The University of
Nottingham

COMP3050. Final Year Project Proposal

Project: Social Media Analytics for Healthcare Surveillance using Text Mining

Supervisor: Heng Yu

Student name: Yuyang Liu

Student ID: 16522049

Student E-mail: zy22049@nottingham.edu.cn

Date: October 16, 2019

School of Computer Science, University of Nottingham Ningbo, China

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 [1]. However, if there is no forecast and hospitals are inadequately prepared for a rush of patients, the reception of in-time treatment will be adversely affected [2]. 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. [6] 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 [7]-[10].

According to [7], social media contains information related to healthcare, individual health issue, symptoms, etc. [8] shows that spikes in flu queries and disease breakout coincide. However, since queries have little or no limitation and even do not need an account, they cannot be regarded as reliable data [1]. 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 [9], including: (1) monitoring diseases, (2) public reaction, (3) outbreaks or emergencies, (4) prediction, (5) lifestyle, and (6) geolocation of disease surveillance [10]. In addition, social media is prompt. According to [2], 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 researchers leveraging Twitter predicted flu outbreaks 1–2 weeks ahead of CDC's surveillance average [4]. [2] also showed Twitter data aligns with CDC's outpatient records. It has been increasingly evident, shown from the research community, that social media data are significantly valuable for healthcare surveillance and prediction.

Previous work relying on social media has successfully proposed some novel methods. [5] and [11] utilize time-series analysis on single geography, while [2] 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 be resulted from the public concern of a celebrity's medical situation [1]. Similar experimental results are shown in [2]. During festivals, the number of tweets decreased, affecting the prediction accuracy. 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, we will start from surveying previous works, and use different techniques such as machine learning, heuristic function, and/or NLP to extract textual implication containing such data, and explore a more feasible solution.

Aims and Objectives:

The general aim of this project is processing, modeling, and analyzing social media data, in order to achieve accurate public healthcare surveillance and prediction. 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:

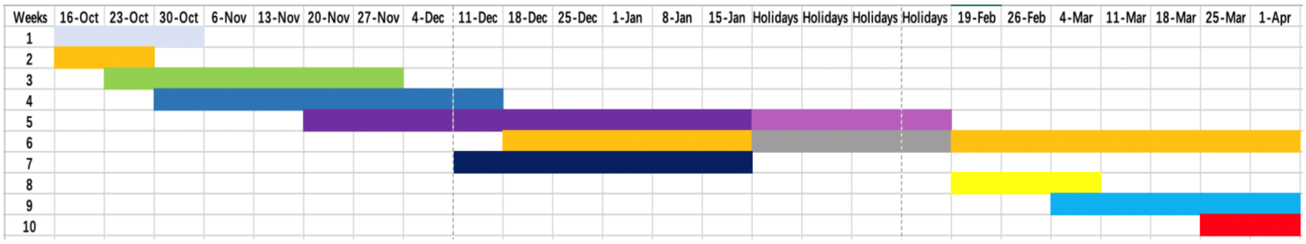
1. Collecting data from a certain social media platform. This data can be either extracted from an existing data set or crawled from that platform. If existing data set can't meet our requirement, our own data set will be made.
2. Filtering data. In this process, a filtering rule should be established. Possible solutions include using NLP techniques to classify raw data based on related verbal list.
3. Designing and implementing algorithms that can successfully predict the outbreak of diseased and its propagation direction with high accuracy. This includes (1) setting a benchmark to evaluate the accuracy; (2) finding a suitable model maximizng the accuracy.
4. Experimental evaluation. 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.

Project Plan:

As mentioned above, this project starts with data collection. Since data are the basis of this project, it is worth spending a propor portion of project 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 steadily in the early stage and reserves enough time for handling contingency. The Gantt chart following shows the preliminary schedule of this project, consisting of 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)



References:

[1] Schmidt, Charles W. . "Trending Now: Using Social Media to Predict and Track Disease Outbreaks." *Environmental Health Perspectives* 120.1(2012):a30-a33.

[2] Elkin, Lauren S , et al. "Network based model of social media big data predicts contagious disease diffusion." *Information Discovery and Delivery* (2017):00-00.

[3] Chen, Yuanfang , et al. "Reality Mining: A Prediction Algorithm for Disease Dynamics based on Mobile Big Data." *Information Sciences* (2016):S002002551630559X.

[4] Alessio, Signorini , et al. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic." *PLoS ONE* 6.5(2011):e19467-.

[5] Sadilek, Adam, H. Kautz, and V. Silenzio. "Predicting Disease Transmission from Geo-Tagged Micro-Blog Data." *Twenty-sixth Aaai Conference on Artificial Intelligence* 2012.

[6] Chen, Min , et al. "Disease Prediction by Machine Learning over Big Data from Healthcare Communities." *IEEE Access* (2017):1-1.

[7] Lee, Kathy , A. Agrawal , and A. Choudhary . "Real-Time Disease Surveillance Using Twitter Data: Demonstration on Flu and Cancer." *Acm Sigkdd International Conference on Knowledge Discovery & Data Mining* ACM, 2013.

[8] Ginsberg, Jeremy . "Detecting influenza epidemics using search engine query data." *Nature* 457(2009).

[9] Denecke, K. and W. Nejdl, "How valuable is medical social media data? Content analysis of the medical web". Information Sciences, 2009. 179(12): p. 1870-1880.

[10] Andreu Perez, Javier, et al. "Big Data for Health." *Biomedical & Health Informatics IEEE Journal of* 19.4(2015):1.

[11] Salathé, Marcel, S. Khandelwal , and L. A. Meyers . "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control." *PLoS Computational Biology* 7.10(2011):e1002199.