University of Nottingham
UK | CHINA | MALAYSIA

Social Media Analytics for Healthcare Surveillance using Text Mining

Submitted Oct 2019, in partial fulfillment of
the conditions for the award of the degree **BSc Computer Science.**

Yuyang Liu
16522049

Supervised by Heng Yu

School of Computer Science
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature ___________________

Date ______ / ______ / ______

I hereby declare that I have all necessary rights and consents to publicly distribute this dissertation via the University of Nottingham's e-dissertation archive.

Public access to this dissertation is restricted until: DD/MM/YYYY

Contents

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and related works

Disease control and prevention is vital for the whole society. Traditional surveillance method adopted by the Centers for Disease Control and Prevention (CDC) is, scrutinizing outpatient records from hospitals and virological test results from laboratories, which notices the disease after it actually occurred [27]. However, if there is no forecast and hospitals are ill-prepared for a rush of patients, the reception of in-time treatment will be affected [11], other severe consequences can also be imagined. Therefore, a robust disease forecast system is needed. To predict the outbreak of disease in advance, massive efforts have been put. [8] monitored the changes of Realistic Contact Networks (RCNs) to predict the dynamic movement of disease, [7] used machine learning with previous illness records to predict the future outbreak, while some researchers tried to build a prediction system based on data from social media. According to [21], social media contains information related to healthcare, individual health issue, symptom. [13] shows that spikes in flu queries and disease breakout coincide. However, since queries has little or no limitation and even don not need an account, they cannot be regarded as reliable data [27]. Other social media platforms such as Twitter and Facebook have proven their value for Big Data analyze. Twitter data has been found to be useful for public health applications [10], including: (1) monitoring diseases, (2) public reaction, (3) outbreaks or emergencies, (4) prediction, (5) lifestyle, and (6) geolocation of disease surveillance [1]. In addition, social media is prompt. According to [11], over 645 million active Twitter users collectively post an average of 58 million tweets (micro-blogs no more than 140 characters long) per day in 2017, and the number is still growing. A practical example is that researcher use Twitter predicted flu outbreaks 1–2 weeks ahead of CDC's surveillance average [28]. [11] also showed Twitter data aligns with CDC's outpatient records. All the information I have read so far proves that such data is valuable in healthcare surveillance.

1.2 Motivation

Previous work relying on social media has successfully proposed some novel methods. [25] and [26] utilize time-series analysis on single geography, [11] provided a generalized solution to identify how contagious diseases diffuse across geographies. However, time-series analyze can be inaccurate in this scenario (they rely merely on the statistics of social media instead of its content). For example, top search about a certain disease could result from a celebrity's illness [27]. Similar experimental result showed in [11]. During festivals, the number of tweets decreased, and the prediction accuracy was affected. Obviously, such results are not robust enough to be applied in the real world. It therefore makes sense to find a new method to enhance the whole system. In this project, I will start from previous works, and use different techniques such as machine learning, heuristic function, NLP to extract textual implication containing in such data and create a more feasible solution.

Chapter 2

Aims and Objectives

The general aim of this project is analyzing social media data to surveille healthcare condition. It can be detailed as follow:

1. Healthcare-related textual information should be extracted and modeled for the purpose of healthcare surveillance.

2. Robust predictive models are required for accurate forecasting of disease outbreaks and hospital emergency visits based on ML techniques.

The key objectives of this project are:

1. Collecting data of a certain social media platform. This data can be either extracted from an existing data set or crawled from that platform. If existing data set can't meet my requirement, my own data set will be made.

2. Filtering data. In this process, a filtering rule should be established, possible solutions include using related verbal list, using NLP to classify raw data

3. Designing algorithm that can successfully predict the outbreak of diseased and its propagation direction with high accuracy, and implementation. This includes (1) setting a benchmark to evaluate the accuracy; (2) finding a suitable model maxing the accuracy.

4. Experiment. Testing and evaluating this model, and compare it with others' work. In this stage, re-implementation may be needed.

5. Result visualization. Since the outbreak is dynamic, visualizing the result can be more readable.

Chapter 3

Design

3.1 Basic assumptions

The whole project is designed based on two basic assumptions: (1) social media data can be used to predict the outbreak of diseases; (2) available data contain sufficient information to show the relationship between input and output.

3.2 Overall design

Treat our algorithm as a Blackbox, the input is the metadata extracted from social media platforms while the output is the prediction of diseases (Figure 3.1).

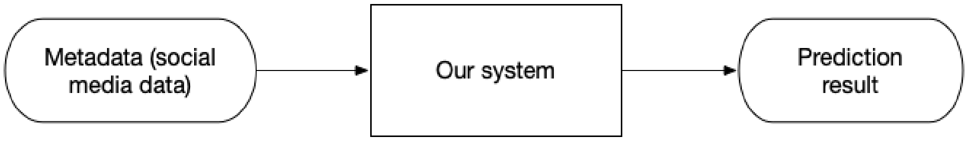


Figure 3.1: Blackbox

The aims of overall design involve building a complete process where the performance is reproducible, the subprocesses can be adjusted and the outcome can be easily understood and visualized. Based on it, we separate the Blackbox into 5 independent components: Data Collector, Preprocess Layer, Prediction Model, Evaluation and Diffusion Modeling Layer, Visualization (undetermined). The overall structure can be seen in Figure 3.2, the square represents component while ellipse represents data.

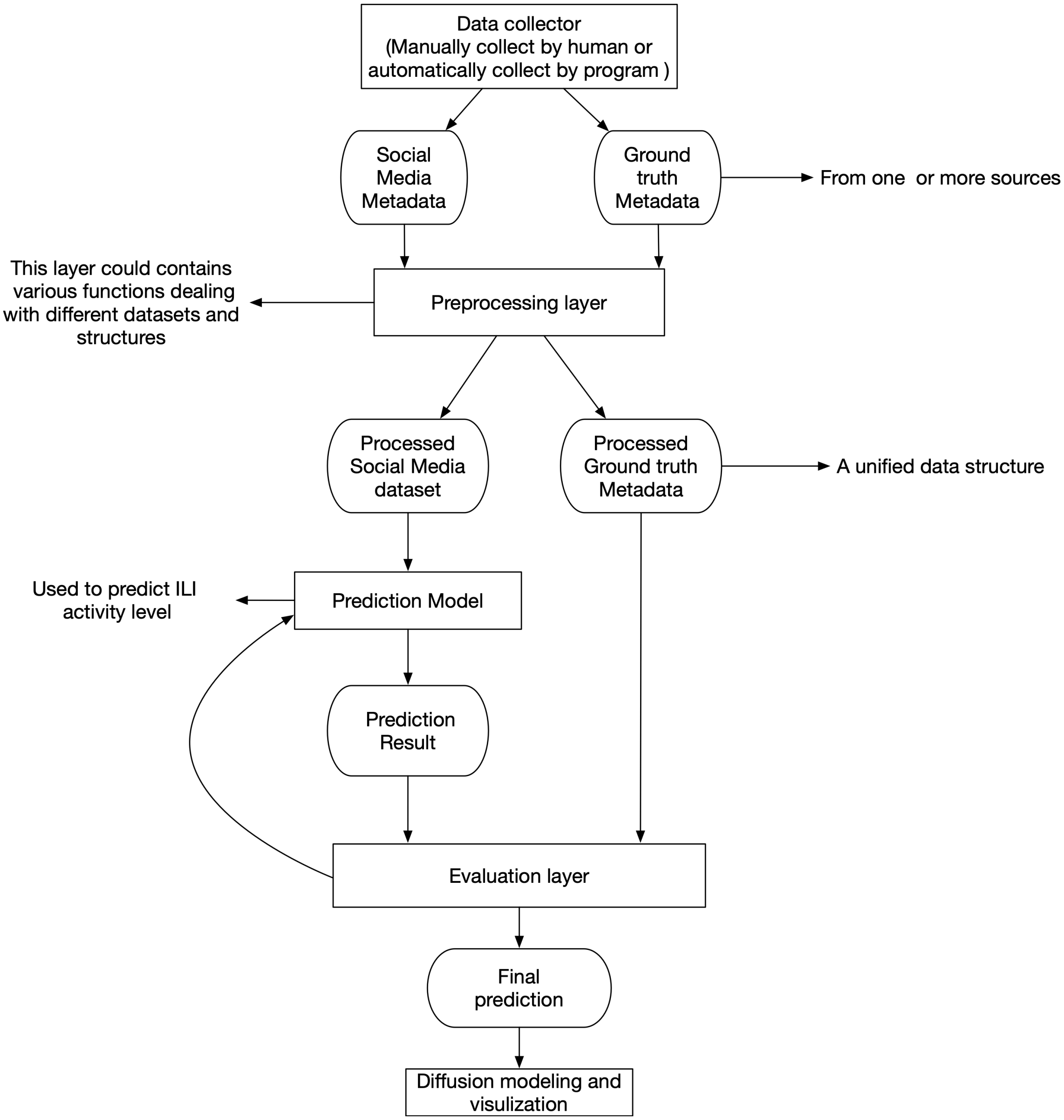


Figure 3.2: Overall design

3.3 Basic components

This section show the preliminary design of our components, including the basic functional requirements of each part and the possible methods we may adopt. Following subsections represent stages(pipelines) of our system respectively, the general procedure design is inspired by [12].

3.3.1 Data collector design

Social media data is the input of the whole system, but according to the different social media platforms and various sources of data (such as extracted from official API or download from open-source dataset), the structure of data and methods of collecting data can be diverse. Therefore, we design an interface which can collect and integrate different metadata. In this project, we focus more on modeling and algorithm design rather than a complete system, therefore, we choose one certain social media platform. In addition, to evaluate the accuracy and performance of our algorithm, we need a ground truth dataset. Here the interface should at least collect these two dataset from at least one source respectively.

Functional requirements of this component are:

1. Collect and store social meida data from at least one source
2. Collect and store ground truth data from at least one source

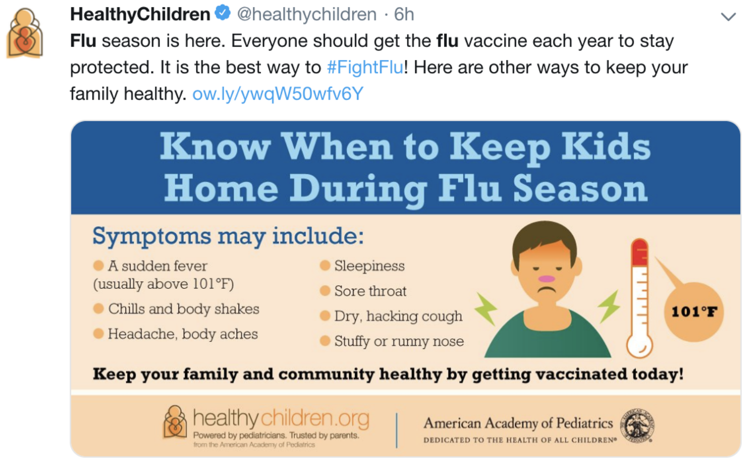
3.3.2 Data preprocessor design

Data collected by stage 1 are metadata, which could be unstructured and irrelevant to this project. We subdivide this step into smaller steps:

1. Unify data structure: if one dataset comes from different sources (for example, facebook data extracted from API and from web spider). To pass these data into functions of later steps, a unified format is required. In addition, dataset could contain information

that won't be used by our algorithm, ignore such information when unify data can save storage space. In our design, both unified structures of social media dataset and ground truth dataset should be implemented (see section 4.3).

2. Text regularization: social media dataset could adopt different coded format (such as ASCII and unicode), here we decide to use utf-8 encoding, which is wildly used in the Internet. The collected data could contain special symbols, unknown characters, URL links and emoji, pictures, videos (See Figure 3.3). In this project, we focus on pure text, thus information such as URL links, pictures and videos will be ignored. Inspired by [34], emoji and some pecial symbols can be transformed in to text based on standard transformation tables. We will adopt such transformation to keep maximum valuable data.



HealthyChildren @healthychildren · 6h

Flu season is here. Everyone should get the **flu** vaccine each year to stay protected. It is the best way to #FightFlu! Here are other ways to keep your family healthy. ow.ly/ywqW50wfv6Y

Know When to Keep Kids Home During Flu Season
Symptoms may include:
A sudden fever (usually above 101°F)
Chills and body shakes
Headache, body aches
Sleepiness
Sore throat
Dry, hacking cough
Stuffy or runny nose
101°F
Keep your family and community healthy by getting vaccinated today!
healthychildren.org
Powered by pediatricians. Trusted by parents.
from the American Academy of Pediatrics
American Academy of Pediatrics
DEDICATED TO THE HEALTH OF ALL CHILDREN®

Figure 3.3: Tweet with URL and picture (screenshot from Twitter)

3. Data filtering: after the structuralization, the dataset can be used in our system. However, not all the data contain information we want. This step will filter out irrelevant data of both social media dataset and ground truth dataset and reserve data that be considered useful. We will set inclusion rules and classifier to filter the data and label them (see section 4.4). The method we adopt to train a classifier is similar to how we build our prediction model. The detailed methods of text tokenization,encoding and

building neural network model can be found in stage 3 of this section.

4. Location extraction: in this project, we need data containing time of creation, geographic information to create our diffusion model. Data without such information can be used to train a classifier in the next step. The time of creation is contained in most sources (all the datasets we search so far provide temporal information). However, based on user's setting, some data don't contain geographic information (users can hide their private information if they wish). In addition, some platforms allow users name their own location (such as Wechat, users can assign personalized name to their location), or use a fake one. Platforms can adopt different standards of placename. Apart from that, even a user provide authentic private location, it still can't be guaranteed that he was in that place when posted tweets. All of such conditions bring noises in the dataset. Use the method we deal with unstructured metadata for reference, here we adopt the same solution: regularize the geographic information, set a standard in our project. In term of customized placename, we will design a function trying to map it to our standard (see Chapter 4).

Functional requirements of this component are:

1. Must unify a data structure of social meida dataset, and can integrate datasets (if more than one sources) into the same structure
2. Must unify a data structure of ground truth datasets, and can integrate datasets (if more than one sources) into the same structure
3. Must regularize all the text in the integrated dataset
4. Must extract geographic information of each data if available
5. Must regularize all the extracted geographic information into a unified format

3.3.3 Model design

The preprocessed social media dataset will be used in this prediction model. In this stage, [11] and [34] use time-series analysis to predict the number of next week's tweets of each state that related to flu, and then map the number to CDC ILI activity level. However, their algorithm needs most-recent weekly tweet counts (last 8 weeks), which means that their algorithm will be affected severely if data is missing or insufficient (explained in their paper). Inspired by this, the output of our prediction is design to the next week's ILI level of states. To overcome the drawbacks of time-series analysis, we decide to mine information related to the condition of diseases and worries of users (undetermined) from the text itself, instead of the counts of relevant tweets, to predict the ILI level. In our preliminary design, we will adopt NLP-based technologies here (see Chapter 5). The steps of this stage are:

1. Text tokenization: neural network can only receive tensors as input, therefore, the first step of this stage is transforming the textual data into tensor, which is called tokenization (a token is a single unit extracted from text) [9]. There are three basic word separation strategies: (1)split text into single words, transform each word into a vector; (2)split text into single character, transform each character into a vector; (3) extract n-gram of words or characters, transform each n-gram into a vector (a n-gram is a set of sequential words or characters), the resulting set is called bag-of-words [9]. Bag-of-words can't record the order of words in the original text, therefore, this method is wildly used in shallow-layer model instead of deep learning model. Extracting n-gram is a feature engineering, which is inflexible and unstable. Here we adopt the first strategy, the feature extraction procedure will done by our deep learning model.

2. Text encoding: the procedure that transform token into vector is called encoding. There are two most common methods. One is one-hot encoding, which assign each token a unique integer i, transform i into a binary vector (only contains 1 and 0) of length N (N is the size of token list), only the ith element is 1, others are 0. This method returns a

high dimensional sparse vector (20000 dimensions or more), since each token takes one dimension. Anthoer encoding method is word embedding, which is learnt from dataset, returns a low dimensional intensive vector. The idea behind this method is that: the geometrical distance of two token should base on their relation (synonyms should have shorter distance than antonyms), and the vector's direction should have sense. For example, the vector of word "king" plus the vector of word "female" should return the vector of word "queen" [9]. Therefore, we can't assign each token a vector randomly. In addition, for different tasks, the embedding space could be diverse, the embedding space used for sentiment analysis may not fit argot detection. In our design, we will try both of these two methods, and for the second, we will train a embedding space based on our dataset.

3.3.4 Evaluation layer design

In our overall design (Figure 3.2), there is a cycle linking this layer and prediction model. But note that it doesn't mean that evaluation will only happen in this stage. In fact, in each step of implementation, we plan to try different possible methods and provide convincing reasons to explain why we finally adopt certain strategies (such as why we collect data from certain sources). Here the evaluation layer is mainly focus on assessment of the final output of our algorithm. It receives two inputs, the prediction result and the ground truth dataset, and output a score of prediction.

Our task is classification, according to [9], there are 8 common methods can be adopted to evaluate the model: (1) confusion matrix; (2) accuracy; (3) precision; (4) recall; (5) F1 score; (6) ROC curve; (7) AUC (Area Under Curve); (8) PR curve. Especially, accuracy and recall are wildly used in class-imbalance problem (our task is class-imbalanced). Following are formulas of (2)(3)(4)(5), where TP, TN, FP, FN represents true positive (the number of cases correctly identified as required), ture negative the number of cases correctly identified as not required, false positive (the number of cases incorrectly identified as required), false negative (the number of cases incorrectly identified as not required), respectively:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN}$$
$$Precision = \frac{TP}{TP+FP}$$
$$Recall = \frac{TP}{TP+FN}$$
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Once the socre (target) is defined, we must adopt a method to assess the result. There are 3 common methods: (1) Hold-out method (test set estimation); (2) K-Fold Cross Validation; (3) Repeated k-fold Validation. The first method works by randomly divided dataset into two mutually exclusive subsets, the training set (often $\frac{2}{3}to\frac{4}{5}$ of the original set) and testing set. It is simple to implement but will be severely affected by the size of subsets. If the training instances is far more than testing instances, the evaluation result is unreliable, but in reverse, the model will lose fidelity. In addition, this method is unsuited to small sample sizes, since it can't make full use of data [9,22]. Second method partition data into k separate subsets of similar size. Each subset will be used as testing data in turns (k times) while left subsets will be used as training data, the final socre is the mean of all rounds. It can be regarded as a kind of hold-out method with the ability to exploit more data and provide higher reliability [22]. The third one is used when the available data are too fewer while high prediction accuracy is required. It repeat the second method and calculate average score [9].

Functional requirements of this component are:

1. It should provide different evaluation indexes (Accuracy, Recall, etc.)to assess the prediction outcome.

2. It should provide different evaluation methods (K-Fold, Hold-our, etc.) to assess the prediction outcome.

3. Must choose a best combination of methods to evaluate prediction result based on ground truth dataset.

4. Must set a baseline (or target) to stop training.

3.3.5 Visualization design

The final result may not clear and meanful for users. Visualization can help users/researchers to figure out the potential information of data/result, such as its feature, pattern, trend and relationship [15]. There are various visualization techniques for different sceniros, purposes and data/input, such as 2D display (bar chart) and 3D display (cloud vapor), in addition, if the prediction is real-time, the visualization could be dynamic. In this project, visualization is used in the last stage, therefore, we can assume the the input is stable and predictable. In addition, the prediction is numerical, according to [15], geometric representing methods could be used, such as scatetr plot, lines etc. The final method will be adopted based on the experimental result.

Functional requirements of this component includes:

1. It should receive the final prediction as input and choose an approach to display the result.

Chapter 4

Data collection and processing

In our system design, there are two general datasets are needed. The first is the social media textual data, which is the most vital to this system, called social media dataset. It contains the implication of potential outbreak of diseases, all the prediction are made based on it. However, to evaluate the accuracy of the prediction, actual records of diseases are needed to serve as benchmark.

4.1 Social media dataset

In this project, we choose one certain social media platform to test our algorithm, that is the Twitter. More specifically, we focus on tweets expressed in English. These choices based on the characteristics of Twitter:

- Providing location: In our design, the posting location of each post is required. And we found that Twitter provides such information. According to research conducted by [14] in 2016, about 1.6 percent of Twitter users opted in Twitter's location-sharing service.

- Availability: Most tweets are available for research. According to [34], around 95% of Twitter users opted in sharing their tweets with public, meaning their tweets can be searched and filtered by keywords without their permission.

- Comparability: In our research, we found that most related works used data from Twitter, which means that Twitter dataset can act as a benchmark to evaluate our algorithm.

- Timely: According to [34], each tweet is received within seconds of their creation.

- High user volume: According to [14], about 21% of American citizens use Twitter.

4.2 Benchmark

In this project, we focus on one certain disease, flu. The benchmark used in this project is Fluview [6], a weekly-update, influenza surveillance report of the U.S. published by Centers for Disease Control and Prevention (CDC) [5]. Such report is a collaborative effort between CDC and its many partners in state, local, and territorial health departments, public health and clinical laboratories, vital statistics offices, healthcare providers, clinics, and emergency departments [5]. CDC maintains a surveillance network called Influenza-like Illness Surveillance Network (ILINet), which collect information on outpatient visits to health care providers for influenza-like illness [5]. We choose CDC's Fluview as our benchmark because of its:

- Reliability: ILINet collects data from about 2600 outpatient healthcare providers across the U.S. weekly [5].
- Accessibility: All the reports of Influenza-like Illness (ILI) can be accessed by public.
- Comparability: The dataset maps the activity of ILI into levels between 1 to 10, which can be used as labels or targets of our training data.

Figure4.1 shows the ILI activity levels across the U.S. in Week 43, 2019. It can be seen that the there are 10 levels divided into 3 categories, each level is assigned a unique color.

4.3 Data collection

This section shows the detailed methods of how we collect data and unify the data structure.

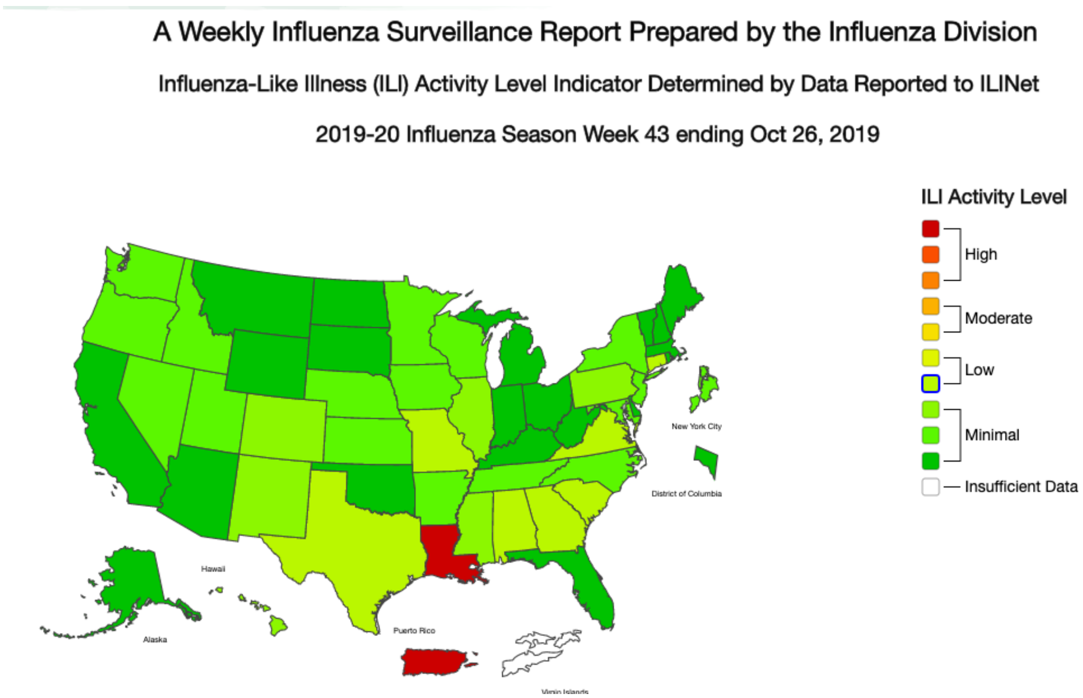
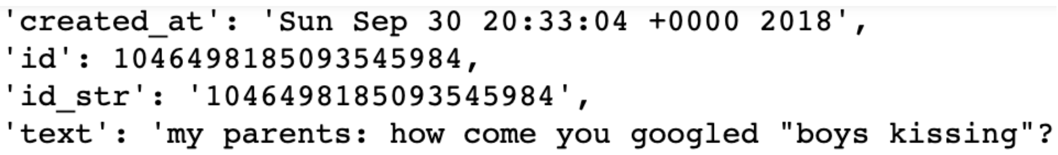

A Weekly Influenza Surveillance Report Prepared by the Influenza Division
Influenza-Like Illness (ILI) Activity Level Indicator Determined by Data Reported to ILINet
2019-20 Influenza Season Week 43 ending Oct 26, 2019
ILI Activity Level
High
Moderate
Low
Minimal
Insufficient Data
New York City
District of Columbia
Hawaii
Alaska
Puerto Rico
Virgin Islands

Figure 4.1: Influenza Season Week 43 ending Oct 26, 2019, Source: Fluview

4.3.1 Twitter dataset collection

In this project, we collect Twitter data from Internet Archive [16], a non-profit digital library of millions of free books, movies, software, music, websites. It contains daily tweets from Feb 2011 to Jul 2019 (Accessed: Dec 08,2019). All the data are Spritzer version (roughly 1% of the whole tweets) grabbed from the general twitter stream The number of tweets collected in Oct 05 2018 is 4273031, in Oct 04 2018 is 4337327, in Oct 01 2018 is 4317376, on average is above 4 million per day. All the tweets are stored in json files. Such data volume is sufficient for our research and its data structure is convenient to use. More important, it contains the information we need, which mentioned in section 4.1. Figure 4.2 shows part of the information those json files contain (geographic and linguistic information are contained but not listed here). In our project, we mainly focus on tweets posted during 2018.

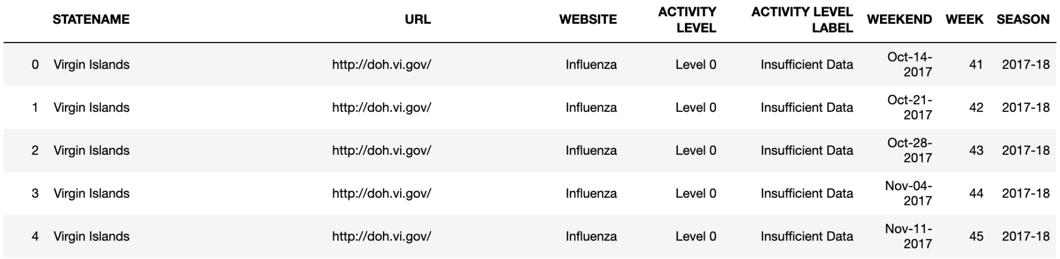


```
'created_at': 'Sun Sep 30 20:33:04 +0000 2018',
'id': 1046498185093545984,
'id_str': '1046498185093545984',
'text': 'my parents: how come you googled "boys kissing"?
```

Figure 4.2: Screenshot of Archive's Twitter data

4.3.2 Fluview dataset collection

This data can be downloaded from official websites of Fluview [6] (Accessed: Dec 08,2019), user can customize the data they want to download (the time span of reports). All the information is stored in a single csv file. Figure 4.3 shows the structure of the data.



	STATENAME	URL	WEBSITE	ACTIVITY LEVEL	ACTIVITY LEVEL LABEL	WEEKEND	WEEK	SEASON
0	Virgin Islands	http://doh.vi.gov/	Influenza	Level 0	Insufficient Data	Oct-14-2017	41	2017-18
1	Virgin Islands	http://doh.vi.gov/	Influenza	Level 0	Insufficient Data	Oct-21-2017	42	2017-18
2	Virgin Islands	http://doh.vi.gov/	Influenza	Level 0	Insufficient Data	Oct-28-2017	43	2017-18
3	Virgin Islands	http://doh.vi.gov/	Influenza	Level 0	Insufficient Data	Nov-04-2017	44	2017-18
4	Virgin Islands	http://doh.vi.gov/	Influenza	Level 0	Insufficient Data	Nov-11-2017	45	2017-18

Figure 4.3: Screenshot of Fluview's report

Our data structure is a cut-down version of it, where "URL" and "WEBSITE" columns are deleted. "WEEKEND" is transformed into the format we used in Twitter dataset.

4.4 Data Preprocessing

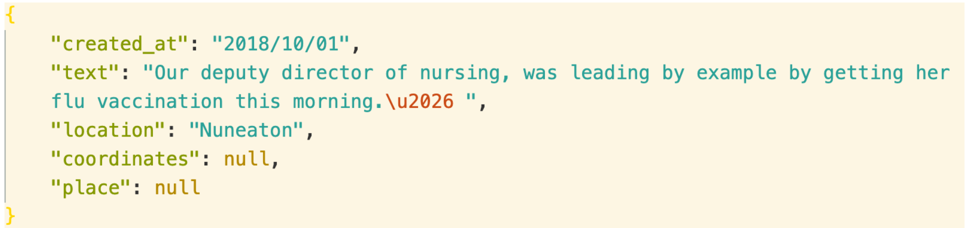
Data preprocessing is the first stage of a typical text classification framework. According to experiment conducted by [32], different combination of preprocessing methods can influence the accuracy of prediction. However, there is no best combination for all tasks. Some strategies can improve classification success of certain tasks while lower that of others. Following sections represent our preprocessing methods designed for this project.

4.4.1 Unify data structure

In section 3.3.1, we mentioned a unified data structure of our datasets. The aim of this stage is to unify and regularize all the metadata before analysis.

The Fluview dataset is well structured (see Figure 4.3), and can be used in our system directly (we used its “STATENAME”, “ACTIVITY LEVEL”, “ACTIVITY LEVEL LABEL” and “WEEK” columns). One point should be noticed here is that the STATENAME used in Fluview dataset can be different from that in Twitter (Twitter’ users can set their own location), therefore, we created a unified name list of states and a function designed to regularize different geographic information.

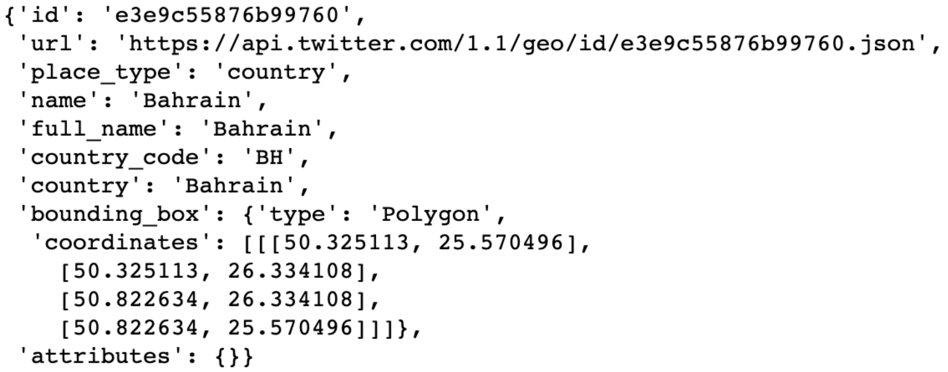
Our social media dataset’s structure is built on Twitter’s official data structure [30], and we only reserve the information we may use. It is a hashmap with 5 keys: “created_at”, “text”, “location” and “coordinates”, “place”. We regularize the time into format Year/Month/Day (2018/10/01) and store it in “created_at” keys, exclude other information (see Figure 4.4). Note that in the metadata, there are massive “deleted” tweets, which contain no textual information, and we removed all such data. According twitter’s official document [30], there are two classes of geographical metadata in Tweet data: (1) Tweet location, which is available when users share location at time of Tweet; (2) Account Location: a free-form character field set in user’s profile and may or may not contain metadata that can be geo-referenced. “location” attribute stores the user-defined placename. “coordinates” attribute provides the



```
{
    "created_at": "2018/10/01",
    "text": "Our deputy director of nursing, was leading by example by getting her
    flu vaccination this morning.\u2026 ",
    "location": "Nuneaton",
    "coordinates": null,
    "place": null
}
```

Figure 4.4: Screenshot of unified structure of social media dataset

exact location of a tweet (in long-lat order) but has no placename and only available when the location is assigned. “place” attribute is always present when a Tweet is geo-tagged, and it contains Twitter “place” with a display name and type. Here we keep all these three keys, and “place” key gets the first priority when we extract location of tweet. Figure 4.5 shows a sample of Twitter “place”. Note that most data don’t contain “location”, “coordinates” and “place” keys, which is expected in section 4.1.



```
{'id': 'e3e9c55876b99760',
 'url': 'https://api.twitter.com/1.1/geo/id/e3e9c55876b99760.json',
 'place_type': 'country',
 'name': 'Bahrain',
 'full_name': 'Bahrain',
 'country_code': 'BH',
 'country': 'Bahrain',
 'bounding_box': {'type': 'Polygon',
  'coordinates': [[[50.325113, 25.570496],
    [50.325113, 26.334108],
    [50.822634, 26.334108],
    [50.822634, 25.570496]]]},
 'attributes': {}}
```

Figure 4.5: Screenshot of Twitter’s place object from our dataset

4.4.2 Manually screening

We have more than 4 million tweets per day in our dataset, it’s obviously not all of them are relevant to our task. Based on our initial design, we only accept tweets written in English. While reviewing on the dataset, we found that there are massive retweets in it (8312846 of 21694196, roughly 38%). Defined by Twitter [30], retweeting is forwarding content wrote by other users (like forwarding an email). Although retweets can be used in some tasks such as sentiment analysis (mainly focus the counts of being retweeted) [23], we don’t think it will contribute to our prediction. As stated by [18], retweets don’t have user’s own view and should be removed in data cleaning. We need tweets that can show the health condition of the user or of people the user cares. In addition, retweets don’t have any geographical information [30], which is the key component in our diffusion modeling. Therefore, we decided to exclude all retweets.

Since Twitter dataset contains massive tweets which are irrelevant to flu, before we analyze the data, further filtration is required. We adopt two steps to filter out irrelevant data, keywords search and supervised classification. Following sections are the details of these two methods.

4.4.3 Keywords search

To filter out irrelevant data, [2] used a simple word look-up of "influenza",which we think may lose massive valuable data. In inspired by [19, 20], we first create a word list related to flu based on Flu Symptoms [4] Cambridge Dictionary [3] and relatedwords.org [24]. The complete word list can be seen in table 4.1, note that not all the words from the sources are added to our list. Professional terms are excluded since they are hardly used in colloquialism. Words that are wildly used in other scenarios (such as chill, cold) and phrases that contain keywords in our list (such as asian influenza) are not added. We then filter tweets according to the list (ignore case). In this step, we initially adopted a relatively lose filtering strategy: accept tweets containing any string in our list. The filtering result shows that the majority of the filtered tweets are irrelevant to diseases. The possible reasons can be that people will use such words even when they are healthy, and some words are substring of other words (such as chill Achill, flu influence).

Source	Word list
CDC	fever, feverish, sore throat, runny nose, stuffy nose, headache, nasal congestion, diarrhea, bluish lips, bluish face, dehydration
Dictionary	flu, catarrh, cough, common cold, influenza, sniffle, snuffle
relatedwords.org	h1n1, h5n1, coughing, cholera, ebola, epidemic, feverous, measles

Table 4.1: Inclusion list

Table 4.2 shows the number of tweets after keywords filtering in the first 5 days in Oct 2018.

Date	Original	Filtered
2018/10/01	4317376	5984
2018/10/02	4349129	5740
2018/10/03	4417333	5415
2018/10/04	4337327	5676
2018/10/05	4273031	5190

Table 4.2: Tweet counts after filtering

4.4.4 Supervised classification

To overcome the drawbacks of first round filtration, further filtration is required. [11] created another exclusion dictionary containing keywords and phrases indicating tweets should not be included in the filtered dataset, such as ”sick and tired”. But their experiment shows the accuracy of their method is roughly 70%, which we think is relatively lower than our requirement. This result can result from their choice of exclusion words and their dataset. Since we want to get a higer filtering accuracy, we decide to use a machine learning based classification method [2]. We labeled our dataset, and trained a binary classifier. Following are our detailed methods:

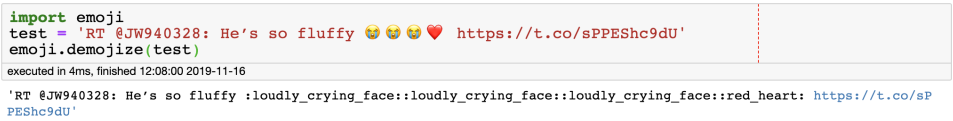
1. Text regularization: As mentioned in our design (section 3.2.2), the original text is unstructured, which can not be put in to our model directly and hard for labeling. In addition, each dataset has its own structure, meaning that there is no common regularization rule can be applied to all tasks. Here our regularization rule is built on the observation of our dataset and on our experience. Followings are common formats we observed and removed (or modified) in our dataset, expressed in regular expression (Python version):

    - retweet: retweeting is a way of forwarding content to others (like forwarding an email). It starts with a “RT ” pattern. As we mentioned in section 4.4.2, we removed all retweets.

    - @ and # Tags: in social media, @ refers to a person/group in a conversation, and the # refers to a topic of conversation. Their regular expression are “+[\S]*” and “#+[\S]*” respectively. In our preliminary filtered dataset, we found that some topics contain words in our inclusion list, such as #ALDUBFever4ever. We exclude them.

- URL links: some tweets contain URL links starting with “http”, “ftp” or “https”, which can’t contribute to our classification. Its regular expression is: “http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+”

- emoji: emojis in our dataset are encoded in unicode, inspired by [34], some of them can be translated into their name based on emoticon dictionaries. Through our search, we find a Python library called emoji(version 0.5.4), which embedded the full emoji list from [31] and can help us to translate emoji into text through function call [29]. Figue 4.6 shows a example how to use this library, the translated emojis are embraced by ‘:’ signs by default. In our implementation, each translated emoji is assigned a prefix ‘emo_’ to identify it, and we separate emojis with a blank space for split convenience.



```
import emoji
test = 'RT @JW940328: He’s so fluffy 😭😭😭❤️  https://t.co/sPPEShc9dU'
emoji.demojize(test)
```

executed in 4ms, finished 12:08:00 2019-11-16

```
'RT @JW940328: He’s so fluffy :loudly_crying_face::loudly_crying_face::loudly_crying_face::red_heart: https://t.co/sP
PEShc9dU'
```

Figure 4.6: Example of using emoji library

- e-mail address: we exclude e-mail address in our data, whose regex(can match most e-mail addresses) is: “[a-zA-Z0-9_.+-]+[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+$”

- html entities: a html entity can be regarded as escape characters used in html (eg. &quot; represents “” sign). We use the Python’s standard library “html” to translate it.

2. Data labeling (classification): as mentioned before, we want to our classifier can tell wether a tweet is relevant to flu, therefore, we adopt a simple binary labeling strategy. We write a piece code and manually labeled the regularized data 1 and 0 (1 means it

indeed relates to flu). While labeling, we found that some tweets can't be told wether they are truly related to flu, such as “Coughing!”, “i'm chilling'. And since in our first filtering (section 4.4.3), we adopt a relatively loose rule, which accepts tweet that containing any string (word) in our inclusion dictionary, no matter the word is indeed a sub-string of another word (eg. flu and influence), the majority of filtered tweets (roughly 83%, 450 of 540) are toally irrelevant to flu. In addition, some tweets are flu-related, but don't show signs of catching a flu, such as “Just had my flu jab!”. All the problems mentioned above increase the difficulty of labeling. While we have more than a billion raw data, we decide to change our initial design and set a strict standard in filtering. Followings are our specific labeling rules:

- tweets that don't contain any exact word in our word list will be excluded (eg. tweet containg “feverfew” and no other keywords won't be accepted even it has string “fever”)
- tweets containing fewer words (no specific number) will be label 0, even they have our keywords insinde (eg. “coughing!”), since they can hardly be classified
- tweets containing keywords but are irrelevant to flu will be labeled 0 (eg. “Monday Chill. Always Chill”)
- tweets talking about flu outbreak happened in the history and past, about catching a flu before (months or years ago) and about flu statistics will be labeled 0 (eg. “Over 80,000 Americans Died of Flu Last Winter, Highest Toll in Years”)
- tweets talking about flu but don't have signs of catching flu or outbreak of flu will be labeled 0 (eg. “Flu vaccine is very safe - risk of serious reaction is less than one in a million, much lower risk than catching the flu”, “It's that time of year again. Here are some tips to help your child get through a cough or cold.”)
- tweets containing symptoms of flu but can't tell the whether such symptoms are result of flu will be labeled 0 (eg. “i've had a headache since yesterday and i wanna

die")

- tweets meeting the requirements above but still cause confusion when classification will be labeled 0 (eg. "All iwanna do is cuddleand not cough every minute")
- only tweets containing real information of getting flu and can easily be recognized will be labeled 1 (eg. "Oh, is it colds and flu? Get well soon, and rest")

All the examples given above come from our dataset. Note that this standard is still not precise enough and can't guarantee all the data are correctly labeled, since the classification result highly depends on individuals who label the data. To minimize such difference brought by manual work, we built a exclusion list based on our deliberate review on the dataset during labeling, and labeled tweet containing word in the list 0.

3. Stop words elimination: stop-words are words that are not regarded as keywords in text mining. Examples are: is, am, did, a, an, the, ect. Exclude stop-words can reduce the length of the text and dimensionality of term space [33]. There are four stop-word removal methods: (1) manually define the removal list; (2) Zipf' law based methods, which removes the most frequent and least frequent words; (3) Mutual information method (MI), a supervised method that can evaluate the discrimination power of a term for classification; (4) Term Based Random Sampling (TBRS), which is based on a term's importance [17].

Chapter 5

Progress

This chapter covers the progress made so far and the remaining tasks needed to be done.

5.1 Project management

This section records the weekly progress of this project so far, starting from Oct 20, 2019. Following is the progress list ranked by chronological order:

- Oct 27, 2019: Searched and read 13 papers related to our project, having a basic idea in mind. Wrote some sketch of functions that will be used in our implementation.

- Nov 03, 2019: Searched and downloaded alternative twitter dataset. Implemented functions that can batch process the downloaded files (unzip the whole directory recursively). Built a preliminary exclusion dictionary used to filter out irrelevant tweets. Implemented functions that can batch filter the unzipped twitter json files. Prepared and past the Gre exam.

- Nov 10, 2019: Searched and collected CDC dataset. Implemented functions that can process the CDC file. Wrote interim report (Data collection section). Searched and read papers related to filtering tweets. Finished coursework 1 of Computer Graphics.

- Nov 17, 2019: Wrote interim report (Data collection and Design section), Implemented functions used to regularize our dataset. Set filtering rules, implemented functions used to label our dataset, and manully labeled 1000 tweets.

- Nov 24, 2019: Found the filtered dataset contains few samples, and can hardly be used to analyze. Therefore, the topic was changed slightly based on the dataset itself. Wrote interim report of Computer Graphics project.

- Dec 01, 2019: Finished coursework2 of Computer Graphics, implemented functions for Computer Graphics project.
- Dec 08, 2019: Finished coursework2 of Computing Ethics, built 3D scene for Computer Graphics project.
- Dec 15, 2019: Finished the Computer Graphics project.

5.2 Conclusion and Future work

So far, our project is at the end of the first stage. We collected the one social media dataset of Twitter and one benchmark dataset from CDC. Our programme is implemented with various functions to process them, including keyword search, unifying data structure, data regularization, etc. However, we found that the data volume after preprocessing may not be able to support our initial design (the percentage of useful data is less than 1/100000, after second round filtering, the number could be much lower), even though we have more than one billion metadata. Therefore, the next work is changing our initial plan slightly based on current dataset to ensure we have convincing data volume for experiment. This could be done by: (1) changing our filtering method, relaxing the filtering criteria; (2) changing the topic (but still related to healthcare); (3) collecting more data. For the second choice, the system design may change slightly based on the final topic. Generally, the remaining tasks includes:

1. Find new topic/dataset/preprocessing method to aquire enough training data.
2. Re-design some components of our system based on the changes
3. Re-implement some functions (mostly related to preprocessing) based on the changes.

4. Implement the remaining components of our algorithm based on the new design (if re-designed)

5. Design experiment methods and implement functions for evaluation.

6. Write final report

Bibliography

[1] Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., and Yang, G.-Z. Big data for health. *IEEE journal of biomedical and health informatics 19*, 4 (2015), 1193–1208.

[2] Aramaki, E., Maskawa, S., and Morita, M. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing* (2011), Association for Computational Linguistics, pp. 1568–1576.

[3] Cambridge Dictionary. https://dictionary.cambridge.org/us/topics/disease-and-illness/colds-and-flu/, 2019. Accessed December 15, 2019.

[4] Centers for Disease Control and Prevention. Flu symptoms & complications. https://www.cdc.gov/flu/symptoms/symptoms.htm, 2019. Accessed December 8, 2019.

[5] Centers for Disease Control and Prevention. U.s. influenza surveillance system: Purpose and methods. https://www.cdc.gov/flu/weekly/overview.htm, 2019. Accessed December 8, 2019.

[6] Centers for Disease Control and Prevention. Weekly u.s. influenza surveillance report. https://www.cdc.gov/flu/weekly/index.htm, 2019. Accessed December 8, 2019.

[7] Chen, M., Hao, Y., Hwang, K., Wang, L., and Wang, L. Disease prediction by machine learning over big data from healthcare communities. *Ieee Access 5* (2017), 8869–8879.

[8] Chen, Y., Crespi, N., Ortiz, A. M., and Shu, L. Reality mining: A prediction algorithm for disease dynamics based on mobile big data. *Information Sciences 379* (2017), 82–93.

[9] Chollet, F. *Deep Learning with Python*, 1st ed. Manning Publications Co., Greenwich, CT, USA, 2017.

[10] Denecke, K., and Nejdl, W. How valuable is medical social media data? content analysis of the medical web. *Information Sciences 179*, 12 (2009), 1870–1880.

[11] Elkin, L. S., Topal, K., and Bebek, G. Network based model of social media big data predicts contagious disease diffusion. *Information discovery and delivery 45*, 3 (2017), 110–120.

[12] Feldman, R., and Sanger, J. *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge university press, 2007.

[13] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature 457*, 7232 (2009), 1012.

[14] Greenwood, S., Perrin, A., and Duggan, M. Social media update 2016. *Pew Research Center 11*, 2 (2016).

[15] Grinstein, U. M. F. G. G., and Wierse, A. *Information visualization in data mining and knowledge discovery.* Morgan Kaufmann, 2002.

[16] Internet Archive. https://archive.org/, 2019. Accessed December 8, 2019.

[17] Jivani, A. G., et al. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl 2*, 6 (2011), 1930–1938.

[18] Kim, Y., Dwivedi, R., Zhang, J., and Jeong, S. R. Competitive intelligence in social media twitter: iphone 6 vs. galaxy s5. *Online Information Review 40*, 1 (2016), 42–61.

[19] Lamb, A., Paul, M. J., and Dredze, M. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2013), pp. 789–795.

[20] Lampos, V., De Bie, T., and Cristianini, N. Flu detector-tracking epidemics on twitter. In *Joint European conference on machine learning and knowledge discovery in databases* (2010), Springer, pp. 599–602.

[21] Lee, K., Agrawal, A., and Choudhary, A. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), ACM, pp. 1474–1477.

[22] Omary, Z., and Mtenzi, F. Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning. *International Journal for Infonomics (IJI) 3*, 3 (2010).

[23] Perdana, R. S., and Pinandito, A. Combining likes-retweet analysis and naive bayes classifier within twitter for sentiment analysis. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC) 10*, 1-8 (2018), 41–46.

[24] relatedwords. Words that are associated with flu. https://relatedwords.org/relatedto/flu, 2019. Accessed December 18, 2019.

[25] Sadilek, A., Kautz, H., and Silenzio, V. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012).

[26] Salathé, M., and Khandelwal, S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS computational biology 7*, 10 (2011), e1002199.

[27] Schmidt, C. W. Trending now: using social media to predict and track disease outbreaks, 2012.

[28] Signorini, A., Segre, A. M., and Polgreen, P. M. 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 (2011), e19467.

[29] Taehoon Kim and Kevin Wurster. emoji 0.5.4. https://pypi.org/project/emoji/, 2019. Accessed December 15, 2019.

[30] Twitter. Tweet objects. https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json, 2019. Accessed December 17, 2019.

[31] unicode.org. Full emoji list, v12.1. http://www.unicode.org/emoji/charts/full-emoji-list.html, 2019. Accessed December 15, 2019.

[32] Uysal, A. K., and Gunal, S. The impact of preprocessing on text classification. *Information Processing & Management 50*, 1 (2014), 104–112.

[33] Vijayarani, S., Ilamathi, M. J., and Nithya, M. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks 5*, 1 (2015), 7–16.

[34] Şerban, O., Thapen, N., Maginnis, B., Hankin, C., and Foot, V. Real-time processing of social media with sentinel: a syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management 56*, 3 (2019), 1166–1184.