ISSN2303-4521
Periodicals of Engineering and Natural Sciences *Original Research*
Vol. 10, No. 1, January 2022, pp.122-137

Evaluation of IBM Watson Natural Language Processing Service to predict influenza-like illness outbreaks from Twitter data

Kanita Karađuzović-Hadžiabdić[1], Rialda Spahić[1], Emin Tahirović[1]
[1]International University of Sarajevo, Computer Sciences and Engineering, Hrasnička cesta 15, 71210 Sarajevo, Bosnia

ABSTRACT

Social media has opened the gates for collecting big data that can be used to monitor epidemic trends in real time. We evaluate whether Watson NLP service can be used to reliably predict infectious disease such as influenza-like illness (ILI) outbreaks using Twitter data during the period of the main influenza season. Watson's performance is evaluated by computing Pearson correlation between the number of tweets classified by Watson as ILI and the number of ILI occurrences recovered from traditional epidemic surveillance system of the Centers for Disease Control and Prevention (CDC). Achieved correlation was 0.55. Furthermore, a 12 week discrepancy was found between peak occurrences of ILI predicted by Watson and CDC reported data. Additionally, we developed a scoring method for ILI prediction from Twitter posts using a simple formula with the ability to predict ILI two weeks ahead of CDC reported ILI data. The method uses Watson's sentiment and emotion scores together with identified ILI features to analyze influenza-related posts in real time. Due to Watson's high computational costs of sentiment and emotion analysis, we tested if machine learning approach can be used to predict influenza using only identified ILI keywords as influenza predictors. All three evaluated methods (Random Forest, Logistic Regression, K-NN), achieved overall accuracy of ~68.2% and 97.5% respectively, when Watson and the developed formula are used as medical experts. The obtained results suggest that data found within social media can be used to supplement the traditional surveillance of influenza outbreaks with the help of intelligent computations.

Keywords: IBM Watson, Infectious Disease Prediction, Public Health, Social Media Analysis, Text Mining

Corresponding Author:

Kanita Karađuzović-Hadžiabdić
International University of Sarajevo
Computer Sciences and Engineering
Hrasnička cesta 15, 71210 Sarajevo, Bosnia
E-mail: kanita@ius.edu.ba

1. Introduction

Over the past decade we have witnessed an explosion of data available on the Internet. Due to the lower error rates compared to the ones characteristic to humans, technological improvements are already starting to change the health system. Early detection of an epidemic can play an important role in providing a timely response to an emergence of an epidemic. Monitoring diseases is not an easy task. Traditional monitoring processes usually retrieve information after the particular disease such as influenza has already spread. As

more and more people search the internet for medical information, and later tend to post their medical conditions on social media, utilizing this information in an intelligent way can alter how medical institutions prepare for outbreaks of infectious diseases. Real-time ILI prediction can be used to identify ILI trends, and by doing so, to discover the relative rise in influenza-related tweets. Intelligence surveillance systems can serve as an early warning signal of new and emerging health topic that is becoming important to the public that has not yet been detected. It may reveal specific health concern that may be on the rise. This information may be crucial to assist healthcare and government officials for a timely preparedness and effective planning.

Social media has brought the availability of big data. Compared to the traditional approaches of collecting public information via public surveys, social media allows us to access valuable public information in real time. On daily basis, millions of user-generated content appears on social media services. This data includes valuable information ranging from public opinion, personal experiences, politics, health information, etc. Approaches such as data crawling methods, enable the retrieval of public posts from social media in a fast and economical way.

A large body of research has been published where authors use data collected from public web search queries or social networking sites to monitor and predict the spread of diseases. As one of the most widely used social media platforms, Twitter has been primarily employed for this type of research [1], [2], [3], [4], [5]. When compared to the traditional ways of collecting data, Twitter provides a faster approach to collecting real-time user data including opinions, sentiment, emotion, activities, etc. Jordan et al. [6] provide state-of-the-art review in using Twitter for public health surveillance

Influenza-like illness (ILI), or more commonly known as flu, is a contagious respiratory illness with flu-like symptoms caused by the influenza virus. The spread of infectious diseases has been further assisted as the global travel increased. An example is the H1N1 virus that appeared in 2009 in North America and has spread worldwide with remarkable speed causing the first influenza pandemic since 1968. A few more pandemics have hit the headlines such as the Ebola, SARS, Zika, MERS, and recently, the Coronavirus pandemic. (Note: this research was performed before the COVID-19 pandemic.) With the rise of the Corona pandemic, the need for research targeted at non-traditional infectious disease surveillance has only increased. Intelligent surveillance systems can help in timely detection of contagious diseases that threaten public health. As the recent developments of the COVID-19 pandemic show, the time factor is of a crucial importance in early detection of an outbreak of a contagious disease. A comprehensive review of influenza detection, prediction and tracking using social networks can be found in Alessa and Faezipour [7], and Al-garadi et al. [8].

Centers for Disease Control and Prevention (CDC) [9] defines syndromic surveillance as "surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response." Usually there is 1–2-week delay between ILI diagnosis and the time when this information becomes part of published ILI statistical data. At the same time, many people do not alert their doctors after the first symptoms but choose to search the web for causes of the symptoms instead, and later post this information on social media. Extracting data from social media and analyzing is an infant technology. It requires sentiment analysis for understanding the context in order to retrieve useful information. Being able to structure information in a meaningful way, a system can be designed to detect the occurrence of a particular disease early and prompt the health officials to take appropriate measures to prevent epidemic outbreaks.

As reported by [10], CDC itself initiated influenza challenge in 2013/2014. The challenge encourages innovative means of forecasting influenza in US, based on weekly surveillance of ILI occurrences. Teams are required to use digital data from sources such as social media, search queries or other forms of data collected from the internet [11]. Initiated challenge also resulted in establishment of the 'Influenza Division' at CDC, where CDC together with the scientific community is engaged on real-world influenza forecasting challenges known as FluSight, [12].

In order to use the technology to extract potentially useful information from the big data hidden within layers of social networks, continuous and joined efforts from computer scientists and influenza domain experts are needed to supplement traditional ILI surveillance. The main objective of this work was to evaluate whether IBM's Watson NLP service can be used in prediction of ILI by analyzing public Twitter posts. Watson is a highly sophisticated computing system developed by IBM. It applies cognitive computing methods in its decision-making process [13], [14]. These methods surpass the traditional computation methods that rely on computing, which requires well-structured data. Since 2011, IBM Watson has invested billions of dollars in cognitive computing with an attempt to use artificial intelligence (AI) in healthcare to help medical practitioners in their day-to-day work. For almost a decade, Watson has been fueled and trained with vast amounts of data with the guidance of human experts to solve problems with varying complexities. One of the most sophisticated Watson services is the Natural Language Understanding (NLU). Watson NLU uses NLP to analyze sentiment, emotion, semantics, entities, syntax, relationships, and categories of raw text. Despite huge investments, paper by Strickland [15] shows that further research is needed. We believe that our publication addresses an important issue raised in [15]. Namely, as reported in Strickland [15], a former IBM employee states that there is very little peer reviewed papers that demonstrate the advantages of using AI in medicine to improve patient care and save health systems money. Furthermore, he also mentions that such publications are scarce, and "none of consequence for Watson". Our research attempts to fill this gap, by evaluating Watsons NLP service in prediction of influenza from tweet messages (i.e., Watson's ability to understand human natural language with application in healthcare). We evaluate how well does Watson predict ILI by computing correlation between positive ILI tweets predicted by Watson's confidence score for 'cold and flu' category and the influenza occurrences reported by real world CDC Data.

We also developed a formula to predict if a tweet qualifies as pertaining to ILI occurrence that can supplement existing healthcare surveillance systems. Details of the formula can be found in Section 2.6, (1). Unlike most approaches, the proposed formula uses IBM's Watson NLP service to include sentiment and emotion into the analysis of Twitter posts. Obtained sentiment and emotion results are further enriched with identified discriminative features facilitated by the proposed method. The idea is to use the additional discriminative features and a custom made functional formula relating these features to the binary outcome that predicts ILI from Twitter posts with high accuracy as judged by the official CDC data.

Machine learning methods have been widely used in predictive modeling and analysis of many diseases including cardiovascular disease, diabetes, liver disease, various cancers, etc. [16], [17], [18], [19]. In order to reduce Watson's computational costs of cognitive computing predictions, we also evaluate utilization of machine learning methods in prediction of influenza using appropriate influenza discriminative features.

The main contributions of the paper are the following:

- Identification of discriminative features to detect ILI from social media.
- Collection of public raw tweets from Twitter and creation of a structured dataset.
- Evaluation of IBM's Watson NLP service for prediction of influenza from public Twitter posts using correlation of weekly influenza occurrences as reported by real world CDC data.

Additional contributions of the paper are:

- Development of a simple functional formula that computes influenza score from Twitter posts that strongly correlates with CDC data. The method uses Watson's sentiment and emotion scores as well as identified ILI discriminative features which can be used to further guide Watson towards a more accurate prediction of influenza.
- Analysis of machine learning approach to predict influenza from Twitter. Once trained, the machine learning methods can predict influenza without the need to use Watson's expensive computations of sentiment and emotion analysis.

2. Materials and methods

2.1. Twitter dataset

Twitter is one of the most popular social media websites. It is a micro blog that contains messages (i.e. tweets) up to 280 characters in length. Public data available on Twitter present an efficient approach for generating data that can be used to detect various trends among its users, including tracking of epidemics.

Only in the USA, millions of tweets are being posted per day. In order to select tweets related to ILI, we first identified relevant keywords most frequently tweeted by users when reporting the incidence of ILI. Initially, keyword selection was done by extracting ILI symptoms from various sources including NHS [20] (www.nhs.uk), CDC [9], WebMD [21] Healthline Media [22] and Google Trends [23]. This resulted in an initial set of 82 ILI related keywords. In order to assess their relevance, tweets that contained selected influenza keywords were collected within the period of one week, starting from 10.04.2018 until 18.04.2018. This resulted in 90,000 tweets. Tweet extraction was done by TweetScraper [24]. As a pre-processing step, repetitive words, non-English tweets, URLs, empty documents, retweets, and unknown symbols were removed to improve robustness.

By careful analysis of the extracted tweets (based on the initial 82 keywords), it was found that the dataset contained a large number of misleading and irrelevant data. Over 1, 000 tweets were manually analyzed. Large number of these tweets were related to circumstances other than influenza, such as mental illness, heart conditions, weather prognosis, etc. (Table 1 shows some examples of candidate ILI keywords). Based on our manual tweet analysis, and on the reported CDC symptoms (fever, cough, sore throat, runny nose, muscle aches, headaches, fatigue, sneezing and chills), many of the initial keywords identified as the reason for extracting unrelated tweets were removed, and we reduced their number to nine of those listed in Table 2. Identified set of keywords was then used for crawling Twitter posts within a 26 week period, starting from week 44 in 2017 to week 17 in 2018. Just like for CDC, our reporting week starts on Sunday and ends on Saturday. After tweet pre-processing, we obtained a dataset of 11,118 tweets across USA from 5thNovember 2017 until 28th April 2018 (i.e. during the main influenza season.)

Table 1. Example of candidate ILI keywords

Example tweet	**Influenza keyword candidate**	**Selected as influenza keyword**
I will not go out today, the temperature is too high.	temperature	no
I had a sore throat like a month ago and I was sick for a week.	sore throat	yes
That is a new mental illness.	illness	no
It is August. It is no longer acceptable to suffer from corps wide stomach flu.	flu	yes
His disease is one of the last stages.	disease	no
He needs treatment immediately.		
I'm sick. Throat feels weird and also runny nose. What a good combination hahahah.	runny nose	yes

Table 2. Identified ILI keywords

flu	influenza	fever
runny nose	sore throat	headache
coughing	sneezing	fatigue

2.2. CDC dataset

CDC offers national, regional and state level outpatient illness surveillance in the USA. It posts data about influenza on a weekly basis by conducting a traditional healthcare surveillance system by tracking the number of patients that are tested at hospitals for influenza virus. CDC presents its results through FluView [25], an interactive graph showing influenza trend by weeks, seasons or virus types. Just like for the collected tweets, influenza related data used from the CDC also range from week 44 (in 2017) to week 17 (in 2018.)

Since CDC reports number of positive ILI occurrences on a weekly basis, the number of positive ILI tweets predicted by Watson and devised ILI formula were also aggregated into weekly reports. Weekly aggregated results of positive ILI tweets were then normalized in order to obtain the percent increase of positive ILI tweets. This was done by dividing each weekly data report by the maximum (peak) of 26 weekly data reports (i.e. over the tested time period). The percent increase of positive ILI tweets is particularly useful as it shows the rise of the number of posts on a topic over a particular time interval (which may even indicate an early alarm for ILI outbreak). Google Trends shows the results of its searches in a similar manner.

This CDC data set was used to compute the correlation between normalized weekly number of positive influenza occurrences as reported by CDC and: a) Watson positive ILI tweets b) positive ILI tweets computed by the developed ILI formula.

2.3. IBM Watson

A considerable amount of information must be handled in NLP, which is best accomplished with the assistance of cognitive computing. IBM's Watson Natural Language Understanding API cloud technology can extract sentiment and emotion values from phrases and compute their values from the collected tweets.

Sentiment analysis is a technique of detecting main states of the text by applying natural language processing. The output of sentiment analysis is a sentiment label (negative, neutral, positive) accompanied by a sentiment score. Negative number indicates negative sentiment, 0 is neutral, and positive number indicates positive sentiment. Watson NLP service classifies human emotion into five categories: anger, fear, joy, sadness, and disgust. In a world of big data, most of the valuable data comes unstructured from social media. Users of social media tend to express their opinion by posting their attitude publicly.

Navigating through vast amount of unstructured data has become a challenge. Sentiment and emotion analysis play a crucial role in detecting the attitude of majority towards a specific topic. Watson also generates confidence score results for ‘cold and flu’ category. The confidence score results are represented in the range from 0 to 1, where 0 indicates Watson is not confident and 1 indicates Watson is highly confident of the output result.

2.4. Feature selection

An important step in machine learning classification is selection of discriminative features (predictors) to be used as inputs into machine learning methods. To avoid high computational costs of Watson's sentiment and emotion analysis we used tweet date and nine identified ILI keywords as inputs. ILI keyword indicators were set to 1 if the keyword appears in a tweet, 0 otherwise.

Outputs of Watson's sentiment and emotion scores including ILI keywords with an example tweet is shown in Table 3.

Table 3. Example tweet with corresponding sentiment, emotion and ILI keywords extracted from it

Analyzed tweet	**“i'm sick throat feels weird and also runny nose. what a good combination hahahah”**														
Sentiment	**Emotion**						**ILI keywords**								
Sentiment Label	**Sentiment Score**	**Sadness**	**Joy**	**Fear**	**Disgust**	**Anger**	**flu**	**influenza**	**fever**	**runny nose**	**sore throat**	**headache**	**coughing**	**sneezing**	**fatigue**
negative	-0.534	0.216	0.025	0.630	0.250	0.155	0	0	0	1	0	0	0	0	0

2.5. Correlation with CDC data

Our main aim was to evaluate Watson's confidence score for 'cold and flu' category for ILI prediction from public Twitter posts. To do this, Pearson correlation coefficients were computed between normalized number of positive ILI tweets as predicted by Watson and the normalized number of ILI diagnosis from the real-world CDC surveillance system data. For this, we evaluated 10 confidence scores (0.5, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90 and 0.95) as threshold values to classify positive and negative ILI tweets. For each threshold value, tweets with score equal to or greater than the selected threshold were considered as positive ILI tweets, and tweets with score lower than the selected threshold value were considered as negative ILI tweets.
Afterwards, we performed three experiments 1) ILI prediction using the proposed ILI formula, 2) ILI prediction using machine learning methods using Watson as a medical expert, and 3) ILI prediction using machine learning methods using ILI formula as a medical expert. The experiments are outlined in Sections 2.6-2.8. The results of these experiments as well as the correlation results are presented and discussed in Section 3.

2.6. ILI Prediction using the proposed ILI formula

We developed a simple ILI formula for influenza prediction that computes ILI score. The formula uses nine identified keywords (Table 2). Since the task of ILI detection from tweets is ultimately an NLP task, using only ILI keywords is not enough to identify influenza. It is important to look at the overall picture beyond a single word analysis. To do this we include Watson's sentiment and emotion in understanding the meanings of messages through context. Sentiment and emotion can add to the power of discernment between twitter users expressing their actual present health condition and them being concerned about contracting the disease in question or being worried in general about the spread of the disease.
To develop the ILI formula the identified ILI features were further categorized into 4 groups: main keywords (flu or influenza), symptoms (i.e., remaining ILI keywords: fever, coughing, sore throat, runny nose, sneezing, headache and fatigue), sentiment (negative, neutral or positive) and emotion (sadness, fear, anger, disgust or joy). Each feature was assigned a weight depending on the severity of its impact.
According to CDC listed symptoms of flu are fever, cough, sore throat, runny nose, muscle aches, headaches, fatigue, sneezing and chills. From the seven selected influenza symptoms, according to CDC, most usual and severe symptom of influenza virus is fever, and thus the highest symptom weight has been assigned to it. CDC prioritizes the flu symptoms based on the severity of their impact. According to CDC and other medical sources (WHO [26], Mayo Clinic [27]), from our seven selected influenza symptoms, most usual and most severe symptom of influenza virus is fever, and thus the highest symptom weight has been assigned to it. Following fever, cough, headache, and fatigue are 'common' flu symptoms. After a careful analysis of a large number of tweets, headache and fatigue have been found to have ambiguous meanings if left alone and are thus weighted with a lower score to avoid misclassification. According to CDC resources, runny nose, sneezing, and sore throat are categorized as symptoms that occur 'sometimes' for a person having a flu. Runny nose and sneezing are also common symptoms to many other diseases and are hence also weighted with a lower score. Due to careful interpretation of a great number of tweets, it is most often that self-reporting on influenza infection would very often include the key term 'flu' and/or state a number of flu symptoms. Thus the main keyword (flu or influenza) is assigned the highest weight. Table 4 shows the final weights assigned to all keywords.
The final ILI outcome computed by the formula (i.e., whether tweet is classified as ILI or non-ILI) will also depend on the threshold value used to discriminate between ILI and non-ILI tweets based on their score. The optimum threshold value will be determined during the evaluation of the proposed ILI formula. It will be determined as the cut-off value which when applied results in the highest Pearson correlation between the number of weekly ILI occurrences reported by CDC and the number of influenza occurrences predicted by ILI score (presented in Section 3.).
The formula computes the ILI score of a tweet as follows:

$$ILIscore = \frac{w_k + \sum_j^n w_{sy} + w_{se} + w_e}{t_1} \quad (1)$$

where $w_k = \{0, 10\}$is the set of possible weights that may be assigned to the main keywords. We assign $w_k = 10$if the main keyword (flu or influenza) appears in a tweet, 0 otherwise. $\sum_j^n w_{sy}$ represents sum of the weights of seven possible symptoms that may appear in a tweet, where$j \in [1, n]$,n is the total number of symptoms and $w_{sy} = \{0.5, 2, 3, 5\}$ is the set of possible weights that may be assigned to a symptom. Note that each symptom is considered at most once in order to avoid the computation of a repetitive symptom. E.g., tweet: "I have a headache, headache, headache." The headache symptom is considered only once; $w_{se} = \{1,2,3\}$. Watson outputs only one sentiment per tweet. We set the weight for negative sentiment to 3, neutral to 2, and positive sentiment to 1. $w_e = \{1, 2, 4, 5\}$. Watson computes and assigns a value for each of the five sentiments (i.e., sadness, fear, anger, disgust and joy). In (1), only emotion with the highest value computed by Watson is assigned a weight, lower values are not considered. We set 5 for 'sadness', 4 for 'fear' and 'anger', 2 for 'disgust', and 1 for 'joy'.

$$t1 = sum\left(\sum_i^k w_k, \sum_j^n w_{sy}, \max(w_{se}), \max(w_e)\right) = (10) + (5 + 3 + 2 + 0.5 + 0.5 + 0.5 + 0.5) + (3) + (5) = 30$$

The maximum score that can be achieved by the formula is 1. Table 5 shows some example tweets with the corresponding ILI score.

Table 4. ILI formula weights

Group	**ILI keyword**	**Weight**
Main keywords	Flu or Influenza	10
Symptoms	Fever	5
	Coughing	3
	Sore throat	2
	Runny nose	0.5
	Sneezing	0.5
	Headache	0.5
	Fatigue	0.5
Sentiment	Negative	3
	Neutral	2
	Positive	1
Emotions	Sadness	5
	Fear	4
	Anger	4
	Disgust	2
	Joy	1

Table 5. Example Tweets and the corresponding ILI score

Tweet	**ILI Score**
I stayed at home today as I have a flu.	0.6; (main keyword:10 ; symptoms: 0; sentiment: 3; emotion: 5)
I have a flu, I feel fatigue the whole day.	0.62 (main keyword:10 ; symptoms: 0.5; sentiment: 3;

Tweet	ILI Score
	emotion: 5)
I have a flu, I feel fatigue and headache the whole day.	0.63 (main keyword:10 ; symptoms: 1; sentiment: 3; emotion: 5)
I did not go to school today as I have a fever.	0.43 (main keyword:0 ; symptoms: 5; sentiment: 3; emotion: 5)
I have a flu, the fever is killing me:	0.77 (main keyword:10 ; symptoms: 5; sentiment: 3; emotion: 5)
I have a flu, the headache is killing me:	0.62 (main keyword:10 ; symptoms: 0.5; sentiment: 3; emotion: 5)
I have a fever, cough, sore throat, runny nose, and a headache, I have been sneezing the whole day and feel extreme fatigue.	0.67 (main keyword:0 ; symptoms: 5 + 3 + 2+ 0.5 + 0.5 + 0.5 + 0.5; sentiment: 3; emotion: 5)
I have a flu, ha ha. I am happy I do not have to go to school tomorrow	0.47 (main keyword:10 ; symptoms: 0; sentiment: 3; emotion: 1)
It's basketball season, looks like everyone has got a basketball fever.	0.4 (main keyword:0 ; symptoms: 5; sentiment: 2; emotion: 5)
It's football season, the flu game is on.	0.47 (main keyword:10 ; symptoms: 0; sentiment: 2; emotion: 2)

Fig. 1, illustrates the data collection process and analysis of Watson NLP service and the developed ILI Formula in prediction of influenza from Twitter posts.

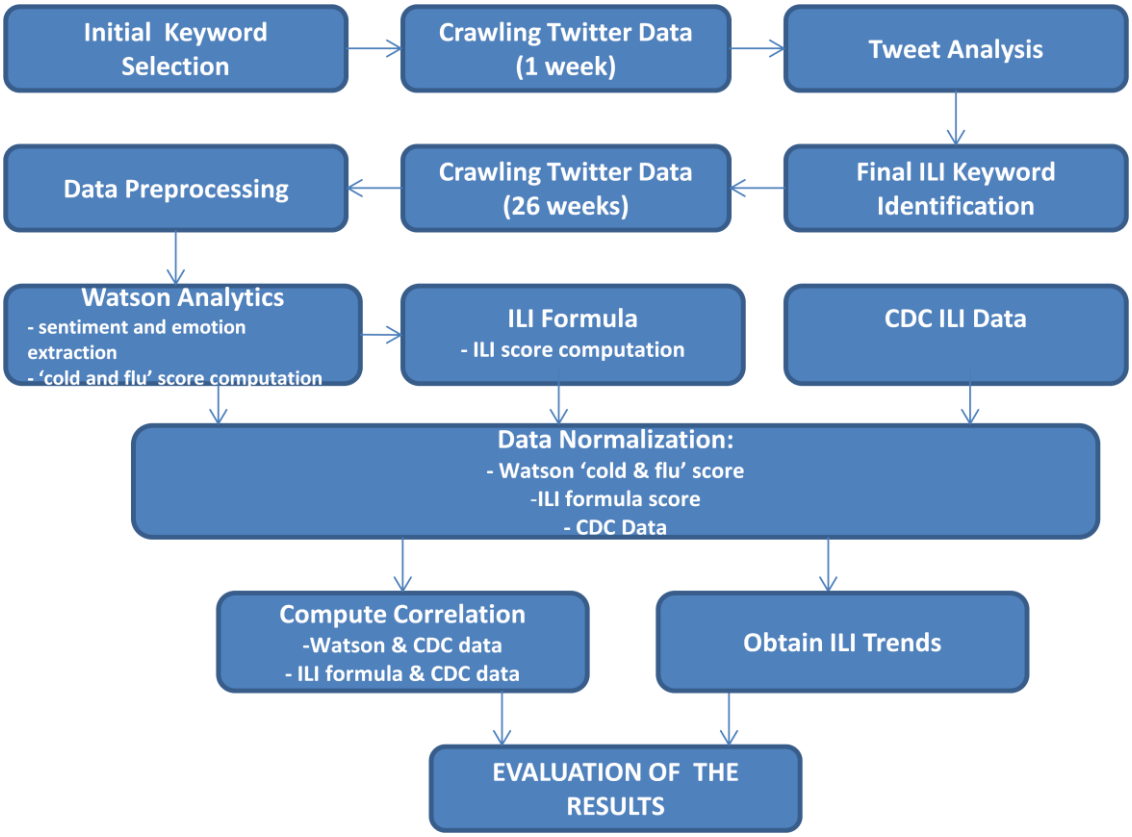

Figure 1. Data collection and analysis of Watson NLP service and the developed ILI formula in prediction of influenza from Twitter posts

2.7. ILI prediction using machine learning methods with Watson as a medical expert

We evaluate three machine learning methods (Random Forest [28], Logistic Regression [29] and K-NN [30]) to predict influenza from Twitter posts. To avoid computationally expensive task of Watson's sentiment and emotion computations, we use only the identified ILI features and tweet date as inputs into machine learning methods.

For each tweet, as the ground truth outcome for machine learning methods, the results of Watson's confidence score for 'cold and flu' category was used. For analysis, from the 10 evaluated confidence scores, we used the score that had the highest Pearson correlation results with the real world CDC data as Watson ground truth threshold value. Even though the cognitive computing predictions of Watson's confidence score are computationally expensive as well, the idea is that if successful, the need for such computations will only be performed during the training phase of machine learning methods. During testing, these methods will use the knowledge acquired in the training phase to make ILI predictions from new Twitter posts. The dataset was randomly divided into training data (66%) and testing data (34%).

2.8. ILI prediction using machine learning methods with ILI formula as a medical expert

In the final experiment, we tested how well do the mentioned machine learning methods perform when the results of the developed ILI formula are used as influenza ground truth, instead of Watson's confidence score (which is the main difference between the two machine learning experiments). Similarly as in the previous experiment, from the evaluated 10 ILI score thresholds, score that had the highest correlation with the CDC data was used as ILI formula ground truth threshold value. ILI keywords and Tweet date were used as influenza predictors.

Figure 2. depicts the data collection process and prediction of influenza from Twitter using machine learning methods with Watson and ILI formula as medical experts (i.e. ground truth).

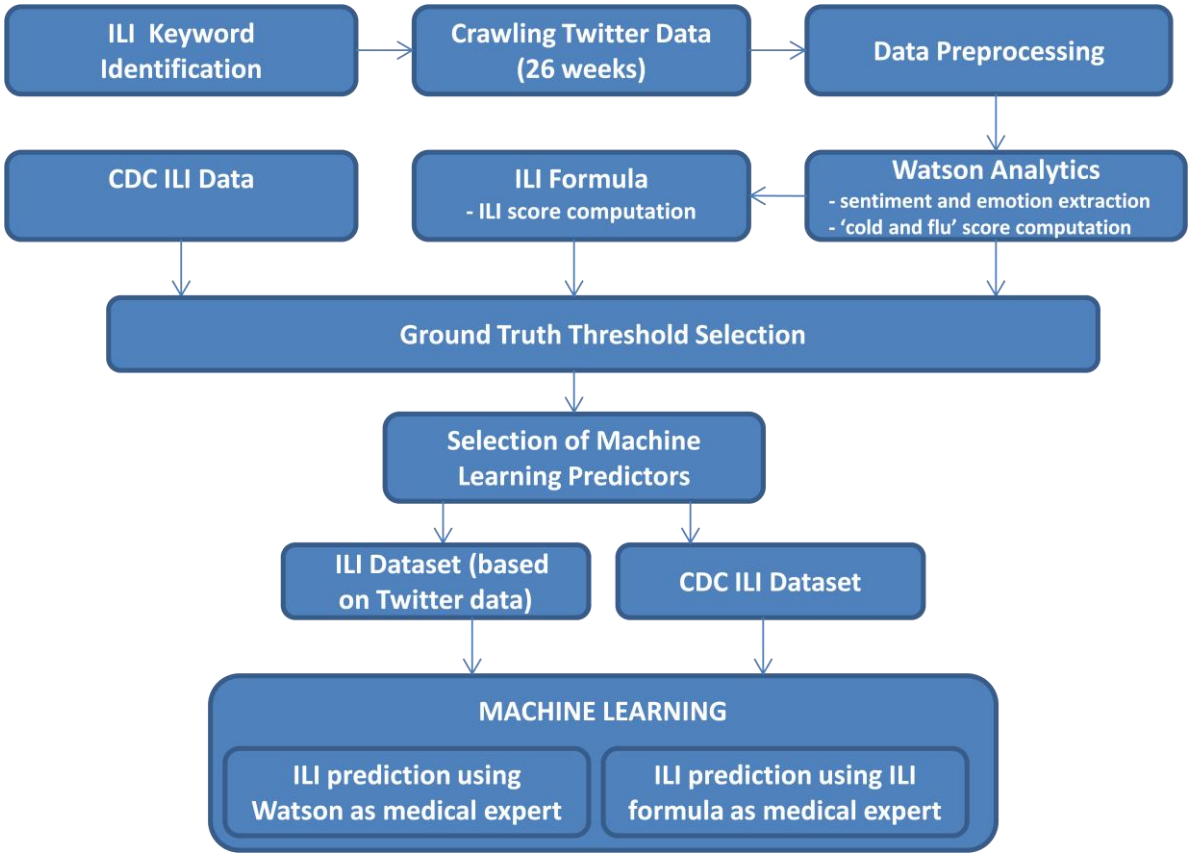

Figure 2. Data collection and prediction of ILI from Twitter posts using machine learning methods with Watson and ILI formula as medical experts

3. Results and discussion

3.1. Correlation results

Table 6. shows Pearson correlation results between (normalized) weekly number of positive ILI occurrences from CDC data and: a) normalized weekly number of positive Watson ILI tweets computed by using confidence scores for 'cold and flu' category, b) normalized weekly number of positive ILI tweets computed by the developed ILI formula.

The results show that out of the ten tested threshold values for both ILI predictors, 0.65 threshold value produces the highest correlation score: 0.55 for Watson and 0.91 for the developed ILI formula. This value has been selected to be used as a threshold ground truth value throughout the performed experiments.

Table 6. Pearson correlation results

Tested threshold values for Watson and ILI formula predictors	**Pearson correlation between CDC and Watson**	**Pearson correlation between CDC and ILI formula**
0.5	0.51	0.75
0.55	0.49	0.86
0.6	0.50	0.87
0.65	**0.55**	**0.91**
0.7	0.50	0.88
0.75	0.28	0.90
0.8	0.27	0.90
0.85	0.26	0.91
0.9	0.28	0.87
0.95	0.54	0.81

Fig. 3, shows the comparison of weekly aggregated and normalized number of positive ILI tweets computed by Watson and the developed ILI formula against the CDC data for the tested period (from 44th week, 2017 until 17th week, 2018). As shown in the figure, prediction of influenza using the ILI formula better follows the real world data then does the Watson's prediction of influenza.

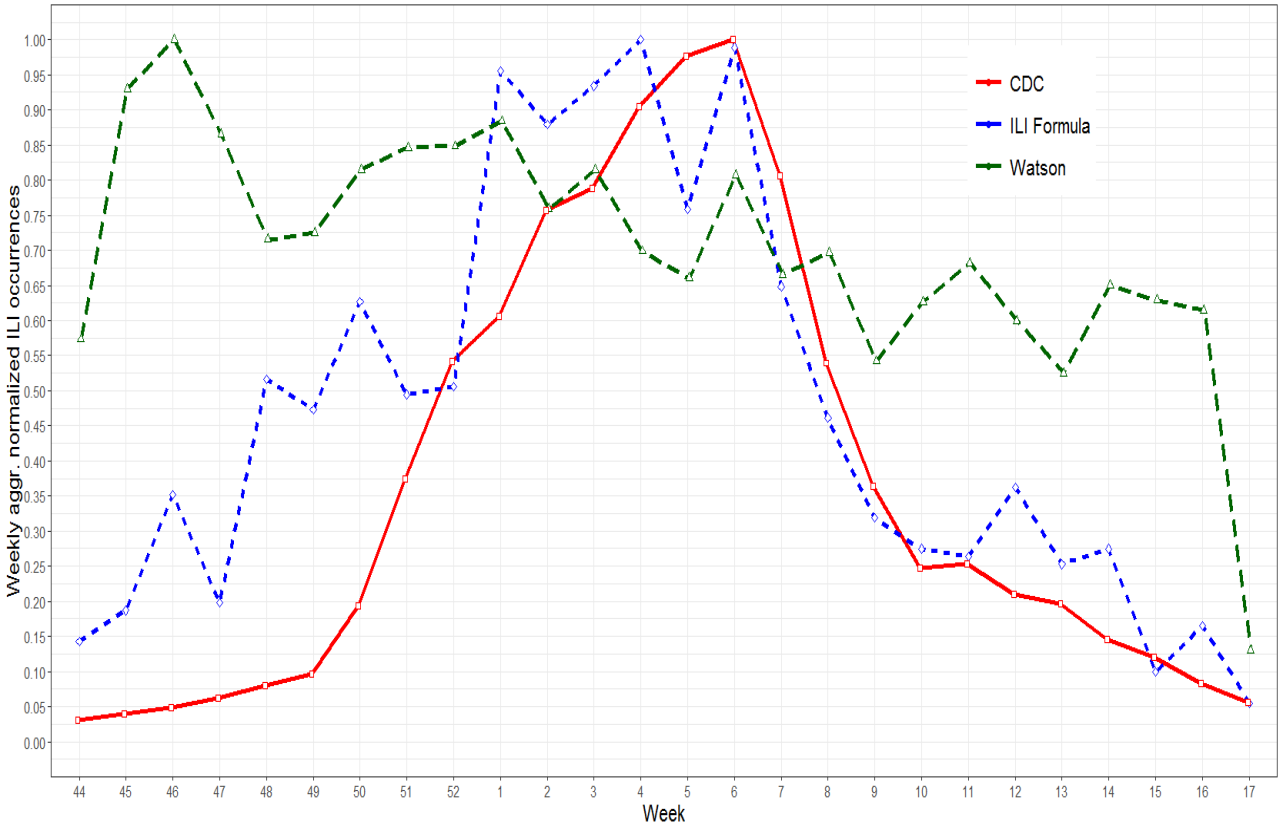

Figure 3. Comparison of weekly aggregated and then normalized number of positive ILI tweets computed by Watson confidence score (dashed green line) and the developed ILI formula (dashed blue line), using 0.65 as influenza threshold value, against the real-world CDC (solid red line) influenza reports over a 26-week period (from 44th week, 2017 until 17th week, 2018).

0.55 correlation has been achieved between the data predicted by Watson and the data observed by CDC. Further analysis of the results indicate that there is a large 12 week discrepancy between the peak occurrence of ILI predicted by Watson (week 46) and ILI peak based on CDC data (week 6). This indicates that Watson confidence score for 'cold and flu' category might not be a reliable tool for tracking influenza using the data collected from Twitter.

A strong correlation 0.91 has been found between ILI formula and the CDC data. Peak value of ILI predicted by the proposed model is in week four (recall that the CDC ILI peak was in week six). Thus, the developed ILI model was able to detect influenza two weeks ahead of the official CDC influenza reports, using real time Twitter data.

Analyzing the tweets presented in Table 5, and applying the computationally optimum ILI threshold score (0.65), it can be seen that even though high weight is assigned to the main keyword 'flu/influenza', this is not enough to classify the Twitter post as ILI positive (even if a tweet also has a negative sentiment and 'sadness' for emotion). For example, the tweet "I stayed at home today as I have a flu", is a non-ILI tweet with 0.6 ILI score. Thus, additional ILI symptom(s) are further required for the tweet to be classified as ILI tweet. On the other hand, "I have a flu, the fever is killing me", and "I have a fever, cough, sore throat, runny nose, and a headache. I have been sneezing the whole day and feel extreme fatigue" are classified as a ILI tweets with 0.77 and 0.67 ILI scores respectively. In addition, the former tweet also contains the main keyword 'flu' and a high weighted ILI symptom 'fever'. The latter tweet does not contain the main keyword flu or influenza but has several flu-related symptoms (i.e., fever, cough, sore throat, runny nose, headache, sneezing, fatigue) which yields a result greater than the experimentally found optimum ILI threshold score (0.65). Furthermore,

both of these tweets have a negative sentiment and sadness for emotion, which contribute to greater than 0.65 overall ILI score. As noted earlier, ILI prediction should not be only limited to ILI keywords. Sentiment and emotion also need to be included in the prediction. Combining ILI keywords with sentiment and emotion helps to understand the linguistical meaning of tweeted text with an aim to detect real ILI occurrences.

3.2. Machine learning results

Performance of machine learning methods was evaluated by using the results of the confusion matrix which contains true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). True positives indicate correctly classified ILI tweets; true negatives indicate correctly classified non-ILI tweets; false positives indicate incorrectly classified ILI tweets, and false negatives indicate incorrectly classified non-ILI tweets. When classified by Watson, from the tested tweet dataset, 68.7% tweets were classified as ILI and 1.3% as non-ILI tweets. When judged by the developed ILI formula, tested tweet dataset contains 10.1% ILI and 89.9% non-ILI tweets. Thus the tested datasets are not well balanced: Watson dataset has a much higher number of ILI than non-ILI tweets. On the other hand, ILI formula dataset has a much higher number of non-ILI than ILI tweets. Because of this, other than the overall prediction accuracy, the best performing method was also selected based on the highest sensitivity and specific rates. These metrics are also the most commonly used performance measures for diagnostic predictors and are defined as follows:

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{FP}+\text{FN}+\text{TN})} * 100 \quad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP}+\text{FN})} * 100 \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN}+\text{FP})} * 100 \quad (4)$$

Accuracy is used to evaluate the overall performance of the classifier. It refers to the number of correctly classified tweets by the algorithm over all of the predictions made. Sensitivity refers to the proportion of tweets that are ILI positive and were predicted by the algorithm as ILI positive tweets. The specificity refers to the proportion of tweets that are ILI negative and were predicted by the algorithm as ILI negative tweets.

As shown in Table 7, when Watson confidence score for 'cold and flu' category with 0.65 threshold is used as a medical expert to predict influenza with tested machine learning methods, all three methods achieved a similar overall accuracy: Random Forest 68.15%, Logistic Regression 68.70%, KNN 68.20%. For all three tested methods, obtained sensitivity results are 98.65% and above. However, the specificity results are very low: Random Forest and KNN 1.18% and Logistic Regression 0%. Only a small number (or 0 in the case of Logistic Regression) of non-ILI tweets have been correctly classified as non-ILI tweets (and are classified as ILI tweets instead). Such misclassification errors may cause false alarms for government and public healthcare services. Thus, the identified ILI keywords cannot be reliably used to predict influenza using the tested machine learning methods and Watson's confidence score for 'cold and flu' category as a medical expert.

When ILI formula is used as a medical expert for influenza prediction, all three tested machine learning methods achieved high overall accuracy (Table 8.): Random Forest and Logistic Regression 97.54%, KNN 97.51%. All methods also achieved very high sensitivity and specificity results. Sensitivity: Random Forest and Logistic Regression 98.96%, KNN 98.70%; specificity results: all three methods achieved 97.38%. These results show when the developed ILI formula is used to train tested machine learning methods with identified influenza keywords (Table 2) and tweet date as discriminative input features, tested machine learning methods were able to classify new tweets with high accuracy.

Table 7. Binary classification results using Watson as a medical expert

	Random Forest	**Logistic Regression**	**K-NN**
Accuracy (%)	68.15	68.70	68.20
Sensitivity (%)	98.65	100	98.73
Specificity (%)	1.18	0	1.18
ROC Area	0.543	0.522	0.543

Table 8. Binary classification results using ILI formula as a medical expert

	Random Forest	**Logistic Regression**	**K-NN**
Accuracy (%)	97.54	97.54	97.51
Sensitivity (%)	98.96	98.96	98.70
Specificity (%)	97.38	97.38	97.38
ROC Area	0.993	0.993	0.993

3.3. Comparison with state of the art

Comparison of classifiers built in this work with state of the art results is summarized in Table 9. State of the art results were obtained from the articles that performed a comprehensive review on the use of the social media for public health research [7,31,32]. From the evaluated methods performed in each study, we report methods that achieved highest results. For the purpose of this work, from the reported articles we selected only those that apply supervised machine learning methods to monitor influenza-like-illnesses using Twitter as a social media platform. As the COVID-19 pandemic is still raging we also include the results of the recent work of Kelin et al. [2021] where Twitter is used for tracking COVID-19 (which similarly to influenza is also a highly infectious respiratory disease.)

When compared with the results of other studies, the results obtained with the ILI formula are satisfactory with very strong correlation of 0.91 with CDC data, and the ability to detect ILI two weeks ahead of CDC influenza reports. Furthermore, we achieved 97.54% overall classification accuracy with Random Forest and Logistic Regression methods. However, when using Watson to predict ILI we obtained a 0.55 correlation with CDC data, a large discrepancy (12 weeks) with the CDC data, and the highest overall classification accuracy of 68.70% achieved with Logistic Regression method.

Bronitowski et al. [33] proposed a SVM classifier to differentiate between actual flu tweets and "chatter" tweets, (i.e. tweets related to flu). They achieved 0.93 correlation with CDC data, and Pearson correlation of 0.88 with the data of the Department of Health and Mental Hygiene of New York City. Alex Lamb et al. [34] developed a model that differentiates between the infection and awareness tweets and reported high correlation of 0.9897 with CDC data. Santos and Matos [2] proposed a two-phase model where they first identify ILI Tweets using several classifiers achieving the highest precision score of 78% with Naïve Bayes and SVM models. In the second phase, they applied linear regression model for monitoring health data using classified tweets and search queries and obtained a correlation of 0.89 with Influenzanet data, a system that monitors ILI activities in Europe. Lee et al. [35] proposed a system based on multilayer perceptron with backpropagation algorithm and achieved a correlation of 0.93 with CDC data for the real time analysis (for the current week prediction model), and a correlation of 0.71 for the one-week ahead forecast model. Kelin et al.

2021 [36] developed a deep neural network based on bidirectional encoder representation from transformers (BERT) for tracking COVID-19 using Twitter data, and obtained F1 score of 76% (precision: 76%, recall: 76%).

Table 9. Summary of the results obtained by state of the art studies

Author	Method	Reported Results
David A. Bronitowski et al.[33]	SVM	Pearson correlation of 0.93 with CDC data.
Alex Lamb et al.[34]	Log-linear model with L2 regularization	Pearson correlation of 0.9897 with CDC data.
Santos and Matos [2]	SVM, Naïve Bayes (NB), Linear Regression	Precision score achieved with SVM and NB: 78%; Pearson correlation of 0.89 with Influenzanet achieved by linear regression.
Lee et al. [35]	ANN: Multilayer perceptron with back propagation	Pearson correlation of 0.93 with CDC data.
Kelin et al. 2021 [36]	Deep neural network based on BERT	F1-score: 76% (precision: 76%, recall: 76%)
IBM Watson	**Logistic Regression**	**Pearson correlation of 0.55 with CDC data; Accuracy: 68% (sensitivity: 100%, specificity: 0%)**
Proposed method with ILI formula	**Random Forest, Logistic Regression**	**Pearson correlation of 0.91 with CDC data; Accuracy: 97.54% (sensitivity: 98.96%, specificity: 97.38%)**

4. Conclusion

There is an increasing amount of research done on predicting the outbreak of diseases using social media data such as Twitter. The main aim of this work was to evaluate if Watson's confidence score for 'cold and flu' category can be used for tracking the spread of influenza in the period of ILI season. To do this we computed Pearson correlation coefficient between weekly aggregate of Watson positive ILI tweets and real-world CDC ILI data. To find the best confidence score threshold for positive Watson ILI tweets, we analyzed the correlation between normalized number of weekly Watson ILI tweets as decided by ten different threshold values and weekly normalized number of CDC reported ILI occurrences.

We also developed and presented a model for prediction of ILI outbreak using the data collected from Twitter. To predict ILI, the model uses the identified ILI keywords as well as the sentiment and emotion results computed by IBM Watson's NLU service. Analysis of the results revealed that the proposed model was able to predict ILI two weeks ahead of CDC. Indeed, CDC publishes influenza reports within two weeks of delay. Whereas there was a large, 12 week, discrepancy between ILI peak predicted by Watson and real world CDC data.

To reduce the computational costs of Watson's cognitive computing, application of machine learning methods to predict influenza from Twitter posts resulted in high accuracy results using ILI formula score as medical

expert. However, when Watson is used as a medical expert, evaluated machine learning methods were unable to successfully predict influenza.

Performed analysis can be used to guide IBM Watson to improve the confidence score of 'cold and flu' category to predict influenza using social network data, such as Twitter, that will better correlate with the real world data by incorporating (or better utilizing) the proposed ILI keywords into its NLU intelligence system. We believe that continued evaluation of existent and established means of surveillance that IBM Watson represents, is necessary for effective and appropriate response to pandemics as the one we are currently witnessing (i.e. COVID-19 pandemic).

As one of the future research directions, the described approach can be further expanded and applied for tracking of other diseases or medical conditions using messages posted by millions of people daily on social media.

References

[1] H. Achrekar, R. Gandh, R. Lazarus, et al., "Predicting Flu Trends using Twitter data," In: *IEEE Conference on Computer Communications Workshop on Cyber-Physical Networking Systems*. Shanghai, China, 10-15, April 2011.

[2] J.C. Santos and S. Matos, "Analysing Twitter and web queries for flu trend prediction," *Theoretical Biology and Medical Modelling*; vol. 11, no. 1, 2014. DOI: 10.1186/1742-4682-11-S1-S6.

[3] F. Wang, H. Wang, K. Xu, et al., "Regional Level Influenza Study with Geo-Tagged Twitter Data," *Journal of Medical Systems* 2016; vol. 40, no. 189, 2016.

[4] H. Hu, H. Wang, F. Wang, et al., "Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network," *Sci Rep*., vol. 8, no. 1, 2018. DOI: 10.1038/s41598-018-23075-1. PMID: 29559649; PMCID: PMC5861130.

[5] S. Molaei, M. Khansari, H. Veisi, et al., "Predicting the spread of influenza epidemics by analyzing twitter messages," *Health and Technology*, vol. 9, no. 4, 2019. DOI:10.1007/s12553-019-00309-4

[6] S. Jordan, S. Hovet, I. Fung, et al., "Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response," *Data (Basel)*, vol. 4, no. 1, 2018. DOI: 10.3390/data4010006

[7] A. Alessa, and M. Faezipour, "A review of influenza detection and prediction through social networking sites," *Theoretical Biology and Medical Modelling vol. 15*, no. 2, 2018. DOI: 10.1186/s12976-017-0074-5

[8] M. Al-garadi, M.S. Khan, K.D. Varathan, et al., "Using online social networks to track a pandemic," *Journal of Biomedical Informatics*, vol. 62, 2016. DOI:10.1016/j.jbi.2016.05.005

[9] Centers for Disease Control and Prevention (CDC), https://www.cdc.gov (accessed May 2018)

[10] C. Viboud and A. Vespignani, "The future of influenza forecasts," *Proceedings of the National Academy of Sciences*, vol. 116, no. 8, pp. 2802-2804, 2019. DOI: 10.1073/pnas.1822167116

[11] CDC Flu Challenge, www.cdc.gov/flu/news/predict-flu-challenge.htm 2013. (accessed February 2019)

[12] CDC Flu Sight, https://www.cdc.gov/flu/weekly/flusight/index.html (accessed February 2019)

[13] R. High, "The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works," RedBooks, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 USA, 2012.

[14] Wang Y and Chiev V. "On the cognitive process of human problem solving," *Cognitive Systems Research*, 2008. DOI:10.1016/j.cogsys.2008.08.003

[15] E. Strickland, ", IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care," in *IEEE Spectrum*, vol. 56, no. 4. pp. 24-31, 2019.

[16] R. Cuocolo, T. Perillo, E. De Rosa, et al., "Current applications of big data and machine learning in cardiology," *Journal of Geriatric Cardiology*, pp. 601-607, 2019.

[17] R. Deo and S. Panigrahi, "Performance Assessment of Machine Learning Based Models for Diabetes Prediction," *IEEE Healthcare Innovations and Point of Care Technologies*, (HI-POCT), Bethesda, MD, USA, pp. 147-150, 2019. DOI 10.1109/HI-POCT45284.2019.8962811
[18] R. Deo and S. Panigrahi, "Prediction of Hepatic Steatosis (Fatty Liver) using Machine Learning," *In Proceedings of the 2019 3rd International Conference on Computational Biology and Bioinformatics (ICCBB '19). Association for Computing Machinery*, New York, NY, USA, pp. 8–12, 2019. DOI:https://doi.org/10.1145/3365966.3365968
[19] J.M. Luna, H.H. Chao, E.S. Diffenderfer, et al., "Predicting radiation pneumonitis in locally advanced stage II–III non-small cell lung cancer using machine learning," *Radiotherapy and Oncology*, vol. 133, pp. 106-112, 2019.
[20] National Health Service (NHS), https://www.nhs.uk (accessed May 2018)
[21] WebMD, https://www.webmd.com/cold-and-flu/flu-guide/flu-symptoms-types(accessed May 2018)
[22] Healthline, https://www.healthline.com/health/cold-flu/early-flu-symptoms#emergency-symptoms (accessed May 2018)
[23] Google Trends, https://trends.google.com/trends (accessed May 2018)
[24] J. Baker "TweetScraper is a Scrapy crawler/spider for Twitter Search without using API," 2018. https://github.com/jonbakerfish/TweetScraper. (accessed May 2018)
[25] Centers for Disease Control and Prevention, FluView. Weekly U.S. Influenza Surveillance Report. CDC, 2018. https://www.cdc.gov/flu/weekly/index.htm. (accessed June 2018)
[26] World Health Organization, www.who.int, (accessed May 2018)
[27] Mayo Clinic www.mayoclinic.org, (accessed May 2018)
[28] Biau G, Scornet E. "A random forest guided tour," *TEST*, vol. 25, 2016. DOI: 10.1007/s11749-016-0481-7
[29] Dreiseitl S, Ohno-Machado L. "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review," *Journal of Biomedical Informatics*, vol. 35, no. 5–6, pp. 352-359, 2002. DOI: 10.1016/S1532-0464(03)00034-0.
[30] Cover T, Hart P. "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27. 1967. DOI: 10.1109/TIT.1967.1053964
[31] G. Aakansha and K. Rahul, "Social media based surveillance systems for healthcare using machine learning: A systematic review," *Journal of Biomedical Informatics*, 2020. DOI: 10.1016/j.jbi.2020.103500
[32] E.O. Oduwa, D.L.I. Beatriz, L. Iain, et al., "A scoping review of the use o Twitter for public health search," *Computers in Biology*, 2020. DOI: 10.1016/j.compbiomed.2020.103770
[33] D.A. Bronitowski, 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, 2013.
[34] A. Lamb A, M.J. Paul and M. Dredze, "Separating fact from fear: Tracking flu infections on twitter," in *HLT-NAACL*, pp. 789-95, 2018.
[35] K. Lee, A. Agrawal, and A. Choudhary, "Forecasting Influenza Levels Using Real-Time Social Media Streams," in *IEEE International Conference on Healthcare Informatics (ICHI) 2017*, pp. 409-414, 2017. DOI: 10.1109/ICHI.2017.68
[36] A.Z. Klein, A. Magge, K. O'Connor, et al., "Toward Using Twitter for Tracking COVID-19: A Natural Language Processing Pipeline and Exploratory Data Set," *J Med Internet Res*, vol. 23 no. 1, 2021. DOI: 10.2196/25314