Influenza prediction from social media texts using machine learning

Sandeep Singh¹, Harjot Kaur¹
Department of Computer Science and Engineering, Guru Nanak Dev University, Gurdaspur Campus, Punjab, India-143521.

E-mail: er.ss1989@gmail.com, harjotkaurohal@rediffmail.com

Abstract. The monitor and prediction of influenza through conventional modes seem obsolete in the Internet era. The flu prediction using machine learning methods is more resourceful for both economically as well as logistically, than traditional methods. The machine learning methods exploit freely available data on social media platforms. The main objective of this paper is to harness the power of publicly available data through machine learning methods. This work involves the scraping of textual data from six social media platforms and training of five machine learning predictors. The paper also proposes a three-layered prediction method based on support-vector-regression for predicting trends of influenza spread. The results obtained from proposed model are validated and analysed against other four machine learning methods. The results reveal that proposed model has better predictive accuracy in terms of least prediction errors among the five methods. Thus, it can provide effective means to control and to prevent the flu outbreak.

Keywords: Flu prediction, Machine Learning, Social Media, Outbreak, Flu trends, Predictors

1. Introduction
The most disastrous concern for public health is an infectious disease. The transmissible nature of disease can result in huge pandemic even when the initial number of cases is less. The same happened with COVID-19, the first case was reported in Wuhan, China, on December 31, last year. Initially, around 100 cases reported in Wuhan only, and then within four weeks the continuous spread has been observed globally encompassing Italy, Spain and USA during the first four weeks [1]. Studies reveal that COVID-19 became the most pathogenic of respiratory and lungs diseases. COVID-19 pathogens are the protein molecules with single stranded RNA, can cause illness in humans ranging from mild fever, common cold, to breathlessness and cardiac arrest [2]. The prevention and cure of COVID-19 affected patients need collaboration of governments, hospitals, and medical institutions worldwide. Centres for Disease Control (CDCs) must report accurate information weekly or fortnightly to locate where and when the spread happens [3]. It has been observed that a new and efficient source of information comes from social media platforms and online news agencies. People from around the globe came to know about COVID-19 outbreak from Facebook, Instagram, Twitter and Google searches [4-5].

The flu trends reported by Google and Facebook are nearly accurate due to cross platform validations performed on it. Tweets and Instagram posts have been continuously reporting includes almost everything ranging from symptoms, location, impact, and travel history to further transmission [6]. Other data sources are online news agencies. The online version of “The Tribune” has over 10 million subscribers as of January 2020, these subscribers submit their online reviews to the news headlines. The user reviews make an information rich data source regarding the outbreak. The two
benefits of this repository are; the data is self-validated by the reviewers through discussions; secondly, unlike printed news which disseminate once in 24 hours, online news are instantly popping up online when any new case bring to surface. The three main news agencies in India (The Tribune, Hindustan Times, and The Hindu) are well established mass-communication institutions (around a century old). Their publication policies and ethics made them trustworthy and reliable [7]. Due to these reasons, we have considered these three news houses for the dataset creation.

2. Related work
Since the SARS (Severe Acute Respiratory Syndrome) outbreak in the year 2002, there were very rare user of Internet media and hence the first study on social media effects came three year down the line. Bergeron Sheri L. et al., in the year 2005 were the first group of researchers who conducted the survey of students on online social media to study the effects of SARS on student community [8]. Similarly, after MERS (Middle East Respiratory Syndrome) in the year 2013, Internet media penetrated enough among the masses, Chen Liangzhe et al., in the year 2014 utilized Twitter data and proposed a temporal model to estimate the flu trends [9], [10]. Further there were many researchers till 2020 who have utilized social media data for predicting flu outbreaks. The summary of such works is shown in the Table 1 below:

| Reference | Description | Dataset(s) | Method(s) | Outcome |
|-----------|-------------|------------|-----------|---------|
| [8]       | Analysed social media effects of SARS outbreak on student community. | Gathered student reviews using web-based flyer tool | Statistical methods using t-test of significance | 92.5% of students consider anxiety due to social media |
| [11]      | Coined a healthcare system for swine flu prediction using machine learning techniques | 100 swine flu cases were manually annotated | Naive Bayes, BPN, kNN | 63.33% of accuracy achieved for NB |
| [10]      | Proposed a temporal model using Tweets to estimate flu trends | Twitter data using Twitter API | Proposed algorithm for flu state prediction | Proposed model found effective in predicting hidden state of flu |
| [2]       | Proposed a classification model for Chinese social media for disease | Text data from Sina micro-blog | SVM, kNN | 94.44% of precision for 3-fold cross validation |
| [12]      | Predicted flu using GIS and machine learning from noisy social media data | GIS, Statistical and SVM are used at regional and national level | 0.671 precision and 0.949 recall for classification |
| [13]      | Conducted influenza prediction using machine learning techniques on Twitter data | Twitter’s data | SVM, GA, and PSO | Predictions obtained from proposed method are significant with least MAPE |
| Reference | Methodology                                                                 | Datasets                                                                 | Model/Techniques                                                                 | Results/Findings                                                                 |
|-----------|----------------------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| [14]      | Mapped random forest method for flu prediction using online textual data.  | Four Datasets: Wiki Logs, Flu Views, Tweets, and Google Trends            | Random Forest and Gradient Boosting                                              | Forecasting flu trends helps in flu seasons                                      |
| [15]      | Proposed a prediction model using ANN for flu prediction                   | CDC’s US dataset, Google Flu trends, Internet Searches, and Tweets      | ANN                                                                             | Proposed model was significant in terms of lowest MSE and RMSE                  |
| [4]       | Applied machine learning methods for flu prediction                         | Instagram posts                                                          | SVM, XGBoost, Random Forest, kNN, Regression, CNN                              | CNN + XGBoost found 0.963 Pearson’s correlated                                   |
| [6]       | Performed data mining of Twitter messages for Sentiment analysis and prediction of flu related Tweets | Twitter’s data                                                          | Data mining for filtering, machine learning for Sentiment Analysis             | Predictions made were significant and validated against NASA data.              |
| [7]       | Coined a prediction model for influenza prediction using existing SVM technique | 7769 news articles from 2013-2018                                       | SVM                                                                             | 86.7% of mean accuracy achieved                                                  |
| [16]      | Deployed AI based model and forecasted influenza events at Chongqing, China | Collected data from Baidu Search, Sina Weibo, and online data from seven Hospital’s | SAAIM, LASSO, and LSTM                                                         | Proposed method was 96.2% accurate in prediction.                               |
| [3]       | Made a framework for flu outbreak prediction                               | Twitter’s data                                                           | Linear Regression                                                               | 96.29% Pearson correlation between CDC’s data and predicted data                |
| [17]      | Prediction of flu pandemic using Wiener Peru Series enabled machine learning | AH1N1 dataset of Wiener Peru Series enabled machine learning            | Wiener Series enabled machine learning                                          | Prediction results were quite similar to official data                           |
| [1]       | Conducted a field review of flu affected COVID-19 patients and databases till they reviewed their historical classics | Social media data from websites and databases till February 12, 2020     | Field survey of flu affected patients followed by statistical analysis          | Chinese herbal formula found significant when observed on 95% confidence interval |

3. Data extraction and pre-processing
The two main sources rich in COVID-19 information are social media platforms and news-websites. The famous social media platforms considered are Facebook, Instagram, and Twitter, while The Tribune (www.tribunedia.com), Hindustan Times (www.hindustantimes.com), and The Hindu (www.thehindu.com) are the news agencies taken for data extraction [7]. Only textual data is extracted from these platforms using BS4 library of Python 3.6. Beautiful-Soup version 4 is a very good library used for web scraping. The textual data is collected in a text file and then regular expression parser (regexpparser) is used for data cleaning. Standard NLTK’s functions are used for chunking, chunking, segmentation, bigram and trigram extraction [3]. Further eight samples are collected from this dataset.
using eight symptomatic keywords. The keywords are “dry-cough, breathlessness, sore-throat, fever, running-nose, headache, drowsiness, and chest-pain”. However, the other keywords like “flu, influenza, sickness, muscle-ache, and outbreak” are also there but for 8 samples, the keywords corresponding to eight symptoms are taken. Social media posts and news mentions are collected from January 01, 2015 to September 30, 2020. The total number of posts collected from these six sources are 2,77,257. The summary of statistics is shown in the Table 2 below:

| Keyword         | FB     | IG     | TW     | TT     | HT     | TH     | Total  |
|-----------------|--------|--------|--------|--------|--------|--------|--------|
| DRY-COUGH       | 974    | 529    | 1243   | 179    | 253    | 219    | 3397   |
| BREATHELESSNESS | 661    | 451    | 874    | 111    | 97     | 123    | 2317   |
| SORE-THROAT     | 842    | 297    | 693    | 129    | 170    | 217    | 2348   |
| FEVER           | 1234   | 4239   | 6734   | 681    | 519    | 539    | 13946  |
| RUNNING-NOSE    | 943    | 442    | 1539   | 80     | 59     | 71     | 3134   |
| HEADACHE        | 327    | 439    | 932    | 39     | 30     | 17     | 1784   |
| DROWSINESS      | 419    | 211    | 853    | 132    | 197    | 122    | 1934   |
| CHEST-PAIN      | 628    | 192    | 632    | 51     | 47     | 56     | 1606   |
| FLU             | 14393  | 18313  | 25370  | 240    | 371    | 299    | 58986  |
| INFLUENZA       | 10403  | 9147   | 16460  | 343    | 472    | 371    | 37196  |
| SICKNESS        | 152    | 69     | 445    | 135    | 170    | 280    | 1251   |
| MUSCLE-ACHE     | 129    | 81     | 288    | 46     | 39     | 77     | 660    |
| COVID-19        | 15863  | 21536  | 48347  | 342    | 475    | 430    | 86993  |
| OUTBREAK        | 9473   | 13433  | 37509  | 413    | 460    | 417    | 61705  |
| **Total**       | 56441  | 69379  | 141919 | 2921   | 3359   | 3238   | 277257 |

There are more than 30% mentions containing word “COVID-19”, while around 22% of the posts mention keyword “OUTBREAK”. Approximately 21% of the total posts use keyword “FLU”, while around 13% of them quotes “INFLUENZA” in their posts. The data extraction along with pre-processing steps explained using the Figure 1.

4. Proposed model
The proposed regression model works at three layers of scanning. The textual data from social media sources is quite rich in terms of flu related information these days. The textual information is mapped to the very first layer of proposed model after vectorization i.e. $M_{tw}$. Further the degree of spread is estimated by the second layer which introduces weighting term along with the mapping obtained from the first layer. The three occurrences for a target word are verified against the ten other combinations of bigrams made with it. The following layer three further considers the combinations from layer two to look for possible trigrams. Each layer of proposed model employs Support Vector Regression through a machine learning classifier termed as (SVM) Support Vector Machine.
Figure 1. Data Extraction and Pre-processing

\[ M_{i,w}^1 = \sum_{l=1}^3 \alpha_l X_{i,l,w} + \varepsilon_w \]  \hspace{1cm} \text{..........................}(1)

\[ M_{i,w}^2 = \sum_{l=1}^3 \beta_l X_{i,l,w-1} + \sum_{j=1}^{10} \gamma_l \omega_{l,j} X_{j,w} + \theta_w \]  \hspace{1cm} \text{..........................}(2)
\[ M_{lw}^3 = \sum_{i=1}^{3} \delta_i X_{lw}^{i+1} + \sum_{i=1}^{3} \mu_i \alpha_{i}^2 + \sum_{j=1}^{10} \gamma_j \omega_{i,j} X_{jw} + \tau_w \]  

Here, \( M_{lw}^1, M_{lw}^2, \) and \( M_{lw}^3 \) are the three layers of proposed model, the two-dimensional matrices \( X_{lw} \) will collect \( i \) occurrences for a word \( w \) about the flu pandemic. \( \alpha, \beta, \gamma, \delta, \mu, \) and \( \sigma \) are the coefficients used for adjusting the influence of matrices \( X_{lw} \). The values of these coefficients lie in the range \((0, 1)\). The omega \( \omega_{i,j} \) determines the weighting factor for every \( i^{th} \) occurrence correlated with \( j^{th} \) coefficient. The weighting factor helps in establishing relationship between three layers of model \( (M_{lw}^1, M_{lw}^2, \) and \( M_{lw}^3) \). The \( e_w, \theta_w, \) and \( \tau_w \) are the three residues in each of the three respective layers of proposed model. These residues apply the balancing effect on the three layers with their values varying from \((-1 \ to \ 1)\). Each layer works on the three \((1 \leq i \leq 3)\) occurrences for a word \( w \). Therefore, at maximum, nine occurrences can be caught till the third layer. The third layer itself validates the second layer (as the second term of third layer comes from second layer) and produces appropriate prediction for flu pattern. The proposed methodology is shown in the Error! Reference source not found.

Word vectorization is done by using GloVe model (Global Vectorization model) representation. The input vectors for training set are labelled using Python’s standard feature extractor from NLTK (Natural Language Toolkit). The Training set is represented by the equation (4) below:

\[ Train(D) = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, \ldots, (x_n, y_n)\} \]  

Where, \( x_i \in R^n \) and \( y_i \in R \).

Here, input words are represented by real number vectors \((x_1, x_2, x_3, x_4, \ldots, \ldots, \ldots, x_n)\) where \( n \) is the sample size. The \( y_i \)’s are the real number labels used for COVID-19 flu prediction.

Since, there is a high dimensionality of features related to flu (cough, running nose, lungs infection, sore throat, fever, headache, chest pain, breathlessness, and many other terms) a non-linear mapping function \( f(x) \) (as shown in the equation (5) below) maps input vectors into multi-dimensional features \( F \).

\[ f(x) = \omega^T \cdot \phi(x) + b \]  

Here, \( \phi: R^n \rightarrow F \) and \( \omega \rightarrow R^m \).

\( F \) is a feature space and \( \omega^T \) is transposed weight matrix represents hyperplane by introducing bias ’b’.

Furthermore, each layer performs optimization by limiting the constraints on training errors as shown in the equation (6) below: i.e.

\[ Training \ error = \min_{\omega, \epsilon, \epsilon'} \left\{ \frac{1}{2} \omega^T \omega + c \sum_{i=1}^{n} (\epsilon_i + \epsilon'_i) \right\} \]  

Where, \( |y_i - (\omega^T \cdot \phi(x) + b)| < \epsilon \), for small values of epsilon \( \epsilon \) which allows \( \epsilon \in (\epsilon_i, \epsilon'_i) \) and \( c \) is a constant term to keep error minimum and \( c < 0 \). \( c \) is responsible for over-fitting when large training error comes with large taken value of \( c \), whereas for small value of \( c \), the corresponding training error would be small, leads to good machine learning for flu prediction.

Due to multi-dimensionality of features, Radial basis kernel function (as shown in equation (7) below) is used in this model to transform linearly inseparable input vectors into linearly separable ones.

\[ RBKF(x_i, x'_i) = \exp \left( -\frac{||x_i - x'_i||^2}{2\sigma^2} \right), \sigma > 0 \]  

Where, \( \sigma \) is the standard deviation and \( ||x_i - x'_i|| \) is the norm (input vector with largest separation from the central vector) for the input vector ‘\( x'_i \)’. The large values for parameter ‘\( \sigma \)’ represent wide input space whereas the small input space corresponds to the smaller value of ‘\( \sigma \)’.
5. Results and discussion
This research focuses on four state-of-the-art machine learning models viz. Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF), and k-Nearest Neighbours (kNN). The predictions made by these four classifiers are validated against proposed model. The performance comparisons are made by using three standard errors (training errors) i.e. MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Square Error), and MSE (Mean Squared Error).

Figure 2. Proposed Methodology, SVM is Support Vector Machines, NB is Naive Bayes, kNN is k-Nearest Neighbors, and RF is Random Forest.
\[
MAPE = \frac{1}{B} \sum_{i=1}^{B} \left| \frac{y_i - \bar{y}_i}{\bar{y}_i} \right| \times 100
\] .................................(8)

\[
MSE = \frac{1}{B} \sum_{i=1}^{B} (y_i - \bar{y}_i)^2
\] .................................(9)

\[
RMSE = \sqrt{\frac{1}{B} \sum_{i=1}^{B} (y_i - \bar{y}_i)^2}
\] .................................(10)

### Table 3. Prediction errors for eight samples

| Sample | Predictor | MAPE  | MSE   | RMSE  |
|--------|-----------|-------|-------|-------|
| S1     | SVM       | 0.7261| 1.2531| 0.9211|
|        | NB        | 0.8432| 2.1357| 1.3531|
|        | RF        | 0.6754| 1.8792| 2.7315|
|        | kNN       | 0.5136| 1.1526| 3.5731|
|        | Proposed  | 0.3254| 0.9843| 1.4662|
| S2     | SVM       | 0.7576| 3.7417| 0.8621|
|        | NB        | 0.6432| 2.7183| 1.0521|
|        | RF        | 0.9821| 0.9221| 1.7372|
|        | kNN       | 1.1315| 1.9213| 2.8188|
|        | Proposed  | 0.5421| 1.2123| 0.7432|
| S3     | SVM       | 1.7592| 7.2153| 0.5987|
|        | NB        | 1.9527| 9.7211| 0.9875|
|        | RF        | 1.2792| 3.2392| 0.7215|
|        | kNN       | 1.3922| 6.7321| 0.9826|
|        | Proposed  | 1.2637| 4.7371| 0.5587|
| S4     | SVM       | 0.8287| 1.0652| 1.3731|
|        | NB        | 0.7831| 1.7431| 1.5332|
|        | RF        | 0.9235| 1.9833| 2.3157|
|        | kNN       | 0.7668| 1.6372| 1.5365|
|        | Proposed  | 0.7215| 1.0521| 1.2168|
| S5     | SVM       | 0.7231| 1.5279| 0.5241|
|        | NB        | 0.8759| 1.7927| 0.0132|
|        | RF        | 0.8269| 1.3235| 0.0118|
|        | kNN       | 0.5876| 1.4334| 0.0883|
|        | Proposed  | 0.5373| 1.4051| 0.0103|
| S6     | SVM       | 0.9989| 2.4362| 1.2721|
|        | NB        | 0.7271| 2.1083| 1.4164|
|        | RF        | 0.6414| 0.3537| 1.3642|
|        | kNN       | 0.4362| 0.3425| 1.5342|
|        | Proposed  | 0.0982| 0.5352| 1.1132|
| S7     | SVM       | 0.7258| 0.8648| 0.1537|
|        | NB        | 0.8944| 1.3741| 0.3735|
|        | RF        | 0.8794| 0.9315| 0.7315|
|        | kNN       | 0.8583| 0.9512| 0.2751|
|        | Proposed  | 0.8012| 0.8532| 0.0842|
| S8     | SVM       | 0.3607| 1.4827| 1.5681|
|        | NB        | 0.8706| 1.4952| 1.7261|
|        | RF        | 0.5472| 1.4632| 0.8642|
|        | kNN       | 0.7068| 1.3739| 0.9982|
|        | Proposed  | 0.2219| 1.2253| 0.5352|
MAPE is preferred over MSE and RMSE because the error comparison for two different models would be done in percentage values which are lesser than mean squared and root mean squared values. Here \((1 \leq i \leq 8)\) for the eight samples taken on the basis of eight feature-symptoms of COVID-19. \(y_i\) is the correct value of output variable and \(\hat{y}_i\) is the predicted value of the output variable for each sample.

The Table 3 shows MAPE, MSE, and RMSE for the eight samples collected against eight symptomatic features. These prediction errors are calculated for the four state-of-the-art machine learning algorithms and SVM based proposed model. The results reveal that the proposed model has more prediction capabilities among the five predictors as shown in Table here. The only few cases (for samples S1, S2, S3, S5 and S6) are found where NB, RF and kNN (respectively) have performed better than our three-layered proposed model.

The smaller values for MAPE, MSE, and RMSE obtained using proposed model reveals that our three-layered framework may be a better predictor for finding the spread of COVID-19 flu.

Table 4. Prediction errors for the whole dataset

| Predictor | Error   |
|----------|---------|
|          | MAPE    | MSE   | RMSE  |
| SVM      | 2.7357  | 11.3344 | 4.5374 |
| NB       | 1.9876  | 11.9216 | 5.9463 |
| RF       | 2.5437  | 9.4374  | 4.2371 |
| kNN      | 3.9431  | 11.6325 | 4.9574 |
| Proposed | 1.7322  | 9.1357  | 3.2649 |

The Table 4 above shows more robustness of the proposed model in terms of flu prediction than other four predictors. Smaller values for MAPE, MSE, and RMSE for three layered proposed predictor reflects the effectiveness of three-layered framework used for flu prediction out of social media mentions.

6. Validation of proposed model

The matplotlib (imported from pyplot) library of Python 3.6 is utilized for generation of line graphs. The Figure 3 to Figure 7 below depicts the comparison of actual and predicted MAPEs for the first 50 observations. It is found that the proposed curve in the Figure 7 is quite close to the actual curve. This further confirms that the prediction capabilities of the proposed model are better among the five predictors. The train-test-split function of sklearn’s cross-validation library is used for cross validation of model. There is a utility function named shuffle from sklearn’s utils package that helps in cross validation, after every epoch, a shuffle is invoked for waddling the training and testing parts followed by the invocation of train-test-split function again. Pre-defined 50 observations regarding flu outbreak are collected randomly from the entire dataset. The MAPE error curve generated from these observations is termed as actual curve. Then, these 50 input vectors are unlabeled through simple numpy array, followed by MAPE curve generation from five predictors. These validation curves are represented in the following five figures (Figure 3 to Figure 7) to validate the predictors.
The Figure 3 and Figure 4 above depict the comparisons of actual curve against the curves obtained from SVM and NB respectively. The respective average values of MAPE for SVM and NB are 2.7357 and 1.9876 obtained from first 50 observations. The higher value of MAPE for SVM makes it trail from NB in terms of flu prediction.
The Figure 5 and Figure 6 above show the comparison of actual curve against the curves obtained from kNN and RF models. The larger number of variations between kNN and actual curve are observed than the variations between RF and actual curve. The average values of MAPE for kNN and RF are 3.9431 and 2.5437 respectively (refer Table 4). Thus, RF is more precise in prediction than kNN.
Figure 7. Predictions made for Proposed Model (line graph generated using matplotlib of Python 3.6)

Beside the few discrepancies the actual curve is almost in synch with the light grey colored curve of proposed model (as shown in Figure 7 above). The certain values of MAPE like 5th, 6th, 13th, 29th, 39th, and 46th are different but the average value for the whole dataset is 1.7322 which is least among the average values for other models (refer Table 4).

7. Conclusion
Our work has presented that social media data is the most valuable resource for predicting influenza spread. We have proposed three layered model for flu prediction using support vector-based classification. The proposed model is validated through first fifty observations on MAPE against actual curve. The results show that proposed model is more accurate in terms of prediction capabilities. The lowest values for MAPE obtained through proposed model provide the insights for appropriate flu prediction and control in the epidemic. The proposed work can be extended in future by hooking a geographical information system in parallel with the three layers. Then it can be used to conduct the flu outbreak prediction at regional level followed by nationwide prediction. The temporal model at regional level may reach every affected person and bring in some course of action for his/her cure. Another offshoot may look forward to tie a healthcare recommender system for providing antidote and therapy to the affected people.

References
[1] H. Luo et al., “Can Chinese Medicine Be Used for Prevention of Corona Virus Disease 2019 (COVID-19)? A Review of Historical Classics, Research Evidence and Current Prevention Programs,” Chin. J. Integr. Med., vol. 11655, no. 100029, pp. 1–8, 2020.
[2] N. Yang, X. Cui, C. Hu, W. Zhu, and C. Yang, “Chinese social media analysis for disease surveillance,” Proc. - 2014 Int. Conf. Identification, Inf. Knowl. Internet Things, IIKI 2014, pp. 17–21, 2014.
[3] 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,” Journal of Medical Internet Research, vol. 21, no. 6. Journal of Medical Internet Research, 2019.
[4] O. Gencoglu and M. Ermes, “Predicting the Flu from Instagram,” Nov. 2018.
[5] N. Spasojevic and A. Rao, “Identifying actionable messages on social media,” in Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015, 2015, pp. 2273–2281.
[6] M. D. Albayrak and W. Gray-Roncal, “Data Mining and Sentiment Analysis of Real-Time Twitter Messages for
Monitoring and Predicting Events,” 2019 9th IEEE Integr. STEM Educ. Conf. ISEC 2019, pp. 42–43, 2019.

[7] J. Kim and I. Ahn, “Weekly ILI patient ratio change prediction using news articles with support vector machine,” BMC Bioinformatics, vol. 20, no. 1, pp. 1–16, 2019.

[8] S. L. Bergeron and A. L. Sanchez, “Media effects on students during SARS outbreak,” Emerg. Infect. Dis., vol. 11, no. 5, pp. 732–734, 2005.

[9] N. Zhu et al., “A novel coronavirus from patients with pneumonia in China, 2019,” N. Engl. J. Med., vol. 382, no. 8, pp. 727–733, 2020.

[10] L. Chen, K. S. M. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, “Flu Gone Viral: Syndromic Surveillance of Flu on Twitter Using Temporal Topic Models,” Proc. - IEEE Int. Conf. Data Mining, ICDM, vol. 2015-Janua, no. January, pp. 755–760, 2014.

[11] B. A. Thakkar, M. I. Hasan, and M. A. Desai, “Health care decision support system for Swine flu prediction using Naïve Bayes classifier,” Proc. - 2nd Int. Conf. Adv. Recent Technol. Commun. Comput. ARTCom 2010, pp. 101–105, 2010.

[12] C. Allen, M. H. Tsou, A. Aslam, A. Nagel, and J. M. Gawron, “Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza,” PLoS One, vol. 11, no. 7, pp. 1–10, 2016.

[13] H. Hu, H. Wang, F. Wang, D. Langley, A. Avram, and M. Liu, “Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network,” Sci. Rep., vol. 8, no. 1, pp. 1–8, 2018.

[14] K. Will, “Predicting Flu Outbreak with Random Forest,” Dissertation, 2017. [Online]. Available: http://www.jarad.me/research/thesis/KatieWill_CC.pdf.

[15] K. Su et al., “Forecasting influenza activity using self-adaptive AI model and multi-source data in Chongqing, China,” EBioMedicine, vol. 47, pp. 284–292, 2019.