Proposal: When Air Quality Prediction Meets Twitter

Weijia Lu
tt18284@bristol.ac.uk
University of Bristol
(supervised by Dr. Peter Bennett and Mr. Joshua Taylor)
Bristol, UK

EXECUTIVE SUMMARY

As the population living in urban areas and the global number of cars continuously increase, air pollution levels keep rising in both developed and developing countries. With air pollution the problem is highly related to human health and material damage, air quality is becoming more and more a concern to citizens. In order to quantify whether the quality of air is good or bad, Air Quality Index(AQI) is used by government agencies, which tends to give citizens a more intuitive view of air quality condition and help citizens do preparation for going out. However, due to the high cost of constructing and maintaining air quality monitoring station, there is a great shortage and skew of air quality data. Therefore, a more economical way becomes essential as an alternative. This project will try to use Twitter data to make prediction on air quality index.
The aim of this project is to construct a model to predict urban air quality index by using Twitter data.
The objectives of this project are to:

- Construct a streamer which will automatically fetch related real-time tweets and do the analysis.
- Fetch at least 100 thousand historical tweets about air quality based on location and time.
- Collect real sensor data and do the preprocessing based on location and time.
- Construct an auto-bot which can interact with engaged users to collect more data without bothering users.
- Collect at least 10 thousand engaging tweets of Bristol air quality and make the visualization.
- Design a model to demonstrate the correlation between Twitter data and real-time air quality index.

Deliverables can be shown as follows:

- A literature survey focusing on the method how other researchers using social media data to make the prediction, especially in air quality field.
- A cloud service fetching Twitter stream and making automatic reply with 24-hour service.
- A database to store the analysis result of air-quality-related tweets.
- A website to visualize citizen's rate and suggestion about air quality in Bristol.
- A poster to engage more people to give their daily rate about Bristol air quality.
- A model to show the correlation between Twitter data and AQI.

This project will reveal how Twitter data should be processed to make the prediction of real-time air quality index. What's more, the project will compare the passive historical Twitter data about air quality and the active format data to test the hypothesis whether the active format data will improve the prediction. In addition, the human computer interaction factors such as using auto-bot will also be tested to prove the hypothesis that auto-bot will improve the prediction. It is likely these results will show a novel way to collect Twitter data and encourage more interaction in the process of obtaining Twitter data instead of just using passive historical data.

1 INTRODUCTION

1.1 Air quality

As the population living in urban areas and the global number of cars continuously increase, air pollution levels keep rising in both developed and developing countries. With air pollution problem is highly related to human health and material damage, air quality is becoming more and more a concern to citizens. In order to quantify whether the quality of air is good or bad, Air Quality Index(AQI) is used by government agencies. The AQI is first put forward in 1968 by National Air Pollution Control Administration. Since then, different AQI standards are developed by government of each country. Illustrated by the case of United Kingdom, the Daily Air Quality Index is recommended by the Committee on Medical Effects of Air Pollution(COMEAP). It has 10 points, from 1 to 10, where a lower value indicates a better air quality condition. The index is based on the concentrations of 5 pollutants monitored by real-time monitors. The index value will give citizens an intuitive vision of air quality condition and help citizens make necessary preparation for going out. However, due to the high cost of monitor base, the number of air quality monitors is limited. What's more, almost all air quality monitors are concentrated in the city center. Take Bristol as an instance, as the figure[] shows, there are only 5 monitors in Bristol area which are all within a radius of 3 miles from the city center of Bristol.

Figure 1: Sensors in Bristol

Such shortage and limitation of distribution of the air quality monitors make the governance measures and policies based on the analysis of real-time air quality monitoring data less feasible. In some developing countries, this situation is more serious. The public can't even make preparation for going out based on immediate air quality data. Therefore, a more economical way becomes essential as an alternative.

1.2 Twitter Prediction

Twitter could be a great alternative way to predict air quality. Twitter who has an active user number of 330 million has great power as a new medium of information sharing. Moreover, users of Twitter intend to talk about their daily activities and to seek or share information, which makes prediction by this data feasible.[1] Nowadays, data of social media like Twitter is used by many researchers to make prediction in various field. Bollen J[2] uses Twitter to predict the stock market. Burnap P[3] demonstrates the possibility to predict election by Twitter data.

However, in the field of air quality, the prediction by using social media data is not such common so far. Both W.Jiang [5] and Wang[15] used Sina Weibo data to predict outdoor air quality index. Both of the researches achieved a great correlation but just easily correlated the volume of message with the air quality index and used just passive data to make prediction. And this project will try to use active-format data(actively engaging Twitter users to tweet about air quality) to improve the robustness of prediction.

1.3 Aims and objectives

The aim of this project is to predict urban air quality by using Twitter data. Based on related work of Twitter prediction on air quality, this project has hypothesis as followed:

(1) Twitter data about air quality can be correlated with real-time air quality index.
(2) Using active format-data by engaging Twitter users will have a higher prediction accuracy than passive data.
(3) Using auto-bot which make conversation with engaged Twitter users will improve prediction accuracy.

The objectives of this project are to:

- Construct a streamer which will automatically fetch related real-time tweets and do the analysis.
- Fetch at least 100 thousand historical tweets about air quality based on location and time.
- Collect real sensor data and do the preprocessing based on location and time.
- Construct an auto-bot which can interact with engaged users to collect more data without bothering users.
- Collect at least 10 thousand engaging tweets of Bristol air quality and make the visualization.
- Design a model to demonstrate the correlation between Twitter data and real-time air quality index.

1.4 Deliverables

- A literature survey focusing on the method how other researchers using social media data to make the prediction, especially in air quality field.
- A cloud service fetching Twitter stream and making automatic reply with 24-hour service.
- A database to store the analysis result of air-quality-related tweets.
- A website to visualize citizen's rate and suggestion about air quality in Bristol.
- A poster to engage more people to give their daily rate about Bristol air quality.
- A model to show the correlation between Twitter data and AQI.

1.5 Added value

This project will achieve the prediction of urban air quality based on Twitter data. Currently, few studies apply data of social media on air quality prediction, most of which are based on Sina Weibo, a Chinese social media platform. If successful, this project will try to reveal how Twitter data should be processed to make the prediction of real-time air quality index. What's more, the project will compare the passive historical Twitter data about air quality and the active format-data by engaging users to rate the air quality to test the hypothesis whether the active format data will improve the prediction. In addition, the human computer interaction factors such as using auto-bot will also be tested to prove the hypothesis that auto-bot will improve the prediction. It is likely these results will show a novel way to collect Twitter data and encourage more interaction in the process of obtaining Twitter data instead of just using passive historical data.

2 LITERATURE REVIEW

2.1 Air Quality

Since 1950, the world population has almost tripled, and the global number of cars has increased by a factor of 20. Moreover, the population living in urban areas has increased from below 30% to 55%. Such factors bring serious urban air pollution problems. With air pollution problem is highly related to human health and wealth damage, air quality is becoming more and more a concern to citizens.

J Fenger[6] describes the history of air pollution and air quality monitoring. Initially the outdoor air pollution was a purely urban phenomenon, then the problems were extensive. Up to the Second World War many people had an ambivalent attitude towards pollution, the records of material damage and impacts on human health and vegetation started to be used as evaluation of the early urban air pollution. After London smog disaster, the number of measuring sites started increasing, a more systematic and official measurement of air pollutants entered the stage of the times.

In order to quantify whether the quality of air is good or bad, Air Quality Index(AQI) is used by government agencies. The AQI concept is first put forward in 1968 by National Air Pollution Control Administration. Since then, different AQI standards are developed by government of each country. As the air pollution is with complex nature that there are about 3000 different anthropogenic air pollutants, a condensed and simplified monitoring data is required to help public reporting and decision makers[7]. Air quality monitoring stations are used to measure the level of such complex ingredient of pollutants before the data is then concentrated into a concise value to represent air quality level. Currently in UK, a value from 1 to 10 recommended by Committee on Medical Effects of Air Pollution(COMEAP) is commonly used as air quality index (as fig2), with a lower value indicating a better air quality condition.

Despite the use of air quality monitoring stations, the air quality measurement still faces a serious problem. Due to the high cost of building and maintaining a monitor station, the number of air quality measurement station is insufficient. Based on investigation by Microsoft Research Asia[8], an air quality monitor station needs a certain size of land, 200,000 USD for construction and 30,000 USD per year for maintenance, as well as the corresponding human resources. What's more, almost all air quality monitors are concentrated in the city center.

Such shortage and limitation of distribution of the air quality monitors will make data lack as the the urban air quality varies non-linearly by location[8]. Therefore, the lack and skew of air quality data will make the governance measures and policies based on the analysis of real-time air quality monitoring data less feasible. In some developing countries, this situation is more serious. Therefore, a more economical way becomes essential as an alternative.

2.2 Twitter Prediction

Twitter is a microblogging service published on 2006, which is growing fast and currently has more than 300 million active users worldwide (as fig3). According to Kwak[9], Twitter has the characteristic of rapid spread and trend focus. Users of Twitter tend to talk about their daily activities, seek or share information, and has connection with other users with similar intentions[1], both of which makes further application based on big data of Twitter feasible.[10] Nowadays, data of social media like Twitter is used by many researchers to make prediction in various field.

Bollen J[2] uses Twitter mood to predict the stock market. It investigates the connection between emotions and decision-making on stock market according to behavioral economics. This paper is one of the several earliest researches on making prediction via Twitter data. And it's 87.6% accuracy in predicting daily up and down changes indicates the feasibility of such prediction. After that, more and more researchers started predicating stock via Twitter data. Ruiz[11] studied the relationship between Twitter activities and stock market under a graph based view. Feldman[12] introduced a hybrid approach based on companies' news articles.

Then the research field extended to more areas than just financial field, among which the social topic became a trend. Tumasjan[13] first published the attempt to use Twitter to forecast election in 2010. Burnap P[3] then demonstrates the possibility to predict election by Twitter data. In 2012, Wang[14] applied the Twitter prediction on crime detection. S. Sachdevg[4] uses Twitter data to model the impact of one wildfire in Northern California.

Air Pollution Banding	Value	Accompanying health messages for at-risk individuals*	Accompanying health messages for the general population
Low	1-3	**Enjoy** your usual outdoor activities.	**Enjoy** your usual outdoor activities.
Moderate	4-6	Adults and children with lung problems, and adults with heart problems, **who experience symptoms**, should **consider reducing** strenuous physical activity, particularly outdoors.	**Enjoy** your usual outdoor activities.
High	7-9	Adults and children with lung problems, and adults with heart problems, should **reduce** strenuous physical exertion, particularly outdoors, and particularly if they experience symptoms. People with asthma may find they need to use their reliever inhaler more often. Older people should also **reduce** physical exertion.	Anyone experiencing discomfort such as sore eyes, cough or sore throat should **consider reducing** activity, particularly outdoors.
Very High	10	Adults and children with lung problems, adults with heart problems, and older people, should **avoid** strenuous physical activity. People with asthma may find they need to use their reliever inhaler more often.	**Reduce** physical exertion, particularly outdoors, especially if you experience symptoms such as cough or sore throat.

Figure 2: UK Air Quality Index

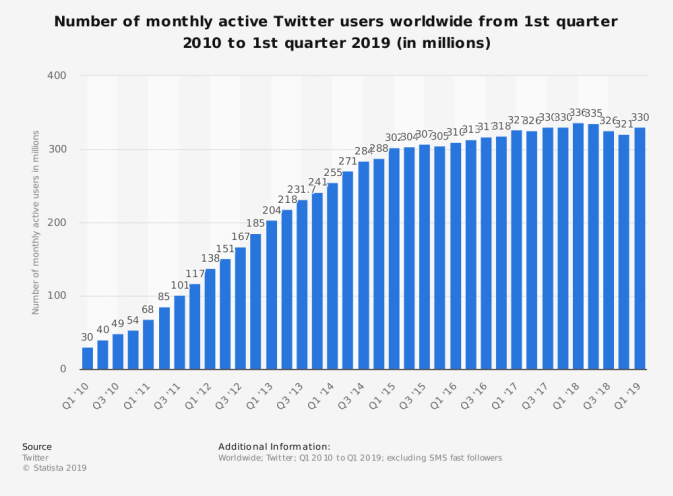

Figure 3: Number of Active Twitter Users Overtime

However, in the field of air quality, the prediction by using social media data is not such common so far. Both W.Jiang[5] and Wang[15] used Sina Weibo data to predict outdoor air quality index. They indicated that there is a close 70% Pearson correlation between the volume of pollution-related messages and air quality index in China. Despite a great correlation found in these researches, there are still three problems to be deal with.

(1) The filter model of message processing and correlation of two researches are based on Sina Weibo which is in simplified Chinese language. There is no research investigating the correlation of English-based tweets and air quality index in UK so far.
(2) The researches before just directly correlated the volume of message with the air quality index without any further analysis such as sentiment analysis, which will make the correlation not robust.
(3) All the data of former researches are historical data fetched from data which is passive. No research tests the feasibility of using active-format data to make the prediction.

Therefore, this project aims to improve these three insufficiency in the area of predicting air quality by twitter.

2.3 Human computer interaction on Twitter

To get air quality data on Twitter by engaging users, the participation level of users and the effectiveness of data become a key thing, which leads out the topic of human computer interaction on Twitter. Edwards C[16] studied the impact of 'botification' of Twitter and concluded the result that human

agent is rated higher than the Twitterbot. Chang[17] studied the effect of hashtag using in tweets. He concluded that hashtag use has become a unique tagging convention to help associate Twitter messages with certain events or contexts which can efficiently embody user participation.

3 PROPOSED APPROACH

3.1 Methodology

The methods of this project can be drawn by the content of objectives.

(1) Automatic data scrapper on cloud.
A Twitter streaming API based streamer module will be run on Oracle cloud 24 hourly to automatically fetch real-time data and do the sentiment analysis and information extraction. Both the original fetched data and the analysis result will then be stored into database automatically.
(2) Historical Air Quality Data.
Such data includes two part, the real sensor historical data and the Twitter historical data. Twitter Search API will be used to get Twitter data and the real sensor data will be downloaded from government website. Both data need to be cleaned and pre-process.
(3) Twitter Bot to interact with engagers.
A Twitter Bot will be designed to make conversation with engagers. Dispatch module based on data distribution in database and analysis of user(such as his/her location) will automatically make questions to engagers to get more needed data. Sentiment analysis and Twitter API will be used to implement the Bot.
(4) Design a engagement topic.
Survey will be conducted to compare different design of engagement topic and a slogan and hashtag will be selected.
(5) Correlation Model.
Topic model or correlation metric such as Pearson Correlation will be used to evaluate the accuracy of prediction on air quality.

3.2 Research Questions

(1) Can Twitter data be used to predict real-time air quality index?
(2) Will active-format data be more effective than passive historical data when taking prediction on air quality?
(3) Will interaction Twitter Bot improve the data?

3.3 Evaluation

(1) An engagement metric will be designed to evaluate the participation of users.
(2) A spatio-temporal based metric will be designed to evaluate the collected air quality data.
(3) A correlation model or metric will be introduced to evaluate the accuracy of prediction.

3.4 Expected Contributions

(1) A model to collect and analysis Twitter data which can effectively make prediction on air quality index.
(2) A pipeline that includes user engagement, data collection and analysis will be constructed to make prediction based on active-format data feasible.

4 GANNT CHART

See Figure 4.

5 RISK ANALYSIS

There are a number of risks that could affect the progress of the project. The likelihood rating is on a scale 1-5 where each point p represents a p/5 probebility of incidence, and similarly severity acknowledges p/5 possibility of the risk have implications on the project outputs. See the risk analysis in Table 2.

REFERENCES

[1] Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we Twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis(pp. 56-65). ACM.
[2] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1-8.
[3] Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 characters to victory?: Using Twitter to predict the UK 2015 General Election. Electoral Studies, 41, 230-233.
[4] S. Sachdeva, S. McCaffrey, and D. Locke, âĂIJSocial media approaches to modeling wildfire smoke dispersion: spatiotemporal and social scientific investigations,âĂİ Information, Communication & Society, vol. 0, no. 0, pp. 1âĂŞ16, Aug. 2016.
[5] W. Jiang, Y. Wang, M.-H. Tsou, and X. Fu, âĂIJUsing Social Media to Detect Outdoor Air Pollution and Monitor Air Quality Index (AQI): A Geo-Targeted Spatiotemporal Analysis Framework with Sina Weibo (Chinese Twitter),âĂİ PLoS ONE, vol. 10, no. 10, p. e0141185, Oct. 2015.
[6] 6. Fenger, J. (1999). Urban air quality.Atmospheric environment,33(29), 4877-4900.
[7] 7. Wiederkehr, P., & Yoon, S. J. (1998). Air quality indicators. InUrban Air PollutionâĂŤEuropean Aspects(pp. 403-418). Springer, Dordrecht.
[8] Zheng, Y., Liu, F., & Hsieh, H. P. (2013, August). U-air: When urban air quality inference meets big data. InProceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining(pp. 1436-1444). ACM.
[9] Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media?. InProceedings of the 19th international conference on World wide web(pp. 591-600). AcM.
[10] Aslam AA, Tsou M-H, Spitzberg BH, An L, Gawron JM, Gupta DK, et al. The Reliability of Tweets as a Supplementary Method of Seasonal Influenza Surveillance. Journal of Medical Internet Research. 2014;16(11):e250. PMC42600664. pmid:25406040
[11] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A. and Jaimes, A. 2012. Correlating financial time series with micro-blogging activity. In Proceedings of the fifth ACM international conference on Web search and data mining (WSDM-2012), 513-522.

[12] Feldman, R., Benjamin, R., Roy, B. H. and Moshe, F. 2011. The Stock Sonar - Sentiment analysis of stocks based on a hybrid approach. In Proceedings of 23rd IAAI Conference on Artificial Intelligence (IAAI-2011).

[13] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185.

[14] Wang, X., Gerber, M. S., & Brown, D. E. (2012, April). Automatic crime prediction using events extracted from Twitter posts. In International conference on social computing, behavioral-cultural modeling, and prediction(pp. 231-238). Springer, Berlin, Heidelberg.

[15] Wang, S., Paul, M. J., & Dredze, M. (2015). Social media as a sensor of air quality and public response in China.Journal of medical Internet research,17(3), e22.

[16] Edwards C, Edwards A, Spence P R, et al. Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter[J]. Computers in Human Behavior, 2014, 33: 372-376.

[17] Chang, H. C. (2010). A new perspective on Twitter hashtag use: Diffusion of innovation theory.Proceedings of the American Society for Information Science and Technology,47(1), 1-4.

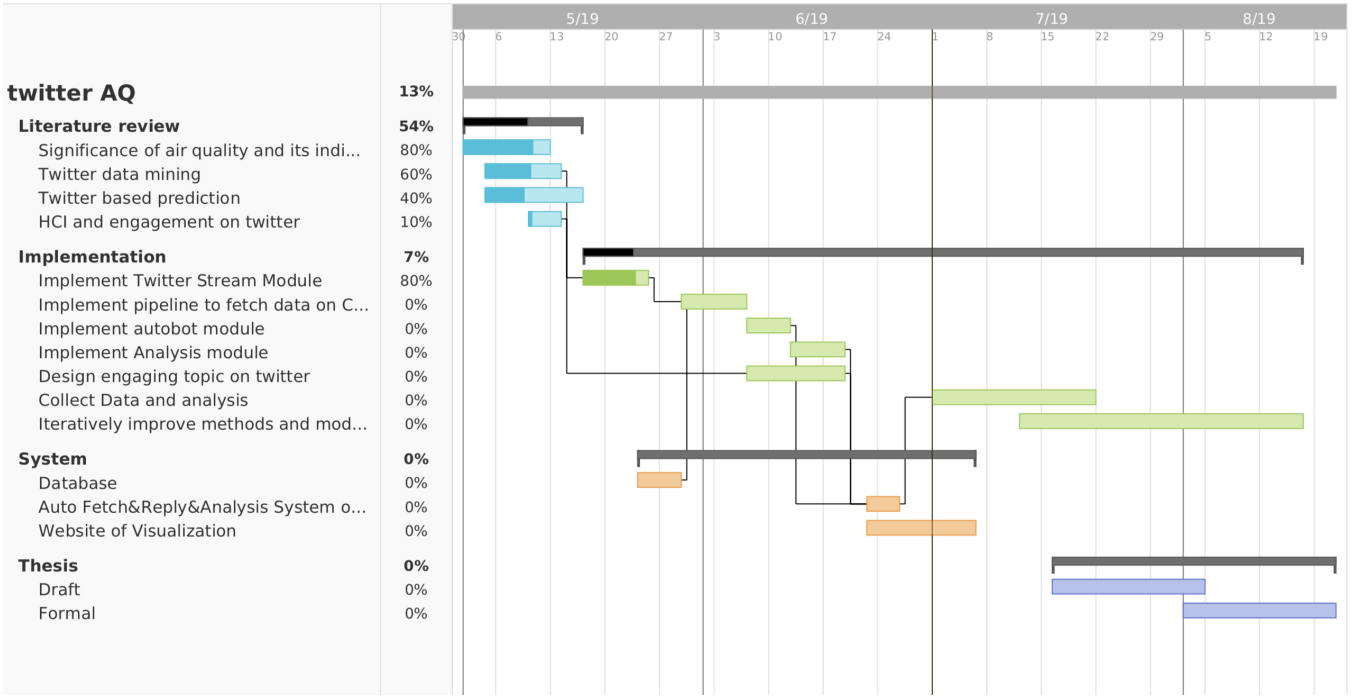

Figure 4: Gantt Chart

Table 1: Risk Analysis

Risk	Likelihood	Severity	Notes and Contingency
Lack of engaging tweets	4	4	Making Post and invite some fame to attend
Correlation Analysis Failure	2	5	More Literature Review and method references
Twitter API limitation	5	3	Use streaming algorithm and apply for advanced API account
Imperfect topic selection of rate engagement	4	2	Iteratively poll and analysis
Not enough time for Visualization Website	5	1	Construct a simpler version of visualization
Not Effective of Twitter Bot Design	4	2	Use more format interaction way rather than relying on sentiment analysis