Using Social Media to Predict Air Pollution during California Wildfires

Sonya Sachdeva
Northern Research Station
U.S. Forest Service
Evanston, IL, USA
sonyasachdeva@fs.fed.us

Sarah McCaffrey
Rocky Mountain Research Station
U.S. Forest Service
Fort Collins, CO, USA
smccaffrey@fs.fed.us

ABSTRACT

Wildfires have significant effects on human populations worldwide. Smoke pollution, in particular, from either prescribed burns or uncontrolled wildfires, can have profound health impacts, such as reducing birth weight in children and aggravating respiratory and cardiovascular conditions. Scarcity in the measurements of particulate matter responsible for these public health issues makes addressing the problem of smoke dispersion challenging, especially when fires occur in remote regions. Previous research has shown that in the case of the 2014 King fire in California, crowdsourced data can be useful in estimating particulate pollution from wildfire smoke. In this paper, we show that the previous model continues to provide good estimates when extended statewide to cover several wildfires over an entire season in California. Moreover, adding the semantic information contained in the social media data to the predictive model significantly increases model accuracy, indicating a confluence of social and spatio-temporal data.

CCS CONCEPTS

• Applied computing → Law, social and behavioral sciences → Sociology

KEYWORDS

social media, wildfire, smoke, air pollution, automated text analysis

ACM Reference format:
Sonya Sachdeva and Sarah McCaffrey. 2018. Using Social Media to Predict Air Pollution during California Wildfires.
In Proceedings of the *International Conference on Social Media & Society*, Copenhagen, Denmark (SMSociety).[1]
DOI:10.1145/3217804.3217946

1 INTRODUCTION

Fires are an essential component in maintaining long-term ecological health in many forest and prairie ecosystems. Yet, along with these ecological benefits, wildfires also pose problems, including air quality impacts for humans. As global climate change elongates the wildfire season and populations living in fire-prone areas increase [1]–[3], smoke from wildfires is becoming a growing public health concern. This necessitates better models for predicting the extent and range of impact of smoke dispersion from wildfire events. The goals of this research were twofold: 1) to assess whether information gleaned from social media sites and online conversation has the potential to fill in estimates of air quality for areas where physical monitoring stations are not located, and 2) to examine these sources of data for essential insights into common conceptions of fire and smoke management.

Citizen science efforts to gather important ecological and environmental data have been in use, at least, since the early 20th century [4]. However, with the advent of internet technology, user-generated content and volunteered geographic information from web sources are increasingly powerful tools in the wake of natural disasters and extreme weather events [5]–[7]. Online communication, whether through micro-blogging sites like Twitter or Weibo or social networking sites like Facebook, are an important means of disseminating information during crisis situations. User-generated data sources can also generate potentially actionable knowledge for managers and policy makers. For instance, using posts about smoke and haze on the Chinese microblogging site, Sina Weibo, has been shown to accurately estimate air quality indicators [8], [9]. Social media data has also been used to map the areas burned by wild and prescribed fire [10].

1.1 The Current Approach

In the current project, our central objective was to assess whether data obtained from the social media site Twitter could be used to ascertain air quality impacts from wildfire events. Social media

[1]

SMSociety '18, July 18–20, 2018, Copenhagen, Denmark
ACM 978-1-4503-6334-1/18/07
https://doi.org/10.1145/3217804.3217946

may be a complementary data source to currently existing air quality data, which are often limited by the location of physical monitoring stations. Harnessing data from micro-blogging sites can be especially useful for fires which often occur in more remote or rural areas where other physical sensors may not be present by utilizing people in the region as on-the-ground monitors. The relatively low-cost and freely available nature of this type of data makes it additionally appealing for land managers and crisis responders who may need to rapidly intervene while maximizing limited budgets.

Secondly, analyzing the semantic content of people's posts on these platforms can provide insight into the socio-psychological dimension of fire and smoke, providing insight into their impact on people residing, working or recreating in affected areas as well as an important source of understanding how people conceptualize wildfire risk.

Although these ideas have been tested before, namely by focusing on a single fire in 2014, the King fire, which burned approximately 100,000 acres in northern California [11], the current framework begins to develop crowdsourced platform from which to predict air quality impacts on a statewide level. In the current work, we cover all major wildfire incidents in California from May 2015 to October 2015. Both models, i.e., a single wildfire model versus state and season-wide fires in 2015, suggest that user-generated data from social media sites can be useful tools in predicting air quality impacts from wildfire events. However, the larger scope and data quantity of the current approach lend some distinct advantages which will be discussed in greater detail below.

2 Materials and Methods

2.1 Social Media Data

Social media data were purchased from Gnip, Twitter's enterprise API platform, on the basis of several key phrases and hashtags, including names of the most destructive wildfires of the season (e.g., wildfire AND smoke, roughfire, valleyfire, etc.). All tweets originated from the state of California between June 1st and October 31st of 2015. The tweets were geo-coded using either the native geographic stamp, when available or extracting the user's profile to extrapolate a location. This resulted in approximately 39,000 tweets with geographic information. As a point of reference, the same analysis in the previous study focusing on a single fire had approximately 700 tweets.

2.2 Air Quality Data

Ground-based monitoring of PM2.5 levels were obtained from the Environmental Protection Agency's (EPA) AirData air quality database. Measurements are collected by monitoring stations nationwide which then send hourly or daily aggregate measures to the EPA's database. The analysis described here used daily mean PM2.5 levels in California, which had a total of 149 monitoring stations.

2.3 Content Analysis

We used the R implementation of STM [12] to derive a structural topic model of tweets from the 2015 California wildfire season. Topic models are a widely used, generative approach to map the semantic content of large text corpora [13]. Topic models have been applied to a number of fields such as health research (e.g., tagging patient records), education research (e.g., quickly identifying commonalities in student-generated text) and political science (e.g., differences in content by party affiliation), to name a few [14]–[17]. Structural topic models allow researchers to examine the prevalence of particular topics in a corpus and examine how it varies based on other factors of interest (such as time, geographic location, personality characteristics, etc.).

3 RESULTS AND DISCUSSION

Our primary objective was to build a spatio-temporally explicit model to assess the correspondence between the frequency of users' tweets about California wildfires (and resultant smoke) and daily mean PM2.5 levels. In order to spatially connect a tweet to a monitoring station, we created variably sized bounding boxes at each degree of latitude and longitude. Each box contained approximately 3800 square miles. Our analysis contained 38 such bounding boxes to give us wide coverage across the entire state of California. Tweets and air quality data were connected in time by using the date of the tweet and the daily PM2.5 report by the EPA. This modeling approach was similar to the one employed by Sachdeva et al. [11] however, the larger sample size in this iteration allowed us to use the semantic content in each tweet to further refine our model.

As noted earlier, the semantic content of tweets was mapped using a structural topic model. Several pre-processing steps were applied to the raw twitter data before it was inputted into the topic model. For instance, all of the words within the entire corpus of tweets were stemmed (i.e., reduced to their root form) and all punctuation was removed, as were standard stop words (e.g., this, that, the, a, is). A model containing 50 topics was used to fit the dataset. Several interesting topics emerged, including many that were similar to the topic model fit to the King fire set of tweets. For instance, there were topics related to the status of various firefighting efforts (composed of words such as "acres", "contained", and "update"), topics related to air quality and smoke ("smoke", "air", "quality", "sky", "haziness", and "smell"), and several topics that showed concern about the safety of firefighters, with thoughts and prayers for firefighters to return safely.

Figure 1. **Word clouds depicting various topics in users' tweets about California wildfires**

However, we also observed topics that showed that Twitter users were looking for ways to help others in their communities that might have been affected by the wildfires ("donate", "relief", "assist", and "effort"). A subset of the 50 overall topics is depicted in Fig. 1.

The topic about smoke was then used to tag each tweet with a flag indicating whether or not it contained any information about smoke. The probabilistic nature of topic modeling implies that a portion of tweets will include smoke by chance (approximately 2%) and not because it is the focus of the tweet. To address this, we used a cutoff of 6%, approximately one standard deviation away from the mean smoke topic prevalence. That is, if a tweet's overall content was made up of more than 6% of information related to smoke, then it was marked as a "smoke" tweet. This flag was included as a variable in our spatio-temporal model assessing the correspondence between the frequency of tweets about smoke, tweets about other wildfire-related topics and air quality data (i.e., daily mean PM2.5 levels). A linear mixed effects model with bounding boxes or quadrats as the random effect was fit to the data. An autoregressive (AR) process was applied to the residuals to account for temporal autocorrelation in the data. A plot of the standardized residuals of the model showed this correction was successful in controlling autocorrelation in the data.

The final model revealed a marginal main effect of frequency of tweets overall, indicating that even general tweets about wildfires had a small, non-statistically significant relationship with mean PM2.5 levels in a geographic region ($p < .10$). However, as shown in Fig. 2, this main effect was qualified by a robust interaction between tweet type (i.e., whether the tweet contained information about smoke or not) and frequency ($p < .05$). In other words, the tweets about smoke were a better predictor of air quality than general tweets about wildfires by an order of magnitude.

3.1 Limitations

The methodology employed here, while yielding important insights into the relationship between tweet content and location, was limited in several ways. First, we have not yet systematically tested whether the size of bounding boxes, i.e., the 3800mi2 unit of analysis used here, affected the results. It is possible that a larger box might introduce more noise into the data and increase the confidence interval of the results. Furthermore, the use of social media data necessitates a discussion of privacy-related issues as users may often not realize that their data are being used by researchers in this way. However, one way that these concerns were mitigated in the current project was by purging all personally identifiable information from the data analysis (i.e., user names). The aggregate nature of the analysis used here further helps in de-identifying particular tweets, as we only assessed patterns within a 3800mi2 region

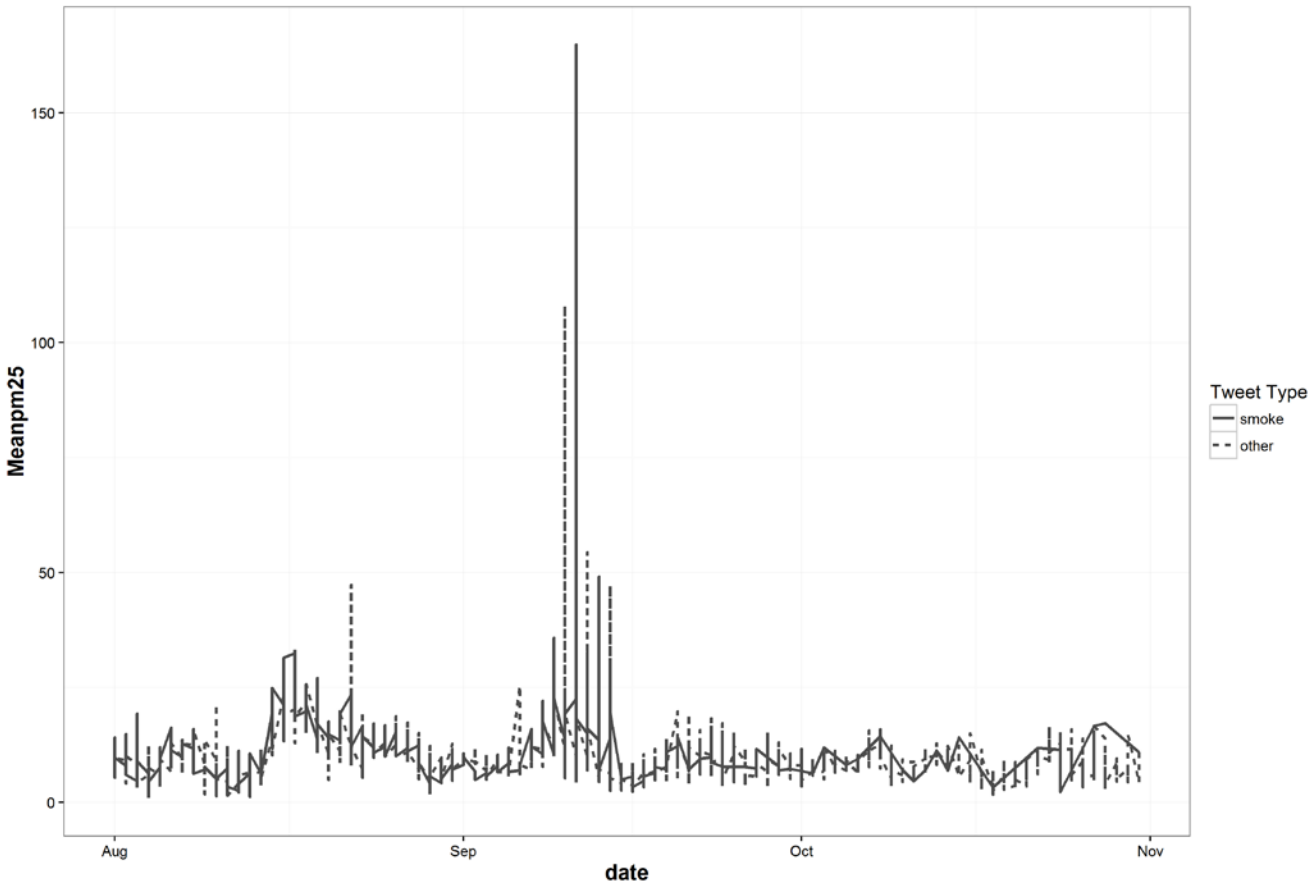

***Figure 2*. Daily Mean PM2.5 levels from August 1st, 2015 to October 31st, 2015 by Tweet type**

.

3.2 Future Directions

As this work is currently in progress, there are several future directions that we wish to pursue. For one, we would like to expand the model to estimate air quality impacts from wildfires across the United States and throughout the globe. As stated earlier, we believe this model could be useful in remote areas where other estimates of air quality are not widely available. In developing areas of the world where wildfires are a significant health concern (e.g., Indonesia), social media conversations about smoke could be a valuable means of estimating air quality and consequently, public health costs. Future work will also aim to incorporate health data within the model to assess increases in emergency room visits due to respiratory conditions, for example, to directly evaluate whether tweets about wildfire and smoke can predict public health impacts.

4 CONCLUSIONS

The results of this study yield three important insights. First, we replicated and extended the crowdsourced model of air quality impacts of the King fire to multiple fires across California, in the summer of 2015. These results demonstrate that social media and crowdsourced data is a viable, low-cost source of information about where and when air quality is affected by wildfire events. Second, we were able to integrate social/semantic information contained in the tweets with the spatio-temporal model and consequently increase the accuracy with which we can predict air quality impacts. Third, the semantic content analysis of the tweets in this study revealed important new topics such as those about people helping one another after wildfires, or looking for ways to get involved in rescue efforts.

These results suggest that our model can be expanded beyond being a predictive tool for air quality impacts to a tool that connects people willing to help with those that require assistance. This could be a crucial need as wildfires increase in the coming years and budgets across management agencies are stretched thinner.

REFERENCES

[1] V. H. Dale *et al.*, “Climate Change and Forest Disturbances,” *BioScience*, vol. 51, no. 9, pp. 723–734, Sep. 2001.

[2] S. H. Doerr and C. Santín, “Global trends in wildfire and its impacts: perceptions versus realities in a changing

world," *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, vol. 371, no. 1696, Jun. 2016.
[3] J. S. Fried, M. S. Torn, and E. Mills, "The Impact of Climate Change on Wildfire Severity: A Regional Forecast for Northern California," *Climatic Change*, vol. 64, no. 1–2, pp. 169–191, May 2004.
[4] J. Silvertown, "A new dawn for citizen science," *Trends in Ecology & Evolution*, vol. 24, no. 9, pp. 467–471, Sep. 2009.
[5] M. F. Goodchild, "Citizens as sensors: the world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, Nov. 2007.
[6] J. D. Kent and H. T. Capello, "Spatial patterns and demographic indicators of effective social media content during theHorsethief Canyon fire of 2012," *Cartography and Geographic Information Science*, vol. 40, no. 2, pp. 78–89, Mar. 2013.
[7] T. Shelton, A. Poorthuis, M. Graham, and M. Zook, "Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data,'" *Geoforum*, vol. 52, pp. 167–179, Mar. 2014.
[8] W. Jiang, Y. Wang, M.-H. Tsou, and X. Fu, "Using Social Media to Detect Outdoor Air Pollution and Monitor Air Quality Index (AQI): A Geo-Targeted Spatiotemporal Analysis Framework with Sina Weibo (Chinese Twitter)," *PLoS ONE*, vol. 10, no. 10, p. e0141185, Oct. 2015.
[9] S. Mei, H. Li, J. Fan, X. Zhu, and C. R. Dyer, "Inferring air pollution by sniffing social media," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2014, pp. 534–539.
[10] K. Endsley and J. McCarty, "Mapping prescribed burns and wildfires from Twitter with natural language processing and information retrieval techniques," *Proceedings of the International Smoke Symposium 2013*, Oct. 2013.
[11] S. Sachdeva, S. McCaffrey, and D. Locke, "Social media approaches to modeling wildfire smoke dispersion: spatiotemporal and social scientific investigations," *Information, Communication & Society*, vol. 0, no. 0, pp. 1–16, Aug. 2016.
[12] M. E. Roberts *et al.*, "Structural Topic Models for Open-Ended Survey Responses," *American Journal of Political Science*, vol. 58, no. 4, pp. 1064–1082, Oct. 2014.
[13] E. Chen, "Introduction to Latent Dirichlet Allocation," 2011. .
[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[15] J. Grimmer, "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases," *Political Analysis*, vol. 18, no. 1, pp. 1–35, Dec. 2010.
[16] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev, "How to Analyze Political Attention with Minimal Assumptions and Costs," *American Journal of Political Science*, vol. 54, no. 1, pp. 209–228, Jan. 2010.
[17] C. Wang and D. M. Blei, "Collaborative Topic Modeling for Recommending Scientific Articles," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2011, pp. 448–456.