, dataset enhancement for analysis, public acceptance of policies, sustainable computing issues, and the development of smart cities.

The papers surveyed in this article present the scope for future research on several topics such as multilingual domain-specific social media mining, enhanced geo-location tagging with advanced demographic analysis, subtle issues such as irony and sarcasm in social media, veracity related to sensitive subjects, crowdsourcing research in conjunction with social media mining etc. We anticipate that addressing such topics for future research can make social media text mining an even more impactful area, with greater benefits to the data science community and various application domains.

Xu Du is a PhD Candidate in Environmental Science and Management at Montclair State University, NJ, since May 2015. He completed his MS in Environmental Science from NJIT (New Jersey Institute of Technology), NJ, USA in 2013 and his BS in Environmental Science from Beijing Normal University, China in 2010. Xu has been a Doctoral Research Assistant in this PhD program, and an Adjunct Lab Instructor in the Department of Earth and Environmental Studies at Montclair State University. He has 8 publications through his PhD research as first author / co-author, some in IEEE venues, with more ongoing. He also has journal and conference papers based

on his MS Thesis from NJIT.

Matthew Kowalski is an award-winning undergraduate student in the Bachelor of Science degree program in Computer Science at Montclair State University, New Jersey. He is pursuing the field of Software Engineering with overlapping areas such as Data Mining and Web Development. His specific current interests are in writing easy-to-read and efficient code, scalable API design, and mobile app development, particularly for Apple's iOS. In April 2019, he presented his work "A Web Portal for Urban Policy Analysis by Data Mining" at the exclusive 2019 Mario M. Casabona Future Scientists Program among thirteen other student-competitors shortlisted through a highly competitive campuswide selection process. In addition to his other achievements, Matthew belongs to the Alpha Lamda Delta honor society.

Aparna Varde is an Associate Professor of Computer Science at Montclair State University (NJ). Her research spans Data Science, AI and multidisciplinary work, on areas such as commonsense knowledge, smart cities and text mining. She has around 100 publications and 2 software trademarks. Her recent honors include a best paper award at IEEE UEMCON 2019, Columbia University, NY (IoT track). She has been a visiting researcher at the Max Planck Institute for Informatics, Germany. She is the founder of the ACM CIKM PhD workshop PIKM and has co-chaired it 5 times, in addition to being a reviewer / PC member for many journals / conferences (e.g. TKDD, WWW). She has been the dissertation advisor of 2 PhD students in Environmental Management; research advisor of many MS, BS students in Computer Science; and external mentor/examiner of 4 PhD students worldwide. She obtained her PhD and MS in Computer Science from WPI (Massachusetts) and BE in Computer Engineering from Univ. of Bombay, India. Please visit `www.montclair.edu/~vardea` for details.

Gerard de Melo is an Assistant Professor at Rutgers University (NJ, USA), where he serves as the Director of the Deep Data Lab. He has published over 100 papers on NLP, AI, and Big Data analytics, with Best Paper/Demo awards at WWW 2011, CIKM 2010, ICGL 2008, and the NAACL 2015 Workshop on Vector Space Modeling. Prior to joining Rutgers, he was a faculty member at Tsinghua University, a Post-Doctoral Research Scholar at ICSI/UC Berkeley, and a doctoral candidate at the Max Planck Institute for Informatics. Notable research projects include Lexvo.org and the Universal WordNet. For more information, please consult `http://gerard.demelo.org`.

Robert W. Taylor is a Full Professor in the Department of Earth and Environmental Studies, Doctoral Faculty Member in the Environmental Science and Management PhD Program, and Coordinator of the MS Graduate Program in Sustainability Science - Sustainability Leadership at Montclair State University, NJ. He specializes in urban sustainability planning which emphasizes development of technologies and policy for smart cities, and strategies for risk management in global cities. He completed a Fulbright Specialist Program in Vietnam where he worked with planners and government officials to develop a transit-oriented development program for Ho Chi Minh City. Through a consultancy agreement (CEMA) with the NJ Board of Public Utilities, he formed a team that negotiated the first PV Solar net-metering agreement with PSEG in NJ. He has advised several graduate students, MS and PhD, at Montclair State University. He has a plethora of publications including book chapters, journals papers and conferences proceedings, in highly reputed venues.

Acknowledgments. Xu Du has been funded as a Doctoral Research Assistant by the Environmental Management PhD program and as an Instructor in the Department of Earth and Environmental Science at Montclair State University. Aparna Varde receives support through the Faculty Scholarship Program and the Doctoral Faculty Program at Montclair State University. She has obtained research grant funding through PSEG and NSF. Gerard de Melo's research at Rutgers University has been funded in part by ARO grant no. W911NF-17-C-0098 (DARPA SocialSim program). Robert Taylor has a grant awarded by the USA Department for Aerial Drone Environmental Research. He is also a Fullbright specialist in Environmental Science.

REFERENCES

BACCIANELLA, S., ESULI, A., AND SEBASTIANI, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Intl. Conf. on Language Resources and Evaluation, LREC*. Valletta, Malta.

DAHAL, B., KUMAR, S., AND LI, Z. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social Network Analysis and Mining, Springer 9,* 24.

DE MELO, G. 2017. Inducing conceptual embedding spaces from Wikipedia. In *Proceedings of WWW 2017*. ACM.

DONG, X. AND DE MELO, G. 2019. A robust self-learning framework for cross-lingual text classification. In *Proceedings of EMNLP-IJCNLP 2019*. ACL.

DU, X., EMEBO, O., VARDE, A., TANDON, N., NAG CHOWDHURY, S., AND WEIKUM, G. 2016. Air quality assessment from social media and structured data. In *IEEE Intl. Conf. on Data Engineering - Workshop on Health Data Management and Mining (ICDE-HDMM)*. Helsinki, Finland, 54–59.

FELLBAUM, C., Ed. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

GANDHE, K., VARDE, A., AND DU, X. 2018. Sentiment analysis of Twitter data with hybrid learning for recommender applications. In *IEEE Ubiquitous Computing, Electronics and Mobile Communications Conference (UEMCON)*. New York, NY, 57–63.

GU, Y., QIAN, Z., AND CHEN, F. 2016. From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies 67*, 321–342.

HOLLANDER, J. AND RENSKI, H. 2015. Measuring urban attitudes using Twitter: An exploratory study. Working Paper WP15JH1, Lincoln Institute of Land Policy, USA.

HUANG, Q., CERVONE, G., AND ZHANG, G. 2017. A cloud-enabled automatic disaster analysis system of multi-sourced data streams: An example synthesizing social media, remote sensing and Wikipedia data. *Computers, Environment and Urban Systems 66*, 23–37.

NUORTIMO, K. 2018. Measuring public acceptance with opinion mining: The case of the energy industry with long-term coal R&D investment projects. *Journal of Intelligence Studies in Business 8,* 2, 6–22.

NUORTIMO, K. AND HARKONEN, J. 2018. Opinion mining approach to study media-image of energy production - implications to public acceptance and market deployment. *Renewable and Sustainable Energy Reviews 96*, 210–217.

PAMPOORE-THAMPI, A., VARDE, A., AND YU, D. 2014. Mining GIS data to predict urban sprawl. In *ACM Conference on Knowledge Discovery and Data Mining (KDD), Bloomberg Track*. NYC, New York, 118–125.

PURI, M., DU, X., VARDE, A., AND DE MELO, G. 2018. Mapping ordinances and tweets using smart city characteristics to aid opinion mining. In *The Web Conference (WWW) Companion Volume*. Lyon, France, 1721–1728.

ROZEVA, A. AND ZERKOVA, S. 2017. Assessing semantic similarity of texts - methods and algorithms. In *Intl. Conf. on Applications of Mathematics in Engineering and Economics (AIP Conf. Proc).* 1–8.

SACHDEVA, S., MCCAFFREY, S., AND LOCKE, D. 2016. Social media approaches to modeling wildfire smoke dispersion: Spatiotemporal and social scientific investigations. *Information, Communication and Society 20,* 8, 1146–1161.

TANDON, N., DE MELO, G., SUCHANEK, F., AND WEIKUM, G. 2014. WebChild: Harvesting and organizing commonsense knowledge from the web. In *ACM International Conference on Web Search and Data Mining (WSDM)*. NYC, New York, 523–532.

TANDON, N., VARDE, A., AND DE MELO, G. 2017. Commonsense knowledge in machine intelligence. *ACM SIGMOD Record 46*, 49–52.

TAYLOR, R. 2012. Urbanization, local government and planning for sustainability. *Sustainability Science: the Emerging Paradigm and the Urban Environment*.

TU-WIEN. 2015. European smart cities, technical report. Vienna University of Technology, Vienna Austria.

WANG, R., MAO, H., WANG, Y., RAE, C., AND SHAW, W. 2018. Hyper-resolution monitoring of urban flooding with social media and crowdsourcing. *Data, Computers and Geosciences 111*, 139–147.

WANG, S., PAUL, M., AND DREDZE, M. 2015. Social media as a sensor of air quality and public response in China. *Journal of Medical Internet Research 17,* 3.

WANG, S., ZHANG, X., CAO, J., HE, L., STENNETH, L., YU, P., LI, Z., AND HUANG, Z. 2017. Computing urban traffic congestions by incorporating sparse GPS probe data and social media data. *ACM Transactions on Information Systems 35,* 4, 1–30.

WU, Y., XIE, L., HUANG, S., LI, P., YUAN, Z., AND LIU, W. 2018. Using social media to strengthen public awareness of wildlife conservation. *Ocean and Coastal Management 153*, 76–83.

ZHAN, X., UKKUSURI, S., AND ZHU, F. 2014. Inferring urban land use using large-scale social media check-in data. *Networks and Spatial Economics 14,* 3, 647–667.

ZOU, L., LAM, N., CAI, H., AND QIANG, Y. 2018. Mining Twitter data for improved understanding of disaster resilience. *Annals of the American Association of Geographers 108,* 5, 1422–1441.