Social Media Analytics for Healthcare Surveillance using Text Mining

Submitted Oct 2019, in partial fulfillment of
the conditions for the award of the degree **BSc Computer Science.**

Yuyang Liu
16522049

Supervised by Heng Yu

School of Computer Science
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature _____

Date _____ / _____ / _____

I hereby declare that I have all necessary rights and consents to publicly distribute this dissertation via the University of Nottingham's e-dissertation archive.

Public access to this dissertation is restricted until: DD/MM/YYYY

Chapter 1

Background information and motivation

Disease control and prevention is vital for the whole society. Traditional surveillance method adopted by the Centers for Disease Control and Prevention (CDC) is, scrutinizing outpatient records from hospitals and virological test results from laboratories, which notices the disease after it actually occurred [10]. However, if there is no forecast and hospitals are ill-prepared for a rush of patients, the reception of in-time treatment will be affected [5], other severe consequences can also be imagined. Therefore, a robust disease forecast system is needed.

To predict the outbreak of disease in advance, massive efforts have been put. [3] monitored the changes of Realistic Contact Networks (RCNs) to predict the dynamic movement of disease, [2] used machine learning with previous illness records to predict the future outbreak, while some researchers tried to build a prediction system based on data from social media.

According to [7], social media contains information related to healthcare, individual health issue, symptom. [6] shows that spikes in flu queries and disease breakout coincide. However, since queries has little or no limitation and even don not need an account, they cannot be regarded as reliable data [10]. Other social media platforms such as Twitter and Facebook have proven their value for Big Data analyze. Twitter data has been found to be useful for public health applications [4], including: (1) monitoring diseases, (2) public reaction, (3) outbreaks or emergencies, (4) prediction, (5) lifestyle, and (6) geolocation of disease surveillance [1]. In addition, social media is prompt. According to [5], over 645 million active Twitter users collectively post an average of 58 million tweets (micro-blogs no more than 140 characters long) per day in 2017, and the number is still growing. A practical example is that researcher use Twitter predicted flu outbreaks 1–2 weeks ahead of CDC's surveillance average [11]. [5] also showed Twitter data aligns with CDC's outpatient records. All the information I have read so far proves that such data is valuable in healthcare surveillance. Previous work relying on social media has successfully proposed some novel methods. [8]

and [9] utilize time-series analysis on single geography, [5] provided a generalized solution to identify how contagious diseases diffuse across geographies. However, time-series analyze can be inaccurate in this scenario (they rely merely on the statistics of social media instead of its content). For example, top search about a certain disease could result from a celebrity's illness [10]. Similar experimental result showed in [5]. During festivals, the number of tweets decreased, and the prediction accuracy was affected. Obviously, such results are not robust enough to be applied in the real world. It therefore makes sense to find a new method to enhance the whole system. In this project, I will start from previous works, and use different techniques such as machine learning, heuristic function, NLP to extract textual implication containing in such data and create a more feasible solution.

Chapter 2

Aims and Objectives

The general aim of this project is analyzing social media data to surveille healthcare condition. It can be detailed as follow:

1. Healthcare-related textual information should be extracted and modeled for the purpose of healthcare surveillance.

2. Robust predictive models are required for accurate forecasting of disease outbreaks and hospital emergency visits based on ML techniques.

The key objectives of this project are:

1. Collecting data of a certain social media platform. This data can be either extracted from an existing data set or crawled from that platform. If existing data set can't meet my requirement, my own data set will be made.

2. Filtering data. In this process, a filtering rule should be established, possible solutions include using related verbal list, using NLP to classify raw data

3. Designing algorithm that can successfully predict the outbreak of diseased and its propagation direction with high accuracy, and implementation. This includes (1) setting a benchmark to evaluate the accuracy; (2) finding a suitable model maxing the accuracy.

4. Experiment. Testing and evaluating this model, and compare it with others' work. In this stage, re-implementation may be needed.

5. Result visualization. Since the outbreak is dynamic, visualizing the result can be more readable.

Chapter 3

Project Plan

As mentioned above, this project starts with data collection. Since data is the basis of this project, it is worth spending more time on data collection and preprocessing. Once the data is processed, re-implementation and algorithm design will be commenced. Ideal due of those stages is before the end of autumn semester (or before the start of spring semester). In the spring semester, the focus will be experiment, evaluation, improvement and writing paper. Such arrangement ensures that the whole project runs in steady at an early stage and reserves enough time for handling contingency. The Gantt chart following shows the preliminary schedule of this project, consisting tasks below (note that events such as exams are excluded in this schedule, some tasks will commence in parallel):

1. Literature search (3 weeks)

2. Complete project proposal (1-2 weeks)

3. Data collection and preprocessing (5-7 weeks)

4. Write interim report (7-8 weeks)

5. Algorithm design and implementation (about 10-13 weeks)

6. Write final dissertation (at most 16 weeks)

7. Re-implement others' work (about 6 weeks)

8. Test and evaluate the performance of the algorithm and compare with previous work (2-3 weeks)

9. Review the whole project and do final adjustment (5 weeks)

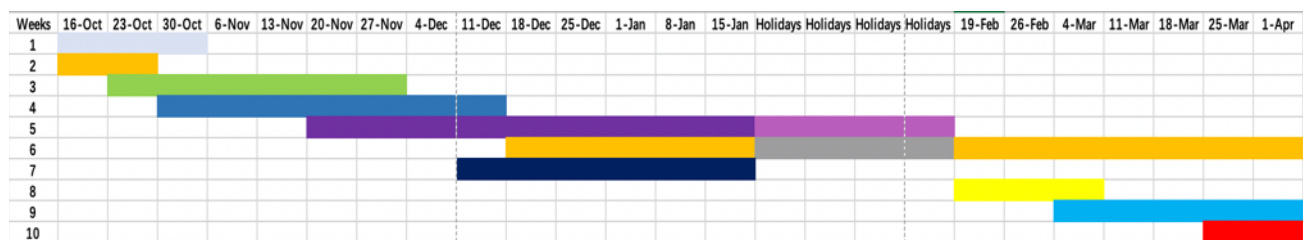10. Prepare for demonstration (2 weeks)

Weeks	16-Oct	23-Oct	30-Oct	6-Nov	13-Nov	20-Nov	27-Nov	4-Dec	11-Dec	18-Dec	25-Dec	1-Jan	8-Jan	15-Jan	Holidays	Holidays	Holidays	Holidays	19-Feb	26-Feb	4-Mar	11-Mar	18-Mar	25-Mar	1-Apr
1																									
2																									
3																									
4																									
5																									
6																									
7																									
8																									
9																									
10																									

Figure 3.1: Schedule

Bibliography

[1] ANDREU-PEREZ, J., POON, C. C., MERRIFIELD, R. D., WONG, S. T., AND YANG, G.-Z. Big data for health. *IEEE journal of biomedical and health informatics 19*, 4 (2015), 1193–1208.

[2] CHEN, M., HAO, Y., HWANG, K., WANG, L., AND WANG, L. Disease prediction by machine learning over big data from healthcare communities. *Ieee Access 5* (2017), 8869–8879.

[3] CHEN, Y., CRESPI, N., ORTIZ, A. M., AND SHU, L. Reality mining: A prediction algorithm for disease dynamics based on mobile big data. *Information Sciences 379* (2017), 82–93.

[4] DENECKE, K., AND NEJDL, W. How valuable is medical social media data? content analysis of the medical web. *Information Sciences 179*, 12 (2009), 1870–1880.

[5] ELKIN, L. S., TOPAL, K., AND BEBEK, G. Network based model of social media big data predicts contagious disease diffusion. *Information discovery and delivery 45*, 3 (2017), 110–120.

[6] GINSBERG, J., MOHEBBI, M. H., PATEL, R. S., BRAMMER, L., SMOLINSKI, M. S., AND BRILLIANT, L. Detecting influenza epidemics using search engine query data. *Nature 457*, 7232 (2009), 1012.

[7] LEE, K., AGRAWAL, A., AND CHOUDHARY, A. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), ACM, pp. 1474–1477.

[8] SADILEK, A., KAUTZ, H., AND SILENZIO, V. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012).

[9] SALATHÉ, M., AND KHANDELWAL, S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology 7*, 10 (2011), e1002199.

[10] SCHMIDT, C. W. Trending now: using social media to predict and track disease outbreaks, 2012.

[11] SIGNORINI, A., SEGRE, A. M., AND POLGREEN, P. M. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one 6*, 5 (2011), e19467.