Faculty of Computer Science and Information Technology

PREDICTION OF HFMD DISEASE OUTBREAK FROM TWITTER

Tay Guo Hong

Bachelor of Computer Science with Honours (Information System)

2019

PREDICTION OF HFMD DISEASE OUTBREAK FROM TWITTER

TAY GUO HONG

This project is submitted in partial fulfilment

of the requirements for the degree of

Bachelor of Computer Science with Honours

Faculty of Computer Science and information Technology

UNIVERSITI MALAYSIA SARAWAK 2019

FORM B

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE PREDICTION OF HFMD DISEASE OUTBREAK FROM TWITTER

ACADEMIC SESSION: 18/19

(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

[ ]	CONFIDENTIAL	(Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)
[ ]	RESTRICTED	(Contains restricted information as dictated by the body or organization where the research was conducted)
[√]	UNRESTRICTED	

Validated by

(AUTHOR'S SIGNATURE)

(SUPERVISOR'S SIGNATURE)

Dr. Suhaila Saee
Lecturer (Information Systems Programme)
Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK

Permanent Address

77-A, Taman Woon, 73000
Tampin, N.S.

Date: 25/4/19

Date: 25/4/19

Note * Thesis refers to PhD, Master, and Bachelor Degree
** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

DECLARATION

I hereby declare that the thesis based on my original work except for the quotations and citations which have been duly acknowledge. I also declare that no portion of the work referred to in this report has been submitted in support of an application for another degree at University Malaysia Sarawak (UNIMAS) or qualification of this or any other university or institution of higher learning.

Tay Guo Hong (54061)

Faculty of Computer Science and

Information Technology

University Malaysia Sarawak

May 2019

ACKNOWLEDGEMENT

I wish to express my sincere gratitude and indebtedness to my course coordinator, Prof. Dr. Wang Yin Chai, supervisor, Dr Suhaila and examiner, Dr. Stephanie Chua Hui Li for their invaluable guidance and support throughout the course of this project. Their invaluable supervisor, continuous encouragement and inspired guidance were instrumental in the successful completion of the project.

I am also very thankful to all my friends and classmates who always helped me whenever I needed help. I heartily thank for their timely suggestion and feedback that had helped me to enrich this project.

Finally, I appreciate the morale support from my family members for they always backup me whenever I faced difficulty.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Word/Phrase
HFMD	Hand, foot, and mouth disease
SVM	Support Vector Machine
API	Application Programming Interface

ABSTRACT

Hand, foot, and mouth disease (HFMD) is a common childhood infection caused by a group of enteroviruses. This research paper describe a work about predicting of HFMD disease outbreak from Twitter. After reviewing the existing work, a proposed pipeline is being introduced. In this project, the data collection methods is extracting Twitter tweets using Twitter API with Python. The extracted tweets is going through pre-processing process. The output from this process is the corpus of HFMD disease. On the other hand, Naive Bayes and SVM algorithm is using in classification of the tweets related with HFMD disease. This is because both Naive Bayes and SVM are baseline algorithm used in text classification. In the end, a visualisation of HFMD Disease Map is presented to visualize the city that suffer HFMD outbreak using geo-located tweet that related with HFMD. Based on the Map visualisation, Malaysia is predicted to face HFMD outbreak in the period between January until March for the coming years. For the classification result, Naive Bayes and SVM provide result with accuracy of 92.8% and 96.7% respectively.

ABSTRAK

Penyakit tangan, kaki, dan mulut (HFMD) adalah jangkitan zaman kanak-kanak yang disebabkan oleh sekumpulan enterovirus. Kertas kajian ini menggambarkan satu kerja mengenai meramalkan wabak penyakit HFMD dari Twitter. Selepas mengkaji semula kerja yang sedia ada, saluran paip yang dicadangkan sedang diperkenalkan. Dalam projek ini, kaedah pengumpulan data mengekstrak Twitter tweet menggunakan API Twitter dengan Python. Tweet yang diekstrak akan melalui proses pra-pemprosesan. Output dari proses ini adalah korpus penyakit HFMD. Sebaliknya, Naive Bayes dan algoritma SVM menggunakan klasifikasi tweet yang berkaitan dengan penyakit HFMD. Hal ini kerana kedua-dua Naive Bayes dan SVM adalah algoritma asas yang digunakan dalam klasifikasi teks. Akhirnya, visualisasi Peta Penyakit HFMD dibentangkan untuk menggambarkan bandar yang mengalami wabak HFMD menggunakan tweet geo-located yang berkaitan dengan HFMD. Berdasarkan visualisasi Peta, Malaysia dijangka menghadapi wabak HFMD dalam tempoh antara Januari hingga Mac untuk tahun-tahun akan datang. Untuk keputusan klasifikasi, Nave Bayes dan SVM memberikan hasil dengan ketepatan masing-masing 92.8% dan 96.7%.

CHAPTER 1 INTRODUCTION

1.1 Introduction

Hand, foot, and mouth disease (HFMD) is a common childhood infection caused by a group of enteroviruses (Liu et al., 2017). It is usually begins with a fever and feels generally unwell. This is followed a day or two later by flat discolored spots of bumps that may blister, on the hands, feet and mouth and occasionally buttocks and groin. There is 51 147 HFMD cases have been detected since January as of 14 August in 2018 (Cavarlho, 2018). Moreover, there is a report said that HFMD situation in Sarawak is currently below warning level (Boon, 2018). The warning level is based on the average number of HFMD cases in past five years (Boon, 2018).

Nowadays, Twitter users have increased rapidly. People nowadays tend to tweet during they sick. For example, parents tweets in Twitter when their child suffering HFMD disease by using #HFMD in Twitter post. Twitter has become one of the famous social media platform for user around the world to share or viral anything happening around them with their friends. For instance, Twitter has over 300 million of active monthly users as of 30 June 2015 (Paul et al., 2015). The evolution in the usage of social media and machine learning reveal us the possibility of utilizing the data source in Twitter for detecting the trend of disease, as long as the researchers are able to mine the data source from Twitter if using data mining technique professionally.

Moreover, several researches that combine the Twitter and illness together have been done so far. For instance, the system by Aramaki et al. extracts the Twitter post related with influenza, conducted factuality analysis using SVM algorithm on extracted Twitter tweets for detecting epidemic influenza (Amaraki et al., 2011). Another existing

research will be research on disease surveillance with location estimation used simple geocoding, geotagged tweets, and user profiles (Ryusei et al., 2018).

Prediction of HFMD disease outbreaks from Twitter will possibly help in closing down the spread of HFMD disease and the government and hospital can make preparation before the outbreak. This project will be carry out by using machine learning algorithm to predict the HFMD disease outbreaks from Twitter.

1.2 Problem Statement

The usage of social media to viral the news or information about disease or health-related information become prevalent nowadays. Hence, there is a potential to mine the valuable information in social media for topic detection purpose. However, there are some difficulties faced by the researchers when they conduct topic detection research in social media mining.

One of the problem faced by researchers is doing location estimation. This is due to some tweets contained two or more location words and many of them are misclassified (Ryusei et al., 2018). Some of the places with special symbols that cannot be removed in pre-processing process may reduce the accuracy of the result as well (Ryusei et al., 2018). Moreover, the location database contains only nationwide known locations or events, for example, “Tokyo Sky Tree” and “Tottori Dune Hill” for the spot name, “Tokyo Game Show” for the event name (Ryusei et al., 2018). Hence, the accuracy may improve if a little popular location can be added.

On the other hand, the researchers also facing difficulties when they doing infectious disease detection. In the research, researchers treat the fundamental children infectious disease as positive case, even if the person related to the user is infected, not the person who posted the tweet sentence (Ryusei et al., 2018). For instance, if we have

a tweet sentence “My son has caught a mumps”, that tweet sentence consider to be positive example (Ryusei et al., 2018). Thus, the accuracy of the research is reduced.

1.3 Objective

The objectives of this project are as below:

i. To investigate the accuracy of machine learning algorithm in predicting outbreak of HFMD disease in worldwide scope.

ii. To build a model to predict the outbreak of HFMD in Malaysia.

iii. To visualise the location estimation based on Twitter tweets using Google Map API.

iv. To build a HFMD corpus in worldwide scope that can be used for other researchers in future.

1.4 Methodology

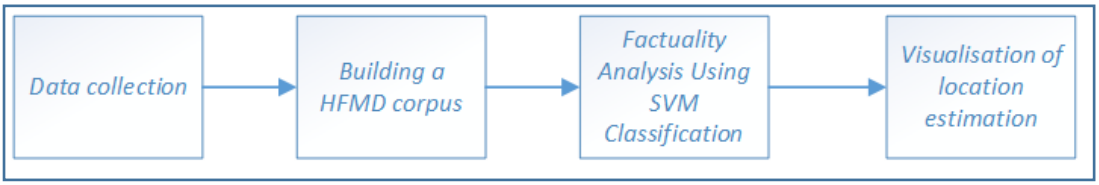

Figure 1.1: HFMD disease prediction methodology

Figure 1.1 shows the methodology of HFMD disease prediction along this project. Every phase will be explained in following paragraphs.

The methodology start with data collection phase, where initially extracting the tweets related with HFMD disease from Twitter and the scope is within Sarawak state.

Then, data pre-processing will be done in order to build a HFMD corpus. This process including data transformation, data normalization, and cleaning noisy and incomplete data from the extracted data. After that, supervised learning approach will

be implemented by using classification technique in model engineering phase. For this project, SVM algorithm will be chosen as learning algorithm from classification algorithms.

Finally, the accuracy of SVM algorithm can be determined based on HFMD prediction in Twitter and a visualisation of location information based on Twitter tweets will be presented.

1.5 Scope

The project scopes are stated below:

i. Only crawl the data in Twitter.
ii. The crawled data has to be related with HFMD.
iii. The crawled data cover the region for world wide.
iv. The data collection period is from 1st January 2019 until 1st April 2019.
v. The crawled tweets are limited to language code with “en” with represent English.

1.6 Significance of Project

The significance of project are stated below:

i. Prediction of the HFMD outbreaks in Malaysia throughout a year.
ii. To build a HFMD corpus in worldwide scope that can be used for other researchers in future.

1.7 Expected Outcome

This project is expected to investigate the accuracy of support vector machine (SVM) in detecting the HFMD outbreak in Sarawak throughout a year. Moreover, the

visualisation of location information based on Twitter tweets are expected to be done as well.

1.8 Project Schedule

This project is takes about a year to complete. The project schedule of FYP 1 is shown in Gantt chart in Appendix A.

1.9 Project Outline

The outline of this project is divided into five chapters as described below:

i. Chapter 1: Introduction

   Chapter one gives an overview and brief introduction about Prediction of HFMD Disease Outbreaks from Twitter.

ii. Chapter 2: Literature Review

   Chapter two focused on the study of existing research for topic detection in data collection and existing pipeline. A pipeline will be proposed based on the review of existing work.

iii. Chapter 3: Methodology

   Chapter three discusses the proposed pipeline used to predict HFMD disease outbreaks from Twitter in detail. Software requirements in this project will be discussed as well.

iv. Chapter 4: Implementation and Result

   Chapter four summarizes the implementation of HFMD disease outbreak detection from Twitter and the result after implementation.

v. Chapter 5: Conclusion

   Chapter five is mainly about the summary of this project.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

This chapter will first introduce the general architecture for topic detection. It is crucial to understand how the pipeline works before further studies being carried out. The review of existing topic detection research will be carried out and presented in this chapter. Next, the data collection in disease detection will be explained. This chapter will continue with reviewing existing pipeline in disease detection. Moreover, finding of existing research will be described. The last part of this chapter will describe a proposed pipeline for HFMD prediction using Twitter.

2.2 General Architecture of Topic Detection

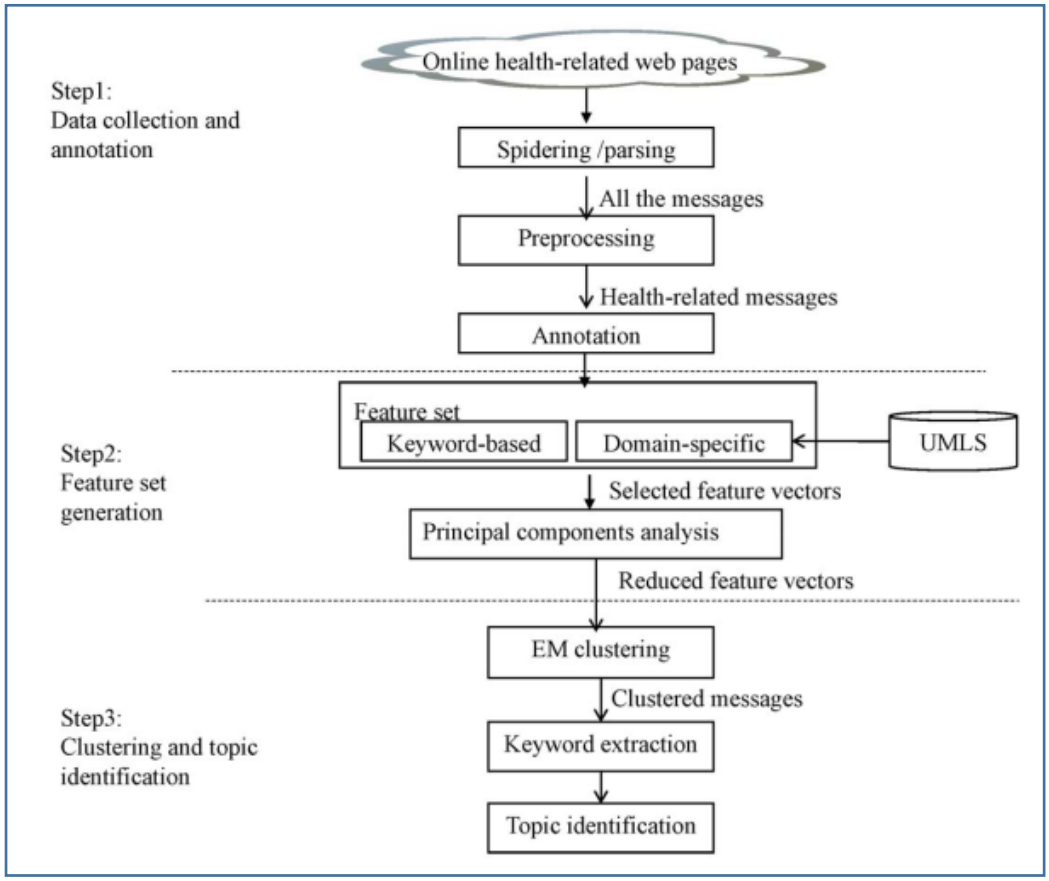

Figure 2.1: General Architecture of Topic Detection (Lu et al., 2013)

It is important to understand the general architecture for topic detection before proceed with further study. Figure 2.1 shows the general architecture of topic detection adopt from health-related hot topic detection in online community research done by YingJie Lu (Lu et al., 2013). This architecture consists of 3 steps, which are data collection and annotation, feature set generation, and clustering and topic identification.

For data collection and annotation step, it is started with determining the data source from online health-related web pages, which is Medhelp.org in this research. After that, the researchers used web crawling software to parse or crawl the web pages from discussion board (Lu et al., 2013). Next, pre-processing of parsed or crawled data have been carried out. Then, all of the messages are annotated into different topic groups according to pre-specified (Lu et al., 2013).

Next, feature set generation is done using two types of feature which are keyword-based and domain-specific respectively. By carry out principal components analysis, feature vectors will be reduced. The extracted feature from the messages were quantified as feature vectors as input for the topic clustering. However, these vectors were characterized by high dimensionality, redundancy and high correlation among individual attributes, which was not favourable for clustering (Witten & Frank, 2005). In order to reduce the high dimensionality, the original attributes were replaced using principal component analysis (PCA).

Finally, clustering and topic identification will be carry out. EM clustering is an algorithm which used in finding maximum likelihood estimates that will maximize the expectation of log-likelihood of the Wishart distribution (Kersten et al., 2005).In this step, EM clustering will be done by using data mining tool. From EM clustering, result obtained is known as clustered message. The clustered message can represent key

phrases with high scores that combined with expert opinions (Lu et al., 2013). Some of the key phrases were selected for topic detection to distinguish different clusters better (Lu et al., 2013).

2.3 Data Collections in Topic Detection

There are a few existing researches that have been done in topic detection. For instance, there is a research done by YingJie Lu on detecting health-related hot topics in online community. YingJie Lu used Medhelp.org as their data source since Medhelp is one of the most famous online health community (Lu et al., 2013). This community consists of over 230 discussion boards related to different diseases. This site attracts also visitors over 12 million per month. The site has also been selected as an experimental data source in previous studies (Yang & Tang, 2010). The aim of this research is to automatically distinguish different health-related topics in online health communities more effectively (Lu et al., 2013). The researchers also wish to investigate the differences between interesting topics among different types of disease discussion boards using text clustering method.

In order to perform data collection, web crawling software named Offline Explorer was used to crawl the web pages from the discussion boards (Lu et al., 2013). Next, extract messages were parsed and extract messages were stored into a database (Lu et al., 2013). Next, the process proceed with pre-processing steps including stop word removal and word stemming (Lu et al., 2013). After pre-processing steps, all messages were independently annotated by two annotators which were professional health experts (Lu et al., 2013). These messages are then categorized into different topic groups according to pre-specified categories (Lu et al., 2013). The crawling process are

related with this project, however there is different between the data source since this related work crawl web pages from discussion board whereas in this project is from Twitter.

Another research done by Eiji Amaraki describes about detecting influenza epidemics using Twitter. The objectives of this research is to describe an SVM based classifier which can filter negative influenza tweets and experiments empirically demonstrate the proposed method to detect influenza epidemics. Amaraki and other researchers collected 300 million tweets, starting from 2008 November until 2010 June, using Twitter API. They extracted only influenza-related tweets by using single keyword of "influenza". They obtained 0.4 million tweets from this operation (Amaraki et al, 2011). Then, they separated the data into two groups which are training data and test data. Training data are 5000 tweets sent in November 2018 whereas the rest are test data (Amaraki et al., 2011). The data collection method used in this research are related to my project.

On the other hand, there is an existing research done by Ryusei Matsumoto concern about visualisation of the occurrence trend of infectious disease using Twitter. The objective of this research is to provide user with real-time-infection maps which automatically created from Twitter tweets (Ryusei et al., 2018). Twitter API is used in this research to acquire tweets. Synonyms or names of seven kinds of infectious diseases also being used. As an example, another name for infectious disease "epidemic parotitis" is "mumps". Usually, "mumps" is used more frequently compared with "epidemic parotitis" (Ryusei et al., 2018). Hence, using synonyms can increase the number of tweets retrieved.

Moreover, a research describes about mining of health and disease events on Twitter is being done by Aditya L. Ramadona. The purpose of this research is to validate a search protocol that related with health terms using real-time Twitter data. These data can used to reveal a pattern of health situation in Indonesia (Ramadona et al., 2016). Phrases and words related with disease symptoms and health outcomes are extracted from Twitter by using Twitter Search API. Only tweets in Bahasa is being filtered because this study is intended to reveal relationships in the setting of Indonesia (Ramadona et al., 2016). There is only 390 tweets collected in Bahasa since retweets is being removed (Ramadona et al., 2016). Since many words occur infrequently, only the 22 highest word frequencies is considered (Referring to Figure 2.2).

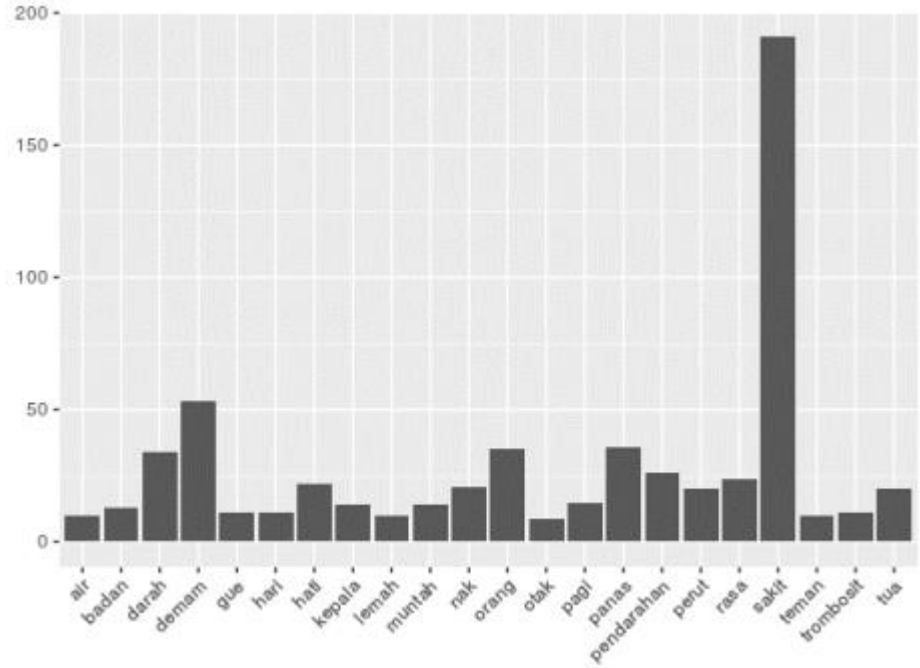

Figure 2.2: Words that appear more than 10 times (Ramadona et al., 2016)

Furthermore, there is also research done by Hideo Hirose and Liangliang Wang explain about prediction of infectious disease using Twitter. This research was carried out to predict the future trends of the disease much earlier (Hirose & Wang, 2012). The