# Machine Learning Assisted Location-based Tweet Approach for Public Health Surveillance

Samip Ghimire
Dept. of Computer Engineering
Nepal College of Information Technology
Kathmandu, Nepal
ghimiresamip91@gmail.com

Bhawana Poudel
Dept. of Information Technologies
NMB Laghubitta Bittiya Sanstha Ltd.
Pokhara, Nepal
bhawana.pudl@gmail.com

Bidur Devkota
Dept. of Information and Communications Technologies
Asian Institute of Technology
Khlong Luang, Thailand
bidur.devkota@gmail.com

*Abstract*— **Rapid response to a health epidemic is critical to reducing the loss of life. Existing methods mostly rely on expensive surveys of hospitals across the country, typically with lag times of one to two weeks for influenza reporting, and even longer for less common diseases. Alternatively, we can analyze social networking sites like Twitter where the public broadcast a wealth of data along with their location. With the increasing accessibility of location-based services, various location-based applications like public health surveillance are becoming a major topic. Here, we purpose a method for disease surveillance by utilizing the public opinions shared on Twitter. We concentrate on a particular problem: how to extract influenza outbreak information from Twitter on a location basis. We propose a method by integrating 'Naive Bayes' with the 'bag-of-words' scheme. This study attempts to uncover the outbreak of an epidemic occurrence during its early phase.**

***Keywords—Influenza, Naïve Bayes, Text classification, Twitter.***

## I. INTRODUCTION

Twitter [1] is a massive social networking site tuned towards fast communication. A large number of active users publish over 500 million 280 characters "Tweets" every day [1]. The feature that Twitter avails to create and share ideas and information instantly without barriers made it an important communication medium for people. Twitter has played a prominent role in socio-political events, such as the Arab Spring [2] and the Occupy Wall Street movement[3]. It has also been used to post damage reports and disaster preparedness information during large natural disasters, such as Hurricane Sandy. As a result, Twitter's data has been coveted by both computer and social scientists to better understand human behavior and dynamics.

The traditional approach employed for flu surveillance includes the collection of Influenza-like Illness (ILI) patients' data from sentinel medical practices. However, it takes time to collect and process data, and there is usually 1-2 weeks' time lag before the data becomes available. Early detection of a disease outbreak is critical because it would allow faster communication between health agencies and the public and provide more time to prepare a response. Hence, this study attempts to examine the applicability of freely available tweets to build an early warning system using state-of-the-art text classification algorithm. The proposed framework is based on the bag-of-words model and Bayes text classification. Oftentimes there are questions regarding the representativeness of streaming tweet samples. A study by Morstatter et.al. titled" Is the sample good enough? Comparing data from Twitters streaming API with Twitters firehose"[4] explored this problem. The results of the investigation advocated that the free Streaming m the type of analysis that is to be performed. It was uncovered that the Streaming API provides a workable set of geotagged tweets.

In this paper, we proposed a more general approach that discovers many different ailments associations from tweets. Our first contribution is to create a data set of influenza-related tweets. For this, we use different keywords related to flu: flu, cough, common cold, sneeze, influenza, fever, runny, stuffy and sore. Our method uses Naïve Bayes, a supervised machine-learning algorithm to separate out flu infection relevant tweet from non-relevant tweets. We exploited the bag-of-words model to classify the tweet as influenza relevant or not. Moreover, the geotagged tweet helps to find the location of the outbreak.

The selection of Twitter in our study is made because it has a simple and well-defined public interface for collecting data. Most of the Twitter data is public in nature unlike other popular social networking sites like Facebook. Cesare et.al. [5] investigated and studied over 60 existing research literature on social media sites. About thirty-nine studies (i.e. 65%) used Twitter and only two on Facebook and two or less on other sites. This is

1 Twitter.com

because different characteristics like message size, metadata, availability, and accessibility have helped Twitter to gain popularity as an important platform of study in the academic community [6].

## II. Related Works

There have been many types of research which utilize twitter data to understand real-world scenarios like hurricane-related studies [3], earthquake studies [7], box-office revenues forecasting [8], public health studies. Several existing influenza surveillances have been done. Google Flu Trends [9] uses google online search queries related to flu-like symptoms to track influenza activities.

D. A. et. al. [10] used a "health" stream, which downloads only tweets containing any of 269 health-related keywords they provided. They used a staged-approach to data filtering. They used binary classification models to identify relevant data for influenza surveillance at each stage. These models indicated whether tweets were relevant to health, relevant to influenza, and indicative of an actual infection. For location filtering, they used GPS information. In addition, they utilize information from the users' public biographic profiles. Novel flu surveillance system was proposed by Lee et. al. [11] that uses twitter data to track U.S. influenza activities in real-time. The system consists of four stages: Data Collection, Data Pre-processing, Data Modelling, and Data Visualization. The data-collecting module continuously downloads flu-related public twitter data using the Twitter streaming API. The pre-processor module extracts tweet texts, time stamps, and user locations and stores them in a database for further analysis. There were three models in the data modelling stage: geographical model, text model, and temporal model. Work [12] done in North-western University in Chicago which provide a real time digital flu surveillance using twitter. It does not provide any analysis of data, instead the past 50 days' tweets are digitally visualized by filtering different type of flu related keywords that are located in USA. It provides Daily Flu Activity, U.S Flu Activity Map, live Tweet Stream, Flu Types, Symptom, Treatments and Most Frequent Words.

S. Doan [13] developed a novel filtering method for Influenza-Like Illnesses (ILI)-related messages using 587 million messages from Twitter micro-blogs. They first filtered messages based on syndrome keywords from the Bio Caster Ontology, an extant knowledge model of nonprofessionals' terms. They then filtered the messages according to semantic features such as negation, hashtags, emoticons, humor, and geography. Signorini et. al. [14] collected and stored a large sample of public tweets beginning April 29, 2009, that matched a set of pre-specified search terms: flu, swine, influenza, vaccine, Tamiflu, oseltamivir, zanamivir, Relenza, amantadine, rimantadine, pneumonia, symptom, syndrome, and illness. Real-time flu and cancer surveillance system [15] consists of all recent tweets that mention the keywords 'flu' or 'cancer'. They have collected over 6 million flu-related tweets generated by more than 3.3 million unique users in the 5.5 months from October 16, 2012. They were interested in investigating the popularity of terms used in three categories: (1) disease types (2) symptoms (3) treatments, and have created a keyword list for each category. Alessa et. al. [16] presented an efficient early warning system that utilizes tweets in order to track disease outbreaks. A weekly flu prediction system was developed using text classification, tweet mapping, and linear regression model. The text classification module utilized sentiment analysis and keyword occurrences. The text classifier was trained using a prelabeled flu-related and unrelated tweet dataset. Next, the flu-related tweets were mapped and passed together with historical official data to a linear regression module for making weekly flu rate predictions. The overall framework demonstrated good efficiency and worked accurately even for new kinds of disease outbreaks. Lu et. al. [17] investigated the problem of monitoring and forecasting influenza using internet data sources at the city level. They proposed an ensemble-based approach for combining information from models based on heterogeneous data sources so as to nowcast and forecast influenza-related activity in the Boston region. This study presented an accurate and near real-time workflow that could predict influenza occurrence ahead of the Boston Public Health Commission reports.

## III. Background

Multinomial Naive Bayes is a specialized version of Naive Bayes that is designed more for text documents. This is a supervised learning technique, in which every new document is classified by assigning one or more labels from a fixed set of predefined classes. A simple Naïve Bayes would model a document as the presence and absence of particular words, however, multinomial Naïve Bayes explicitly models the word counts and adjusts the underlying calculations. In this approach, a document is an ordered sequence of word events, drawn from the same vocabulary V [18]. In the proposed method, we assume that the lengths of the document are independent of class. Similarly, a Naive Bayes assumption is made such that the probability of each word event in a document is independent of the context and position of the word in the document.

The selection of Naive Bayes is due to its simple and easy implementation, it can work well with small dataset and it is less prone to noise [19]. Other algorithms like decision tree is more prone to

overfitting, random forest is less prone to overfitting but needs large dataset [20]. Neural network requires complex computation and also needs large dataset [21].

In bag-of-word model, a text such as a sentence or a document is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity. It is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier [22]. In a bag-of-words approach for flu classification, each tweet is modeled as an unordered collection of words selected from one of two probability distributions: one representing influenza relevance 'flu' and one representing non-relevance ("noflu"). The set of words is generated from all tweets downloaded to the dataset. One bag is filled with words found in flu-related tweets, and the other with words found in non-related tweets. While any given word is likely to be somewhere in both bags, the "flu" bag will contain flu-related words significantly more frequently, while the "noflu" bag will contain more words related to the non-relevant data.

## IV. PROPOSED METHOD

We propose a supervised classification model that separate tweets indicating influenza infection from those that does not indicate influenza. The overall procedure for analyzing and processing the twitter stream is illustrated in Fig. 1. The proposed method uses the Multinomial Naïve Bayes and bag-of-word approach for the classification. Geotagged tweets are further used for understanding the occurrence of influenza in different locations.

At first, the tweet stream is collected from free Twitter API, which has availed twitter data to external parties for developing products and services. Table I lists the different available methods for obtaining the twitter data. These API provides tools to access ongoing "streams" of publicly shared opinions at no cost. Thus, such streams are collected using Tweetinvi c# library[2] in the proposed method. The tweet payload contains various information like tweet timestamp, user identifier, tweet identifier, tweet and user location data, etc. Important data like tweet text, timestamp, and location coordinates i.e. longitude and latitude are extracted from the tweet payload stored for further processing. Any tweet which does not contain explicit location coordinates is ignored.

A subset of the tweets thus retrieved is manually classified as relevant or non-relevant. The part of the sentiment analysis for determining the relevancy of the tweet is performed in this stage. In order to apply machine learning algorithms, these tweets are then preprocessed, i.e., tweets are tokenized, stemmed, filtered and are stored for training. The first step of the preprocessing known as tokenization is the process of breaking down a text corpus into individual elements that serve as input for various natural language processing algorithms [23]. Fig. 2. shows an example of a simple but typical tokenization step that splits a sentence into individual words, removes punctuation, and converts all letters to lowercase. Thus, in the proposed method, tweets are first broken down into individual elements, as shown in Fig. 2. The tweet 'got terrible flu' is divided into individual elements, 'got', 'terrible', and 'flu', which is thus a tokenization process. In the next step, the stemming which is developed by Martin F. Porter in 1979 [19] is computed, where each word in the tweet is transferred to its root form. Taking the same figure as an example, the word 'got' is transferred to its root form 'get'. Thus, with a similar procedure, all the tweets downloaded are thus breakdown into individual elements, after which each word is transformed into root-form known as stemming.

After the preprocessing step, Multinomial Naïve Bayes is applied to each tokenized and stemmed (relevant and non-relevant) tweets for which the bag-of-word model is exploited in the proposed method. As described in section III, the bag of words contains two bags representing words and their frequency count within each relevance and non-relevance tweets, thus forming corresponding relevance and non-relevance bags. Thus, a single bag contains each word corresponding to the (relevant or non-relevant) tweets and the frequency of that word within that bag. The Bayesian flu filter assumes that the tweet is a pile of words that have been poured out randomly from one of the two bags, and uses Bayesian probability to determine which bag it is more likely to be in.

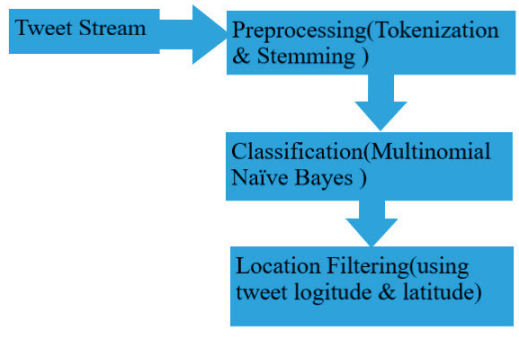


Figure 1. System Architecture

2 https://github.com/linvi/tweetinvi

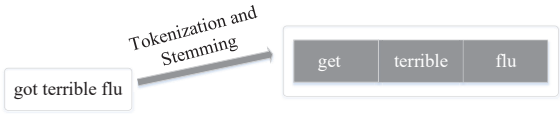


Figure 2. Text tokenization and stemming.

TABLE I
OVERVIEW OF TWITTER APIS

| API | Description |
|---|---|
| Free Streams | 1% of tweets freely available worldwide. |
| Paid Streams | Firehose data, i.e. limitless data streams, purchasable via providers like GNIP (https://gnip.com/). |

Thus, for the tweet classification during testing, firstly the prior probabilities of each bag are calculated, which are estimated from the training corpus. Then the conditional probability for each word of the tweet occurring in both bags are estimated. At the final stage, the posterior probability for each tweet is computed from which the bag with higher posterior probability is selected as the required class.

Prior probabilities of the relevant class

$$P(r) = \frac{N_r}{N} \quad (1)$$

Prior probabilities of non-relevant class

$$P(nr) = \frac{N_{nr}}{N} \quad (2)$$

Here $N$, $N_r$, and $N_{nr}$ denote number in tweets, number of tweets on relevant class, number of tweets in non-relevant class, respectively. To classify a new tweet, the conditional probability of token (or word) of tokenized tweet given a relevant class or non-relevant class is calculated.

Likelihood of 'w' given relevant class:

$$P\left(\frac{w}{r}\right) = \frac{count(w,r)+1}{count(r)+V} \quad (3)$$

Likelihood of 'w' given non-relevant class:

$$P\left(\frac{w}{nr}\right) = \frac{count(w,nr)+1}{count(nr)+V} \quad (4)$$

The posterior probability of tweet ($t$) being of the relevant class:

$$P\left(\frac{r}{t}\right) = p(r) * \prod_1^n P\left(\frac{w_n}{r}\right) \quad (5)$$

Posterior probability of tweet (t) being of non-relevant class:

$$P\left(\frac{nr}{t}\right) = p(nr) * \prod_1^n P\left(\frac{w_n}{nr}\right) \quad (6)$$

Here, $t$ refers to tweet, $w$ refers to the word of the tweet, $n$ refers to a number of words in tweet and $r$, $nr$ refers to relevant and non-relevant classes and $V$ refers the vocabulary count. Finally, the class of tweet can be found out by comparing posterior probabilities. The new tweet $t$ is labeled with a class that achieves the highest posterior probability.

Furthermore, tweet with the location feature allows users to add location information. The users who choose to add a location to their tweets will be able to add their location information to new tweets that they post. These tweets are known as geotagged tweets, which gives the location of the user.

## V. EXPERIMENTAL RESULTS

We collected 143,325 public tweets for testing data from January 17, 2019, to January 25, 2019, that contains some keyword related to flu: flu, cough, common cold, sneeze, influenza, fever, runny, stuffy and sore. The number of tweets used for training was 573,300. The proportion of training dataset to testing dataset was 80:20. Out of these testing tweets, 5666 were geotagged. These tweets were classified into flu relevant and non-relevant tweets using Naïve Bayes. After analysis on tweets using Naïve Bayes, we found 72343 tweets that were relevant to the flu. Out of 72343, 2846 tweets were geotagged as shown in Table II.

It is found that the highest number of flu relevant tweet was recorded on the 18th of January and lowest on 22 January. The daily distribution of tweets is shown in Table III. Twitter location feature gives longitude and latitude of a user-provided with a tweet. This information is available only if a twitter user selects his/her location to be visible to the public in his/her twitter profile. Thus, geotagged tweet gives the location of flu-infected people. The trend of user activity on twitter with geotagged tweets helps to surveillance the flu infection on the region.

TABLE II
TWEETS DOWNLOADED OVER A PERIOD OF TIME

| | |
|---|---|
| Total numbers of tweets | 143325 |
| Tweets with location | 5666 |
| Total Tweets relevant to flu | 72343 |
| Tweets relevant to flu with location | 2846 |

TABLE III
GEOTAGGED TWEETS DOWNLOADED DURING THE STUDY PERIOD

| Date | No. of tweets | No. of geotagged tweets |
|---|---|---|
| 17th January | 9555 | 357 |
| 18th January | 14862 | 572 |
| 19th January | 9574 | 391 |
| 20th January | 8530 | 313 |
| 21st January | 12695 | 469 |
| 22nd January | 4041 | 215 |
| 23rd January | 4492 | 163 |
| 24th January | 4373 | 177 |
| 25th January | 4221 | 189 |

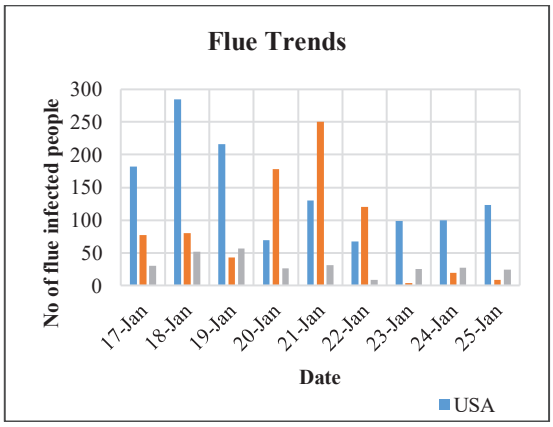


Figure 3. Comparison of several people infected from the flue on a different location on a different date.

TABLE IV
FLU RELEVANT TWEETS IN DIFFERENT COUNTRIES

| Country | No. of flu relevant Tweets |
|---|---|
| USA | 1271 |
| Indonesia | 779 |
| United Kingdom | 279 |
| Brazil | 96 |
| Malaysia | 79 |

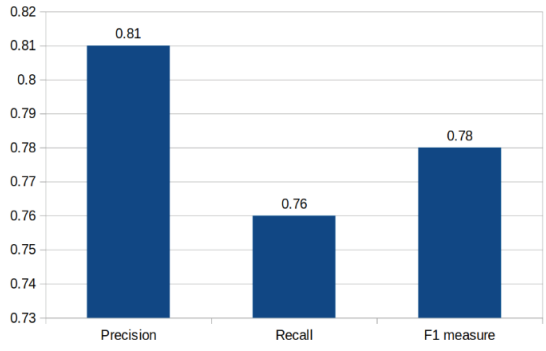


Figure 4. Precision, recall and F1 metrics of the proposed method

Moreover, the comparative study of top three countries based on flu relevant tweets over the period is shown in Fig. 3. On 18 January, a total of 285 tweets relevant to flu are recorded in the USA, which is the maximum on a single day for a country. It is followed by 250 tweets on 21 January recorded from Indonesia. On the UK, a maximum number of tweets on a single day is 56 on 19 January.

The analysis of flu relevant tweets for different countries shown in Table IV. The country-wise distribution of geotagged tweets shows that a number of flu-infected people were high in the USA followed by Indonesia. It shows that 1271 numbers of geotagged tweets are relevant to flu from the USA over a period of 9 days.

Further, we analyzed the performance of the system using precision, recall and F1 score, which is shown in Fig. 4. As shown in Fig. 4, the proposed method achieved the precision of 0.81 and recall of 0.76 with an F-measure of 0.78, which indicates the efficiency of the method.

## VI. CONCLUSION

We have demonstrated that public health information can be extracted from social media. Generally, users of microblogging site like twitter tweets about their illnesses, which helps to find out the epidemic. Thus, the number of epidemic related tweets increases when there are a greater number of infected people. We find out that twitter can be reliable and real-time source for detecting epidemic that helps to reduce the time to detect flu outbreaks. Therefore, it can be said that the increase in epidemic related tweets indicates an outbreak. In addition, the geotagged tweet helps to find the location of an outbreak. Thus, analysing tweets of twitter can be very helpful to find the outbreak or public health information.

Several things jump out as needing some attention. First, the independence assumption of Naive Bayes is very strong, and it works pretty well. It would be good to supply with more sophisticated linguistic features (n-grams, synonyms, etc.) that will improve its accuracy. Perhaps more importantly, given the informal syntax and number of spelling mistakes in twitter messages, it is likely that a more sophisticated pre-processing stage could improve the quality of the analysis. Finally, an exploration of other methods like Support Vector Machines and Neural Networks as classifiers is bound to be interesting. In addition, the manual sentiment analysis that classifies the relevancy of the tweet could be improved by utilizing deeper learning network with a larger dataset, including additional factors such as correctness on the tagged location.

## REFERENCES

[1] "Home / Twitter," *Twitter*. [Online]. Available: https://twitter.com/home. [Accessed: 08-Sep-2019].
[2] "Arab Spring - HISTORY." [Online]. Available: https://www.history.com/topics/middle-east/arab-spring. [Accessed: 08-Sep-2019].
[3] D. Carr, "How Hurricane Sandy Slapped the Sarcasm Out of Twitter," *Media Decoder Blog*, 31-Oct-2012. .
[4] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.
[5] N. Cesare, C. Grant, and E. O. Nsoesie, "Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices," *ArXiv*, vol. abs/1702.01807, 2017.
[6] M. Burghardt, "Tools for the Analysis and Visualization of Twitter Language Data," *10plus1: Living Linguistics*, 2015.
[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," p. 10, 2010.

[8] S. Asur and B. A. Huberman, "Predicting the Future with Social Media," *Applied Energy*, vol. 112, pp. 1536–1543, Dec. 2013.
[9] "Google Flu Trends." [Online]. Available: https://www.google.org/flutrends/about/. [Accessed: 08-Sep-2019].
[10] D. A. Broniatowski, M. J. Paul, and M. Dredze, "National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic," *PLoS ONE*, vol. 8, no. 12, p. e83672, Dec. 2013.
[11] K. Lee, A. Agrawal, and A. Choudhary, "Real-Time Digital Flu Surveillance using Twitter Data," p. 9.
[12] "Real-Time Digital Allergy Surveillance." [Online]. Available: http://pulse.eecs.northwestern.edu/~kml649/flu/Treatments.php. [Accessed: 08-Sep-2019].
[13] S. Doan, L. Ohno-Machado, and N. Collier, "Enhancing Twitter Data Analysis with Simple Semantic Filtering: Example in Tracking Influenza-Like Illnesses," *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*, pp. 62–71, Sep. 2012.
[14] A. Signorini, A. M. Segre, and P. M. Polgreen, "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic," *PLoS ONE*, vol. 6, no. 5, p. e19467, May 2011.
[15] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using Twitter data: demonstration on flu and cancer," p. 4.
[16] A. Alessa and M. Faezipour, "Preliminary Flu Outbreak Prediction Using Twitter Posts Classification and Linear Regression With Historical Centers for Disease Control and Prevention Reports: Prediction Framework Study," *JMIR Public Health Surveill*, vol. 5, no. 2, p. e12383, Jun. 2019.
[17] F. S. Lu *et al.*, "Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis," *JMIR Public Health Surveill*, vol. 4, no. 1, p. e4, Jan. 2018.
[18] Z. S. Harris, "Distributional Structure," in *Papers in Structural and Transformational Linguistics*, Z. S. Harris, Ed. Dordrecht: Springer Netherlands, 1970, pp. 775–794.
[19] S. Raschka, "Naive Bayes and Text Classification I - Introduction and Theory," *arXiv:1410.5329 [cs]*, Oct. 2014.
[20] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," vol. 9, no. 5, p. 7, 2012.
[21] D. Xhemali, C. J. Hinde, and R. G. Stone, *Naïve Bayes vs. Decision Trees vs. Neural Networks in the classification of training web pages*. Loughborough University, 2009.
[22] S. Xu, Y. Li, and Z. Wang, "Bayesian Multinomial Naïve Bayes Classifier to Text Classification," in *Advanced Multimedia and Ubiquitous Engineering*, 2017, pp. 347–352.
[23] L. Michelbacher, "Multi-word tokenization for natural language processing," *Mehrworttokenisierung für maschinelle Sprachverarbeitung*, 2013.