Predicting Prevalence of Influenza-Like Illness From Geo-Tagged Tweets

Kewei Zhang[*†]
kewei.zhang@
uqconnect.edu.au

Reza Arablouei[†]
reza.arablouei@
csiro.au

Raja Jurdak[†*]
raja.jurdak@
csiro.au

[*]School of Information Technology and Electrical Engineering, University of Queensland, St.Lucia QLD, Australia
[†]CSIRO Data 61, Pullenvale QLD, Australia

ABSTRACT

Modeling disease spread and distribution using social media data has become an increasingly popular research area. While Twitter data has recently been investigated for estimating disease spread, the extent to which it is representative of disease spread and distribution in a macro perspective is still an open question. In this paper, we focus on macroscale modeling of influenza-like illnesses (ILI) using a large dataset containing 8,961,932 tweets from Australia collected in 2015. We first propose modifications of the state-of-the-art ILI-related tweet detection approaches to acquire a more refined dataset. We normalize the number of detected ILI-related tweets with Internet access and Twitter penetration rates in each state. Then, we establish a state-level linear regression model between the number of ILI-related tweets and the number of real influenza notifications. The Pearson correlation coefficient of the model is 0.93. Our results indicate that: 1) a strong positive linear correlation exists between the number of ILI-related tweets and the number of recorded influenza notifications at state scale; 2) Twitter data has promising ability in helping detect influenza outbreaks; 3) taking into account the population, Internet access and Twitter penetration rates in each state enhances the prevalence modeling analysis.

Keywords

Classification; data mining; disease modeling; public health monitoring; regression analysis; Twitter

1. INTRODUCTION

Public health surveillance is an essential mission of every government. In the current era of big data, data-driven epidemics modeling and surveillance system has drawn unprecedented attention.

In Australia, epidemics of seasonal influenza are one of the major public health concerns. Seasonal influenza strains circulate at peak during each winter. During the first half of 2015, there were more than 30,000 influenza cases notified [5] when the number of flu notifications reached the highest in history during the same time period. Besides, public health data are traditionally collected via surveys and by aggregating statistics obtained from healthcare institutions. Such data collection processes are usually costly, slow, and retrospective.

Recently, analyzing data collected from *Twitter*, a microblogging social network, has shown promise in assessing the prevalence of flu [9]. However, modeling disease spread and distribution with Twitter data involves several challenging tasks. First of all, detecting tweets that contain expression of disease symptoms requires natural language processing (NLP), which is an active research field with plenty of open challenges [12]. Moreover, health-related tweets are relatively scarce [9] making their detection within a large corpus of tweets a highly unbalanced classification problem. Zuccon *et al.* [21] investigated the suitability of statistical machine learning approaches in detecting ILI-related tweets automatically. Their results show that the optimal f-score, which is the harmonic mean of precision and recall, is only up to 0.736 among most of the state-of-the-art approaches. Considering the limited likelihood of users mentioning their health condition in Twitter, only relying on classification techniques for obtaining ILI-related tweets can induce large errors and lead to a biased epidemic model.

In this paper, we analyze a large database of 8,961,932 tweets from Australia collected in 2015 for studying the disease spread and distribution of influenza-like illness epidemics. We propose modifications to the algorithm proposed in [16] to improve the ILI-related tweets classification performance. We also take into account the Internet and Twitter penetration rates at each state to normalize the results. Afterwards, we establish a state-level model between the Twitter data and the true influenza notification data and also perform temporal and spatial analysis for exploring how well can Twitter data capture the feature of disease spread and distribution. Furthermore, we identify the limitations of our study as well as the opportunity for further study on utilizing Twitter data for public health surveillance.

The remainder of the paper is organized as follows. Section 2 presents related work. Section 3 gives some general statistics about the dataset we use and provides the methodology of the experiment design. Section 4 presents the experiment results and discussions. Section 5 elaborates on the limitations of the work. Section 6 provides conclusions and ideas for future work.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3051150

2. RELATED WORK

In the area of social media data mining, Twitter data have been used in many studies and provided valuable insights into various research fields including demographics estimation, public opinion reflection, real-time event monitoring, and public health surveillance. For example, Sakaki *et al.* proposed an algorithm to monitor earthquakes on the basis of tweet text features [17]. Tumasjan *et al.* showed the feasibility of tracking public political opinion and predicting the election results by analyzing the relevant tweets [20].

Culotta *et al.*'s work identifies similar correlation in Twitter data with Google Flu Trend after experiments of tweet keywords generation, selection, document filtering methods, and regression method comparison [9]. Prieto *el al.*'s work focuses on using Spanish and Portuguese tweets to estimate the community health with various maladies such as flu, depression, and eating disorders [14]. Moreover, Paul and Dredze apply an Ailment Topic Aspect Model (ATAM) over a large number of tweets to discover the mentions of various ailments, such as allergies, depression, cancer, etc. to model syndromic surveillance [13].

Sadilek *et al.* model disease epidemics by analyzing the interactions of online user activity and human mobility patterns using geo-tagged tweets [16]. They propose a semi-supervised cascade-based approach for detecting ILI-related tweets. Then they model the spread of influenza by analyzing the co-location of "sick" post users and his or her surrounding Twitter users. Our work proposes modifications to the ILI-related tweets detecting part of Sadilek *et al.*'s work, which is an iterative labeling and training approach along with classification result validation, to improve the performance of the classification algorithm.

In Jurdak *et al.*'s work [15], the authors demonstrate that the Twitter data can be considered as a reliable source for studying the human mobility patterns. Their research also provides insights into the potential of using the Twitter data for public health studies.

3. METHODOLOGY

In this section, we first describe the dataset we use. Then, we discuss how we modify the classification approach to achieve a better performance. In addition, the methodology of temporal and spatial mapping of ILI-related tweets in Australia and regression model for estimating the flu notifications from ILI-related tweets are further illustrated.

3.1 The Data

Twitter posts, also known as tweets, which can be up to 140 characters long, form the basis of our work. Within each tweet, users can add the hash-tag symbol (#) before a relevant keyword or phrase to categorize their tweets and use emojis to express their emotions. According to recent Twitter statistics, there are approximately 320 million Twitter users all over the world [7], 2.8 million of them being from Australia [6].

A collection of tweets obtained by CSIRO is our major data source. With the help of Twitter Streaming API[1], a large dataset of geo-tagged tweets within Australia for the entire year of 2015 has been generated by a year long collecting process. The data is stored in MongoDB [2], a cross-platform document-oriented NoSQL database. MongoDB features include the characters of big data storage, index support, straightforward queries and higher speed than traditional relational databases [11], which make interaction with data easier and more efficient.

Table 1: Tweet JSON Fields

Fields	Format	Description
text	"I'm so freaking sick :("	Tweet message
created_at	"Fri Apr 13 11:56:04 +0000 2012"	Time of posting
user.id	"id":552638416	User Id
coordinates	{"type": "Point", "coordinates": [-33.927753, 150.899351]}	Geo-location of device when tweeting
place.full_name	"full_name": "Sydney, New South Wales"	Place information associated with Tweet

All collected tweets are represented is JSON format. In our work, we only consider five particular fields as listed in Table 1. Table 1 provides a more concise description of the required JSON fields using a real tweet example.

After some basic data cleaning, the database contains 8,961,932 tweets posted by 225,641 unique Twitter users. Among all tweets, 3,469,190 of them are posted with precise location coordinates. Nearly every tweet is associated with a "place" field, which is location information that already existing on the Twitter server database. This field, as a coarse location information, can either be automatically assigned or manually allocated by the users. Our work considers this data field as a complement of the geo-enabled tweet database.

3.2 Detecting Illness-Related Tweets

Our primary task is to identify tweets that indicate the authors are infected at the time of posting. Based on the findings from related works [9], [16], the problem of detecting illness-related tweets is expected to be an unbalanced classification problem with scarce data points. In our work, we propose modifications to the classification algorithm in [16] and apply a semi-supervised cascade learning approach to learning Support Vector Machine (SVM) [8] classifiers with a large area under the precision-recall (PR) curve. It is worth to mention that the area under the PR curve is a more valuable evaluation method in our scenario, as the imbalance of the problem will generate a constant large area under the receiver operating characteristic (ROC) curve. The classifiers are trained to distinguish "sick" tweets (ILI-related tweets) and "other" tweets (non-ILI-related tweets) in the tweet database.

The prerequisite of learning such classifiers is to obtain a high-quality set of labeled training data. We employ an iterative process to achieve this. The training process is shown in Figure 1 and the classification process is shown in Figure 2. Within the mechanism, two different SVM classifiers, denoted by *Cs* and *Co*, are trained using *scikit-learn* Python library[2], which label the tweets as either belonging to the class "sick" or the class "other". The classifier *Cs* is highly

[1]https://dev.twitter.com/streaming/public

[2]http://scikit-learn.org

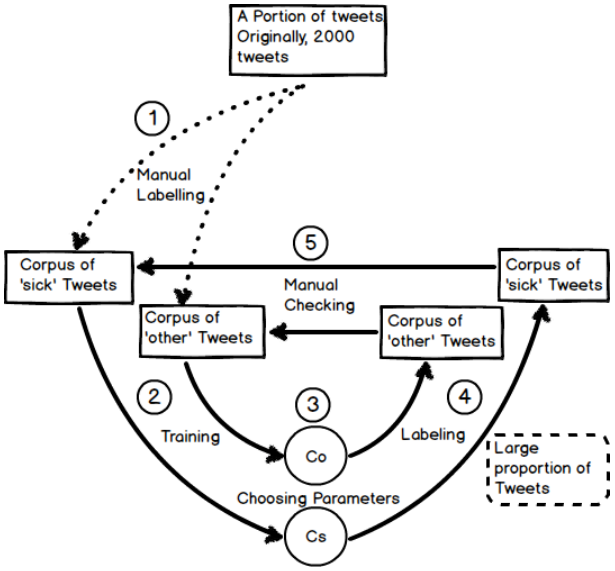

Figure 1: Training the SVM classifiers

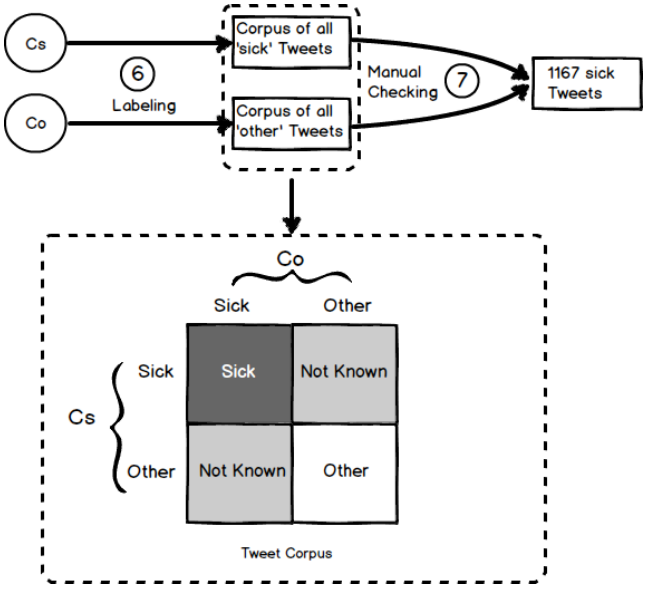

Figure 2: Classification stage

penalized for including false positives (mistakenly labeling an "other" tweet as a "sick" one) and the classifier *Co* is highly penalized for including false negatives (classifying a "sick" tweet as "other").

In each training iteration, two parameters, class weight and the *C* parameters, which influence the performance of the classifiers, are carefully selected through experiments. We fix one parameter and vary the other within a wide range of values to observe the changes in precision, recall, false-positive error rate, and false-negative error rate. The parameters leading to the highest precision and lowest false-positive error rate are chosen for *Cs* while the parameters that give the highest recall and the lowest false-negative error rate are chosen for *Co*. Meanwhile, manual checking validations are included in both training stage and classification stage because those are essential steps for classifying the ILI-related tweets accurately. Step by step instructions for the training and classification processes are discussed in the next paragraph and shown in Figures 1 and 2.

Initially, a small portion of tweets, which is around 2000, has been labeled manually resulting in 36 ILI-related tweets and 1974 non-ILI-related tweets (1). With the labeled dataset, *Cs* and *Co* are trained (2) and examined with a various range of values for the parameters. Parameters that result in the best classification performance are selected (3). Then, a larger tweet corpus is introduced and labeled using *Cs* and *Co* (4). The trained classifiers assign labels to the tweets. We further manually check the tweets and add them to the previous labeled tweets corpus as reforming the basis of training data for next round of classifier training (5). After finalizing the training of *Cs* and *Co*, both classifiers are used for labeling the entire tweet database (6). Any tweet may be labeled as "sick" or "other" by both classifiers or either one of them. Therefore, in the final step (7), we manually check those tweets labeled with different labels by the two classifiers, which is represented by the "not known" part in Figure 2.

For features, all unigram, bigram, and trigram word tokens are considered in our work. For instance, a tweet message "I got the flu" is represented by the following feature vector:

$$(i, get, flu, i\,get, get\,flu, i\,get\,flu)$$

Before tokenization, all texts are converted into lowercase and punctuations and stopwords are stripped. However, hash-tags and emojis are retained as they may stand for authors health condition. We use the term frequency-inverse document frequency (TF-IDF) [18] features to represent tweet data with the help of the tokenization package[3] from the CMU and the scikit-learn library. The TF-IDF numerically represents all terms, which counts word appearances offset by the frequency of words in the corpus.

Our approach employs SVMs with the linear kernel to solve the associated high-dimensional feature space problem, which has been shown to perform well under such circumstances [13]. To overcome the class imbalance problem, where the ILI-related tweets are much fewer than the non-ILI-related tweets, the experiments are designed to optimize the area under the PR curve, which is demonstrated to be more meaningful when dealing with such unbalanced scenarios [10] compared to ROC curve.

3.3 Analysis

Before modeling, we aim to understand to what extent Twitter data can capture the key features of state-level influenza prevalence both on spatial and temporal dimensions. With this objective, we design some experiments with the true influenza notifications data, which is obtained from Influenza Specialist Group (ISG) [1] and Queensland government health department websites [3], as a benchmark.

[3]https://github.com/brendano/ark-tweet-nl

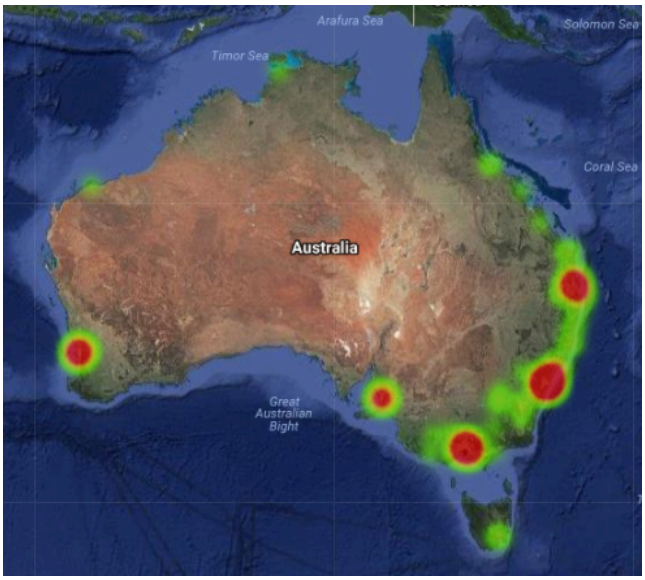

Figure 3: A heatmap of the ILI-related tweets in Australia. Most ILI-related tweets are located at the coastal areas and around the capital in each state.

Figure 4: Hospital and health service (HHS) regions in Queensland (QLD) [3]

3.3.1 *Spatial Analysis*

In the spatial analysis, we first assign all ILI-related tweets to their respective locations with respect to the "geo" and "place" fields obtained from JSON-format tweets using *geopy* Python library[4]. A heat map generated by all ILI-related tweets in Australia is shown in Figure 3. It is evident that most of the sick users are located in those areas along the east coast with high population density. Meanwhile, the number of those target users located in capital of states, such as Perth and Adelaide, is much more than those in other areas. In the state-level analysis, we sort the "sick" tweet numbers and the number of flu notifications in each state according to the population and calculate the associated Pearson correlation coefficient to evaluate the linear relationship between the two examined values, "sick" tweet numbers and true notification numbers.

Meanwhile, we also perform a regional level analysis. We choose the Twitter data and true notifications data from the state of Queensland (QLD) and locate each tweet within its corresponding hospital and health service regions (HHS), as shown in Figure 4. Similar to the state-level case, we are interested in discovering the correlation between the tweet data and the true flu notification data by sorting them with regards to population and calculate the Pearson correlation coefficient.

3.3.2 *Temporal Analysis*

Temporal analysis is conducted by comparing the number of ILI-relate tweets and true notifications in a monthly level.

A bout of flu typically lasts one to two weeks, and flu symptoms usually start within one to four days after infection [19]. In order to identify the infected individuals precisely, multiple sick tweets posted by the same user within one week are seen as duplicate tweets and only counted once in the analysis.

Internet access and social media usage rate are different among the states and territories. For example, residents of Australian Capital Territory and Victoria are more likely to have access to the Internet compared to those living in Northern Territory or Queensland. In order to reduce the potential bias induced by these disparities, we modify our "sick" tweet numbers by weighting them according to the Internet access rate as well as Twitter penetration rate at different states and territories. We obtain the usage rate information from Australian Sensis Social Media Report [4].

3.4 Modeling Influenza-Like Illness Prevalence

In order to establish a state-level model, a linear regression model is fitted with the number of annual ILI-related tweets as the independent variable and the true illness laboratory notifications as the dependent variable. The number of influenza notifications in each state is estimated by:

$$\hat{y} = B_0 + B_1 x$$

where B_0 is the intercept, B_1 is the regression slope coefficient, x is the number of ILI-related tweets, and $\hat{y}$ is the estimated number of influenza patients.

Internet access and Twitter penetration rate parameters are then introduced to eliminate the bias that caused by different Internet and social media usage rate in each state. Accordingly, the independent variable x is calculated by:

[4]https://pypi.python.org/pypi/geopy/1.11.0

Table 2: Classifier Performance

	accuracy	precision	recall	f-score	fp-rate	fn-rate	PR area
Cs	98.3%	82.2%	77.3%	79.6%	0.2%	27.9%	82%
Co	96.2%	78.3%	88.5%	83.0%	2.9%	11.4%	84%
Cf	91.5%	74.3%	95.2%	83.5%	9.6%	4.8%	81%

$$x = \frac{N}{i*t}$$

where i is the Internet access rate and t is the Twitter penetration rate.

To better evaluate the regression model, the Pearson correlation coefficient analysis and t-test are carried out. The t-test conducts a hypothesis test to determine whether there is a linear relationship between the independent variable and the dependent variable. In the t-test, the null hypothesis is that the slope is equivalent to zero (H_0), and the alternative hypothesis states that the slope is not equal to zero (H_1):

$$H_0 : B_1 = 0$$
$$H_1 : B_1 \neq 0$$

The associated p-value tests the null hypothesis. If the generated p-value is lower than a given significance level (normally 0.05), the null hypothesis can be rejected with high confidence.

We also carry out a confidence interval analysis, which can help identify the probable area where the best-fit regression line lies.

4. PERFORMANCE EVALUATION

In this section, experimental results for each stage of our work are displayed and elaborated along with analysis and discussions.

4.1 Classification Results

In the training stage, we fix the parameters of the classifiers after five training iterations with 1,585,918 tweets as the classifiers do not perform better with more training iterations. The average of 10-fold cross-validation performance of the SVM classifiers ,*Cs* and *Co* as well as *Cf*, are presented in detail in Table 2.

In our work, the number of the ILI-related tweets is expected to be limited. Therefore, a 74% precision for classifier *Cf* can induce a large error in the dataset. From Table 2 we can observe that the accuracy is high for all three classifiers because of the existence of a large amount of non-health-related tweet (true negative). However, in our experiments the accuracy and precision of *Cf* decline while the recall improves. A relatively large false positive rate shows that Cf has mistakenly labeled many non-health-related tweets as "sick" tweets. In order to obtain a more precise ILI-related tweet dataset, we employ both classifiers *Cs* and *Co* for tweet labeling and manually check the correctness of labels of tweets that are given different labels by the two classifiers.

After labeling and manual checking, 1167 tweets posted by 896 unique users are found to be ILI-related. We then remove the duplicate tweets posted within a week by the same user. This leaves us with 1027 ILI-related tweets from Australia.

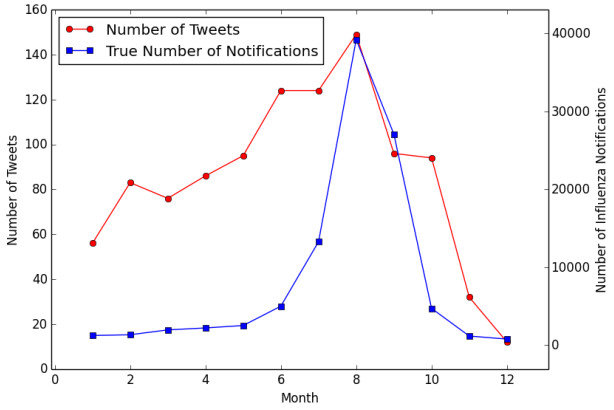

Figure 5: Monthly temporal analysis

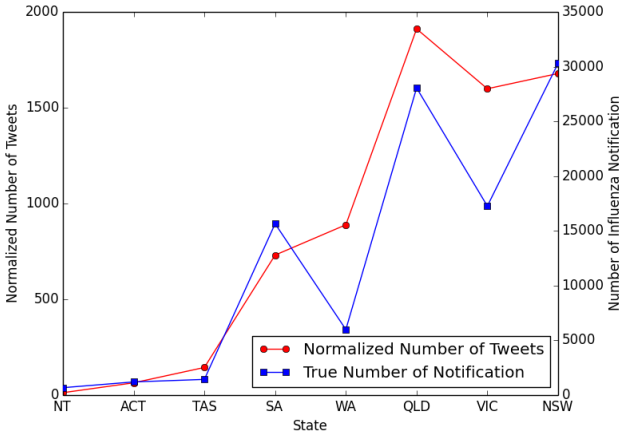

Figure 6: State-level spatial analysis

Compared to the size of entire 2015 tweet database, the number of sick tweet authors is relatively small. Assuming that the data obtained from ISG can cover all individuals in Australia, considering 100,586 laboratory-confirmed influenza cases in 2015 with the Australian population of 24 million, the ratio of influenza infected population within a year is around 0.0042. If we apply this ratio to 225,641, the number of unique users in entire tweet database, the result is around 944, which is close to the detected number of sick users.

4.2 Temporal and Spatial Analysis

4.2.1 Influenza Outbreak

From temporal analysis, Figure 5 shows that both ILI-related tweet data and true influenza notification data reach the peak in August, which is during high flu season in Australia. This indicates that Twitter data can potentially help detect an influenza outbreak in the time series. However, despite a rapid increase in the number of flu notifications

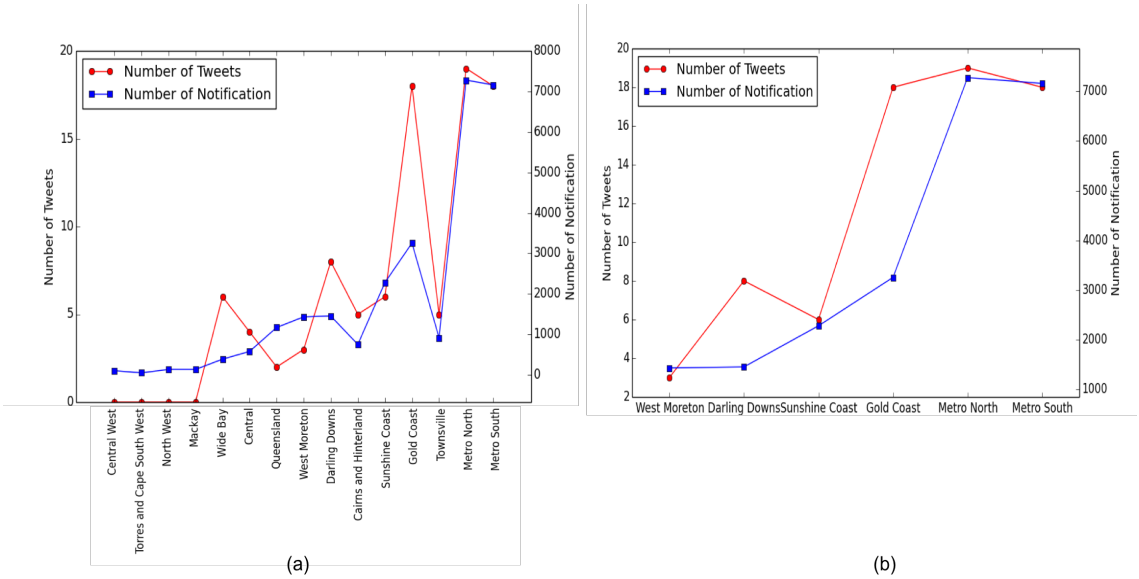

Figure 7: Regional spatial analysis

from June to August, the sick tweet number increases moderately within the same period. It is evident that there are around 40,000 notifications in August and less than 5,000 notifications in May. However, the Twitter data shows 150 ILI-related tweets in August and around 100 in May. Considering true notifications as a benchmark, we would expect the number of tweets in August to be around 8 times of that in May rather than only 0.5 times more. The discrepancies between ILI-related tweets data and true influenza notification data may result from the limited prevalence of mentioning health conditions and this result also shows that it is hard to reveal the severity of the influenza spread in a temporal dimension.

4.2.2 *State-Level Linear Correlation*

We sort the "sick" tweets and true notifications according to the populations of the states and normalize the tweet data with Internet access and Twitter penetration rates as shown in Figure 6. Twitter data appears to have similar variation trends to true notification data. For instance, although there is a high population density, Internet access rate, and Twitter penetration rate in Victoria compared to Queensland, the Twitter data correctly identifies more influenza infections in Queensland. Statistically, there is also a high correlation coefficient between Twitter data and true notification data, which is around 0.94. This indicates that the Twitter data can capture the key features of state-level influenza prevalence on an annual level with a linear relationship.

4.2.3 *Regional Analysis*

At the regional level, we allocate tweets to each encapsulated hospital and health service (HHS) region in Queensland and sort the number of ILI-related tweet and true notification data in each HHS area by population, as shown in Figure 7 (a). As the region names from left to right are in ascending population order, we can see that there are no ILI-related tweets posts in Central West, Torres and Cape South West, and North West. This stands for the population size in those areas being quite small, and the number of Twitter users who constantly tweet is also less. However, there is a relatively large number of ILI-related tweets in Wide Bay and Darling Downs given a small number of true influenza notifications. After further analysis, we find that, as there is a limited number of ILI-related tweets in those regions, Twitter data can be easily influenced by some unwell Twitter users that post frequently.

Interestingly, in regions with higher populations such as Cairns, Sunshine Coast, Gold Coast, Townsville, and Brisbane Metro, Twitter data shows some similar variation trends to the influenza notifications. Based on these observations, we limit our study to the regions around Brisbane city, as shown in Figure 7 (b). The number of ILI-related tweets and true influenza notifications shows a reasonable linear relationship with a correlation coefficient of 0.835. However, Twitter data in Gold Coast seem to overestimate the influenza cases. This may be because Gold Coast is a famous tourist destination and has more younger people which enhances the Twitter usage.

These analysis shows that, we may need to take the nature of cities into account regarding the Twitter usage behavior when studying regional disease distribution. However, owing to the limited Twitter usage and low likelihood of mentioning health conditions in tweets, the number of detected ILI-related tweets may not be sufficient to support regional analysis in Australia.

4.3 Influenza Distribution Modeling

4.3.1 *Regression Analysis*

Finally, we fit a linear regression model to estimate influenza prevalence using the generated Twitter dataset. As shown in Figure 8, the linear regression model is generated with the slope of 83.88 with a Pearson correlation coefficient of 0.875 and p-value of 0.011.

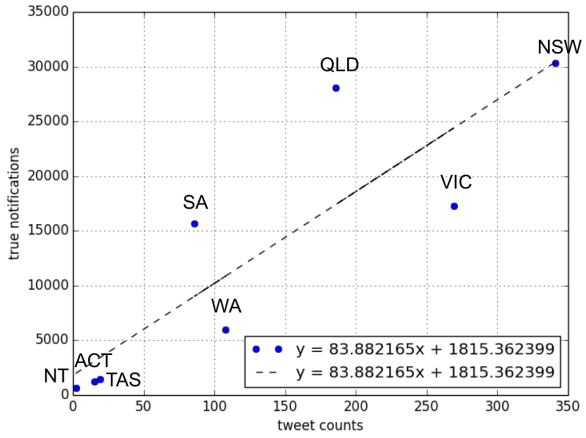

Figure 8: Linear regression with original sick tweet data amount

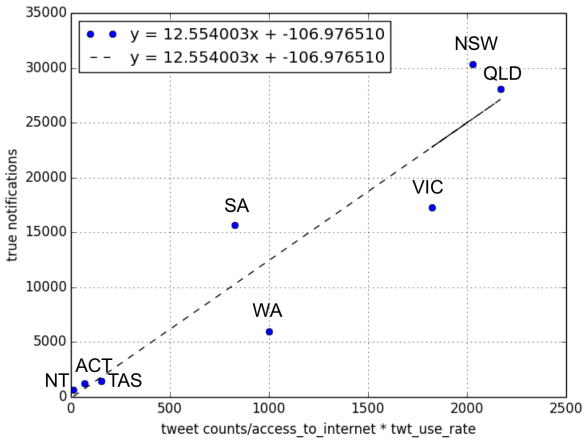

Figure 9: Linear regression after taking into account Internet access rate and Twitter penetration rate in each state

After taking Internet access and Twitter penetration rates into consideration as weighting parameters, a better-fitted model has been generated with a slope of 12.55. A higher correlation coefficient of 0.93 and p-value of 0.017 suggest a state-level linear relationship between the number of ILI-related tweets and true influenza notifications, as seen in Figure 9, which shows the promise of estimating influenza prevalence using Twitter data.

In Figure 10, the confidence intervals generated by sample data points indicate the area where there is a 95% probability that the true best-fit line for the regression lies. The prediction interval indicates that for any specific value of the number of ILI-related tweets (X), weighted by Internet access and Twitter penetration rates, there is a 95% probability that the real value of Y (a number of true influenza notifications) is within this interval where slope varies from 6.02 to 22.19. The positive slope interval indicates a strong positive linear correlation between the two variables.

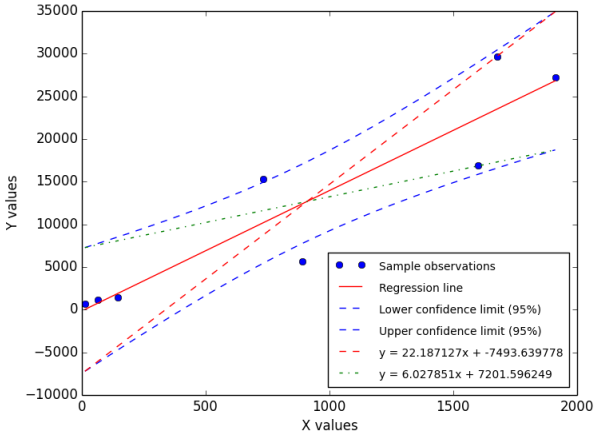

Figure 10: Confidence interval and prediction interval area

4.3.2 *Influence of Population, Internet Access, and Twitter Penetration Rates*

The improvement between linear regression models depicted in Figures 8 and 9 shows that Internet access and Twitter penetration rates are important factors during modeling. During the experiments, we also discover that the number of ILI-related tweets has a strong linear correlation with the population of each state. Although the number of tweets is limited, the Pearson correlation coefficient is 0.99. We then present the data points of ratios between the number of ILI-related tweets and population in each state in Figure 11. Excluding the data point representing Northern Territory (NT) as an outlier, we find out that although each state differs regarding the Twitter user behavior and the population size, there are similar ratios between tweets data and the population. The average ratio of those other seven states is around $4.2{*}10^{-5}$ times, which means when we know the population in a state, the number of "sick" Twitter users is around $4.2 * 10^{-5}$ times of the population.

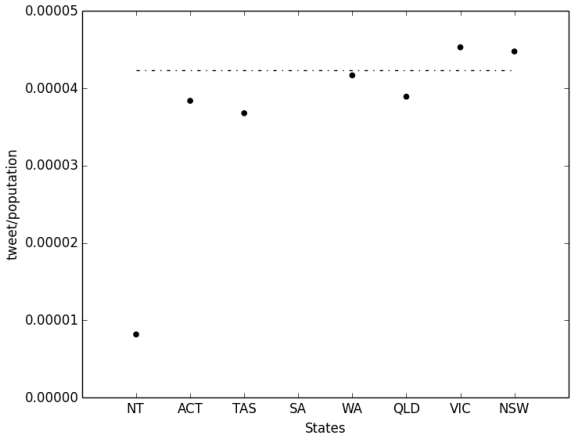

Figure 11: Ratio between number of ILI-related tweets and population in each state

5. LIMITATIONS

This work is mainly limited by the scarcity of the tweets, especially illness-related ones, which may have three main causes. First, according to Sensis social media report 2015 [4], only 17% of Australians are using Twitter, which ranked as the 5th most use social media platform in Australia. Meanwhile, the likelihood of users commenting on health condition in social media is relatively low. Second, user's online behavior may change during an adverse health condition. For example, some users may not want to tweet when they are suffering from illness while others might. People may be more interested in talking about politics, sports, and everyday life, etc. via Twitter. Third, the considered tweet database only contains geo-tagged tweets, which is a small portion of all tweets in Australia.

The laboratory confirmed influenza notifications are also incomplete as many patients may not seek medical treatment when they catch a cold. Meanwhile, the linear regression model is relatively simple in state-level influenza modeling. However, based on the scale of our work, where there are only two variables - Twitter data and true notification data, linear regression is a suitable model in this study.

Meanwhile, our work assumes a similar likelihood and frequency of tweeting by people of different ages and socio-economic backgrounds. However, Twitter is currently more popular among younger generations, which means the presented results and models are younger generation specified.

With respect to our approach to detecting the ILI-related tweets, manual checking steps may restrict the scalability of our learning method when applied to larger datasets.

6. CONCLUSIONS AND FUTURE WORK

Our work proposes effective modifications to the state-of-art approach in detecting illness-related tweets with the purpose of reducing the errors of its classifiers. Along with iterative manual checking for validation, we introduce Internet access and Twitter penetration rates in our modeling to compensate for their discrepancies among the states. We conduct the state-level and the regional-level analysis and show that although the number of tweets is limited, Twitter data is useful in spatial and temporal disease prevalence modeling.

Our analysis results show that Twitter data is a reasonable proxy for detecting disease outbreak and possesses strong linear correlation with real-world influenza notification data. Finally, a linear regression model is established with a correlation coefficient of 0.93 and a p-value of 0.017. A strong positive linear regression model strongly suggests that Twitter data can capture the key features of state-level influenza prevalence and has a good potential in disease spread modeling.

In future work, we will consider introducing other data sources such as public transportation data, Twitter follower relationships, and tweet geo-location changes as features to model influenza prevalence and spread. At the same time, we will attempt to identify the effects of user connections and human movement on disease spread using data from Twitter and other social media. Meanwhile, we will also focus on temporal modeling to identify data correlations during various time spans such as different months and seasons.

7. REFERENCES

[1] Influenza specialist group. `http://www.isg.org.au`.

[2] Mongodb. `https://www.mongodb.com`.

[3] Queensland health. `https://www.health.qld.gov.au/clinical-practice/guidelines-procedures/diseases-infection/surveillance/reports/flu/default.asp`.

[4] Sensis social media report. `https://www.sensis.com.au/about/our-reports/sensis-social-media-report`.

[5] Australian influenza surveillance report and activity updates, 2015.

[6] Social media statistics australia, 2015.

[7] By the numbers: 170+ amazing twitter statistics, 2016.

[8] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297.

[9] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA, 2010. ACM.

[10] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[11] S. Kumar, F. Morstatter, and H. Liu. *Twitter Data Analytics*. Springer, New York, NY, USA, 2013.

[12] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. Natural language processing: an introduction. *J Am Med Inform Assoc*, 18(5):544–551, Jul 2011. 21846786[pmid].

[13] M. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health, 2011.

[14] Á. M. C. F. O. J. Prieto VM, Matos S. Twitter: A good place to detect health conditions. *PLoS ONE 9(1)*, page e86191, 2014.

[15] J. L. M. A. M. C. D. N. Raja Jurdak, Kun Zhao. Understanding human mobility from twitter. *PLoS ONE 10(7)*, page e0131469, 2015.

[16] A. Sadilek, H. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *In Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[18] G. Salton, E. A. Fox, and H. Wu. Extended boolean information retrieval. *Commun. ACM*, 26(11):1022–1036, Nov. 1983.

[19] I. Strauch. How long does the flu last?, 2015.

[20] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment, 2010.

[21] G. Zuccon, S. Khanna, A. Nguyen, J. Boyle, M. Hamlet, and M. Cameron. Automatic detection of tweets reporting cases of influenza like illnesses in australia. *Health Information Science and Systems*, 3(1):S4, 2015.