Data Mining on Twitter for Improving Public Health Safety

Alfred George Francisco Milan III
University of California, Santa Cruz

Introduction

Despite how technology has evolved to the point where people are dependent on social media, social media could alleviate societal problems, such as the imminent threat of diseases to public health in the 21st century and GPS probe data's low sampling frequency. Having Twitter as a data source allows researchers to accurately get real-time data regarding social events and its impact to traffic congestion and apply that to collecting data for flu outbreaks. With the emergence of social networking sites, Twitter has serves as an early warning and response system and as a more effective option to collect traffic congestion data.

Results

Figures 4 and 5 signify that the correlation between the two signals are strong and their near-identical shapes.

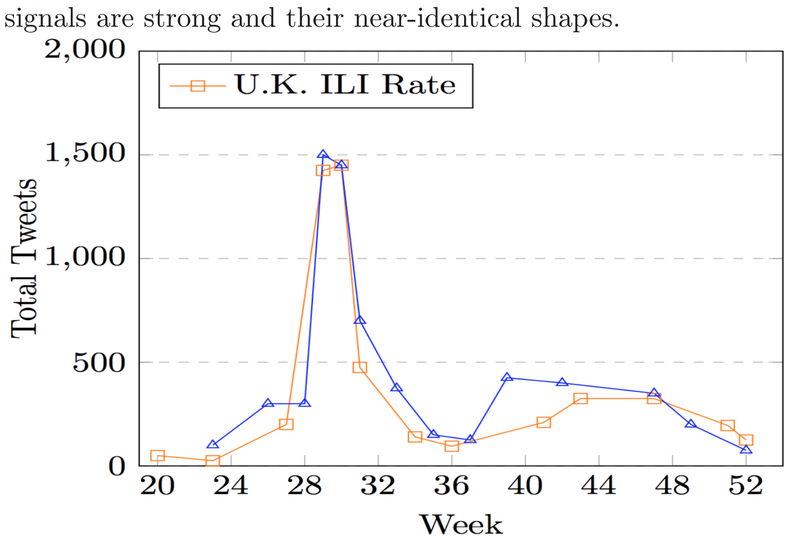

Figure 4: The cross-correlation plot between Twitter and the ILI reporting in the U.K.

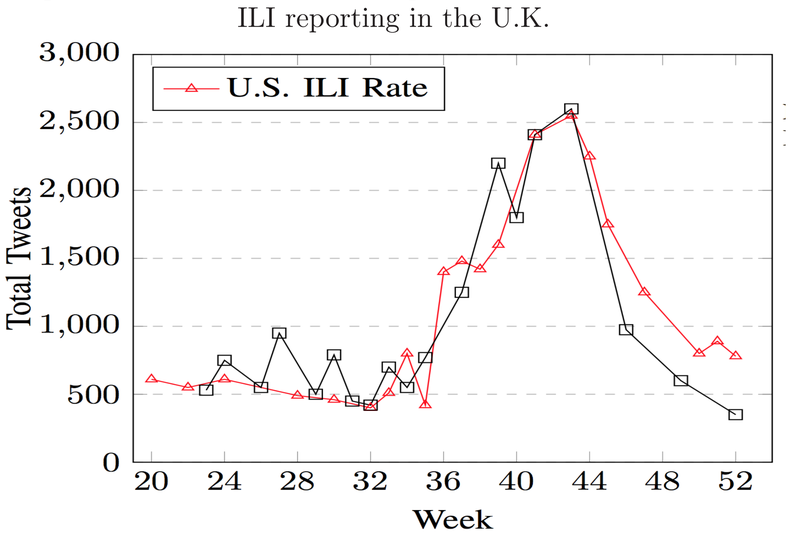

Figure 5: U.K.'s ILI rates vs. Self-Reported Tweets.

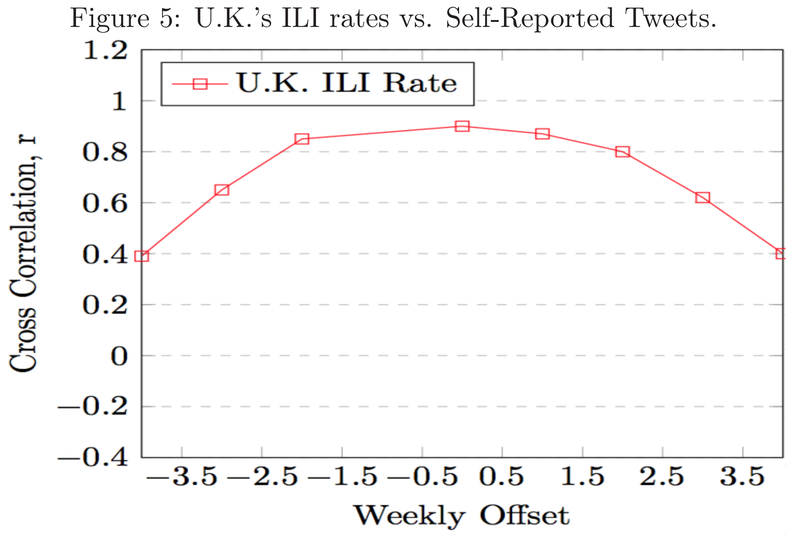

Figure 6: The cross-correlation plot between Twitter and the ILI reporting in the U.K.

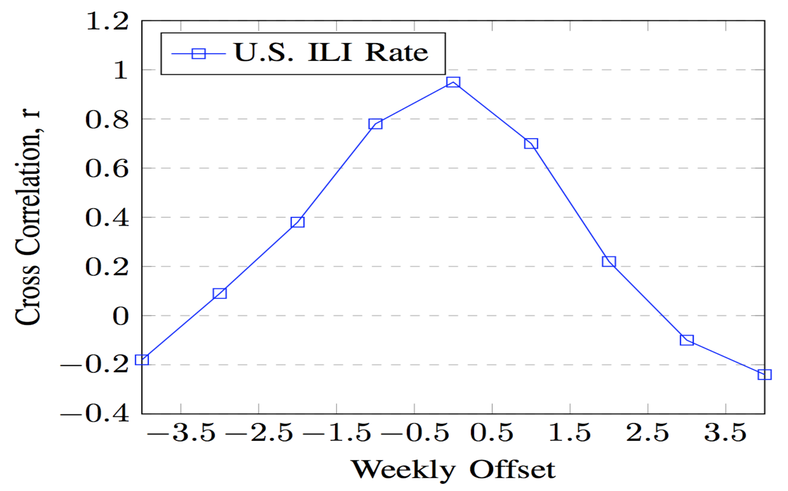

Figure 7: U.K.'s ILI rates vs. Self-Reported Tweets.

As we can see in Figures 6 and 7, the weekly offsets for both the U.K. and U.S. have near-symmetric shapes. Using equation 1, there is only 3.8% error among a sample of 1,000 tweets from social event extraction and geo-coding tweets alone, thus showing that Twitter is a rather reliable source of information.

Background

One studies how probe GPS data and Traffic related Twitter data are different since the latter contains the exact social event, the specific location, and the direction of the traffic meanwhile the former lacks such valuable information.

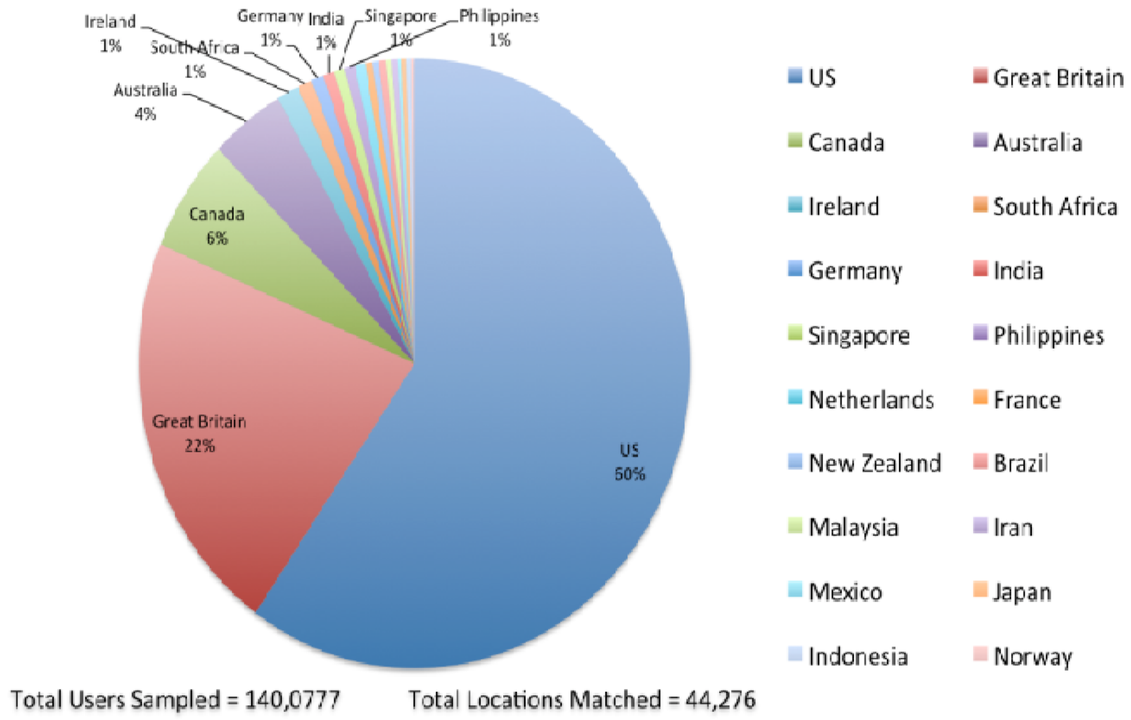

Figure 1: Illustration of probe GPS data and social media data for traffic monitoring.

Using as a comparison to the Twitter data, official health agencies' surveillance data in the United Kingdom must be examined in order to get a sample of 140,770 users whose locations are used as a basis of Twitter users' demographics, as shown in Figure 1.

Experiment

For the Official Traffic Twitter Accounts, 10 Twitter Accounts have been identified to retrieve tweets that clearly shows the content needed to process the road segment, traffic event category, and time, as shown in Figure 2, as well as classifying tweets with the term pandemic in figure 3. One can predict the intensity of traffic due to those social events by using a two-dimensional Gaussian model to measure the social event's impact on the nearby road segments based on their Euclidean distance:

$$I(r_i, se_j) = f(loc_{r_i}; loc_{se_j}, \sum_{se_j}) = \frac{1}{2\pi|\sum_{se_j}^{1/2}|} \times e^{-1/2(loc_{r_i} - loc_{se_j})T} \times \sum_{se_j}^{-1}(loc_{r_i} - loc_{se_j},$$

Meanwhile, we compare the said data with the percentage of self-reporting flu tweets that we calculate from the number of individuals who self- report each day with offcial surveillance data from the U.K. Health Protection Agency, CDC website, and ILINet.

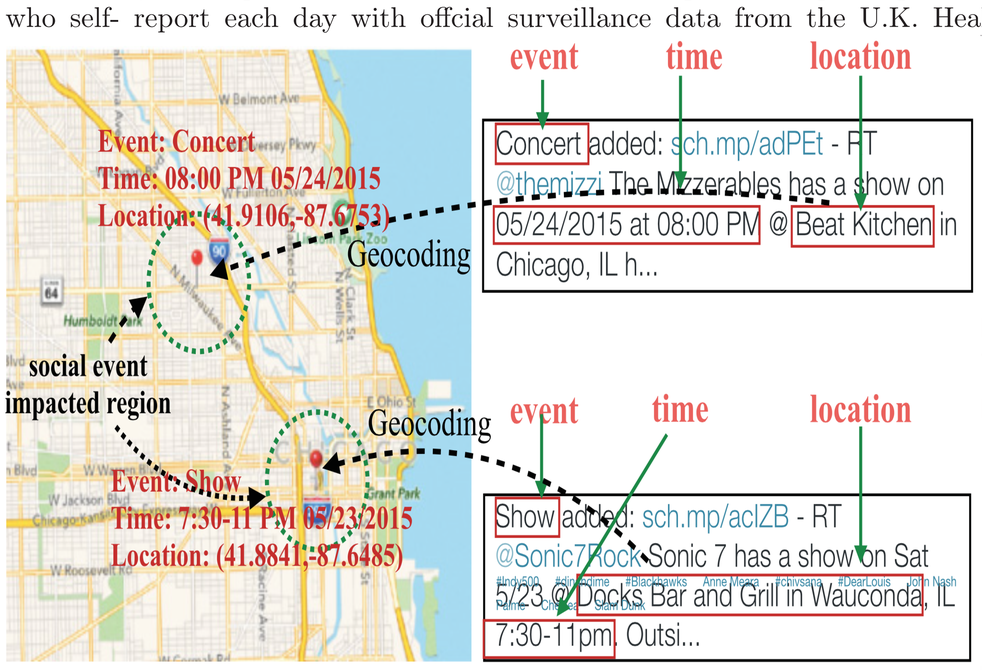

Figure 2: Social Events extraction and geo-coding

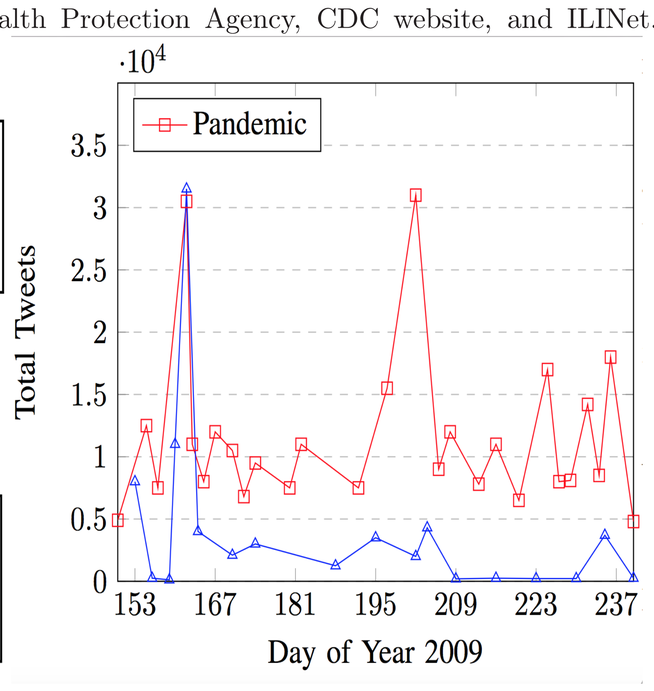

Figure 3: Tweets that have the term *pandemic*

Eq. (1) gives the normalized cross-correlation ratio between various signals from Twitter and the official RCGP U.K. surveillance (collated by HPA) data is calculated to validate the correlation between the two. We also collect a weekly aggregation of Twitter data for the said ratio.

$$r = \frac{\sum_t (x(t) - x) \times (y(t - i) - y)}{\sqrt{\sum_t (x(t) - x)^2 \times \sum_t (y(t - i) - y)^2}} \quad (1)$$

The various values of r for weekly offsets between i = 4 and i = 4 are displayed at figure 4. The similarity of two signals against a moving time lag is computed to get the said correlation ratio.

Conclusion

Since Twitter is used to retrieve valuable data from traffic and health surveillance, the time period of getting information has proven to be shorter and the amount of information to be far greater. We are successful into applying part of a novel framework to effectively compute urban traffic congestion and flu outbreaks by using social media data. Thus, public health and traffic authorities would be more convinced to use Twitter as an accurate complement to their various data sources. We may exploit African-American, Latino, and millennials' social media activity in future work as means of improving the data collected.

References

- Kostkova, P., Szomszor, M., and St. Louis, C. 2014. swineflu: The use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. ACM Trans. Manage. Inf. Syst. 5, 2, Article 8 (July 2014), 25 pages. DOI:http://dx.doi.org/10.1145/2597892
- Senzhang Wang, Xiaoming Zhang, Jianping Cao, Lifang He, Leon Stenneth, Philip S. Yu, Zhoujun Li, and Zhiqiu Huang. 2017. Computing urban traffic congestions by incorporating sparse GPS probe data and social media data. ACM Trans. Inf. Syst. 35, 4, Article 40 (July 2017), 30 pages. DOI: http://dx.doi.org/10.1145/3057281