Mining Twitter Data For Influenza Detection and Surveillance

Kenny Byrd
School of Computer Science
Carleton University
Ottawa, Canada
kenny.byrd@carleton.ca

Alisher Mansurov
Sprott School of Business
Carleton University
Ottawa, Canada
alisher.mansurov@carleton.ca

Olga Baysal
School of Computer Science
Carleton University
Ottawa, Canada
olga.baysal@carleton.ca

ABSTRACT

Twitter — a social media platform — has gained phenomenal popularity among researchers who have explored its massive volumes of data to offer meaningful insights into many aspects of modern life. Twitter has also drawn great interest from public health community to answer many health-related questions regarding the detection and spread of certain diseases. However, despite the growing popularity of Twitter as an influenza detection source among researchers, healthcare officials do not seem to be as intrigued by the opportunities that social media offers for detecting and monitoring diseases. In this paper, we demonstrate that 1) Twitter messages (tweets) can be reliably classified based on influenza related keywords; 2) the spread of influenza can be predicted with high accuracy; and, 3) there is a way to monitor the spread of influenza in selected cities in real-time. We propose an approach to efficiently mine and extract data from Twitter streams, reliably classify tweets based on their sentiment, and visualize data via a real-time interactive map. Our study benefits not only aspiring researchers who are interested in conducting a study involving the analysis of Twitter data but also health sectors officials who are encouraged to incorporate the analysis of vast information from social media data sources, in particular, Twitter.

Keywords

Social media, flu, data mining, cold symptoms, public health surveillance, visualization tool, sentiment analysis.

1. INTRODUCTION

With over 320 million monthly active users currently, Twitter — a popular social network — is producing the enormous volumes of data for extracting and revealing useful insights applicable to many areas of life. In Healthcare, the social media platform has already demonstrated to be an important source of information for such purposes as searching symptoms and treatments of a certain disease, predicting its spread, monitoring its volumes, and evaluating the effectiveness of taken measures against it. Studies have successfully shown that information collected from tweets of Twitter users can be used to detect the outbreak of pandemics, as well as to track the spread of seasonal influenza epidemics [6, 7]. Moreover, academics agree that influenza detection models based on information from social media sites are more effective and accurate than traditional approaches based on gathering activity data from medical documents [8, 14].

Despite the growing popularity of Twitter as an influenza detection source among researchers, public health sector practitioners do not seem as intrigued by the opportunities of leveraging the insights from the social media data to track diseases. To illustrate, FluWatch report, a service used by the Government of Canada to inform the population about the influenza detections and activity in Canada, does not include analysis of information extracted from Twitter. It relies on data retrieved from such sources as laboratory reports of positive influenza tests in Canada (National Microbiology Laboratory), sentinel physician reporting of influenza-like illness (ILI), provincial/territorial assessment of influenza activity based on various indicators, including laboratory surveillance, ILI reporting, and outbreaks, influenza-associated paediatric and adult hospitalizations, WHO, and other international reports of influenza activity [16]. Although the sources of information used by FluWatch are reliable, they may be limited in terms of scale and prediction power. Compared to the FluWatch information, Twitter data represents a more general population and can serve as a leading indicator of influenza detection since it includes tweets from the users who are not sick but experiencing the symptoms that can perhaps be treated. Therefore, the authorities may be missing valuable insights by excluding Twitter from the sources of information for influenza surveillance.

Lack of formalized tools that incorporate analysis of the data obtained from Twitter creates a need to show the potential of the social media in delivering insights that cannot be substituted by more traditional sources of data for influenza detection and surveillance. The above motivation for our work is coupled by our observation that a very limited number of studies shows the complete process of Twitter analysis for influenza detection including data mining and extraction, classification of tweets, sentiment analysis, disease detection and prediction, and real-time presentation of results for surveillance purposes. In this paper, we aim at addressing the following research questions:

1. Can Twitter messages (i.e., tweets) be reliably classified based on influenza related keywords?

SEHS'16, May 14-15, 2016, Austin, TX, USA

DOI: http://dx.doi.org/10.1145/2897683.2897693

2. Can the spread of influenza be predicted with high accuracy?

3. Can we build a real-time surveillance of influenza in selected cities?

The distinction between our paper and the existing literature is that we put together a more complete procedure combining all steps from data mining to presenting the results on a custom-made map to monitor disease cases. Moreover, we will show that FluWatch report which is used by the Government of Canada is missing lots of important components in their investigation by not incorporating Twitter data analysis in their influenza report. We believe that our study could be of great value both to aspiring researchers who are looking to learn valuable lessons of data mining, analysis, and visualization and to government officials who would like to see the potential of Twitter data in providing valuable insights when tracking the spread of influenza.

The rest of the paper is organized as follows. Section 2 discusses previous related work on the analysis of Twitter data. Section 3 describes our methodology including data collection, extraction, analysis and visualization. Section 4 presents the results of our study. Finally, Section 5 concludes the paper and discusses future research directions.

2. RELATED WORK

A number of research studies have been conducted to demonstrate the potential of Twitter data in providing valuable insights into user population by aggregating and analyzing millions of their tweets. Researchers have analyzed tweets to reveal political sentiment [**?**], measure public health trends [1], study linguistic variation [10], and detect earthquakes [19]. Studies have also shown that Twitter content can be used to predict outcomes and expectations in many areas of life. Bollen at al. [5] examined text content of daily Twitter feeds and achieved high accuracy in predicting daily up and down changes in the closing values of Dow Jones Industrial Average (DJIA). Similarly, Asur et al. [3] showed that analysis of message contents in Twitter can help to successfully forecast box-office revenues for movies.

Twitter has also drawn interest from the public health community for providing insights on general health-related information and predicting the spread of certain influenza. Vance et al. [22] investigated the pros and cons of using Twitter, along with other social media websites such as MySpace, Facebook, and Youtube, to spread public health information among young adults. They found that pros are low cost and rapid transmission, while cons can be blind authorship, lack of source citation, and presentation of opinion as fact. Ginsberg et al. [12] proposed a different approach for estimating the spread of flu and argued that the relative frequency of certain search terms are good indicators of the percentage of physician visits and established a linear correlation between the known cases of sickness and the visits. Extending their work, Lampos and Christianini [13] performed a correlation analysis for tweets including the word "flu" and related symptoms with historical data. Ritterman et al. [17] applied the same method to a more specific case of influenza tracking by combining prediction markets and Twitter to predict H1N1. Later, De Quincey and Kostkova [9] gathered Twitter messages during the H1N1 pandemic for analysis. The same approach was used by Chew and Eysenbach [7] who conducted a more thorough investigation of Twitter to reveal the general public understanding of the 2009 H1N1 pandemic. In an interesting study in terms of its approach, Scanfeld et al. [21] manually reviewed tweets that indicated incorrect antibiotic use to analyze public perception of antibiotics.

In general, strategies for detecting influenza using Twitter can be classified into the following four methods that are based on applying natural language processing (NLP) approach for analyzing content from Twitter posts:

- Supervised classification of words representing "flu" and related constructs [8];
- Unsupervised models for disease discovery using ailment methods such as Ailment Topic Aspect Model [15];
- Keyword counting, tracking geographic illness propagation [18];
- Combining tweet contents with the social network location information [2].

In this paper, we use a combination of the above mentioned approaches with the implementation of a real-time Twitter streaming search tool which is designed to fetch data at regular time intervals using a crawler. One of the previous works that utilize a real-time twitter crawler is developed by Achrekar et al. [1] who introduce Social Network Enabled Flu Trends (SNEFT) architecture as a uninterrupted data collection engine that combines the detection and prediction capability on Twitter in discovering real-time flu trends. Our approach, which we discuss in Section 3, is also based on a similar method.

3. METHODOLOGY

To address our research questions we first collected Twitter data, cleaned it up by applying various filters, performed sentiment analysis and geocoding, and developed an application to leverage Twitter data for monitoring the spread of influenza.

3.1 Data Collection

From October 27th, 2015 to November 30th, 2015, we collected all tweets in English using the Twitter Streaming API with location filters[1]. A Java application was setup to connect to the Twitter Streaming API using the Twitter4J Java library[2]. Twitter4J allows users to easily connect to the Twitter API, inserting user authentication consumer keys, and access tokens. Then by creating a status listener the user can have a real-time feed of Twitter data from the Twitter stream.

Once this was setup the Twitter filter parameters were added to the Twitter4J instance using a new filter query. Locations filters are a comma-separated list of longitude, latitude pairs specifying a set of bounding boxes. Only geolocated Tweets falling within the requested bounding boxes are included in the real-time Twitter feed[1]. The bounding boxes of Ottawa and surrounding areas were added to the filter query. To determine the bounding boxes, a Google Maps Developer tool was used to calculate the bounding box using an input address of each city[3]. This geocoding tool re-

[1]https://dev.twitter.com/streaming/overview/request-parameters/
[2]http://twitter4j.org/en/index.html/
[3]http://www.mapdevelopers.com/geocode_bounding\

turned the longitude, latitude pairs we needed to add to the filter query. An example bounding box for Ottawa would be {*-76.353916, 44.962733*}, {*-75.246598, 45.246598*}. Other city bounding boxes that were included were Toronto, Kingston, and Montreal in Ontario, Canada, as well as Syracuse, NY, USA. A few of these bounding boxes were also extended into other nearby cities, thus expanding the overall collection area as seen in Figure 1.

Once a user posted a tweet within the bounding box location the Java application connected to the Twitter Streaming API would receive the tweet and write it out to a "tweet" table in the PostgreSQL database[4]. Along with the tweet text itself, we downloaded the date and time when the tweet was published, the user who posted the tweet, the location of the user (based on their fixed profile location), and the geotag location of the exact position they tweeted from (if provided). For the 35-day period we collected tweets everyday in near real time, barring one short term disruption due to technical Internet connectivity issues. During this time period, we collected a total of 1,848,130 tweets based on our location filters, thus creating our initial dataset.

Details of the technical setup. For setting up the Twitter Stream Java application environment, a Raspberry Pi 2 Model B[5] was purchased to host the Java application. A Raspberry Pi is a mini-computer that features an extremely low power draw, small form factor, no noise, solid state storage, Internet connectivity, and HDMI output making it a very attractive small and lightweight server[6]. The device came pre-installed with Raspbian, a distribution of the highly popular Linux operating system distribution known as Debian. Raspberry Pis can have many different operating systems (OS) installed[6], yet for our research purpose Raspbian was well suited. Next, to get our Twitter Stream Java application up and running we installed the newest Java JDK, as well as the PostgreSQL database environment[7]. With this setup, we were able to collect tweets from the Twitter Streaming API.

3.2 Data Extraction

Once our initial dataset had been created, we had a large database of nearly 2 million tweets based on only our specific locations of interest. We then extracted all the relevant tweets related to influenza. This two-step process was necessary, mainly because the Twitter Steaming API does not allow users to combine multiple filters using an `AND` clause. For example, each additional filter can be applied using an `OR` logical conjunction. Therefore, if a keyword and location filter were both added, the stream would produce all tweets containing the specific keyword or any tweets found in the given location and not tweets containing the specific keyword in the given location.

One workaround for this is to perform a keyword search once the status is received directly in the Twitter Stream Java collection application and filter tweets prior to writing them to the database. This would be necessary if disk space constraints were in place, but since the PostgreSQL database table was of a minimal size (approximately 150 Mbs) even with nearly 2 million tweets this was not a concern. Another benefit of not filtering by keywords before writing to the database is creating a more general tweet database that can be used in future research.

To begin reducing noise in the dataset and filtering the tweets, we first performed a few basic SQL queries to search for tweets containing only the keyword "sick". After analyzing these tweets, we determined that we could best filter the data and extract the most relevant tweets by using three keywords such as "sick", "cough", and "flu". By doing this, we reduced our initial dataset down to 4,696 relevant tweets that contained 81 unique locations, and ultimately provided us with our working dataset. The working dataset includes 318 tweets with the exact geolocation tags. The most popular user locations are Ottawa with 438 tweets, Toronto with 1,554, Montreal with 238 and Syracuse, NY with 1,532 tweets, respectively.

3.3 Sentiment Analysis

Each tweet in the working dataset needed to be classified into one of three sentiment polarities: *positive*, *negative*, or *neutral*. Example tweets expressing these are: "*my doctor gave me that flu shot and I can tell my body is working right now*" – rated as positive, "*When flu got you so sick you feel like ya dying but gotta keep moving on, smile and wave. I feel like shit lol SOS*" – rated as negative, and "*@free981 Flu shot for kids has nasal spray option now: Ontario Medical Officer Dr. William's flu advice: https://t.co/uC6x9kr7El*" – rated as neutral. Since the dataset of 4,696 was too large to classify manually within a reasonable time frame, we used a machine learning approach to identify the expression of sentiment in tweets. This process involved choosing a machine learning algorithm that would maximize the accuracy of the sentiment analysis of each tweet.

We compared three standard classification algorithms: Naive Bayes, Maximum Entropy, and a Dynamic Language Model classifier. The Naive Bayes classifier was implemented using Stanford CoreNLP[8] — a suite of core NLP tools, which offers an integrated toolkit with a good range of grammatical analysis tools, fast and reliable analysis of arbitrary texts, and has the overall highest quality text analytics currently available. The Maximum Entropy classifier was implemented using the Apache OpenNLP machine learning toolkit[9], and the Language Model classifier was implemented using the LingPipe toolkit[10].

Machine learning approaches, regardless of the algorithm used, all require a training dataset. The Stanford CoreNLP library allows developers to easily load existing training models using a Maven dependency in any Java Maven project[11]. Anyone can contribute to improving these models used for the sentiment analysis by using a simple web application[12]. The training dataset used for sentiment analysis using the OpenNLP document categorizer[13] was a user generated input flat file, containing 100 tweets, each having the category

_box.php/
[4]http://www.postgresql.org/
[5]http://www.amazon.ca/Raspberry-Pi-100437-Model-1GB/dp/B00T2U7R7I/
[6]http://www.zdnet.com/article/raspberry-pi-11-reasons-why-its-the-perfect-small-server/
[7]http://raspberrypg.org/2015/06/step-5-update-installing-postgresql-on-my-raspberry-pi-1-and-2/
[8]http://stanfordnlp.github.io/CoreNLP/
[9]https://opennlp.apache.org/
[10]http://alias-i.com/lingpipe/
[11]http://mvnrepository.com/artifact/edu.stanford.nlp/
[12]http://nlp.stanford.edu:8080/sentiment/labeling.html/
[13]http://technobium.com/sentiment-analysis-using-opennlp-document-categorizer/

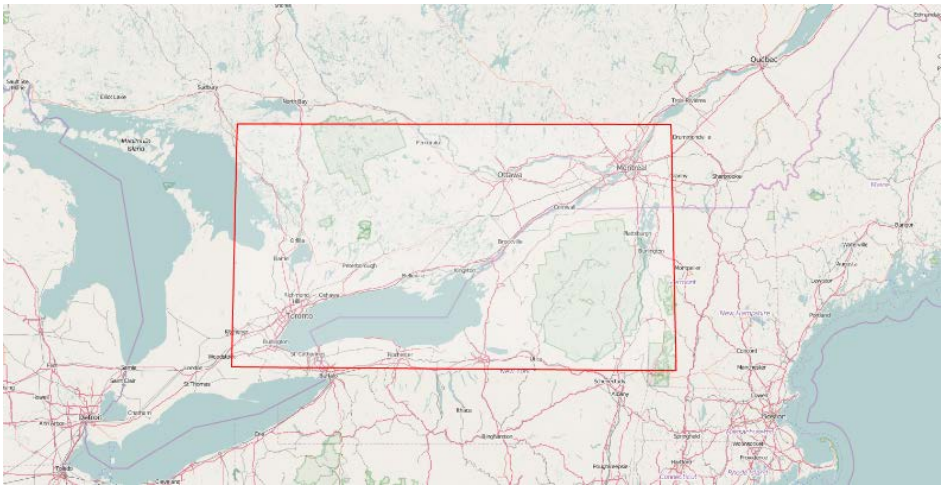

Figure 1: Overall bounding box collection area used for Twitter Streaming API.

(1 for positive and 0 for negative) and the tweet text. LingPipe[14], similar to the Stanford NLP library, also has readily available training models that can be loaded in by adding them to a subdirectory in the Java project folder.

To evaluate the best classifier, we looked at the percentage of 100 randomly selected tweets from the working dataset that the classifiers could rate accurately. A tweet sentiment was considered accurate if the sentiment polarity predicted by the classifier matched the opinion assigned by a manual rating of the sentiment. The randomly selected tweets were selected using the random number generator[15] that offers true random numbers that are determined using atmospheric noise. Using the minimum value of 1 and the maximum number of tweets in the working dataset (4,696), the random number generator was used to generate 100 unique tweet numbers. The tweet number was then looked up in the working dataset and then manually evaluated to determine its sentiment.

The Stanford CoreNLP sentiment analysis tools produced the highest accuracy of determining the sentiment polarity from this subset of tweets, with an accuracy of approximately 70%. Since the Stanford CoreNLP library has the overall highest quality text analytics currently available as previously described, we decided to use it as our classifier. The classifier would have scored an even higher accuracy rating but the randomly selected subset contained an abnormally large amount of vague and short tweets that made classification difficult. In particular, since the working data set was filtered by one of the keywords "sick" which would usually result in an overall negative sentiment, but when given a short tweet it may be hard to determine what the tweet is actually expressing. For example, the tweet *"@ambiguousoup sick yak dude"*, contains the keyword "sick" but is very vague, and may be slightly difficult for even a manual classification depending on the meaning of the word "yak" contained in the tweet.

We achieved the highest accuracy for sentiment analysis by pre-processing the text before having the machine learning algorithms predict the sentiment. The challenge of trying to classify tweets is related to the fact that each tweet can only have a maximum of 140 characters. This limitation encourages slang and otherwise poorly written phrases within the body of a tweet. Thus, most tweets contain spelling and grammar errors as well as emoticons. This general lack of context in a single tweet combined with poorly expressed sentiment means that it is unreasonable to expect a 100% accuracy from the machine learning algorithms [**?**] or even by the manual rating. Pre-processing is necessary because it can extract relevant features while sanitizing inputs and providing a more lexical input which consequentially improves performance.

Pre-processing methods used include detecting URLs, removing stop words, stemming words, identifying negative words, and removing erratic casing of letters. To remove common stop words (most common words used in a language) such as "a, an, and, of, he, she, it", we used Apache Lucene's StopFilter[16] and its English stop words list. Words were stemmed using Apache Lucene's PorterStemFilter[17]. For repeating letters in words, when two identical letters were detected, both were kept. If more than two identical repeating letters were detected, they were replaced with one letter. Negative words that were useful for sentiment analysis (e.g., nothing, never, none, cannot) were substituted with "negtoken". We also removed all words that contain non-letter symbols and punctuation. This included the removal of `@` and `http links`. Retweets were also removed. The last few simple string replacements were done using regular expressions in Java.

In this section we have not only described our sentiment analysis process, but also demonstrated the steps taken to ensure high accuracy of the process. As the results show, the the performed checks are efficient in reliably classifying Twitter messages based on related keywords.

3.4 Geocoding

Twitter allows users to publish their location manually to his or her profile. This location string does not have to be an actual location and accepts any entry. This location remains

[14]http://alias-i.com/lingpipe/web/models.html/
[15]https://www.random.org/
[16]https://lucene.apache.org/core/4_0_0/analyzers-common/org/apache/lucene/analysis/core/StopFilter.html/
[17]https://lucene.apache.org/core/4_0_0/analyzers-common/org/apache/lucene/analysis/core/StopFilter.html/

intact and does not change based on the user's actual GPS location. When users tweet, they have the option to include their precise GPS location from their GPS enabled device via iOS or Android applications. If users are tweeting using their web browsers and have enabled location services, they can attach a location (such as a city or neighbourhood) of their choice to the tweet[18]. Many users do not post tweets with an accurate location revealing their current physical location due to various privacy concerns. Even photos taken on a user's mobile device and uploaded to Twitter can contain hidden location data [11]. For these reasons, the majority of tweets in our working dataset did not contain a precise location (only 318 out of the 4,696).

For the location data to be useful to us, we needed to resolve the user's profile location strings into informative locations. Our goal for this study was to resolve these locations strings to the city level. Since our working dataset contained 81 unique location strings we chose to resolve the locations of only the tweets that fell within one of the four most popular tweeting cities found in our dataset (Ottawa, Toronto, Montreal, and Syracuse). These locations are known to be actual locations and this allowed us to easily resolve 3,762 out of the 4,696 tweets (80%) to an exact location (excluding the already geotagged tweets). To resolve these unique location strings we used the Google Maps API geocoder[19] to find the latitude and longitude coordinates for each city.

In order to have distinct coordinates for each tweet so that all the tweets, for example in Ottawa, did not cluster into the exact spot in the mapping application, we applied the following technique. For each tweet that contained a geotag resolved by the Google Maps API, the bounding box for that location string was calculated. Using the maximum X and Y coordinates of the bounding box a random set of coordinates was generated using the following equation: $x = rand()\,(maxX - minX) + minX, y = rand()\,(maxY - minY) + minY$, where x, y are the latitude and longitude points and $rand()$ is the random number generator function in Java[20]. The generated set of coordinates was then tested to see if it was contained within the bounding box[21].

3.5 Client-Server Mapping Application

To visualize our data, we chose to implement a web based application with the following distinguished features. First, it comes with a graphical user interface that visualizes the Twitter user information using a mapping context. Second, it provides a set of filters allowing users to focus on specific keywords, mined Twitter information with exact geolocation tags, and specific tweet sentiments. Users are able to switch between various base maps from multiple data sources such as Esri's Light Gray, National Geographic maps, Bing Maps, and OpenStreetMap. Users can also enable their desired filters. The web application will query the server, process the response and then populate the mapping application with the Twitter data matching the query. Users are then able to pan and zoom the map to view tweets in any populated area. Once a user clicks on a tweet icon, a pop-up will be shown that displays the Twitter user name who published the tweet, the data and time the tweet was published, the user's current Twitter profile picture, the location tag from the user's profile and the overall sentiment rating of the tweet body.

The web mapping application was implemented using OpenLayers 3, a free, Open Source pure JavaScript library for displaying map data in most modern web browsers. It has no server-side dependencies, and is constantly being improved by the open source community. It allows developers to display map tiles from a variety of data sources, implements client-side tile caching for superior rendering improvement and can be vastly customized and extended to meet user needs. The web mapping application communicates with a Java server that contains a connection to the tweet database via AJAX calls to a REST endpoint to process the users query. For the purpose of our demo the Java server retains a cached list of the working dataset of tweets to optimize queries. The server will simply parse the request from the client side mapping application, and return the matching tweets as a JSON string. The web mapping application then processes the JSON string and creates map features from them which contain the information that is shown when each feature's icon is selected.

Future improvements to allow a real-time feed of Twitter data can be added easily. Using the Twitter Stream Java collection application, tweets can be received and then written to the database. Once these tweets are stored in the database, they can be available and searchable by the web application using the various user filters. A date and time filter could then be added to the client side mapping application to allow users to specifically query a set of tweets given the specified data range. A time-slider could also be implemented to allow users to quickly visualize the spread of influenza on a day-to-day basis.

Table 1: Descriptive statistics of the tweets.

	# of tweets	Percentage
Total tweets collected	1,848,130	100%
Tweets extracted using the keywords	4,696	0.25%
Tweets with exact geotags	318	0.02%
Ottawa, ON	438	0.02%
Toronto, ON	1,554	0.08%
Montreal, ON	238	0.01%
Syracuse, NY	1,532	0.08%

4. RESULTS

In this section we evaluate our methodology, present our findings, and discuss insights learned. We begin by presenting the results for extracted tweets and then seek answers to our research questions.

Table 1 displays the descriptive statistics for the number of tweets obtained in each step of the data extraction and classification process. It also presents the distribution of tweets across cities showing that the number of tweets is considerably higher in Syracuse and Toronto than in Montreal and Ottawa mostly due to the larger population in the first two cities. Figure 2 demonstrates that tweets conveying negative sentiment prevail those of positive one by almost three times indicating that tweets containing the selected keywords are more likely to be posted by users who are actually sick.

Our research questions that we defined in Section 1 are

[18]https://support.twitter.com/articles/78525/
[19]https://developers.google.com/maps/documentation/geocoding/intro/
[20]http://stackoverflow.com/questions/21205273/randomly-generating-coordinates-inside-a-bounded-region/
[21]https://github.com/sromku/polygon-contains-point

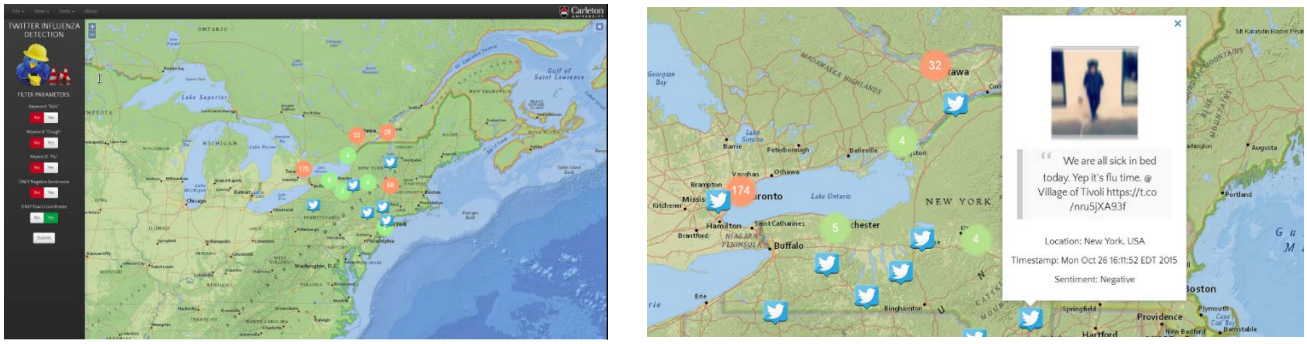

Figure 3: Screenshots of our application for monitoring the spread of influenza by leveraging Twitter data.

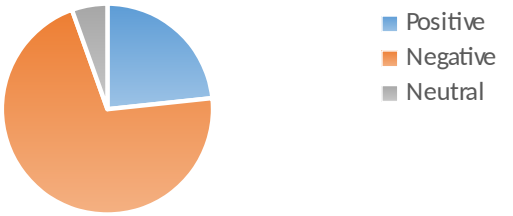

Figure 2: Sentiment analysis: the volume of tweets conveying positive, negative and neutral sentiment.

related to each other and their order is associated with their logical sequence. Therefore, we first attempt to answer whether self-reported Twitter messages containing information about users' common cold, influenza and general symptoms of not feeling well can indicate whether or not the users were affected by the influenza. As described in Section 3 we obtained 4,696 relevant tweets by filtering tweets containing the following three keywords: "sick", "cough", and "flu". Manual review of some randomly selected tweets shows that Twitter messages containing those words can indeed be used to determine users' influenza status. For example, @Athagaross stated - *"This flu has me breathing like Darth Vader"*; @duncaninla writes - *"We are all sick in bed today. Yep, it's flu time"*; or, Cerise_Chua shares *"Like come on life... I JUST got over the flu and I'm just able to breath again... Really had to do this to me already??"*. It can be clearly observed from the above statements that Twitter messages contain information about whether the user is sick or not. Some tweets, such as *"Thanks flu like symptoms for depriving me of my eating rights"* by @_sbest, do not show whether the user is sick or not but are used in predicting the spread of the influenza. There are also tweets that contain one, two or all three of the selected keywords yet do not give any indication of whether or not the user is affected by the influenza. Even though such tweets have less relevant information content for the purposes of our study, we keep them in our final sample because they may be still useful in predicting the spread of the influenza.

Next, to determine whether Twitter messages be reliably classified based on influenza related keywords, we evaluated the reliability of our sample selection process and assess the accuracy of sentiment analysis techniques. Our sample selection reliability check is based on the level of rigour used when deriving the final sample of 4,696 collected tweets from 1.85 million tweets in the initial dataset. Similar studies with a time-frame between 4–8 weeks generally collect 2–4 million tweets to produce 1 or 2% of the tweet for further analysis. In our case this ratio is only 0.25% primarily due to more stringent data cleaning process which assured high relevance of our working dataset. Another important factor that we used to determine the reliability of our tweets classification process is the manual accuracy comparisons of three standard classification algorithms. As we discussed in Section 3, the highest accuracy of 70% was obtained when using Stanford CoreNLP sentiment analysis tools. We believe the obtained level of accuracy along with the rigorous sample selection procedure can warrant the high reliability of tweets classification based on influenza related keywords.

Finally, we determine whether or not it is possible to monitor, in real-time, the spread of influenza in selected cities. As mentioned in Section 3, using such tools as Twitter4J, the Twitter Streaming API, a Raspberry Pi 2, and PostgreSQL, one can specify a set of bounding boxes for the area of interest and extract tweets that meet the selection criteria in real-time. Then, for each tweet in our final sample, using the Google Maps API and tweets' geotags, we showed that it is possible to determine the coordinates containing within the bounding box (see Figure 3 (left)). In the last step, we showed how one can visualize the data tweets data using OpenLayers 3[22] — a JavaScript library to load, display, and render maps from multiple sources. Figure 3 (right) presents a snapshot of our client-server mapping application which shows the distribution, count, description of tweets in the selected locations. The dynamic map updates information in real time and can display information based on selected parameters such as a selected keyword, sentiment, and coordinate. Actual tweets can be seen by hovering the mouse on any of the boxes containing the Twitter symbol.

5. CONCLUSION AND FUTURE WORK

In this paper, we have been able to demonstrate the possibility of Twitter influenza spread surveillance in a given time and area. We have proposed an approach with a set of methods which can be employed to efficiently extract data from Twitter, classify tweets based on their sentiment characteristics, identify Twitter users who are affected by the influenza in selected cities, and display the results on a data visualization tool. The workflow and tools used in our analysis can be of great benefit to future researchers who are interested in deriving meaningful insights from aggregate information contained in Twitter messages. We see the main contribution of this paper in showing that Twitter can be

[22]http://openlayers.org/

reliably used to predict and monitor influenza spread by the government organizations in Canada. Since our motivation stems from the lack of formalized tools that incorporate analysis of data obtained from Twitter, we set the goal to show the potential of the social media analytics in delivering insights that cannot be substituted by traditional sources of data for influenza. To achieve this goal, we first demonstrated that Twitter messages can be reliably classified based on influenza related keywords. Then, we showed the results of our tests that the spread of influenza can be predicted with high accuracy. Finally, we illustrated that there is a reliable way of real-time influenza surveillance in predetermined cities. The robustness checks performed in our analysis ensure that the results obtained are both valid and reliable. For example, to achieve the highest accuracy for sentiment analysis, we pre-processed the text before having the machine learning algorithms predict the sentiment. Even though the topic selected in our study was influenza, our approach may be applicable to a wide range of diseases and syndromes. Therefore, we encourage government specialists to incorporate the analysis of vast information from social media sources, particularly Twitter, in their influenza spread report FluWatch.

There are two directions of research on the topic that future studies should further investigate. First, predicting influenza spread with high accuracy still remains as one of the areas that require attention. Future studies should look into ways of increasing predictive power of existing models to deliver better forecasts and timely response to possible outcomes. Second, prior studies, including this work, have been limited to analyzing disease spread in only selected cities. Future research should include the analysis of more territories and counties to provide valuable insights about influenza spread across the continent or worldwide.

6. REFERENCES

[1] Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., and Liu, B. Predicting flu trends using twitter data. In *IEEE Conference on Computer Communications Workshops* (2011), pp. 702–707.

[2] Asta, D., and Shalizi, C. Identifying influenza trends via twitter. In *NIPS Workshop on Social Network and Social Media Analysis: Methods, Models and Applications* (2012).

[3] Asur, S., Huberman, B., et al. Predicting the future with social media. In *Proc. of the International Conference on Web Intelligence and Intelligent Agent Technology)* (2010), vol. 1, pp. 492–499.

[4] Barbosa, L., and Feng, J. Robust sentiment detection on twitter from biased and noisy data. In *Proc. of the 23rd International Conference on Computational Linguistics* (2010), pp. 36–44.

[5] Bollen, J., Mao, H., and Zeng, X. Twitter mood predicts the stock market. *Journal of Computational Science 2*, 1 (2011), 1–8.

[6] Chen, L., Achrekar, H., Liu, B., and Lazarus, R. Vision: towards real time epidemic vigilance through online social networks: introducing sneft–social network enabled flu trends. In *Proc. of the 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond* (2010), p. 4.

[7] Chew, C., and Eysenbach, G. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one 5*, 11 (2010), e14118.

[8] Culotta, A. Detecting influenza epidemics by analyzing twitter messages. *arXiv preprint arXiv:1007.4748* (2010).

[9] de Quincey, E., and Kostkova, P. Early warning and outbreak detection using social networking websites: The potential of twitter. In *Electronic healthcare*. Springer, 2010, pp. 21–24.

[10] Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. A latent variable model for geographic lexical variation. In *Proc. of the Conference on Empirical Methods in Natural Language Processing* (2010), pp. 1277–1287.

[11] Friedland, G., and Sommer, R. Cybercasing the joint: On the privacy implications of geo-tagging. In *HotSec* (2010).

[12] 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), 1012–1014.

[13] Lampos, V., and Cristianini, N. Tracking the flu pandemic by monitoring the social web. In *2nd International Workshop on Cognitive Information Processing (CIP)* (2010), pp. 411–416.

[14] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. From tweets to polls: Linking text sentiment to public opinion time series. *Proc. of the International AAAI Conference on Web and Social Media 11*, 122-129 (2010), 1–2.

[15] Paul, M. J., and Dredze, M. You are what you tweet: Analyzing twitter for public health. pp. 265–272.

[16] Reyes, F., Aziz, S., Macey, J., Li, Y., Winchester, B., Zabchuk, P., Wootton, S., Huston, P., and Tam, T. Influenza in canada. *Canada Communicable Disease Report 33*, 09 (2007).

[17] Ritterman, J., Osborne, M., and Klein, E. Using prediction markets and twitter to predict a swine flu pandemic. In *1st International workshop on mining social media* (2009), vol. 9.

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

[19] Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of the 19th international conference on World wide web* (2010), pp. 851–860.

[20] Salathe, M., and Khandelwal, S. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Comput Biol 7*, 10 (10 2011), e1002199.

[21] Scanfeld, D., Scanfeld, V., and Larson, E. L. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control 38*, 3 (2010), 182–188.

[22] Vance, K., Howe, W., and Dellavalle, R. P. Social internet sites as a source of public health information. *Dermatologic clinics 27*, 2 (2009), 133–136.