

Measuring and Forecasting Influenza Outbreaks Using Twitter

Technical Report

Tianyu Chen¹, Yuhan Zeng², Rui Zhang¹

¹Department of Computer Science, Indiana University

²Department of Chemistry, Indiana University

chen512@indiana.edu, {rz20, yuhzeng}@iu.edu

April 28, 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 The Twitter API

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 Online Twitter Archives

There are multiple online pre-grabbed archives of Twitter streams for general research purposes. In our experiment, we use the pre-grabbed online streams from Archive Team, which are packed into tarballs and ready to download ¹.

¹<https://archive.org/details/twitterstream>

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: <i>recent</i> (only recent results), <i>popular</i> (only popular ones) and <i>mixed</i> (all tweets)
count	No	The number of tweets returned
until	No	Tweets created before a specific date

The dataset is in JSON format and the fields are as described in section 2.2.1. The data is annotated with date in the directory hierarchy. However, the tweets contained in the dataset are worldwide without passing through any filter. We perform data cleansing to ensure that the preprocessed dataset only contains the tweets that we are interested in and the number of the tweets are large enough for a meaningful result and relatively small enough that it is still within our processing ability.

The data cleansing and preprocessing techniques are explained in detail in section 3.1.

2.2.3 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 Methodology

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.

3.1.1 Data Collection and Preprocessing

As is mentioned in 2.2.2, we collect Twitter streams from the Archive Team online archive. Then we filter the tweets through a preprocessing script that sorts out the tweets whose `location` field contains either the full name of one the states in the United States or the abbreviation.

We store the processed tweets in pickle² binary object serialization format for further processing. One tweet object is a dictionary that contains 4 fields - week, date, location and text. Week and date represent the time at which the tweet is generated. We care about the week number since the CDC reports are organized in weeks. Location are extracted from the `location` field of the original tweet. Text contains the plain text content of the tweet, on which further natural language processing is performed to determine whether it is flu related.

²<https://docs.python.org/3/library/pickle.html>

Listing 1: A sample preprocessed tweet

```
{ 'date ': '2017-10-01',
  'location ': 'LA (Lower Alabama)',
  'text ': "RT @ClayTravis: Wow, Budweiser's getting so many calls
            demanding they drop NFL sponsorship that their official
            phone bank addresses it. Cal...",
  'week ': 39}
```

We collect and process the tweets from October, 2017 to December, 2017 (12 weeks). There are over 200,000 tweets every day that are identified to be located within the U.S.

3.1.2 Manual Labelling

We manually label a random portion of the dataset. The label represents whether the corresponding tweet is related to influenza (0 or 1) and serves as a groundtruth.

Since the portion of flu correlated tweets is small and it is time consuming to look at all of them, we first narrow down the scope using a set of heuristics. The heuristics are a series of keywords that fall into 3 categories - flu names, symptoms and treatments.

Table 2: Heuristics keywords

Category	Keywords
Names	flu influenza h5n1 h1n1
Symptoms	cough fever headache sore throat sorethroat sneez vomit runny nose runnynose fatigue
Treatments	tamiflu tylenol theraflu advil mucinex benadry corenza

The manually labelled data is used for keywords extraction.

3.2 Implementation

3.2.1 Keyword Extraction

To provide a better insight of a Twitter based trend predictor and to reduce the data dimensionality, we text mine the keywords from the datasets collected using a methodology described in 3.1. We used χ^2 -test, which assesses the independence of each word against the relevance label. The rationale behind the test is that, suppose a word occurs in a probability $P(W)$ and a tweet is flu relevant in a probability of $P(R)$, if a word is flu irrelevant, then we

expect the independence of joint probability (i.e the probability of this word occurring in a informative tweet) to be $P(W, R) = P(W) \times P(R)$. The χ^2 parameter is calculated by:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where c is the degree of freedom, O_i is the observation and E_i is the expected value. An advantage of χ^2 -test is its fast performance compared to the Wilcoxon-rank test that appears in our proposal, which requires exhaustively building random forests to estimate the significance of each word. By setting up a $p - value$ threshold, we can skip the words that are not strongly associated with the relevance labels.

3.2.2 Predictive Model

We employ a neural network (NN) that involves a multilayer regressor to predict the flu trend. We combine the ILI-related physician visit percentage with the term frequencies of the keywords in each tweet as input and produce the influenza statistics in the following week as output. It is worth mentioning that our approach introduces a keywords list to filter the tweets so that the dimensionality of dataset can be significantly reduced. This is important to improve the efficiency of the predictor especially when we have such a big pool of tweets. The training and testing samples are prepared by the preprocessing scripts. The `sklearn.neural_network.MLPRegressor` from machine learning library `scikit-learn` 0.19.1 is used to build the neural network.

3.3 Evaluation Strategy

Aside from a direct comparison of flu activities between prediction and real statistics data, we also measure the performance of the predictive model in 3 evaluation metrics including Pearson correlation, mean squared error (MSE), and mean absolute error (MAS).

Pearson Correlation Pearson Correlation (r) is a measure of the linear correlation between two variables X and Y [21].

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

Mean Squared Error Mean Squared Error (MSE) is a measure of the difference between predicted and real values [22].

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

Mean Absolute Error Mean Absolute Error (MAE) is also a measure of the difference between predicted and real values [23], which takes absolute value instead of square. It may be more desirable than MSE in certain circumstances such as easy interpretation and handling large errors [24].

$$MAE = \frac{\sum_{i=1}^n |\hat{y} - y_i|}{n}$$

4 Results

4.1 A Small Dataset

To develop a reliable method to extract the keywords for flu trend prediction, we first study the same datasets used by Lamb et al. [25]. The keywords automatically generated by our methods are evaluated manually, to ensure a good coverage of relevant information while keep a small feature size.

4.1.1 2012 Twitter Dataset

Lamb et al. [25] collect and manually label 4,760 tweets in the year of 2012 [25]. Due to Twitter guidelines, the authors only release the *ids* of the tweets without attaching the original text contents. Unfortunately, the Twitter APIs have a strict limitation of the tweets that one can obtain under a standard developer account.³ - only 20 tweets can be downloaded. Considering such a restriction, we use the pre-grabbed twitter dataset mentioned in 2.2.2 for further studies.

4.1.2 Keyword Extraction

A few downloaded tweets are shown in Table 3. It is worth mentioning that all tweets shown contain the words “flu”. However, only a portion of the tweets are relevant to influenza outbreaks. For example, “getting swine flu vaccine within 4 weeks” is no more than an advertisement, while “@mmWine matty, you have the bird flu! Depends, what cha got going on?!? & 1” implies that the party concerned is potentially infected and is thus relevant. The subtle difference between relevant and irrelevant content poses a challenge for a reliable prediction, especially using a bag-of-word model.

³Twitter developer platform <https://developer.twitter.com/>

Tweets	relevance
RT @EWJJr: Difference between bird flu & swine flu: For bird flu you get tweetment. For swine flu you get oinkment. /That's so bad it's good	0
Illinois getting swine flu vaccine within 4 weeks http://bit.ly/11chvS	0
RT @EWJJr: Difference between bird flu & swine flu: For bird flu you get tweetment. For swine flu you get oinkment. /That's so bad it's good	0
RT @WatchBirds Bird News: Missoula waterfowl tested for bird flu http://bit.ly/17V9or	0
i know it's not "swine flu"! no urge to to poop in the mud and roll around in it. NOT "bird flu"- no poopin' on windshields or statues.	0
Swine Flu - How worried are you? - Take our poll now and check out how others feel! http://bit.ly/A07Zq - Vote Today!	1
@killerbarbie3 nah u good... if u did i wouldnt reply to u lol, I cant afford to get the Bird Flu! lol	1
Even @BarackObama got the Twitter Bird Flu!!	0
@TiernanDouieb It's bird flu mate - it's mutated with twitter.	0
HA @ twitter catching the bird flu	0
Ohhhh... I got Twitter bird flu too... I match @teresakopec	0
@AMAZINGBOSSUP bird flu goin around on twitter	0
RT @jryanlaw: Ohhhh... I got Twitter bird flu too... I match @teresakopec //Did you sneeze on your hand or your elbow young lady???	0
bird flu!	0
RTxoxonosheenRT @laurawalker86 my twitter account has bird flu LOL ***lol I've recovered***—and mine is a pretty colour :)	0
@mmWine matty, you have the bird flu! Depends, what cha got going on?!?	1
The approach of winter and flu has you worried? Worry no more Click Me for Health Now http://bit.ly/Tl8iC	1
RT @wickedpoptart: ooooooh nooooo lolRT @knaught09: @wickedpoptart @bluegrasspundit A case of bird flu?	0
@Star_Qualities? AHHHHH you have bird flu.... Flap your wings so you can fly high	0
Realizing how stoked I really am to go to Norway on the 18th. Really stoked. do they still have bird flu over there?	1

Table 3: A Twitter content sampler (2012)

We generate a list of keywords by χ^2 -test based on various statistics thresholds (p -value), as is shown in Table 4. This keywords list does not even include “flu”, which is due to the fact that all tweets contain the keyword of “flu”. We conclude that by using a small p -value such as 0.01, it simple eliminates too many candidates. While using a larger p -value such as 1.0 we can generate a large number of candidates, some words that do not make sense such as “cha” and “u” are also included in the result, which can be a potential issue especially in large datasets when we attempt to use small subsets of the words in order to effectively decrease the dimensionality of the dataset. As a result, we decide to use p -value = 0.05 to balance the number and the relevance.

p-value	Keywords
0.01	[]
0.05	[worried]
0.1	[afford, approach, cha, check, click, depend, feel, health, poll, realize, reply, stoke, today, u, vote, winter, worried, worry]

Table 4: Keywords for the 2012 Twitter dataset

4.2 Flu Trend Prediction on the Small Dataset

After implementing the keyword extraction algorithm as a way of feature selection, we propose to employ neural network (NN) with regression to predict the flu trend. The groundtruth of flu activities in the United States is obtained from CDC report on the percentage weighted influenza-like illness (%-weighted ILI). We intend to predict the flu trend (whether the number of patients increases or decreases) as well as the %-weighted ILI.

4.2.1 2015-2016 Twitter Dataset

The two datasets are obtained from online resources in a blog post [26], which contain a random sample of 1,000 tweets collected between September 2015 to April 2016. Each tweet in those datasets were also labelled as either “relevant” or “irrelevant”, according to the fact whether the one who posts the tweet need a flu shot or not [26]. This datasets could be a good starting point to develop our framework, since all the tweets had been labeled and the overall datasets size is relatively small.

4.2.2 Keyword Extraction

We also extract a list of keywords on the 2015-2016 Twitter dataset using the previously described method, as is shown in Table 5. Herein, the keywords are classified into groups according to their meaning and implication. It is worth mentioning that our method is able to capture the relevant words such as those in “medical” and “flu” groups.

Class	Keywords
adv.	barely, probably, exactly, literally
emotion	bitch, stupid, brave, ouch, wow
flu names	bird, cold, hurt, influenza, sick, sore, swine
glossaries	appointment, doctor, health, blood, needle, physical, stomach, symptom, shot, virus
time	ago mid, minute, morning, season, time, today, tomorrow, yesterday
verb.	catch, cry, deserve, feel, guess, hate, help, hit, know, like, pass, pump, stay
noun.	arm, bed, case, clock, hop, job, news, outbreak, people, place, luck, man, shoulder, work
other	en, contra, de, la, plus, right, u

Table 5: Keywords for the 2015-2016 Twitter datasets

We process the tweets using the bag-of-words (BoW) model into a numerical vector space and sum the term frequency (TF) of each keyword in every week. The keyword extraction not only decreases the dimensionality of the dataset, but also improves the quality of the features. Such a conclusion is drawn from the most frequent words before and after keyword extraction, which are shown in Figure 1.

4.2.3 Trend Prediction

We employ a multi-layer regressor to predict the percentage of weighted influenza-like illness (%-weighted ILI, y) of all patients. To improve the prediction accuracy, the input (X) contains not only the tweet vectors (from bag-of-words model and keyword extraction), but also the %-weighted ILI in the *previous* week, which serves as a way to assess the historical influence and as a starting point for flu trend prediction. It is also worth mentioning that

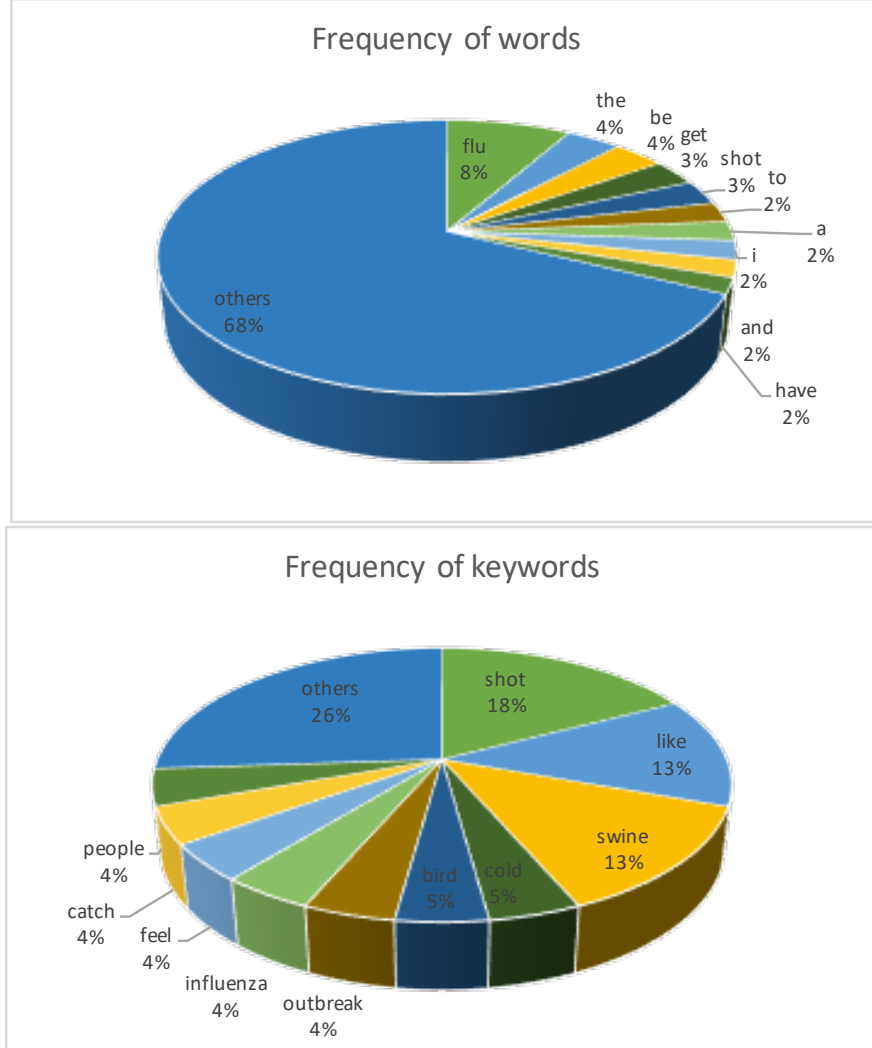


Figure 1: The term frequency in the 2015-2016 Twitter datasets.

we can actually use the %-weighted ILI from anytime in the past, for example, the last three weeks or even add some noise to the factor, like a small random perturbation on the percentage, which better simulates the reality.

To evaluate the effectiveness of neural network on prediction, we process the dataset per week as is described in the methodology section 3, and randomly pick 50% of the weeks for training while hold the remaining weeks for testing (1:1). After parameter optimization, we chose to set **lbfgs**, a quasi-Newton method, for weight optimization. We use the logistic sigmoid as the activation function and add two hidden layers (200 and 100 hidden units respectively) [27]. The predicted trend and the groundtruth are both shown in Figure 2.

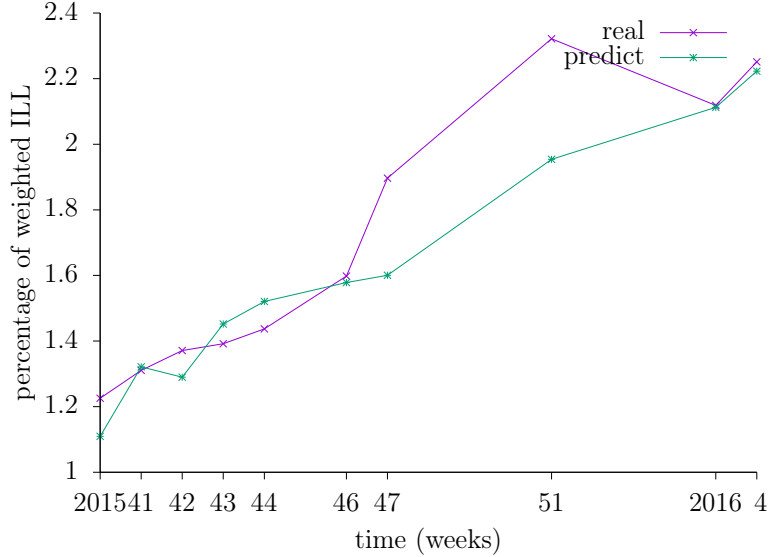


Figure 2: 2015-2016 flu trend timeline

4.3 Flu Trend Prediction on a Larger Dataset

To evaluate our method on real world data, we make use of a relatively large dataset that contains millions of tweets. We randomly sample a small portion of the tweets for manual labelling based on whether the content is relevant to flu outbreaks or not. Keywords are subsequently generated by χ^2 -test. The whole dataset is then processed with the keywords list as is described in the methodology section 3. We use the previously mentioned neural network (NN) developed for flu trend prediction and compare the predicted results with the ones on the CDC report, which serve as groundtruth for our study.

4.3.1 2017 Twitter Dataset (Large)

The dataset used in this section is obtained from the Archive Team Twitter stream [28] 3.1, which is a randomly collected JSON format dataset from general Twitter stream. The original dataset is extremely large (over 1 GB tarball for a single day) and simply impractical to use them as a whole. We cherry-pick 3 months - October, November, and December in 2017. There are 7,001,557 recorded tweets in October, 5,198,021 in November, and 6,739,178 in December, which are huge numbers. To simplify the computing task, we perform the flu trend prediction on each month separately. We further randomly divide each month's tweets into 20 equally sized *chunks*, which multiplex the number of training samples. After keyword extraction we get 279 keywords generated from 5,346 manually labeled tweets in October, 247 keywords from 4,276 tweets in November, and 404 keywords from 6,368 tweets in December respectively.

4.3.2 Trend Prediction

Month	k-fold	MSE	MAE	R^2
Oct.	1	0.0017	0.0328	0.9518
	2	0.0008	0.0210	0.9797
	3	0.0007	0.0217	0.9814
	4	0.0006	0.0187	0.9843
	5	0.0084	0.0504	0.7792
Nov.	1	0.0007	0.0228	0.9858
	2	0.0007	0.0210	0.9891
	3	0.0009	0.0275	0.9794
	4	0.0007	0.0228	0.9876
	5	0.0035	0.0300	0.9244
Dec.	1	0.0166	0.1014	0.921
	2	0.0068	0.0663	0.9976
	3	0.0068	0.0691	0.9959
	4	0.0144	0.0995	0.9927
	5	0.7834	0.3818	0.6080

Table 6: The performance of neural network prediction

Leveraging the multilayer regressor, we get a flu trend timeline prediction as is shown in Figure 3, 4 and 5 by a 66:33 split on the original dataset. We also conduct a measurement on the mean squared error (MSE), mean absolute error (MAE), and Pearson correlation (corr) by a 5-fold cross-validation, as is shown in Table 6. It is apparent that our predictor generally performs well except for the last week of December, which is consistent with our cross-validation results (Table 6). It also shows a high R^2 around 0.9 and a low MSE/MAE < 0.1 , which suggests a promising application of our keyword extraction and neural network prediction model in a general real-world flu trend prediction problem.

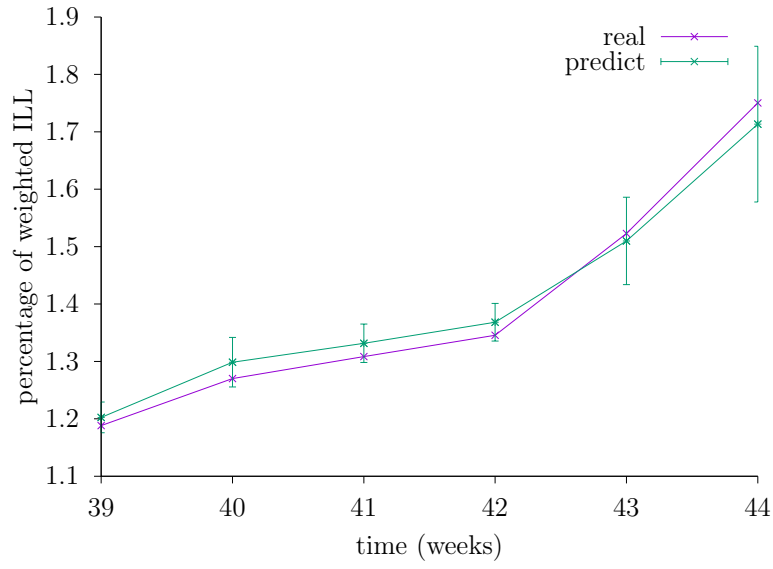


Figure 3: The flu trend timeline of Oct. 2017

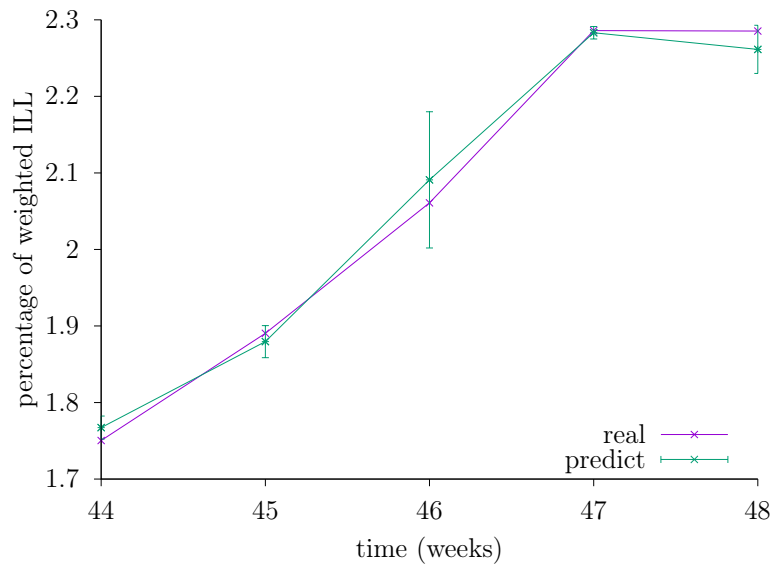


Figure 4: The flu trend timeline of Nov. 2017

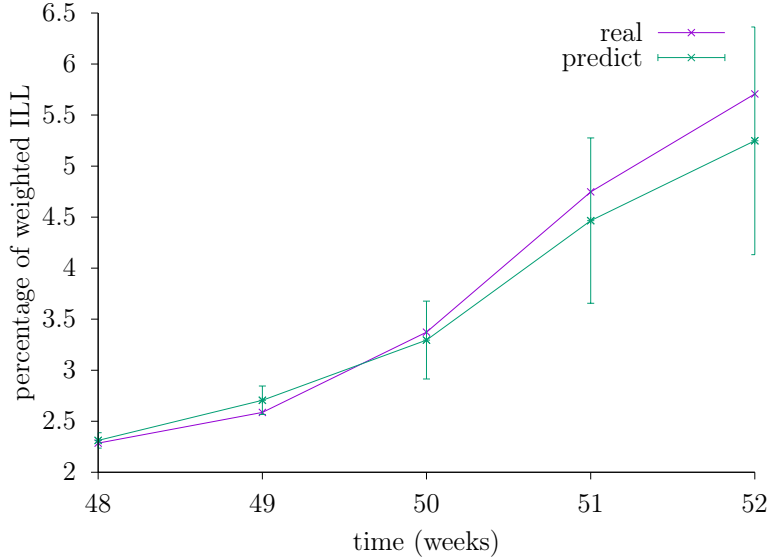


Figure 5: The flu trend timeline of Dec. 2017

5 Conclusions

We have developed a twitter-stream text mining and machine learning approach to predict flu outbreaks. The proposed method is promising in that it forecasts flu outbreaks in real time. Our method shows a high correlation between the Twitter based prediction and the groundtruth from CDC report. The current method heavily relies on the effectiveness of keyword extraction. The quality of keyword generation depends on the quality of manual labelling as well as the accuracy of the association between words and the relevant information. It is also worth mentioning that a Twitter stream dataset can be extremely large, which poses a significant challenge to processing and mining. In the project we have manually labelled 15,990 tweets and handled 18,938,756 JSON records.

For future work, We can look at employing other natural language processing (NLP) techniques other than simple keyword based analysis, since the context of a word or sentence is equally important.

6 Code/Data Release and Individual Tasks

6.1 Code/Data Release

We release our code at <https://github.iu.edu/rz20/B565-DataMining>. The Python source code is at the `src` directory, where `preprocess` contains the scripts for data pre-

processing, labelling for data labelling and mining for training and prediction.

The dataset is stored at <https://bit.ly/2r6hle9> (Google Drive IU).

6.2 Individual Tasks

We have had regular group meetings every week to synchronize the progress of the project, during which each team member has contributed their thoughts.

Tianyu Chen As the team leader, Tianyu worked on preprocessing and dataset handling. He processed the Twitter dataset to suit keyword extraction. He helped Rui improve the code and Yuhan for manual labelling. He also performed the experiments of keyword extraction.

Yuhan Zeng Yuhan manually labelled more than 10,000 flu-relevant tweet candidates, which is one of the most critical part of this project. She also worked on keywords extraction and Twitter-based flu trend prediction. She also contributed to the writing of this report.

Rui Zhang Rui was responsible for the development of keyword extraction and the neural network predictor. He got preliminary results on the small datasets. He also performed the experiments on flu trend prediction on the larger datasets and analyzed the results. He wrote the result section of the report.

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 health*, 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. World-wide 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] Pearson correlation coefficient — Wikipedia, the free encyclopedia. [Online; accessed 1-March-2018].
- [22] Root-mean-square deviation — Wikipedia, the free encyclopedia. [Online; accessed 1-March-2018].
- [23] Mean absolute error — Wikipedia, the free encyclopedia. [Online; accessed 1-March-2018].
- [24] Mae and rmse, which metric is better. [Online; accessed 1-March-2018].
- [25] Alex Lamb, Michael J. Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. June 2013.
- [26] Using machine learning to analyze twitter for real time influenza surveillance. [Online; accessed 1-March-2018].
- [27] Scikit mlpregressor documentation. [Online; accessed 1-March-2018].
- [28] Tweets archives. [Online; accessed 1-March-2018].