Measuring and Forecasting Influenza Outbreaks Using Twitter

Tianyu Chen[1], Yuhan Zeng[2], Rui Zhang[1]

[1]Department of Computer Science, Indiana University
[2]Department of Chemistry, Indiana University
chen512@indiana.edu, {rz20, yuhzeng}@iu.edu

March 3, 2018

1 Objectives and Significance

Infectious diseases are one of the major causes of morbidity and mortality, among which influenza is a ubiquitous epidemic [1]. Commonly known as "the flu", it is caused by influenza virus and can be a highly infectious disease whose symptoms include runny nose, (high) fever, sore throat, headache, coughing and fatigue[2]. Although there are numerous studies about predicting and forecasting the trend of influenza outbreaks, there are some limitations still, specifically of their source of data. Related studies commonly leverage influenza-like illness (ILI) and severe acute respiratory infections (SARI) data, which is defined by WHO[3]. Take the *Influenza Observation and Forecast* system that is part of Columbia Prediction of Infectious Disease (CPID) maintained by Columbia University [4] for example, it only tracks the data from a total of 38 out of 195 sovereignties in the world today. Furthermore, its frequency of prediction can be limited by the publication of the corresponding data source.

To mitigate the problems above, we propose an epidemic outbreak prediction technique based on analyzing the data gathered from popular social media like Twitter. Compared with the traditional approach, a noteworthy advantage of social media information is its ubiquitousness and timeliness - end users are producing such data wherever and whenever they are. By performing a keyword-based text mining based on an anonymized tweet dataset and utilizing the geographic information as a supplement, we can forecast whether a influenza outbreak is expected in near future.

2 Background

2.1 Predicting the Flu

Influenza, often called "the flu", is an acute respiratory illness that infects approximately 10 - 15% of the people around the world each year. It can be deadly. For example in 2009, pandemic H1N1 (*"swine flu"*) is estimated to result in 12,469 deaths in the United States [5]. The early detection of influenza is of paramount importance to timely contain the spread of the illness. In the United States, Centers for Disease Control and Prevention (CDC) report an influenza-like illness (ILI) index weekly based on the data collected from medical practices in a surveillance network. However, this report process is almost entirely manual and there is a typical 1 - 2 week delay due to the lag in clinical data acquisition [6]. Given the increasing demand for epidemic monitoring and forecasting in real time, web and social media (such as Google [7] and Twitter [8]) emerge as a new data source [9].

2.2 Twitter Text Mining

Twitter, an online micro-blogging (limited to 280 characters) and social network service, has an average of 330 million active users worldwide [10] each month. It could be a promising data source for the early detection of ILI due to its timeliness, public availability, and popularity [11]. Paul et al. [12] shows that influenza-related tweets were positively correlated (r: 0.37 - 0.81) with existing government surveillance data in 10 English-speaking countries.

2.2.1 APIs

Table 1: A list of Twitter APIs

Parameter	**Required**	**Function**
q	Yes	The query string, 500 chars max
geocode	No	Specifying the latitude/longitude as well as the radius
lang	No	ISO 639-1 language code, restricting the results
locale	No	Specifying the language of the query string
result_type	No	Has 3 values: *recent* (only recent results), *popular* (only popular ones) and *mixed* (all tweets)
count	No	The number of tweets returned
until	No	Tweets created before a specific date

The Twitter API is an interface that takes a query as input and returns a JSON containing relative tweets and metadata as output. In Table 1 some parameters that we are interested in are listed with their functionalities.

2.2.2 Related Research

A variety of data mining and machine learning techniques have been applied to Twitter data analysis and influenza trend prediction. Santillana et al. [13] combine multiple ILI activity estimates into a single prediction of ILI using machine learning ensemble approaches, which produces accurate weekly ILI predictions for up to four weeks ahead of CDC's ILI reports. Santos et al. [14] show that by changing data pre-processing and feature extraction and selection the Naive-Bayes classifier and regression model built to predict the flu can be adapted to languages other than English. Aron Culotta [15] proposes a document classifier to filter the *misleading* messages, which significantly reduce error rate in simulated false alarm experiments. Li et al. [16] build up a real-time ILI reporting system based on Flu Markov Network (Flu-MN), an unsupervised Bayesian algorithm based on a 4 phase Markov Network. Chen et al. [17] propose both an unsupervised model and an improved supervised weakly model that captures the hidden states of a user from her tweets and estimates the flu trend. Aramaki et al. [18] describe a support vector machine (SVM) based classifier that shows high correlation (0.97) at the outbreak and early spread. Achrekar et al. [6] show that text mining significantly enhances the correlation between the Twitter and CDC reported ILI rates.

In particular, Lee et al. [19] reports a real-time flu surveillance system leveraging data mining on tweets. However, their approach only accounts for a few most *frequent* words that co-occurs with a disease name, which seems naïve in that the model overlooks *infrequent* keywords related to flu outbreak. They update their system in 2017 [20] using multilayer perceptron. Compared to their previous work, it shows an improved accuracy. However, only promoting tweets (ads) and the repeated ones are filtered.

3 Proposed Approach

3.1 Data Collection

Social media like Twitter and Facebook plays an important role in our daily life today. Organizations such as Centers for Disease Control and Prevention (CDC) collect weekly influenza test results from public health laboratories and make infection prediction with an inevitable delay of up to two weeks. For a more timely and efficient prediction of influenza,

we collect a vast number of tweets that fall into a particular time span and extract related data from them. Owing to the popularity and timeliness of Twitter posts, they can be used to assess health status or predict the spread and fluctuation of an epidemic in a certain population in real-time.

To retrieve the dataset, we plan to use the Twitter APIs (introduced in 2.2.1) to collect text tweets during a certain time span and in a particular geographic space. A few keywords such as "flu", "influenza" and "H1N1" will be used to restrict the content of tweets to be collected. However, since not all tweets collected with the keywords can be considered as an indication of real-time illness, the raw data needs to be filtered to further remove irrelevant entries. We plan to train a bag-of-words (BOW) classifier using regression to predict whether a Twitter post containing certain keywords of interest indicates that the Twitter user shows signs of influenza infection. In addition, redundant information such as re-tweets, duplicate tweets (in terms of content) and tweets by a same user within a time threshold will be cleaned from the dataset.

3.2 Methodology

3.2.1 Keyword Extraction

To provide a better insight of a Twitter-based influenza trend forecast, we propose to extract the keywords by performing text mining on the dataset collected using the methodology described in 3.1. We propose a random forest (RF) based word sampling method to calculate the significance of a keyword.

Random Forest A random forest is an ensemble of decision trees grown from the bootstrap dataset - the bootstrap means independently and uniformly sampling a subset of data with replacement. In each decision tree, the attribute is the presence of selected word (either 0 or 1) and the output is the influenza trend (i.e. %-weighted ILI).

The Gini index is used to split a node:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

The Gini score computed over all decision trees will then be used to rank the word.

Word Sampling In this step, the Wilcoxon rank-sum test [21] will be used to compare the Gini score of a word with that of a random generated "noise word" to determine the significance of the word of interest. The null hypothesis (H_0) is the mean Gini score of

a given word is the same as that of the random "noise word". As for the "noise word", a number of words will be picked randomly from a dictionary and selected only when it has no *prediction power*. A significance threshold θ will be chosen to remove all the trivial words (i.e. the words whose p-value is greater than θ). Additionally, we plan to evaluate the correlation between the words and the influenza trend using χ^2-test, which could assist keyword selection in case the sampling method described above fails.

3.2.2 Predictive Model

The proposed predictive model is based on the one descibed in K. Lee et al., 2017 [22]. As a supplement, we plan to combine the ILI-related physician visit percentage with keyword frequencies appeared in different tweets as in the original paper. The influenza statistic after a week is produced as output. A major improvement over the original approach is that by introducing a filter before the layered perceptron, those apparently irrelevant tweets can be filtered in advance using heuristics. This can improve the efficiency of the perceptron especially when the number of the tweets on which our forecast is based is relatively large.

3.3 Evaluation

We propose 3 evaluation metrics to measure the performance of our method. These measures will be employed in text mining and also in the predictive model part if applicable.

Pearson Correlation Pearson Correlation (r) is a measure of the linear correlation between two variables X and Y [23].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}((y_i - \bar{y})^2}}$$

Root Mean Squared Deviation Root Mean Squared Deviation (RMSD) is a measure of the difference between predicted and real values [24].

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y} - y_i)^2}{n}}$$

In keyword extraction, it can be benefitial to observe the RMSD of keywords over a certain period of time t.

Precision Precision is a fraction where the number of the documents relevant to the user's information is divided by the total number of documents. By definition we have:

$$PPV = \frac{TP}{TP + FP}$$

3.4 Expected Outcome

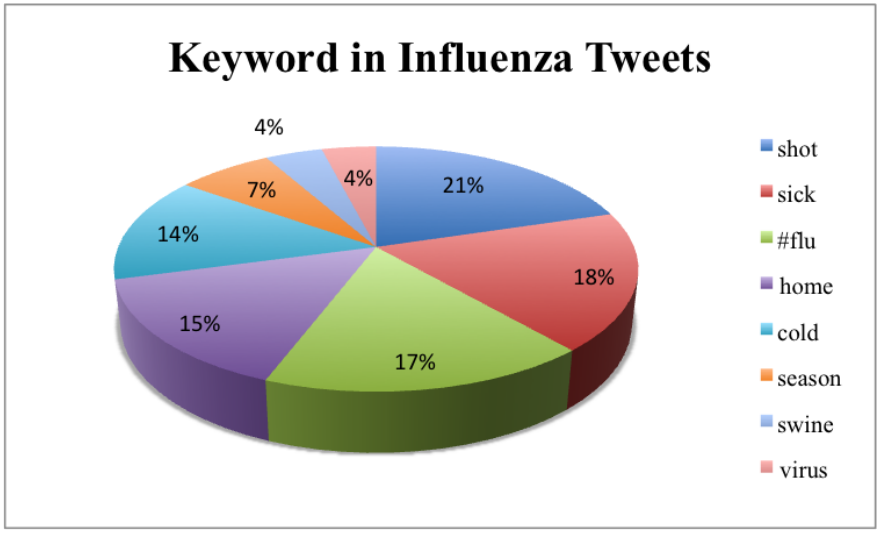

Figure 1: The example of expected outcome for keywords extraction

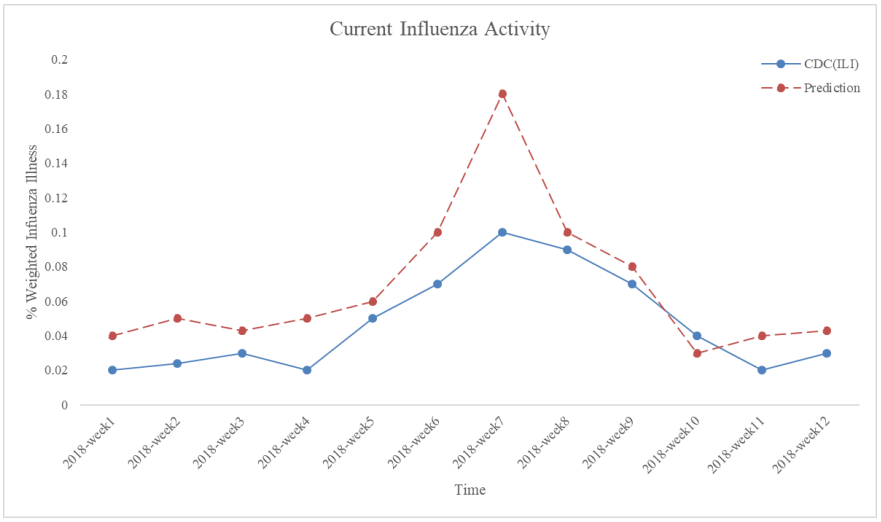

Figure 2: The example of expected outcome for Predictive Model

We expect to extract a list of keywords from the tweets with a high correlation with the actual flu trend. In terms of the visualization, we will plot the frequency of keywords (similar to the one shown in Figure 1). As a backup option, we may perform text mining using Twitter Latent Dirichlet Allocation (LDA) model for trend prediction, where there are

many open-source libraries available for implementation. However the issue with the LDA model is that the words generated may not signal an actual flu outbreak.

We expect to predict the flu trend 1 - 2 weeks ahead of the release of CDC reports. To visualize the prediction, we will plot our predictive model in the current flu activity (i.e., percentage of ILI-related physicians visit) in weekly time-series (similar to the one shown in Figure 2. As a fallback option, we could use support vector machine (SVM) classifier as reported by Aramaki et. al [18]. However, SVM tends to overfit the flu trend and may need a longer training time due to the size of the Twitter dataset.

4 Individual Tasks

Member	Task Summary
Rui Zhang	Develop and perform experiments
Yuhan Zeng	Develop and perform experiments
Tianyu Chen	Process and analyze data, doing documentation

Table 2: Task allocations

Rui Zhang Rui Zhang is responsible for system development as well as performing experiments to collect data and analyze the result.

Yuhan Zeng Yuhan Zeng, leveraging her data science skills, is responsible for data processing and analysis. She is also responsible for the storage and access of the Twitter dataset. She will also ensure the correctness of the implementation by testing.

Tianyu Chen Tianyu Chen will lead the code development. He is also responsible for testing and assisting other team members.

Depending on the actual project progress, additional task or work may be assigned to each team member.

5 Expected Outcome

We expect that our predictive model can generate a similar influenza outbreak trend with the CDC report, measured by the 3 evaluation approaches mentioned in 3.3.

References

[1] Kyle S Hickmann, Geoffrey Fairchild, Reid Priedhorsky, Nicholas Generous, James M Hyman, Alina Deshpande, and Sara Y Del Valle. Forecasting the 2013–2014 influenza season using wikipedia. *PLoS computational biology*, 11(5):e1004239, 2015.

[2] Influenza. Influenza — Wikipedia, the free encyclopedia, 2018. [Online; accessed 1-March-2018].

[3] World Health Organization et al. Global epidemiological surveillance standards for influenza. *World Health Organization*, 2014.

[4] Columbia prediction of infectious diseases. Online; accessed 1-March-2018.

[5] Dc dc estimates of 2009 h1n1 influenza cases, hospitalizations and deaths in the united states. Online; accessed 1-March-2018.

[6] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Twitter improves seasonal influenza prediction. In *Healthinf*, pages 61–70, 2012.

[7] Gary King David Lazer, Ryan Kennedy and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203, 2014.

[8] Avinash Gandhe Harshavardhan Achrekar and Ross Lazarus. Predicting flu trends using twitter data. *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. IEE*, 2011.

[9] Mikler AR Corley CD, Cook DJ and Singh KP. Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public healt*, 7(2):596, 2010.

[10] Number of monthly active twitter users worldwide from 1st quarter 2010 to 4th quarter 2017 (in millions). Online; accessed 1-March-2018.

[11] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. *Proceedings of the first workshop on social media analytics*, page 115, 2010.

[12] David A. Broniatowski Michael J. Paul, Mark Dredze and Nicholas Generous. Worldwide influenza surveillance through twitter. *In AAAI Workshop: WWW and Public Health Intelligence.*, 2015.

[13] André T. Nguyen Mark Dredze Michael J. Paul Elaine O. Nsoesie Santillana, Mauricio and John S. Brownstein. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10):e1004513, 2015.

[14] Jose Carlos Santos and Sergio Matos. Analysing twitter and web queries for flu trend prediction. *Theoretical Biology and Medical Modelling*, 11(1):S6, 2014.

[15] Aron Culotta. Detecting influenza outbreaks by analyzing twitter messages. *arXiv preprint arXiv*, page 1007.4748, 2010.

[16] Jiwei Liand Claire Cardi. Early stage influenza detection from twitter. *arXiv preprint arXiv*, page 1309.7340, 2013.

[17] KSM Tozammel Hossain Patrick Butler Naren Ramakrishnan Chen, Liangzhe and B. Aditya Prakash. Syndromic surveillance of flu on twitter using weakly supervised temporal topic models. *Data Mining and Knowledge Discovery*, 30(3):681, 2016.

[18] Sachiko Maskawa Aramaki, Eiji and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using twitter. *Proceedings of the conference on empirical methods in natural language processing*, page 1568, 2016.

[19] Ankit Agrawal Lee, Kathy and Alok Choudhary. Real-time disease surveillance using twitter data: demonstration on flu and cancer. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, page 1474, 2013.

[20] Ankit Agrawal Lee, Kathy and Alok Choudhary. Forecasting influenza levels using real-time social media streams. *Healthcare Informatics (ICHI), 2017 IEEE International Conference on. IEEE*, page 409, 2017.

[21] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80, 1945.

[22] Kathy Lee, Ankit Agrawal, and Alok Choudhary. Forecasting influenza levels using real-time social media streams. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*, pages 409–414. IEEE, 2017.

[23] Pearson correlation coefficient — Wikipedia, the free encyclopedia. [Online; accessed 1-March-2018].

[24] Root-mean-square deviation — Wikipedia, the free encyclopedia. [Online; accessed 1-March-2018].