Identifying spread of influenza by social network data analyzing on Twitter

Jeonghwa Kang, 6109806

Computer Science (BSc)

Dr. Vasile Palade

School of Computing, Engineering and Mathematics, Coventry University

2nd May 2017

Department of Computing
COVENTRY UNIVERSITY

300COM / 303COM Declaration of originality

I Declare that This project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed:  Date: 30th April 2017

Please complete all fields.

First Name:	Jeonghwa
Last Name:	Kang
Student ID number	6109806
Ethics Application Number	P51480
1st Supervisor Name	Vasile Palade
2nd Supervisor Name	Saad Amin

This form must be completed, scanned and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

Certificate of Ethical Approval

Applicant:

Jeonghwa Kang

Project Title:

Identifying spread of influenza by social network data analyzing on Twitter.

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:

24 February 2017

Project Reference Number:

P51480

Acknowledgement

I would like to show my gratitude to my supervisor Dr. Vasile Palade for without him this dissertation work could not have been possible. I would like to take this chance to thank Dr. Palade for his generous advice, care and guidance throughout my dissertation. Without his inspiration, motivation, moral support and suggestions, this work could not have been possible at all. This dissertation is entirely due to his interest.

I would also like to thank Abdulaziz Alayba who provided support and advice during the research. Without his generosity for support, this study would not have been possible and I would like to thank him for taking out precious time from his busy schedules to communicate with me regarding my dissertation.

I would also like to extend my gratitude to Dr. Peter Every who is a project leader of all the 3rd year students of Computer Science. I am thankful for his moral support, guidance and comments throughout my study.

Abstract

News channels are traditional communication channels which have limitations to providing spontaneous information about spread of diseases unlike social media namely, Twitter. The present project work proposes a framework by social media mining real-time influenza data from Twitter to identify spread of influenza viruses and visualize severity of the diseases on a physical map of UK. The users of Twitter work as the indicators which provide relevant data about the flu by posting personal experience, warnings, location of a tweet. The framework includes steps such as – data collection, data filtering, data processing (using sentiment analysis and manual processing) and data visualization as a final step. The framework can be validated by evaluating the past events for future reference. Furthermore, it has potential to be further developed into a complete system to identify the progression of such diseases and warn people ahead of time. The warnings may be sent to news broadcasting stations for pro-active action. The outcome of the research work presents visualized data of influenza on a physical map separated by regions in gridlines. The flu data are generally concentrated in the regions where there are highly populated cities possibly due to increased occurrence of human contacts that facilitate transmission of influenza viruses.

INDEX WORDS: Twitter, influenza, data visualization, social network data analysis, data mining, disease detection

Introduction

The aim of the work described in this project is to identify spread of influenza based on thorough analysis of collected data from social network. Flu is an infectious disease caused by influenza viruses which often results thousands of human lives being lost every year around the globe. It is a complex biological phenomenon which also results financial and human losses throughout the world every year. Precise identification and prediction of these types of disease is based on many complicated physical and environmental parameters hence, the process is quite complex. There are many number of approaches in the literature based on statistical and scientific analysis. Social media mining technique is one useful method to be used for predicting the spread of these infectious biological hazards. Once making predictions on such diseases become more precise and accurate enough, it would, in no doubt, save great number of valuable human lives along with medical costs. This project uses search API algorithm via Python to collect tweets and make predictions on spread of influenza after thorough analysis of historical and real-time data. The collected data are processed further to extract geographical coordinates of each tweet to identify the exact location of where it came from and the processed data are finally visualized on a physical map to help our understanding in the direction of spread of influenza. The objective of this research is to visualize severity and progression of influenza viruses throughout UK and to come up with relevant conclusions based on the findings. This research also aims to help certain target audiences benefit from the outcome of the study by allowing them to be proactive in preparing against the spread of influenza so that they can minimize the possible cost of damages in finance, health or even death. The target audiences of this research are as follow:

Citizens – Early detection of spread of influenza would benefit the citizens in general. Early predictions may warn citizens to avoid using public transportations such as trains, buses, taxies etc. to avoid getting infected by flu during the critical time. The disease, once infected, is likely to affect negatively on one's financial status and even valuable life based on the severity of the infection. If the predictions are accurate enough to allow people to take simple preventive actions against such diseases, it would save many lives as well as medical costs in the future.

Schools – are where staffs need to protect vulnerable population of children and locals. Early predictions would allow people to take preventive actions against such infection. Basic protective measures, for instance, providing vaccines at early stage would increase the number of people saved from the infections. The predictions could also help to close local schools temporarily before the actual spread of flu occurs so that additional spread of diseases in local areas can be prevented.

Organizations – may also benefit from early predictions of disease by providing vaccines for their employees and advising individuals to receive preventive health care ahead of time. These actions, if taken correctly, may prevent decrease in productivity, process of work and loss of employee which are all valuable assets that generate profits to an organization. Furthermore, early predictions will allow organizations to take early actions against such viruses to ensure that their employees are safe. This would be a great opportunity for organizations to improve their employees' loyalty hence, improving goodwill of the organization in general.

Hospitals – Early predictions on influenza viruses may allow hospitals to better prepare for treating patients ahead of time. These actions include preparing for enough number of vaccines, doctors and medical rooms to treat infected patients. Trajectory graph may play an important role in planning delivery routes for distribution of vaccines throughout the affected area. In that way, local hospitals will less likely to suffer from shortage of vaccines for treating their patients. This would not just save people's lives but also the process of work and precious time which equals to money.

A structure of solution in achieving the objectives and aims of the project is defined as follows:

1) Identifying web-based data sources

Since this project is about extracting and analyzing big data from Twitter to identify spread of influenza, selecting the relevant types of data source is crucial. There are many useful techniques that can be used to extract data from social network and these include search API and streaming API. Furthermore, publicly available data sources may be found from

news providers (e.g. websites, RSS feeds etc.) and other social networking sites (e.g. Facebook).

2) Collecting relevant data (tweets)

The focus of this project must be narrowed down to certain events and geographical locations. To do this, data which relate to specific social events and geographical locations should be collected via either Search API or Stream API with appropriate selection of keywords that characterizes the target dataset. The choice of keywords should be selected carefully (not too specific nor too general) to decrease the probability of noise in the dataset. Furthermore, posted time of each tweet must be restricted during the process of collection. For example, it would be less accurate to make assumptions on the spread of influenza based on the tweets that were posted two years ago, Hence, setting up the time scale, say, tweets from the past week would be more reliable source to be used to make future assumptions.

3) Data processing of the collected data

The collected dataset will likely to contain noise which is not useful and irrelevant to the study. It is crucial to remove noises to save computation power and time in general. One way to process data is by considering dataset that contain English tweets only. Secondly, it is possible to obtain the geo-coordinates of each of the collected tweets which can then be counted and sorted per the region. Among the collected data, there are spam tweets which must be removed via blacklists as well. Retweets should also be removed as they are duplicate data. Once all these irrelevant data are removed, we can conduct sentiment analysis which categorizes each tweet into positive, negative and neutral clusters based on the polarity rate of the content. Tweets that are labelled as negative are the only ones to be considered for analysis therefore, tweets that fall into positive and neutral categories can be ignored.

4) Plotting the collected data as figures on a physical map of UK

After fully processing and analyzing the data, it is possible to sort the data based on the geo-location of each tweet. For example, tweets posted from Coventry would have different geo-location compared to those posted in London. After sorting the tweets, it is then

possible to count the total number of tweets that were posted in each specific geographical area. Using the total count of tweets collected from each region, we can plot the numbers on a map with an indicator. This indicator could be in any type of symbol such as circle, triangle and so on. The larger the number of tweets, the bigger the indicator on the map will be.