Research Collection School Of Computing and Information Systems

School of Computing and Information Systems

1-2016

A study on Singapore haze

Bingtian DAI
Singapore Management University, btdai@smu.edu.sg

Kasthuri JAYARAJAH
Singapore Management University, kasthurij.2014@smu.edu.sg

Ee-Peng LIM
Singapore Management University, eplim@smu.edu.sg

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

Shriguru NAYAK
Singapore Management University, shrigurun@smu.edu.sg

Citation

A Study on Singapore Haze

Bing Tian Dai   Kasthuri Jayarajah   Ee-Peng Lim   Archan Misra
Shriguru Nayak
School of Information Systems
Singapore Management University
{btdai, kasthurij.2014, eplim, archanm, shrigurun}@smu.edu.sg

ABSTRACT

In 2015, Singaporean have experienced one of the worse air pollution crises in history. With datasets from a well-known photo sharing social network, we analyze how this haze affects Singaporean's daily life. We will share our preliminary results in this paper.

CCS Concepts

•**Human-centered computing** → **Social networks; Social network analysis;** •**Applied computing** → *Law, social and behavioral sciences;*

Keywords

Online Social Networks, Social Media Analyses, Singapore Haze, Behavior Studies

1. INTRODUCTION

The 2015 southeast asia haze is an ongoing air pollution crisis affecting several southeast asia countries including Singapore, Indonesia, Malaysia, Brunei and part of Thailand. Indonesia, being one of the major victims and the instigator at the same time, failed to stop people doing so-called *Slash-and-burn* practice, which then caused the months-long forest fire. The smoke has quickly spreaded to Indonesia's neighboring countries, e.g., Singapore, and triggered health problem and affected people's daily lives.

In this paper, we mainly focus on studying *how this haze affects Singaporean's behaviors*, from the `Instagram`[1] posts we collected. `Instagram` is a mobile photo-sharing social network which allows people to take picture spontaneouly. These photos may describe where users are, and what objects are being captured. Therefore, it is possible to find our answers from `Instagram` photos to answer the aforementioned question.

---

[1]instagram.com

There are two types of features in `Instagram` posts, meta-data-based and content-based. For example, meta-data-based features may include *where* and *when* the photos are taken. Content-based feature can be any feature describing the picture itself, e.g., if the picture is a scenery picture or a selfie.

Our analyses are also carried out based on these two types of features. We first analyze the meta-data-based features from Section 3 to Section 5, and then analyze the content-based features in Secture 6.

2. DATASETS

We collected `Instagram` data posted within the Singapore area or posted by Singapore users since the beginning of 2015 through its public API. In order to study haze that affects Singapore and other southeast asia countries in the month of September, we chose to compare data we collected in March 2015 with that collected in September 2015.

Not every `Instagram` post comes with geographical location. We therefore further selected the two subset of posts with specified locations. One subset contains 30 dyas of geotagged `Instagram` posts in March (from March 1st 0000hrs to March 30th 2359hrs), and another also contains 30 days of geotagged `Instagram` posts in September. We call them the March dataset and September dataset respectively. Their basic Statistics are shown in Table 1. Note that all locations come with latitudes and longitudes, but only some locations come with specified venue identifiers and venue names in `foursquare`[2], as specified by users.

Table 1 shows that the number of `instagram` posts is more in March than in September, but the number of users who posted in March are fewer. The number of specified locations in March are a lot more (1.6 times) than the specified locations in September, which is to be discussed with great details in Section 3.

	March	September
Number of Instagram posts	974,153	869,845
Number of Users	164,851	179,147
Number of Specified Locations	89,168	55,501

Table 1: Basic Statistics about March and September Datasets

---

[2]foursquare.com

(a) Percentile heat map of Instagram counts in March

(b) Grid cells with zero count in March but non-zero count in September

(c) Comparison heat map of Instagram counts in March against September

(d) Percentile heat map of Instagram counts in September

(e) Grid cells with zero count in September but non-zero count in March

(f) Comparison heat map of Instagram counts in September against March
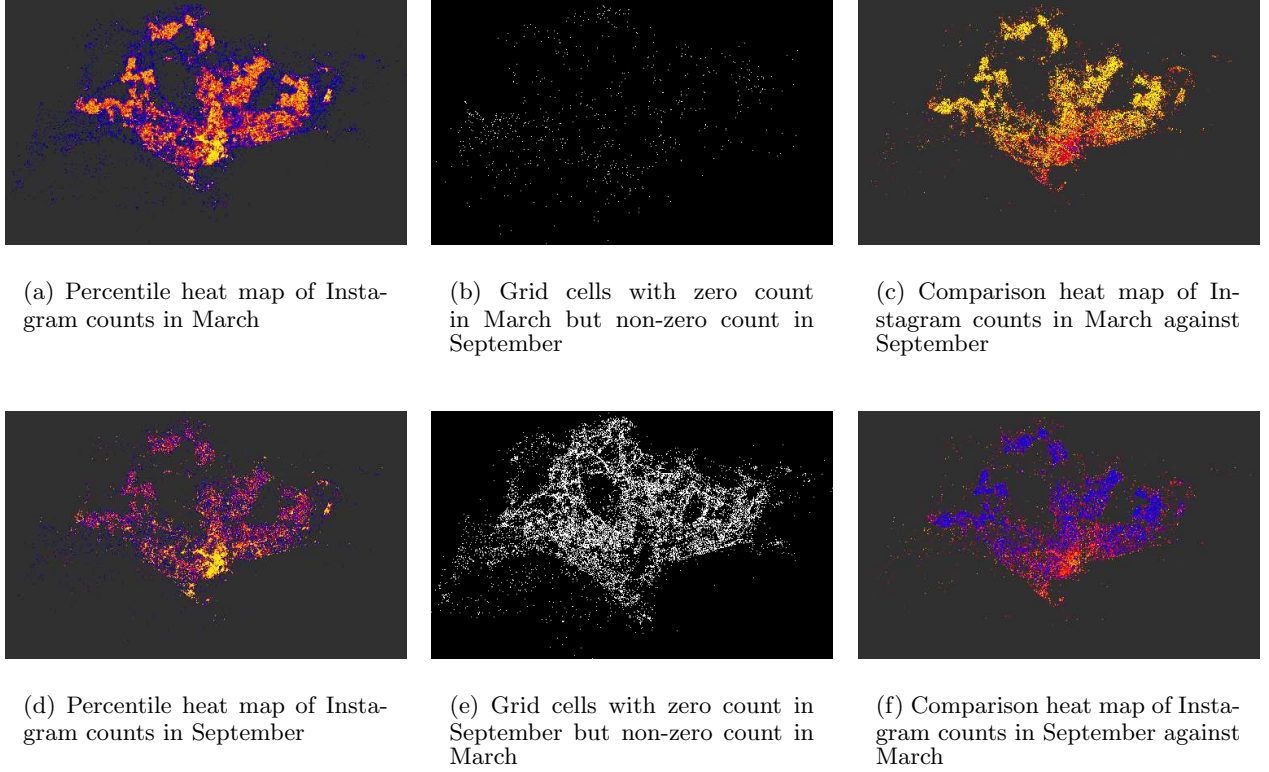
Figure 1: Heat maps of Instagram posts

3. GEOGRAPHICAL ANALYSES

We first analyze the locations of Instagram posts. As not all locations are specified, we are going to first present the volume of the Instagram posts. We first partition our Instagram posts using their latitude-longitude locations into grid cells with length equals to $\frac{1}{1000}$ degree or approximately 111 meters. It takes 494×304 grid cells to cover the entire Singapore.

Each grid cell is assigned a color as shown in Figure 1(a) and Figure 1(d), representing the count of Instagram posts within that specific grid cell. As such counts are extremely heterogeneous, instead of visualizing the counts directly, we visualize the percentiles of these counts. The blue grids are on the low side while the yellow ones are on the high side.

It is very obvious that the heat map for the September dataset is more seggregated in smaller region as shown in Figure 1(d) with high count of posts concentrated in the downtown area. The March heat map however shows high counts convering many more areas in Singapore. This tells us that the haze has changed people's behaviors. They visit downtown more often than days without haze. It is probably because there are more air conditioned buildings and better connected buildings with underground walkways and subway stations in the downtown area, and people feel better there because air conditioners shield them from the hazy air. On the other hand, other parts of Singapore do not enjoy as much protection and convenience.

We next try to answer the questions: "Are there locations people avoid during the hazy days?" and "Where are these

location?". We then compare the Instagram counts of March with that of September. As shown in Table 1, the number of specified locations in September is much less than that in March. We suspect this is the same case for the grid cells, i.e., the number of grid cells with non-zero counts are much fewer in September than in March. Thus we visualized the grid cells with zero counts in March but non-zero counts in September in Figure 1(b), and the grid cells with zero counts in September but non-zero counts in March in Figure 1(e).

It is as what we expected that, such grid cells in March are much more sparse than such grid cells in September. The zero-count grid cells in September almost cover the entire Singapore. This clearly shows that, the amount of ourdoor activities has been tremendously reduced during the hazy days.

Next, we further contrast the instagram posting behavior of March and September. We compare the counts of Instagram posts in March to the counts in September by assigning blue for grid cells with counts less than or equal one tenth of the counts in the other month, provided both counts are not zero, yellow for grid cells with counts more than or equal ten times of the counts in the other month, and red for grid cells with almost equal counts. Figure 1(c) and Figure 1(f) have confirmed with Figure 1(a) and Figure 1(d) that people visit downtown much more often during the hazy days.

3.1 Top Indoor and Outdoor Locations

We now pick some specific popular indoor and outdoor locations to conduct a direct comparison of posting activi-
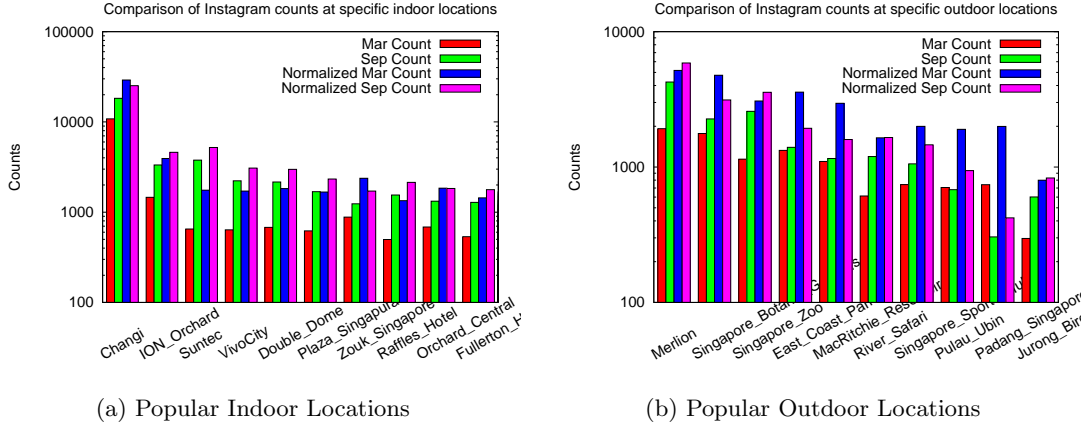
(a) Popular Indoor Locations

(b) Popular Outdoor Locations

Figure 2: Instagram counts at popular indoor and outdoor locations

ties of March and September. This hopefully will help us to further substantiate our earlier findings. We picked top 10 locations by our own judgement on whether it is an indoor or an outdoor place, ranked by popularity. For example, Marina Bay Sands is a mixture of indoor and outdoor locations, thus we did not put it in either list.

At first, we found that, the `Instagram` counts of September have increased from the count of March at almost every location, as shown by the left two vertical bars at each location in Figure 2. Note that we have earlier stated, locations are marked by latitudes and longitudes, but not every locations comes with location ID and name. So when we inspect the `Instagram` counts at one specific location, we actually discard the `Instagrams` that could belong to this location, but not specified by the location ID. We compared the number of Instagrams with location-ID specified in our March and September dataset in Table 2: the probability that a location in September comes with a location ID is almost the double of that in March. So then we normalized the counts by dividing the counts by their respective ratio, which are shown by the right two vertical bars at each location in Figure 2. With this normalization, the counts in September at indoor places are generally increased from March, while the counts in September at outdoor places are generally decreased, which matches with what we have found in Figure 1.

	March	September
# of Instagrams with LatLng	974,153	869,845
# of Instagrams with Location ID	361,889	629,610
Ratio	37.15%	72.38%

Table 2: Ratio of locations with location-ID specified in March and September datasets

4. TEMPORAL ANALYSES

We now analyze `Instagram` counts against different time of a weekday or a weekend. Figure 3 shows the average counts for March weekday, March weekend, September weekday and September weekend. As the total number of In-
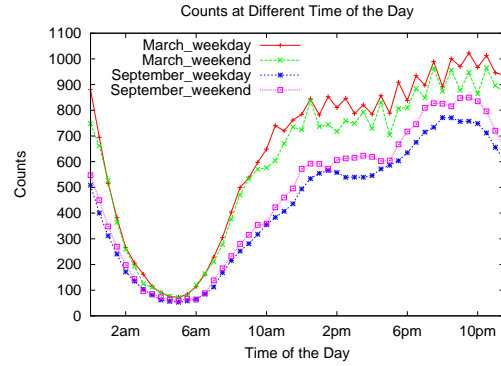

Figure 3: Average Instagram counts during a day

`stagram` counts in September is less than that in March, as shown in Table 1. It is not surprising that, both March weekday curve and March weekend curve are above September weekday curve and September weekend curve. All curves reflected the intensities of entertainment activities: as more people wake up in the morning, `Instagram` becomes more active, and then the intensity remains almost the same in the afternoon until the evening, when people start to have a lot of entertainments. Interestingly, the March weekday curve is above the March weekend curve while the September weekend curve is above the September weekday curve. We would suggest there might be some events in March which stopped people from entertainment, and the haze did not actually stop people from going outside, or at least did not stop the `Instagram` users from going outside, maybe their hanging-out places have been changed from outdoors to indoors.

5. ANALYSES AGAINST PSI

Before we proceed to the content analysis part, we will like to analyze how `Instagram` counts change with *PSI*, which is the Pollutant Standards Index.
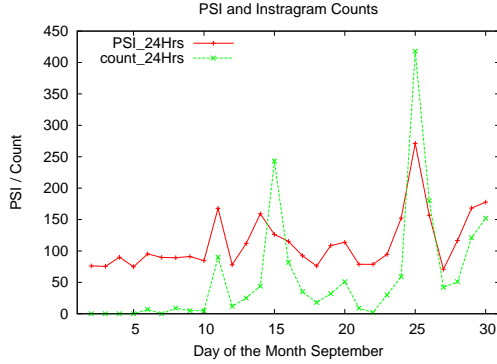
24-hour PSI are obtained from Singapore's official mete-

Figure 4: Comparison between PSI and Instagram counts

orology agent – NEA[3], which are the average PSI over 24 hours.

In order to compare `Instagram` counts with PSI, we first count the number of `Instagrams` which contain either tag "sghaze" or "sghaze2015". There are 2 and 1792 `Instagrams` for the March and September dataset respectively. We thus do not compare PSI and `Instagram` counts for the March dataset.

Figure 4 plots the daily PSI and `Instagram` counts. Both curves reveal peaks on 11th, 15th and 25th of September, which are the dates with severe haze. The two curves match with each other very well. Therefore, the daily number of `Instagram` counts could be used for the prediction of daily PSI.

However, when we break the 24-hour `Instagram` counts into hourly counts, the PSI does not match with `Instagram` counts well. This is due to the fact that the `Instagram` count is first dominated by the time of the day, e.g., users post more during the daytime and evening, and less in the night and early morning, as shown in Figure 3. While `Instagram` count goes down to zero in the early morning, the PSI is not significantly lower than the PSI during the day time. It is therefore impossible to model more fine-grained PSI by `Instagram` count than 24-hour PSI.

The Limitation of Meta-Data Based Analyses. However, there are limitations of meta-data based analyses. For example, we can only relate the PSI with the subset of `Instagram` with either tag "sghaze" or tag "sghaze2015". If we take the whole set of Singapore `Instagrams`, there is no such correlation. Therefore, our current approach relies on the tags users give. If a hazy photo is not properly tagged, our approach will miss this photo in our analyses. Without identifying haze-related photos from content, it is almost impossible to zoom in to the number of haze-related photos. Using the number of the entire set of photos can never be an accurate estimator. In the next section, we will present some simple but effective ways to identify haze-related photos.

6. CONTENT ANALYSES

In the previous sections, we presented our observations derived from meta data associated with user posts including

---

[3]nea.gov.sg

where an image was posted from, *when* it was posted and *who* posted it. However, the content of the image itself is yet another source of rich information about the haze situation. First, we observe that among the posts tagged with either "sghaze" or "sghaze2015", many are not representative of the haze situation – for example, in many cases, we observe people posting images from indoors (e.g., while having lunch), selfies, stock photos (e.g., memes) – although the tag links the post to the haze situation, the post itself is irrelevant. This is a key limitation in our current work in that we are likely to over-estimate the number of posts related to haze. Second, we observe that many of the images are "hazy". We posit that by estimating the amount of haziness visible in the image, the intensity of the haze situation could be inferred. In the following subsections, we provide early insights, discuss open questions and challenges, in (1) filtering out irrelevant images for better accuracy, and (2) estimating the degree of haziness from visual content.

6.1 Extracting relevant images

In previous sections, we use the whole corpus of posts that were hash-tagged as pertaining to haze. However, closer inspection reveals that not all such posts are related to the haze event. Hence, a pre-processing step is required to extract such images. The most common form irrelevant images are "selfies". As a first attempt in understanding the proportion of images that maybe deemed as irrelevant, we employ the state-of-the-art face detection technique based on HAAR feature-based cascade classification [15], to estimate the proportion of selfies in the corpus.

We use the OpenCV implementaton of the pre-trained face classifier [4] for detecting face pixels and express each image as a proportion of face pixels to the image size. We observe that at least 6% of images in our dataset were detected to have a dominant face in the picture. Further investigation revealed that images that had faces wearing masks (as protection against haze) were not detected. Although selfies, in general, provide less information value, we also note here a particular case where selfies could in fact be useful in understanding the haze situation – selfies where people wear face masks to protect them from the haze. Although none of the existing face detection techniques capture faces wearing masks, it is indeed possible to train a set of HAAR classifiers with images of faces with masks on. An alternative approach is to use perceptual hash based image similarity in combination with supervised classification.

In addition to selfies, other categories of irrelevant images exist: indoor images (e.g., users posting pictures of food) which are less useful since the haze hampers outdoor activities predominantly and stock photos (e.g., memes). It remains an open question to what extent automatic extraction of relevant images is feasible.

6.2 Haze estimation from images

In this section, we explore the possibility of estimating haze intensity from images. With the growing concern for air pollution in major cities, recent works that attempt to estimate the degree of haziness have emerged in the areas of Computer Vision and Image Processing since recently [9, 8]. We use the technique described in [9] which estimates the degree of haziness based on the pixel level content of

---

[4]http://docs.opencv.org/master/d7/d8b/tutorial_py_face_detection.html

Figure 5: Examples of classified instances: (a) a hazy image taken at night misclassified as "haze-free", (b) a selfie with only a small proportion of the image with haze misclassified as "haze-free", (c) a correctly classified hazy image.

the image, accounting for both the overall darkness and the contrast of the image.

In Figure 6, we plot the CDF of the haziness score. We observe that more than 33% of the images in our dataset had a haze degree of at least 0.4 and about 23% of the images received score greater than 0.7. Similar to the classification in [9], we categorize images with a score more than 0.7 as containing "thick haze" (we refer to this subset of images as **very hazy**) and below 0.4 as "haze-free" (we refer to this subset of images as **haze-free**). We randomly sampled 100 images each from the **hazy** and **haze-free** datasets and report the false negatives and false positives in identifying hazy pictures, in Table 3.

Dataset	False-Positives
very hazy	24/100 (24%)
haze-free	19/100 (19%)

Table 3: Accuracy of haze estimation of randomly sampled images.

In further inspecting the images that were erroneously classified as **very hazy**, we find that most of these images were found to be B&W (the technique in [9] only works on RGB images) images and stock images which were mostly grayscale and/or of low-contrast. Among those images that were misclassified as **haze-free**, but were in fact representations of the haze, a significant proportion were taken during the night (the technique in [9] associates haze with lighter pixels). Further, images where the haziness is visible only a small proportion of the image were also misclassified as **haze-free**. In Figure 5, we provide examples of correct and incorrect classifications.

Although the results are still preliminary, it is noteworthy that more complex methods such as those discussed in [8] report better accuracies. An interesting future direction of our work is to identify whether a correlation between visual haziness and actual haziness (based on PSI readings) exists and to what extent. With such correlations, it may become possible to estimate the PSI level at better temporal and spatial granularities.
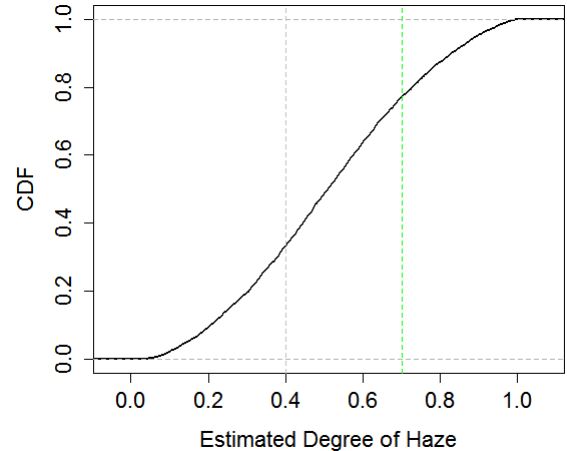
7. RELATED WORK

Figure 6: CDF of the estimated scores of the degree of haziness. Images with a score of at least 0.7 are classified as very hazy.

We describe related work, both in the general topic of event detection using social media streams, and social media content analtyics.

Event Detection using Social Media Streams: The topic of identifying events from social media-generated content (e.g., Twitter feeds) has been studied extensively. Typically, such events are detected based on the anomalous *volumes* of tweets related to specific topics, the associated meta data (hash tags, location data, etc.) and the content of the post (e.g., Tweet). Atefah et. al. [2] provide a comprehensive survey on event detection using social media streams. Typically, approaches such as TwitterMonitor [10] and TopicSketch [17] focus on detecting trending topics based on hash tags. In contrast, Twitcident [1] first specifies which events to be monitored using a broadcast channel, and then mines incoming Twitter streams to extract additional information related to such events. Alternatively, Walther [16] use both meta data and the Tweet content in detecting spatially localized events (e.g., house fires or parties) – a combination of topic and semantic analysis of the content is used in identifying events. We differ from this line of works in two ways: (1) we focus on the implications of the event (e.g., reduc-

tion in outdoor activity due to haze) and the feasibility of using social media to understand the intensity of the event (e.g., how volume of posts correlate with the intensity of haze level), and not on the problem of detecting real world events, and (2) we use visual social media where the mode of information sharing is fundamentally different. Our work complements the analysis of the haze situation in [13].

Social media content analtyics: Text-based analyses of social media content (e.g., Tweets) have received wide attention. The content of posts have been analyzed for a wide range of applications including, but not limited to, political opinions and predictions [7, 14], understanding user and product sentiments [4, 11], and content recommendations [12]. However, we note that works that analyze user-generated images are severely limited. In [5], one of the first works to study Instagram data, the authors focus on how users understand and use the "Instagram medium" through spatio-temporal visualizations. Further, in [6], a preliminary study of the type of image content posted by Instagram users is presented. In particular, they classify posted images under eight categories (e.g., selfies, food, gadgets, etc.) and model user types based on the category of images the users post. Recently, in [3], a study on how the filter used on Instagram images affect user engagement is presented. To the best of our knowledge, none of the existing work in visual social media study image content for the understanding of physical events (such as the haze situation).

8. CONCLUSION

In this preliminary study, we use posts from Instagram, across two time periods in singapore, to understand the implications of haze on people's mobility (visits to places, in particular) in the case of both residents, as well as, tourists. Analyses of this nature are important due to the growing concern of air pollution and its consequences on the health of the residents and the economy, at large, of those regions affected by the recurrence of the haze. In addition to using metadata (location from which an image was posted from on Instagaram) and the volume of posts, we further investigate the possibility of using the "content" of the image itself for better understanding of the user-generated content.

9. REFERENCES

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st international conference companion on World Wide Web*, 2012.

[2] F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. *Comput. Intell.*, 31(1):132–164, 2015.

[3] S. Bakhshi, D. A. Shamma, L. Kennedy, and E. Gilbert. Why we filter our photos and how it impacts engagement. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[4] M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications*, 40(16):6266–6282, 2013.

[5] N. Hochman and L. Manovich. Zooming into an instagram city: Reading the local through social media. *First Monday*, 18(7), 2013.

[6] Y. Hu, L. Manikonda, and S. Kambhampati. What we instagram: A first analysis of instagram photo content and user types. 2014.

[7] A. O. Larsson and H. Moe. Studying political microblogging: Twitter users in the 2010 swedish election campaign. *New Media & Society*, 14(5):729–747, 2012.

[8] Y. Li, J. Huang, and J. Luo. Using user generated online photos to estimate and monitor air pollution in major cities. In *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service*, ICIMCS '15.

[9] J. Mao, U. Phommasak, and H. Shioya. Detecting foggy images and estimating the haze degree factor. In *Journal of Computer Science and Systems Biology 7(6):226–228*, 2014.

[10] M. Mathioudakis and N. Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, 2010.

[11] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326, 2010.

[12] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, 2009.

[13] P. K. Prasetyo, M. Gao, E. Lim, and C. N. Scollon. Proceedings of the 5th international conference on social informatics (socinfo). Lecture Notes in Computer Science, pages 478–491, 2013.

[14] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 2010.

[15] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511, 2001.

[16] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *Advances in Information Retrieval*, pages 356–367. 2013.

[17] W. Xie, F. Zhu, J. Jiang, E.-P. Lim, and K. Wang. Topicsketch: Real-time bursty topic detection from twitter. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 2013.