



A Social Media Platform for Infectious Disease Analytics

Yang Hong^(✉)  and Richard O. Sinnott^(✉)

School of Computing and Information Systems, University of Melbourne,
Melbourne, Australia
hongyang47198@gmail.com, rsinnott@unimelb.edu.au

Abstract. The effect of seasonal epidemics and potentially pandemics represents a significant issue for public health. In this context, early warnings and real time tracking of the spread of disease is highly desirable. In this paper, we address the problem of detecting disease outbreaks through an automated, scalable Cloud-based system for collecting, tracking and analyzing social media data. Specifically, the focus here is targeted to three prevalent diseases (flu, chickenpox and measles) across three Australian cities using data from the Twitter micro-blogging platform. The epidemics related tweets are extracted using an ensemble learning classifier consisting of a combination of Support Vector Machines, Naïve Bayes and Logistic Regression and comparing the results with the Google Trend data to assess the effectiveness of the overall approach.

Keywords: Twitter · Big data · Machine learning · Infectious disease

1 Introduction

Short dynamic micro-blogging messages from services such as Twitter are now in widespread use. Millions of users utilize a real-time communication platform for posting and sharing a variety of dynamic topics from the environment, medicine, sports and politics. The total number of monthly active Twitter users is over 328 million¹ and the total number of Tweets sent per day has now surpassed 500 million² in 2017.

Twitter is an ideal data resource for detecting the trends related to many diverse events: population-wide sentiment analysis [1], prediction of elections [2], disaster detection [3] and public health applications amongst many others [4]. One major advantage of Twitter is its real time nature. This offers many benefits for analysis and prediction of infectious disease outbreaks such as flu, measles and chicken pox. The goal of this study is to verify if Twitter data can be used for tracking the trend and spread of infectious diseases.

In this context, one key problem is how to filter the noise coming from tweets with similar terms but irrelevant content [5]. Many noise filters have been developed and presented in previous works. Song and Lee provided a graph based filtering strategy.

¹ <http://twitter.com>.

² <https://www.omnicoreagency.com/twitter-statistics/>.

They focused on ensuring that noise features should not be connected with desired features or connected through weak relationships [6]. de Oliveira presented another lightweight stream filter [7].

They proposed a grammar-rule independent method called filter stream named entity recognition (FS-NER) to identify meaningful tweets by analyzing the characteristic and context of words in the tweet.

Another approach that is often used is to find the sentiment polarity of tweets based on machine learning approaches. The approach typically involves tweet-based text analysis comprising two significant methods: extracting sensitive information about infectious diseases according to a list of key words related to diseases such as fever, headache, rash and cough, and subsequently employing Twitter sentiment analysis using machine learning and natural language processing (NLP) algorithms.

Machine learning and NLP can be applied for classifier generation through the features present in the 140 characters (text) of tweets. When a new tweet is obtained, the classifier is able to determine the sentiment polarity (positive or negative mood). The success of the combination of both aspects achieves higher accuracy by removing (filtering) irrelevant tweets. Google Trends provides another source of data that is often used as the basis for social media verification. The processed tweets and Google Trend data are compared to identify their similarity as the evidence for the verification [9].

In this paper we investigate the relationship between social media data and Google Trend data with specific focus on infectious diseases. We present an efficient and scalable system design leveraging a Cloud-based system architecture based on the Australian National eResearch Collaborative Tools and Resources (NeCTAR – www.nectar.org.au) Research Cloud.

The rest of paper is organized as follows. Section 2 covers related background work. Section 3 describes related tools and presents an overview of the system architecture. Section 4 focuses on the data collection modules. Section 5 discusses data processing with specific focus on the classifier used for automated tweet analytics. Section 6 presents the analysis and discussion of the result. Finally, Sect. 7 provides conclusions and areas of future work.

2 Related Work

The global population increasingly uses the web. Individuals now commonly use Google to search for health related issues. Google Trends represent a simple way to identify patterns of web usage and health issues that might arise in particular locations, e.g. people searching for “flu remedies”. Similarly, the rise of social media during the last years has greatly increased the amount of content generated by the public. Of particular relevance here is public-generated data using location-based services associated with many mobile devices that allows identification of the precise location of data that is generated.

Social media analytics in the health domain is not new. Zaldumbide et al. focused on an architecture and realisation for real time health events [9]. The work was limited due to the global nature of Twitter and the challenges of high velocity data set issues when dealing with globally significant resources such as Twitter. Chew et al. [10]

analysed H1N1 related data from Twitter over a 6-month period in 2009 and the interaction/overlap with official emergency events. They selected H1N1 because it was the first major (global) event in the Web 2.0 era. They compared their data with official information from the Centre for Disease Control (CDC) and Public Health Agency of Canada. Despite the fact that the behaviour obtained in their research was very close to the official information, they concluded: “These numbers may give us a sense of population demographics; however, those who tweet about H1N1 may not necessarily be representative of the Twitter population, and the Twitter population is not representative of the general population [10].” Finally, they suggested that an improved approach could be performed using geo-location [11].

Aramaki et al. [12] used a different approach to obtain a better accuracy in their results. They recognised the fact that a tweet containing “influenza” or “flu”, may not mean that the twitter user has got flu; they may be just talking about the topic. To tackle this, they classify tweets into negative (suspicious, questions or news) and positive (the most accurate tweets). They compared several classifier methods such as Logistic Regression, Naive Bayes, Nearest Neighbour and Support Vector Machine (SVM). They based their studies comparing data from the Infection Disease Surveillance Center in Japan. They found that 42% of their analysed tweets were negative. This high value directly impacted the final results. For their analysis, they used an SVM-based classifier. They noted that their method outperformed the Google Trend-based methods applied to the same area of Japan.

St Louis et al. [13] focused on using Twitter to predict disease outbreaks? Almost all public health agencies rely on classic sources of information to monitor outbreaks such as doctors and hospital reports. One issue with these reports is that despite the fact they can be very accurate, they have a significant lag behind the actual events – diseases can spread in real-time and lags of knowledge capture can be critical, especially in pandemic situations. On the other hand, efforts of projects such as <http://www.healthmap.org/> often have a large amount of information, but their results cannot be formally verified with other sources.

Ji et al. [14] analysed tweets from a different point of view. They focused on how concerned people are about disease outbreaks. They used sentiment classification on tweets based on a two-step approach including classifying Tweets as personal or negative. First they divided Tweets into personal and non-personal and then classified Tweets into negative and neutral and applied machine learning-based classifiers for tweet classification. When comparing data with official information sources however they did not achieve high degrees of similarity.

Khan et al. [15] focused on early detection of outbreaks before they transform into full-blown pandemics. They identified three kinds of tweet generators: media and news that posts, tweets related to publicity or advertisement of medicines and tweets created by users that contain key words. The scope of their study was to analyse the last group of tweets to identify true reports of symptoms in the person tweeting. They applied NLP by training a subset of data and applying pre-processing techniques. They used unigram and bigram models to compare their proposed model to obtained a precision of 88.7% compared to other models.

Lamos et al. [16] focused on reducing the impact of an epidemic disease such as the flu. They analysed tweets from the United Kingdom for a period of 24 weeks. They

compared this data with official data from the Health Protection Agency. In this work, they used textual analysis and not just a count of occurrences of words, e.g. flu. They obtained a high correlation but noted that the results could be more accurate if they applied methods to remove media hype and discussions around health events.

Achrekar et al. [17] tried to go further by attempting to predict flu trends. In their work, they regarded Twitter users as sensors. They removed retweets and tweets from the same user, and labelled tweets with data related to influenza-like illness (ILI). In order to collect data from Twitter, they use the search API using keywords such as flu, swine flu and H1N1. They compared their results with data from the Center for Disease Control and Prevention and obtained a strong correlation with a Pearson's correlation coefficient of 0.9846.

Signorini et al. [18] explored the analysis and tracking of H1N1 activity including symptoms and medications. Using the timestamp and geo-location information stored in tweets, they used Google Maps to show tweet distributions and their temporal dependencies.

Doan et al. [19] used tweets collected over 36 weeks. The number of tweets collected in that period of time was over 587 million from approximately 24.5 m users. They used this data to compare data from CDC's U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet)³. This institution compiles more than 25 million patient registers per year. They used two methods to filter the information. The first method was based on four keywords: flu, cough, headache and sore throat. With this method, they got a Pearson's correlation coefficient of 0.95. They compared various methods but one key conclusion they reached was that the keywords list was directly related to the official information, i.e. an increase or decrease in keywords correlates directly with the accuracy of official information.

Hirose et al. [20] proposed a method to forecast influenza outbreaks through Twitter, again using ILINet³ as the official data source used for comparison. They filtered tweets into two groups: positives and negatives. Negatives are tweets contained symptom words but not really expressing influenza reports, e.g. "Football fever this weekend!!!". On the other hand, they identified that positive tweets are more likely connected to real symptoms. They identified that performing multiple linear regressions could improve the accuracy of the prediction.

Jin et al. [21] revealed in their study that rumours can spread across Twitter, e.g. how rumours such as "Health officials might inject Ebola patients with lethal substances". These tweets spread quickly in countries like Ghana, Nigeria and Kenya. They analysed three cases: one case from Dallas, another from New York and the case of Dr Spencer case who volunteered in Guinea and became infected with the Ebola virus. They calculated the time of response for retweets on a specific topic and created a ratio table. They identified a shorter time for Responses on News Stories than for Twitter rumours.

³ Influenza-Like Illness Surveillance Network (<http://www.cdc.gov/flu/weekly/>).

3 System Architecture

The platform for infectious disease analytics utilized the Australia-wide NeCTAR Research Cloud. NeCTAR offers access to over 30,000 physical servers across multiple availability zones. All Australian academics and students can request and acquire servers and storage for their own research needs. The overall architecture of the platform utilized in this paper is shown in Fig. 1.

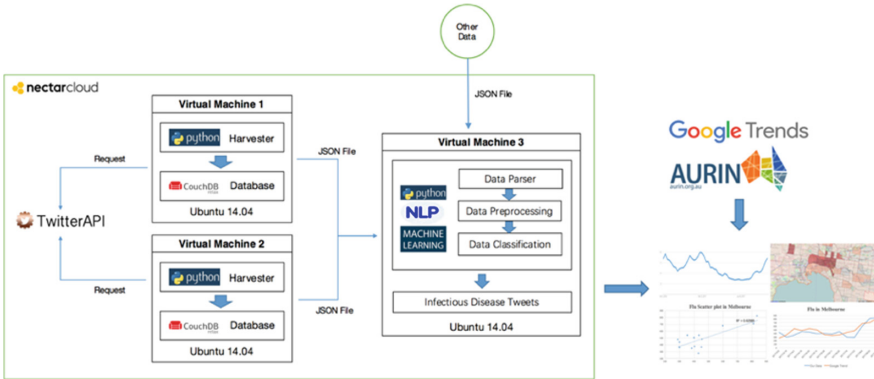


Fig. 1. System architecture.

The system includes three modules: a data collection module, a data processing module and a data analysis module. The first two modules were implemented directly on the NeCTAR Research Cloud. Two virtual machines were used to run Twitter harvesting scripts. Each harvester sends requests to the Twitter server using an API provided, and saves the resultant (JSON) data in a noSQL database (CouchDB). The collected data was integrated and processed by the text analysis components in the third virtual machine. Finally, the processed results were sent to the third module for data analysis.

In the architecture, the virtual machines were based on Ubuntu 14.04 operating system with 8 GB RAM, 2 vCPU and 60 GB memory storage. One virtual machine was required to do more computational tasks so was allocated 16 GB RAM, 4 vCPU and 120 GB memory. It is noted that further virtual machines can be allocated on the fly, e.g. when further data needs to be collected.

4 Data Collection Module

Data was collected from three resources for different purposes. To establish the relationship between tweets and infectious disease in Australia, it was essential to obtain tweets with geo-location (lat/long), i.e. based on uses tweeting with the location-based service (GPS) on their phone activated. It is the case that the majority of Twitter data does not include a geo-location (typically around 15% of tweets), hence it was required to obtain many more tweets and filter those without a geo-location. To aid this work,

the University of Melbourne has established a Twitter resource comprising over 100 m tweets – many of which are geo-located. Furthermore, when obtaining tweets, it is possible to restrict the areas of interest through the setting of a bounding box. The bounding boxes for three major cities of Australia were chosen in this work (Melbourne, Sydney and Adelaide). The platform can readily be repurposed by setting new bounding boxes, e.g. for new cities or indeed new countries.

A further data resource was also utilized⁴. This data provides researchers with huge amounts of annotated tweets suitable for sentiment analysis. This included a training dataset and testing dataset. Since the number of positive tweets and negative tweets is the same in the training data, one of the advantages of this dataset is that it was not necessary to solve dataset imbalance issues. In order to generate a suitable classifier, we applied machine learning algorithms to the training data and compared the performance with different model parameters.

The total amount of data used in this work comprised 6,764,090 tweets (18.63 GB) with 3,160,086 tweets from Melbourne, 1,795,170 tweets from Sydney, 208,834 from Adelaide and 1,600,000 from the rest of the world (Tables 1, 2 and 3).

Table 1. Tweets collected from harvester.

City	Size (GB)	Tweets	Time period
Melbourne	8.07	1,945,588	Mar 5 th 2017 – Sep 24 th 2017

Table 2. Tweets collected from UniMelb Twitter resource.

City	Size (GB)	Tweets	Time period
Melbourne	3.99	1,214,498	July 28 th 2014 – June 13 th 2016
Sydney	5.66	1,795,170	July 28 th 2014 – June 13 th 2016
Adelaide	0.67	208,834	July 28 th 2014 – June 13 th 2016
Total	10.32	3,218,502	

Table 3. Tweets collected from Sentiment140.

Data type	Size (KB)	Polarity	Tweets
Training data	238,800	Positive	800,000
		Negative	800,000
Test data	74	–	14,076

⁴ <http://help.sentiment140.com/home>.

5 Data Processing Module

A raw tweet contains considerable amounts of information – a 140-character tweet typically contains 9 kb of associated metadata. Hence, it is typically necessary to parse the data and reduce the data size. In this work, only a few fields were required for infectious disease analysis. These included:

- *text* field, which forms the main body of the tweet. It contains a 140-character long set of terms.
- *created_at* field indicates the Coordinated Universal Time when the user posted the tweet.
- *coordinates* field provides the GPS location of the tweet [8]. Many tweets omit this information if the location-based network on the device is turned off. Nevertheless, tweets can often be approximately geo-located based on the metadata associated with the tweet, e.g. if the profile mentions Melbourne then it can be associated (retrieved) from a harvester requesting tweets from Melbourne.

Once tweets are harvested and parsed, they are stored in a Cloud-based database. It is also noted that a unique identifier is associated with each tweet. This identifier assists in data operations and ensures that no duplicate tweets are stored.

In order to generate the appropriate data for subsequent classification, a sequence of text processing procedures is applied to process the text field. This includes tokenization, conversion to lower case, stop word removal and word normalization. The preprocessed data also removes non-English language word, punctuation and emoji.

Using the resultant processed data, the next phase involved extracting illness-related tweets related to infectious diseases. The approach taken combined two aspects: keyword and sentiment analytics. The key words list adopted were taken from the public data of Centers for Disease Control and Prevention (CDC). The text fields of tweets were checked if they contained relevant key words (flu, measles, chicken pox) to

Table 4. Infectious disease key words.

Disease type	Key words
Flu	Flu, cough, runny...
Measles	Measles, rash, chill...
Chicken pox	Pox, blister, bump...

Table 5. Infectious disease key words

Tweets	Polarity	Key words	Target tweets
Have a nice day	Positive	No	No
I don't like this food	Negative	No	No
I had a cold. Feel terrible	Negative	Yes	Yes
Crazy sushi fever!	Positive	Yes	No

decide if they were related to the corresponding infectious disease. Table 4 provides part of key word in the work. In addition, sentiment analysis was used to establish the tweet polarity: positive or negative sentiment. Positive tweets were discarded since it can be assumed that tweets related to infectious disease should express negative sentiment. In order to build an effective sentiment classifier, various experiments were carried out. Potential infectious disease tweets contain key words that have negative sentiment. Table 5 provides examples used for tweet selection.

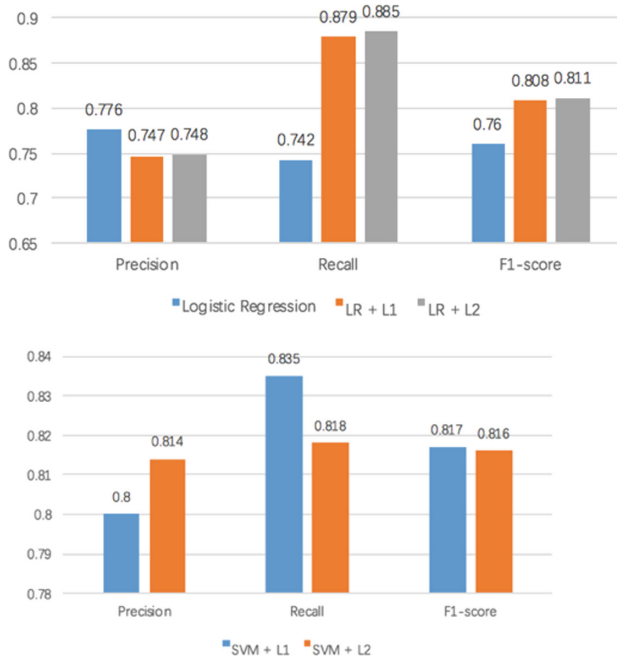


Fig. 2. The performance of the model with L1 and L2 penalty

A set of tweets was used to generate the classifier used for sentiment analysis. A Bag of Word approach was used to extract the features from processed training data. The text model after fitting had a feature space with 777,666 features. The feature vector then was applied to all of the data, whereby the data sharing the same vector had the same dimension. We evaluated the data with test dataset.

In undertaking this we compared three algorithms: Logistic Regression (LR), Support Vector Machine (SVM) and Naïve Bayes (NB). We considered the effect of L1 penalty and L2 penalty with regards to the performance of LR and SVM.

The evaluation of L1 and L2 penalty for LR is described in Fig. 2. Non-regularization LR is used as the baseline. The recall increased from 18.4% with L1 to 19.2% with L2. This indicates that both L1 and L2 reduce false negative errors and greatly improve positive tweet classification. However, their precision drops slightly, which means that the penalty parameter is more sensitive to a positive polarity rather than negative polarity.

The enhancement of the F1-score proves that the penalty parameter is able to improve the model performance. L1 is generally considered more powerful for handling data with sparse features. Comparing L1 and L2, the F1-score of L2 is 1.8% higher than L1 in our experiment, which indicates that L2 actually outperforms L1 with our dataset. Therefore, we chose L2 as the penalty for LR model when building the ensemble classifier. However, unlike LR, the experiment in SVM exhibits different results. The L1 and L2 penalty are evaluated when the penalty term parameter is 0. Although L2 shows a more balanced evaluation, the F1-score of L1 was slightly higher than L2. This indicates that L1 has better performance than L2. Therefore, we choose L2 as the penalty type for the SVM model.

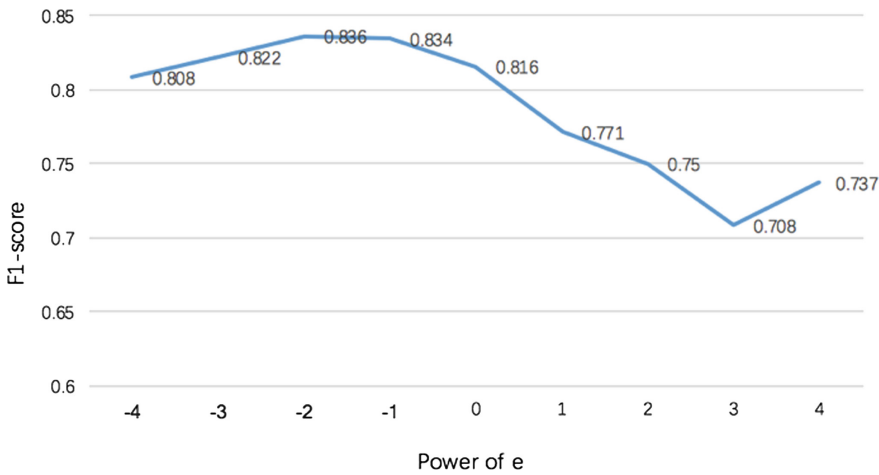


Fig. 3. Penalty term parameter on performance of SVM

The investigation of the penalty term parameter of SVM is displayed in Fig. 3. This parameter plays a significant role in the performance of the model, since the parameter can determine how much the penalty will affect the weight choice in the SVM model. A very large parameter value can make the penalty close to zero, which means it nearly has minimal influence on weight. A small value of parameter has the opposite effect. In Fig. 3, the x-axis represents the power of e , whilst the y-axis represents the F1-score of the model. As seen, the model has the best performance when the power of e is -2 . This indicates that the parameter has a small value between 0 and 1. Moreover, we find that the correlation of performance and the parameter is non-monotonic. Hence, the figure shows that the performance of power 4 is higher than power 3.

Figure 4 shows the evaluation of the different machine learning classifiers. Even after adding the L2 penalty parameter, LR still has the worst performance (0.811). The SVM classifier has the best performance (0.836) when considering a single classifier where the F1-score is slightly greater than NB (0.819). The ability of NB and SVM for predicting positive tweets and negative tweets is more balanced than LR. The percentage of positive tweets and negative tweets is unknown in advance. To address

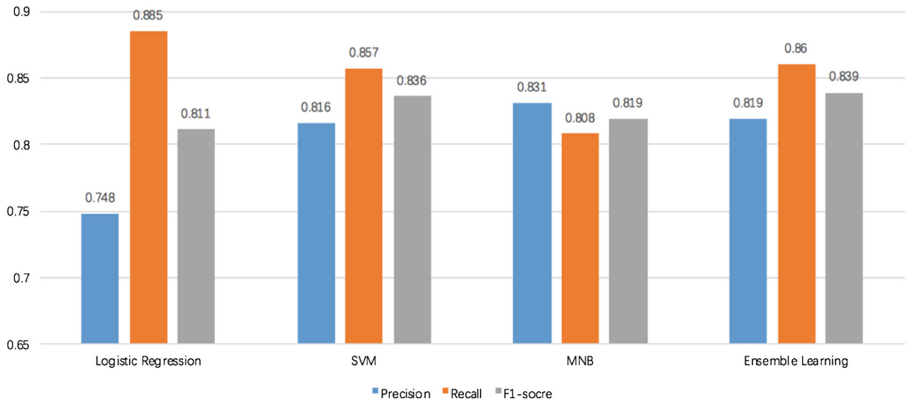


Fig. 4. Comparison of machine learning classifiers

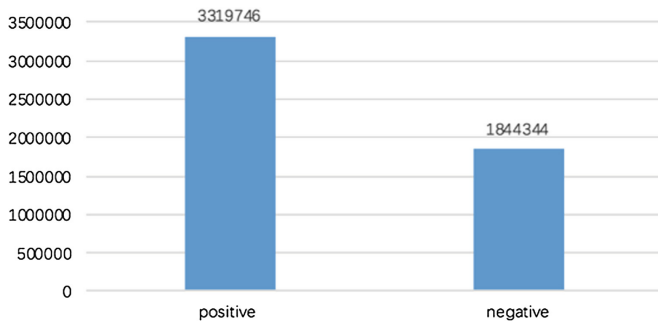


Fig. 5. Overall amount of positive and negative tweets in the dataset

this, we build an ensemble classifier to reduce the impact of an unbalanced dataset. This method considers the characteristics of the three classifiers. It is noted that ensemble learning with hard voting suffers from lower precision than MNB, with F1-score (0.839) higher than SVM (0.836). This indicates that the combination of classifiers is more powerful than any single classifier. It is also noted that the ensemble classifier has balanced tendency to positive or negative sentiment hence the classifier for both polarities is reliable. Figure 5 shows the percentage of positive and negative tweets in the dataset. The figure illustrates that our dataset includes 64.3% positive tweets and 35.7% negative tweets.

6 Results and Analysis

In this section, we analyze and compare our data against a more official data source. Our analysis focuses on three specific infectious diseases: flu, chicken pox and measles. In order to compare this data with Google Trend data, our starting method is grouping our data into time ranges. Specifically, data is categorized into every half-month data

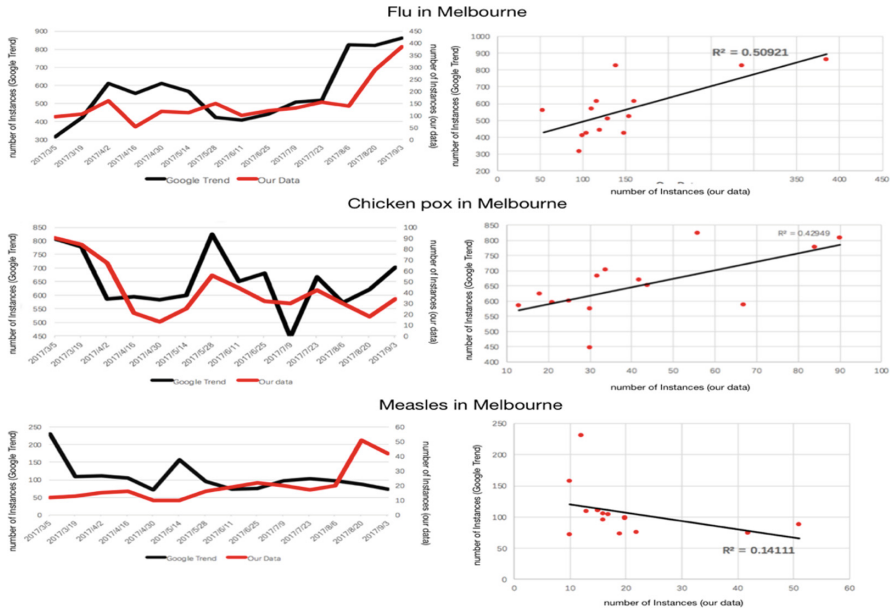


Fig. 6. Infectious disease events in Melbourne

for Melbourne and Sydney. Since we only have a (relatively) minimal data from Adelaide, this is categorized into 28-day periods.

The analysis of results of each disease needs to be rescaled before comparing against Google Trend data. We present a scatter plot and calculate the Pearson Coefficient to indicate the degree of correlation. Moreover, we try to find the interval with 50 units along x-axis that can cover the majority points in scatter plot. From this we calculate the average value of points in the interval, which can subsequently be used as a possible alarm value for an infectious disease outbreak in a given city (Fig. 6).

As seen, for flu in Melbourne, the two lines in the graph both increase from June 11th and keep rising to a peak on September 3rd. It is noted that this period reflects the season change to winter. The scatter plot shows the majority of points are distributed around 139 instances. If there are more flu tweets found over this value at certain time point, it is possible that a flu outbreak is to occur. For chicken pox in Melbourne, two instances both exhibit a downtrend between March 5th to April 2nd. After this decline, the peak of Google Trend is shown on May 28th, which is same as the peak of our data. Points in the scatter plot of chicken pox are random distributed so that we cannot decide a given threshold value. The measles instances do not show any similarity with Google Trend data, i.e. the scatter plot shows a negative correlation between two social media data and Google Trend.

For flu in Sydney, the two graphs show the same downtrend from August 1st to September 12th due to the effect of the season change. The majority of points are distributed in the range of 80 to 130. This interval contains 10 points and their average values are 110, which can be used as a threshold value to indicate a potential outbreak

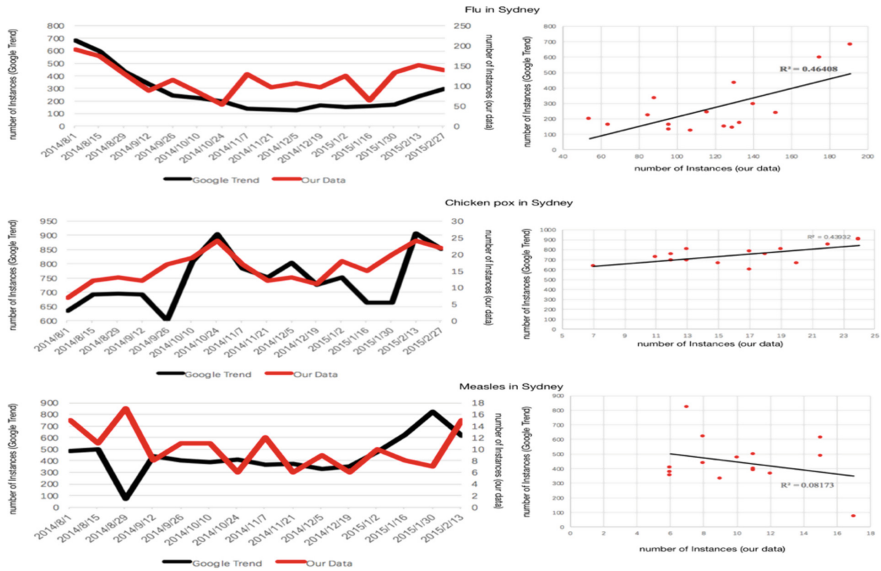


Fig. 7. Infectious disease events in Sydney

of flu in Sydney. For chicken pox, both our data and Google Trend had a peak on October 24th and February 13th. We found 24 tweets with 905 instances respectively on the first peak and 24 tweets with 908 instances on the second peak. However as with as Melbourne, the measles data in Sydney does not show any similarity between the two instances (Fig. 7).

For flu in Adelaide, the two instances show the same trend from August 1st to October 24th. However, the peak of our data occurred on November 21th, which is not indicated in Google Trends. Another similarity is that both charts begin to increase from December 19th. Moreover, points are randomly distributed in the scatter plot. For chicken pox in Adelaide, two instances show a degree of similarity in the two intervals. Although we found that the majority of points are centered at around 24 instances, this value cannot be regarded as a possible alarm value. The main reason is that the size of observed Chicken pox events in Adelaide was too small. For measles in Adelaide, the two instances have a degree of similarity from August 1st to August 29th and from December 19th to January 16th (Fig. 8).

As shown in Table 6, flu has a relatively high degree of correlation in Melbourne, Sydney and Adelaide, which demonstrates that our data is consistent with the Google Trend data for flu. Therefore, we may reasonably assert that tweets can be used as an indicator of flu outbreaks. Similarly, we see that infectious disease data fluctuates along with the changing of the season, i.e. both our data and Google Trend data peak in winter.

Chicken pox also shows a reasonable correlation with a moderate correlation in Sydney and a higher degree of correlation in Melbourne and Adelaide. However,

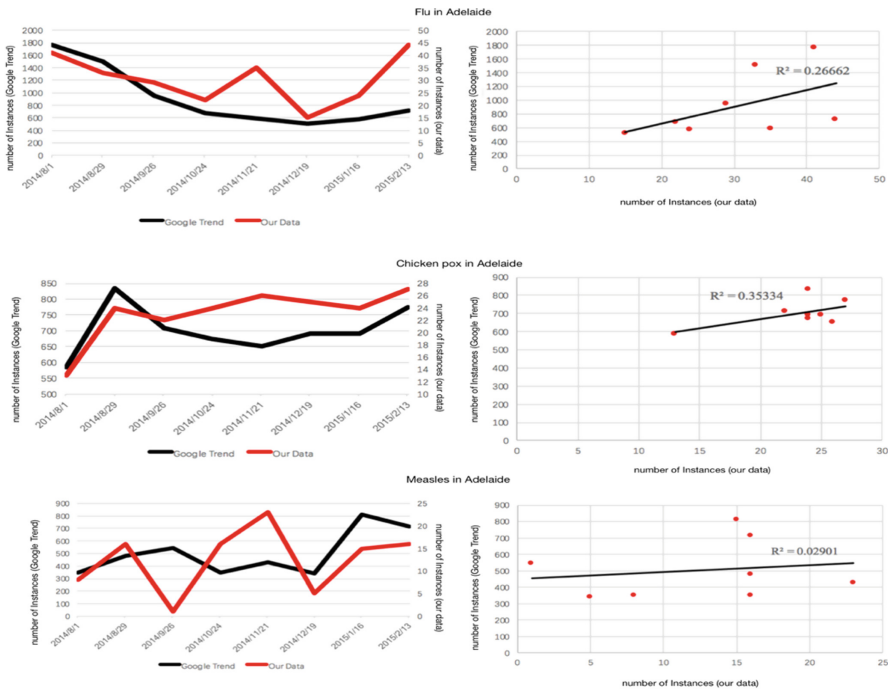


Fig. 8. Infectious diseases data in Adelaide

Table 6. Summary of correlation degree

City	Disease type	R^2	Pearson coefficient	Correlation degree
Melbourne	Flu	0.50	0.71	High
	Chicken pox	0.43	0.66	High
	Measles	0.14	-0.38	Moderate
Sydney	Flu	0.46	0.68	High
	Chicken pox	0.44	0.50	Moderate
	Measles	0.08	-0.44	Moderate
Adelaide	Flu	0.27	0.65	High
	Chicken pox	0.35	0.55	High
	Measles	0.03	0.17	Low

unlike flu, there is no strong evidence that indicates that the peak of chicken pox has a direct relationship with the season.

Measles does not show any correlation between our data and Google Trend data. Indeed, for Melbourne and Sydney, the correlation is negative, which is opposite to expectations. Furthermore, measles shows no change with the changing of the season.

7 Conclusions and Future Work

This project mainly has explored whether social media is capable of tracking and analyzing the trend of the infectious diseases. Unlike other approaches, we combine both sentiment analytics and event detection. This offers enhancements that have not been explored in other works.

We compared our data against Google Trend data. The performance of flu and chicken pox displayed a high correlation with tweet data, which suggests that Twitter data can be used to track flu and chicken pox outbreaks. However, there was no similarity observed for measles. Flu is a seasonal disease. Since there is a difference in geo-location and weather between the three cities, the time period for outbreaks of flu was exhibited during the winter but the specific time was slightly different. Unlike flu, outbreaks of chicken pox and measles are more irregular.

In the future, other aspects will be taken into account to increase the performance of the model. First, we will use more data to train our classifier. Learning more diverse types of data is the most efficient and straight way for model improvement. Second, we will more carefully tune the parameters in each of machine learning model. Third, we will come up with improved methods for sentiment classification, e.g. employ word vectors to extract the features in the text and build neural networks to train the model.

References

1. Liu, B.: Sentiment analysis and opinion mining. *Synth. Lect. Hum. Lang. Technol.* **5**(1), 1–167 (2012)
2. Tumasjan, A., et al.: Predicting elections with twitter: what 140 characters reveal about political sentiment. In: *ICWSM 2010*, vol. 1, pp. 178–185 (2010)
3. Sakaki, T., Makoto, O., Yutaka, M.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: *Proceedings of the 19th International Conference on World Wide Web*. ACM (2010)
4. Callison-Burch, C., Dredze M.: Creating speech and language data with Amazon’s mechanical turk. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. Association for Computational Linguistics (2010)
5. Szomszor, M., Kostkova, P., de Quincey, E.: #Swineflu: Twitter predicts swine flu outbreak in 2009. In: Szomszor, M., Kostkova, P. (eds.) *eHealth 2010. LNICST*, vol. 69, pp. 18–26. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23635-8_3
6. Song, J., Lee, S., Kim, J.: Spam filtering in twitter using sender-receiver relationship. In: Sommer, R., Balzarotti, D., Maier, G. (eds.) *RAID 2011. LNCS*, vol. 6961, pp. 301–317. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23644-0_16
7. de Oliveira, D.M., Laender, A.H., Veloso, A., da Silva, A.S.: FS-NER: a lightweight filter-stream approach to named entity recognition on twitter data. In: *Proceedings of the 22nd International Conference on World Wide Web*, pp. 597–604. ACM, May 2013
8. Burton, S.H., et al.: “Right time, right place” health communication on Twitter: value and accuracy of location information. *J. Med. Internet Res.* **14**(6), e156 (2012)

9. Zaldumbide, J.P., Sinnott, R.O.: Identification and verification of real-time health events through social media. In: IEEE International Conference on Data Science and Data Intensive Systems, Sydney, Australia, December 2015
10. Chew, C., et al.: Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One* **5**(11), e14118 (2010)
11. Byun, C., et al.: Automated Twitter data collecting tool for data mining in social network. In: Proceedings of the 2012 ACM Research in Applied Computation Symposium. ACM (2012)
12. Aramaki, E., et al.: Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2011)
13. St Louis, C., et al.: Can Twitter predict disease outbreaks? *BMJ*, 344 (2012)
14. Ji, X., et al.: Monitoring public health concerns using Twitter sentiment classifications. In: 2013 IEEE International Conference on Healthcare Informatics (ICHI). IEEE (2013)
15. Khan, M.A.H., et al.: A robust and scalable framework for detecting self-reported illness from Twitter. In: 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom). IEEE (2012)
16. Lampos, V., et al.: Tracking the flu pandemic by monitoring the social web. In: 2010 2nd International Workshop on Cognitive Information Processing (CIP). IEEE (2010)
17. Achrekar, H., et al.: Predicting flu trends using Twitter data. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE (2011)
18. Signorini, A., et al.: 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), e19467 (2011)
19. Doan, S., et al.: Enhancing Twitter data analysis with simple semantic filtering: example in tracking influenza-like illnesses. In: 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB). IEEE (2012)
20. Hirose, H., et al.: Prediction of infectious disease spread using Twitter: a case of influenza. In: 2012 Fifth International Symposium on Parallel Architectures, Algorithms and Programming (PAAP). IEEE (2012)
21. Jin, F., et al.: Misinformation propagation in the age of Twitter. *Computer* **47**(12), 90–94 (2014)