Enhancing Air Quality Prediction with Social Media and Natural Language Processing

Jyun-Yu Jiang[†], Xue Sun[†], Wei Wang[†] and Sean Young[‡].
[†]Department of Computer Science, University of California, Los Angeles, CA, USA
[‡]Department of Family Medicine, University of California, Los Angeles, CA, USA
{jyunyu, cynosure, weiwang}@cs.ucla.edu, sdyoung@mednet.ucla.edu

Abstract

Accompanied by modern industrial developments, air pollution has already become a major concern for human health. Hence, air quality measures, such as the concentration of $PM_{2.5}$, have attracted increasing attention. Even some studies apply historical measurements into air quality forecast, the changes of air quality conditions are still hard to monitor. In this paper, we propose to exploit social media and natural language processing techniques to enhance air quality prediction. Social media users are treated as social sensors with their findings and locations. After filtering noisy tweets using word selection and topic modeling, a deep learning model based on convolutional neural networks and over-tweet-pooling is proposed to enhance air quality prediction. We conduct experiments on 7-month real-world Twitter datasets in the five most heavily polluted states in the USA. The results show that our approach significantly improves air quality prediction over the baseline that does not use social media by 6.9% to 17.7% in macro-F1 scores.

1 Introduction

In recent centuries, industrialization has considerably changed human society by providing a stimulus to economic growth and improved life quality. However, the advancement is accompanied by the increase in air pollutant emissions and risks to public health. As a consequence, predicting real-time air quality information (AQI), such as the concentration of $PM_{2.5}$, has attracted more and more attention. Air quality prediction may help the government and society to better protect their citizens from potentially harmful effects of poor air quality.

To forecast AQI, one of the most conventional approaches is to exploit historical air quality and treat the task as a time series prediction problem (Genc et al., 2010; Zheng et al., 2015). However, the air quality information can be too sophisticated to be predicted by only past AQI without any additional knowledge. For example, other environmental factors like humidity and temperature can affect the air quality when real-world events like wildfires may also play a role. To learn the additional information, most of the relevant studies collect data from additional sensors like images (Jiang et al., 2011) and ground sensors (Zheng et al., 2015). Nevertheless, these sensors are expensive in not only installation but also maintenance. As a result, exploiting sensors for air quality prediction may be too costly for most of the cities.

To learn additional knowledge without physical sensors, one of the most effective approaches is to leverage the wisdom of the crowd on the internet. For example, 81% of the adults in the USA spend on average two hours on social media and collectively publish 170 million tweets[1] every day on their feelings and observations (Wu et al., 2018). In other words, social media users can be considered as "social sensors" to perceive environmental changes and real-world events. Although social sensing has been applied to detect or predict several real-world events, such as influenza surveillance (Santillana et al., 2015; Dredze, 2012; Achrekar et al., 2011) and earthquakes (Sakaki et al., 2010, 2013), none of them focuses on predicting the air quality information. Note that although Jiang et al. (2015) and Wang et al. (2017) exploit social media to infer AQIs at current or past time, they cannot predict the future air quality. Moreover, the AQIs in these previous studies usually have considerable fluctuations, under which circumstance users tend to publish related posts,

[1]For simplicity, the posts published on social media are called *tweets* in this paper.

which makes the inference task much more manageable than general cases. In general cases, air quality changes gradually most time, which may be not sufficiently documented in social media. For instance, in California, more than 80% of the changes in air quality conditions are between good and moderate.

In this paper, we aim to leverage social media for air quality prediction. Our approach consists of three stages, including (1) tweet filtering, (2) feature extraction, and (3) air quality prediction. In the first stage, all of the incoming tweets are filtered by geographical locations and keywords extracted from statistical and topical modeling. After filtering the tweets, a convolutional neural network is applied to extract the individual feature vector for each tweet with a max-over-time pooling layer. A max-over-tweet layer is then proposed to aggregate the feature vectors of all tweets as the social media features for predicting air quality using a fully-connected hidden layer to combine with historical measurements. Finally, experiments conducted on 7-month large-scale Twitter datasets show that our approach significantly outperforms all comparative baselines.

2 Air Quality Prediction with Social Media and NLP

Following the previous studies (Zheng et al., 2015), we model the problem as a multi-class classification task. According to the Environmental Protection Agency [2] (EPA) in USA, AQIs can be categorized into six classes as shown in Figure 1. Note that more than 99% of daily AQIs in the USA are similar and falling in the first two classes so that the classification task is more laborious than predicting numerical AQIs. Given a location l and a time t, the corpus $D(l,t)$ is defined as the N tweets published by any user located at the location l at time t. $a(l,t)$ denotes the AQI value in the location l at time t while the historical measurements $H(l,t) = a(l,t), a(l,t-1), \cdots, a(l,t-T+1)$ provide AQIs at T time points. Given the corpus $D(l,t)$ and the historical measurements $H(l,t)$ at location l at time t, our goal is to predict the corresponding class y of the AQI at the next time point $t+1$.

Framework Overview. Figure 1 illustrates the proposed three-stage framework. In the first stage, the incoming tweets are filtered to remove irrelevant information. In the second stage, representative features are extracted from filtered tweets and historical measurements. In the last stage, we predict the category of air quality with a hidden layer and a softmax function.

AQI	Level of Concern
0-50	Good
51-100	Moderate
101-150	Unhealthy for Sensitive Groups
151-200	Unhealthy
201-300	Very Unhealthy
301-500	Hazardous

Table 1: Categorization of AQI from EPA.

2.1 Stage 1: Tweet Filtering

In most of the cities, the majority of tweets should be irrelevant to air quality because users are less likely to discuss air quality situations unless there is a dramatic change. Hence, we need to filter tweets before using them for air quality prediction. Following the previous work (Shike Mei and R.Dyer, 2014), we use three groups of keywords for filtering tweets, including **(1) environment-related terms** like *smog* released by EPA, **(2) health-related terms** like *choke* provided by the National Library of Medicine[3], and **(3) significant terms** including the most significant 128 words correlated to high AQIs in χ^2 statistics (Schütze et al., 2008).

The incoming tweets are filtered by the aforementioned keywords in the three groups. The tweets containing at least one of these keywords are likely to be relevant to the topics about air quality. We denote the corpus of relevant tweets as $D'(l,t)$. The features extracted from relevant tweets are expected to be more robust.

2.2 Stage 2: Feature Extraction

To extract features from text data, the effectiveness of convolutional neural networks (CNNs) has been demonstrated in many studies (Kim, 2014). In this paper, CNNs with max-over-time pooling are applied to derive the representation for every tweet. We then propose *max-over-tweet pooling* to aggregate tweet representations across all relevant tweets as the corpus representation. Finally, the features can be acquired by concatenating the

[2]EPA: https://www.epa.gov/

[3]https://www.nlm.nih.gov/medical-terms.html

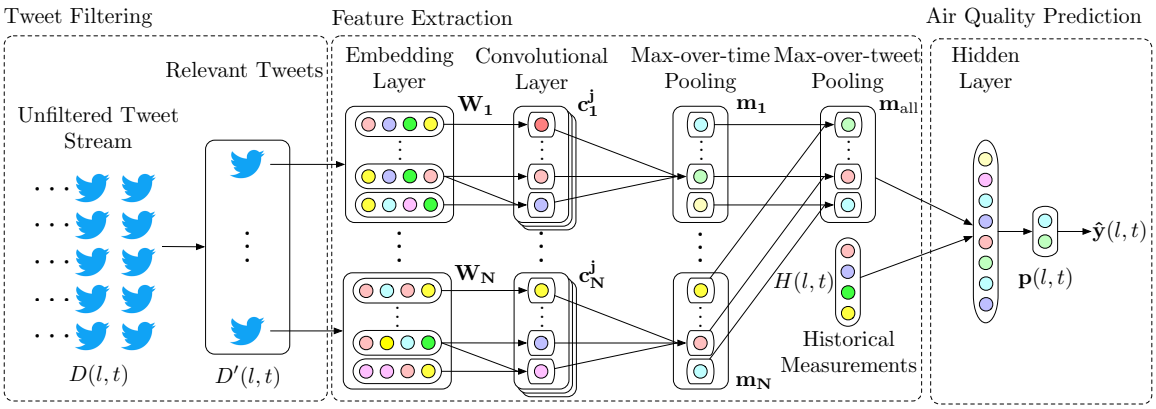

Figure 1: The framework of the proposed approach.

corpus representation and the historical measurements for prediction.

Tweet Representation. A tweet $\boldsymbol{w_i}$ can be represented by a matrix $\boldsymbol{W_i} \in R^{d \times |\boldsymbol{w_i}|}$, where d is the dimension of word embeddings; and $|\boldsymbol{w_i}|$ is the number of words in the tweet. As shown in Figure 1, a CNN with $d \times k$ kernels extracts the n-gram semantics of k contiguous words. Note that the row dimension of kernels is identical to the word embedding dimension to jointly consider the overall embedding vector. The convolution with the j-th kernel produces a numerical vector $\boldsymbol{c_i^j}$, which is then aggregated by max-over-time pooling (Collobert et al., 2011; Kim, 2014). As a result, the representation of a tweet $\boldsymbol{m_i}$ can be derived by chaining the pooled results of all kernels.

Corpus Representation. Since relevant tweets in the corpus can be myriad and not fixed, we need to aggregate various representations into an ultimate representation for the whole corpus. Here we propose max-over-tweet pooling to derive the corpus representation. The layer of max-over-tweet pooling reads all tweet representations and aggregates them by deriving the maximum value for each representation dimension. More precisely, a dimension of the representation can be treated as the sensor about a particular topic while the max-over-tweet pooling layer attempts to find the maximum sensor value among the sensor values of all relevant tweets. Finally, the max-over-tweet pooling layer can derive the corpus representation $\boldsymbol{m}_{\text{all}}$ by considering all tweet representations.

After determining the corpus representation $\boldsymbol{m}_{\text{all}}$, the final features $\boldsymbol{x}(l,t)$ for air quality prediction can be constructed by concatenating $\boldsymbol{m}_{\text{all}}$ and the historical measurements $H(l,t)$. As a consequence, the final features incorporate the knowledge of existing observations and the crowd power on social media.

2.3 Stage 3: Air Quality Prediction

To address the air quality prediction, we apply a fully-connected hidden layer to estimate the logits of all classes. More precisely, the logits $\boldsymbol{z}(l,t)$ can be computed as $\boldsymbol{z}(l,t) = F(\boldsymbol{x}(l,t))$, where $F(\cdot)$ is a fully-connected hidden layer with L hidden units; the dimension of $\boldsymbol{z}(l,t)$ is identical to the number of classes in air quality categorization. Then the probabilistic score for each class can be obtained with a softmax function (Goodfellow et al., 2016) when the prediction can be finally determined as the class with the highest score. Finally, the whole system can be computed and trained in an end-to-end manner and optimized by the cross-entropy loss (Goodfellow et al., 2016).

3 Experiments

3.1 Experimental Settings.

Data Collection. For social media data, we exploit the Twitter developer API[4] to crawl 1% of general English tweets published in the USA with location tags from November 17, 2015, to June 12, 2016. Each of the crawled tweets is associated with the corresponding county and state. EPA releases daily AQIs for every county in the USA publicly, which serve as the historical measurements and the gold standard.

Experimental Datasets. We conduct experiments to predict daily air quality conditions for locations fine-grained to the county level. More specifically, each of the samples can be represented by a tuple (l,t), where l is a county in the USA; t is a date

[4] https://developer.twitter.com/en/docs.html

Dataset	CA	ID	IN	IL	OH
Overall tweets	85.3M	1.2M	9.2M	23.2M	31.7M
Relevant tweets	11.8M	0.07M	0.5M	1.0M	1.4M
Training tuples	7,435	1,175	2,990	1,804	3,647
Validation tuples	1,487	235	598	361	729
Testing tuples	1,483	235	599	361	730

Table 2: Statistics of five experimental datasets. The relevant tweets refer to the remaining tweets after the stage of tweet filtering.

with crawled tweets. For each tuple, the historical measures are the AQIs in the previous seven days as seven numerical features. Five experimental datasets are then constructed with the data of the five most polluted states according to the annual report from America Health Ranking[5], including California (CA), Idaho (ID), Illinois (IL), Indiana (IN), and Ohio (OH). The overall datasets are further partitioned by time into a 30-week training dataset, two 5-week datasets for validation and testing. As a result, Table 2 shows the statistics of five experimental datasets. Note that more than 90% tweets are filtered as irrelevant tweets in the stage of tweet filtering. It also shows the necessity of filtering irrelevant tweets that can probably be noises for air quality prediction.

Implementation Details Our approach is implemented by Tensorflow (Abadi et al., 2016) and trained by the Adam optimizer (Kingma and Ba, 2014) with an initial learning rate 10^{-3}. After parameter tuning, λ is set to 10^{-3} while the number of hidden units in the hidden layer L is 128. The dimension of the word embeddings is 300. All of the activation functions in the model are set to exponential linear units (ELUs) (Clevert et al., 2015). For CNNS, 96 kernels with different sizes from 2 to 4 are applied to obtain a 96-dimensional representation for each relevant tweet in the corpus.

Baseline Methods. Because we are the first study using social media to predict air quality situation, there are much few available methods. Even though some studies (Jiang et al., 2015) claim the capability of inferring ongoing AQIs with social media, they apply strong restrictions to derive features for highly polluted cities so that they are incapable of tackling most of the cases in our experiments. In the experiments, we compare with two baseline methods as follows: **(1) Prediction with only AQIs (PAQI):** To understand the base performance, PAQI predicts the air quality conditions with only historical measurements. The knowledge of social media is ignored for this baseline method. **(2) Bag-of-words Features (BOW):** To demonstrate the effectiveness of extracted features, we replace the extracted features with conventional bag-of-words features as a baseline method. Note that all baselines apply a neural network with a hidden layer for prediction.

[5]https://www.americashealthrankings.org

3.2 Experimental Results

For evaluation, micro- and macro-F1 scores are selected the evaluation metrics. Table 3 demonstrates the performance of the three methods. Micro-F1 scores are generally better than macro-F1 scores because the trivial cases like the class of good air quality are the majority of datasets with higher weights in micro-F1 scores. PAQI is better than BOW although BOW uses the knowledge of social media. It is because BOW features involve all irrelevant words so that the actual essential knowledge cannot be recognized. Our approach significantly outperforms all baseline methods in almost all metrics. More precisely, our approach improves the air quality prediction over PAQI from 6.92% to 17.71% in macro-F1 scores. The results demonstrate that social media and NLP can benefit air quality prediction.

In addition to the unbalanced datasets based on the categorization of EPA, we also conduct the experiments with relatively balanced datasets to show the robustness of our proposed approach. More specifically, the categorization is refined to four classes with finer windows of AQIs, including: $[0, 25)$, $[25, 50)$, $[50, 75)$, and $[75, \infty)$. Figures 2 and 3 illustrate the Micro- and Macro-F1 scores of PAQI and our approach in the refined datasets. The experimental results show that the improvements are consistent with the experiments in unbalanced datasets of extreme air quality prediction. It also demonstrates the robustness of our proposed approach.

4 Conclusions and Discussions

In this paper, we propose a novel framework for leveraging social media and NLP to air quality prediction. After filtering irrelevant tweets, a CNN derives a feature vector for each tweet with max-over-time pooling. We also propose the novel max-over-tweet pooling to aggregate the feature vectors of all tweets over numerous hid-

Dataset	Method	Micro Average			Macro Average		
		Prec.	Rec.	F1	Prec	Rec.	F1
ID	BOW	0.807	0.829	0.809	0.687	0.619	0.631
	PAQI	0.816	0.728	0.757	0.611	0.677	0.617
	Ours	**0.863**	**0.811**	**0.828**	**0.691**	**0.776**	**0.714**
IN	BOW	0.792	0.786	0.786	0.508	0.508	0.501
	PAQI	0.847	0.682	0.737	0.567	0.649	0.548
	Ours	**0.855**	**0.849**	**0.852**	**0.640**	**0.652**	**0.645**
IL	BOW	0.775	0.802	0.791	0.506	0.499	0.484
	PAQI	0.834	0.686	0.737	0.580	**0.666**	0.566
	Ours	**0.844**	**0.847**	**0.845**	**0.646**	0.638	**0.640**
OH	BOW	0.744	0.780	0.760	0.515	0.512	0.510
	PAQI	0.800	0.683	0.724	0.569	0.622	0.562
	Ours	**0.813**	**0.813**	**0.815**	**0.629**	**0.627**	**0.627**
CA	BOW	0.647	0.683	0.660	0.495	0.488	0.485
	PAQI	0.826	0.725	0.745	0.700	0.772	0.694
	Ours	**0.830**	**0.786**	**0.798**	**0.728**	**0.786**	**0.742**

Table 3: The overall classification performance of the baseline methods and our approach. All of the improvements of our approach (ours) over PAQI are significant with a paired t-test at a 99% significance level.

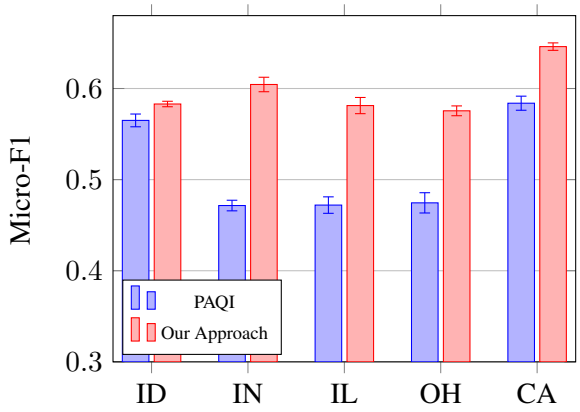

Figure 2: Micro F1 scores with four-class categorization. All of the improvements of our approach over the baseline method are significant with a paired t-test at a 99% significance level.

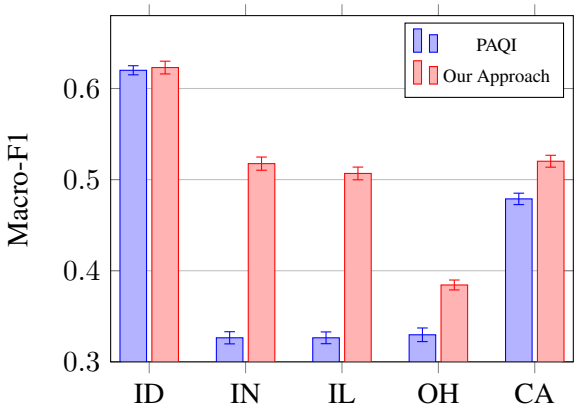

Figure 3: Macro F1 scores with four-class categorization. All of the improvements of our approach over the baseline method are significant with a paired t-test at a 99% significance level.

den topics. Finally, the corpus representation can be taken into account to predict air quality with historical measurements. The results of extensive experiments show that our proposed approach significantly outperforms two comparative baseline methods across both balanced and unbalanced datasets for different locations in the USA. This is because: (1) Most noisy and irrelevant tweets are effectively filtered in the stage of tweet filtering; (2) The convolutional neural network and the proposed max-over-tweets are able to extract essential knowledge about air quality prediction from myriad tweets in social media; (3) There are some limitations on only using historical measurements, such as the capability of recognizing real-world events.

Acknowledgement

We would like to thank the anonymous reviewers for their helpful comments. The work was partially supported by NIH U01 HG008488, R01 A132030, and NSF DGE-1829071.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283.

Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707. IEEE.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Mark Dredze. 2012. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84.

D Deniz Genc, Canan Yesilyurt, and Gurdal Tuncel. 2010. Air pollution forecasting in ankara, turkey using air pollution index and its relation to assimilative capacity of the atmosphere. *Environmental monitoring and assessment*, 166(1-4):11–27.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*, volume 1. MIT press Cambridge.

Wei Jiang, Yandong Wang, Ming-Hsiang Tsou, and Xiaokang Fu. 2015. 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*, 10(10):e0141185.

Yifei Jiang, Kun Li, Lei Tian, Ricardo Piedrahita, Xiang Yun, Omkar Mansata, Qin Lv, Robert P Dick, Michael Hannigan, and Li Shang. 2011. Maqs: a personalized mobile sensing system for indoor air quality monitoring. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 271–280. ACM.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2013. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931.

Mauricio Santillana, André T Nguyen, Mark Dredze, Michael J Paul, Elaine O Nsoesie, and John S Brownstein. 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10):e1004513.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press.

Jing Fan Xiaojin Zhu Shike Mei, Han Li and Charles R.Dyer. 2014. Inferring air pollution by sniffing social media. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 534–539. ACM.

Yan-dong Wang, Xiao-kang Fu, Wei Jiang, Teng Wang, Ming-Hsiang Tsou, and Xin-yue Ye. 2017. Inferring urban air quality based on social media. *Computers, Environment and Urban Systems*, 66:110–116.

Tailai Wu, Zhaohua Deng, Zhanchun Feng, Darrell J Gaskin, Donglan Zhang, and Ruoxi Wang. 2018. The effect of doctor-consumer interaction on social media on consumers health behaviors: Cross-sectional study. *Journal of medical Internet research*, 20(2).

Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. 2015. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2267–2276. ACM.