# Human-centric Computing and Information Sciences

August 2021 | Volume 11

www.hcisjournal.com

KIPS
Korea Information Processing Society

KIPS CSWRG
Korea Information Processing Society
Computer Software Research Group

HCIS | Human-centric Computing and Information Sciences

**RESEARCH(Research Manuscript)** **Open Access**

Hum. Cent. Comput. Inf. Sci. (2021) 11:33
https://doi.org/10.22967/HCIS.2021.11.033
Received October 31, 2020; accepted December 30, 2020; published August 30, 2021

# AI Based Forecasting of Influenza Patterns from Twitter Information Using Random Forest Algorithm

Vimal Shanmuganathan[1], Harold Robinson Yesudhas[2], Kaliappan Madasamy[1], Abdulellah A. Alaboudi[3], Ashish Kr. Luhach[4], Noor Zaman Jhanjhi[5,*]

## Abstract

Nowadays, people are highly addicted to social media or any other social platform, and there is no one without the Internet or Android mobile devices. Therefore, social media is considered to be a large dataset repository. For gathering information on the Internet life, Twitter is one of the largest data repositories where users can share information through tweets ($\#Hashtags$). Early detection of influenza through Twitter information enables a timely response to an influenza pandemic. This paper proposes a machine learning methodology to detect the flu (influenza) virus spreading among people mainly across Asia. The proposed work applies a classification based on the random forest algorithm, a regulated AI strategy for analyzing the text data in the dataset and finding the accuracy level based on its classification. The computed matrix through the proposed technique is used to predict the accuracy with the training and responses using substitution function. The bootstrapping system utilizes the execution model to estimate the tree-based clustering model. The Twitter dataset was collected with many positive tweets, and the text cleaning was enabled with various classifiers to enhance the prediction. Subsequently, the text analysis processing was carried out using multiple plots, such as box plot, scatter plot, lexical dispersion plot, term frequency, and plotting. The proposed system uses the concept of geospatial data analytics for predicting the flu-affected region. Further, the proposed system enhanced the accuracy level of the flu prediction in the given dataset.

## Keywords

Tweetluenza-Flu Prediction, Social Media Content Mining, Classification, Random Forest Algorithm, Geospatial Data Analytics

# 1. Introduction

In recent years, numerous infectious viruses emerged and quickly developed and spread among individuals without appropriate consciousness of the infection [1]. Almost 3–5 million of people



***Corresponding Author:**Noor Zaman Jhanjhi(Noorzaman.jhanjhi@taylors.edu.my)
[1]Department of Computer Science and Engineering, Ramco Institute of Technology, Rajapalayam,Tamilnadu,India
[2]School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India
[3]Department of Computer Science, Shaqra University, Shaqra, Saudi Arabia
[4]Papua New Guinea University of Technology, Lae, Papua New Guinea
[5]School of Computer Science and Engineering (SCE), Taylor's University,Subang Jaya, Malaysia

severely suffering from viral infections are reported each year [2], which prompts to the real existence stringing for the individuals. One of these viral infections is influenza (flu) [3]. Flu complications may include viral pneumonia, emergence of infectious microscopic organisms, pneumonia, and sinus infections, which significantly affect adults [4]. As a result, contacting people through social media is an appropriate approach, as Internetlife is an essential asset [5]. Twitter is the most well-known micro blogging platform and is one of the multilingual long-distance interpersonal communication administrations on which customers collaborate with a brief message known as a tweet. Normally, there are a total of 500 million tweets every day and around 6,000 new tweets every second [6]. Thus, medical-related data can be effectively removed by using tweets to find the well-being tweets of individuals [7]. Furthermore, it is recognized that social media is a tool for expressing oneself in the third person [8]. As a result, detecting every single clinical problem is challenging [9].

In this work, geospatial data analytics is used to obtain a geographical perspective on flu infection spreading in Asia, especially in India. Geospatial data analytics help us anticipate the districts affected by influenza virus. This Forecasting Influenza methodology can be employed to recognize the risk factor of flu infection from web-based life assets, such as Twitter. This architecture allows a Twitter content-based infection examination and provides recognizable proof of causes and preventive proportion of flu-related issues. The Forecasting of Influenza reconnaissance process model has been used for identifying the spread of flu sickness among the public.

The objectives of this study are as follows:

- Collect data, preprocessing tweet information, and extract important highlights based on tweet settings. The Forecasting Influenza procedure is established to know the risk factors of flu infection through Twitter.
- Use geospatial data analytics to obtain geospatial information on the spread of flu infection.
- Use experimental results and conduct analysis to determine significant improvements in the precision results and map plotting through the tweets' spatial analysis.

The remainder of this paper is organized as follows. Section 2 presents the flu forecasts and current techniques. Section 3 discusses data preprocessing, including unsupervised learning and algorithm specifics. Section 4 presents the experimental analysis, which includes geospatial data analytics with various predictions and risk factors. Finally, Section 5 concludes the paper and presents points for development in the future.

# 2. Related Works

One reason for the exponential development of organized and unstructured information is the expanding ubiquity of different online lives and blogging destinations [10]. This tremendous amount of information can be utilized to shape bits of knowledge and carry out important activities in an open domain. Thus, Twitter, an online life stage, has gained the interest of specialists for offering important bits of knowledge tocommon life[11, 12]. In this paper, we focused on the recovery of health-related tweets using a hash tag-based approach and Twitter confirmation bundles [13, 14]. Furthermore, we isolated prevalent illness data from well-being tweets and conducted area-based analysis to identify recurrent zones where a particular disease is more prevalent [15, 16].

Tag-based approach adds to the understanding of the current circumstances occurring and infection spreading in different locations [17]. With Health being the domain, the retrieval of tweets only from verified customers who can chat on Health domain is categorized as client-based retrieval. It is known as the primary level of arrangement used. The other level is the hash label method, which is used to perform a first unigram-based order to obtain data associated with multiple diseases, such as cancer, tuberculosis, diabetes, polio, and human immunodeficiency virus (HIV), and their geographical locations [18, 19].

Hashtag, which functions like a meta-tag like #keyword [10]. Tweets are limited to 140 characters, in

light of the objectives of Twitter's short message servicetransport structure [20]. Because tweets can be continuously passed on to enthusiasts, anyone can see tweets on Twitter, whether or not. In the same manner, Twitter uses an open-source web structure called Ruby on Rails (RoR). The application programming interface (API) is open and available to application creators [21].

Twitter information has been involved in enhancing the surveillance of seasonal influenza and medicinal drugs [22]. The pandemic is analyzed based on Twitter information [23]. Twitter and other social media platforms play a significant role in the uptake of vaccine worldwide [24]. The antimicrobial tool has been constructed to utilize Twitter information to provide a quality-based evaluation [25]. The geographically based Twitter information is used for delivering the regional level influenza [26]. The sociological perspective is implemented for Twitter analysis for developing health through technology [27]. A communication network model has been created for identifying Twitter information [28–32].

# 3. Proposed Work

## 3.1 Data Gathering and Preprocessing

As Twitter is the most established social networking website, it consists of various online journals from different locations worldwide. Rather than downloading the entire online journals, we focus on a single point, which is the influenza, and extract all pages related to it as content documents using a mining device. Thus, to remove tweets, we execute the Twitter API code using R studio [33]. The Twitter search application interface code needs to be executed using the R software [34]. To possess the interface to Twitter tweets, the association must be found out to the Twitter site. We scan the tweets and spare them to the CSV record. We then prepare the information. R bundles must be implemented first via R's introduce order and imported into R using the library command to perform information pre-handling on the Twitter API process. Data processing is performed when text processing is done [35]. If the hashtag #influenza is provided to R package, precise data (influenza flu) are obtained. Thus, quality sentence explanations are acquired [36]. It says about the preparation of the gathered information. Once we obtain information using Twitter, the next stage typically include the following steps: quiet pre-preparing, control, cleaning, designing, and separating [37]. Preprocessing strategies are important for improving outcomes and extract tweets that are considered to be untested information and pre-process them. Fig. 1 presents the data preprocessing architecture in a step-by-step manner to apply the algorithm for the sparse data.

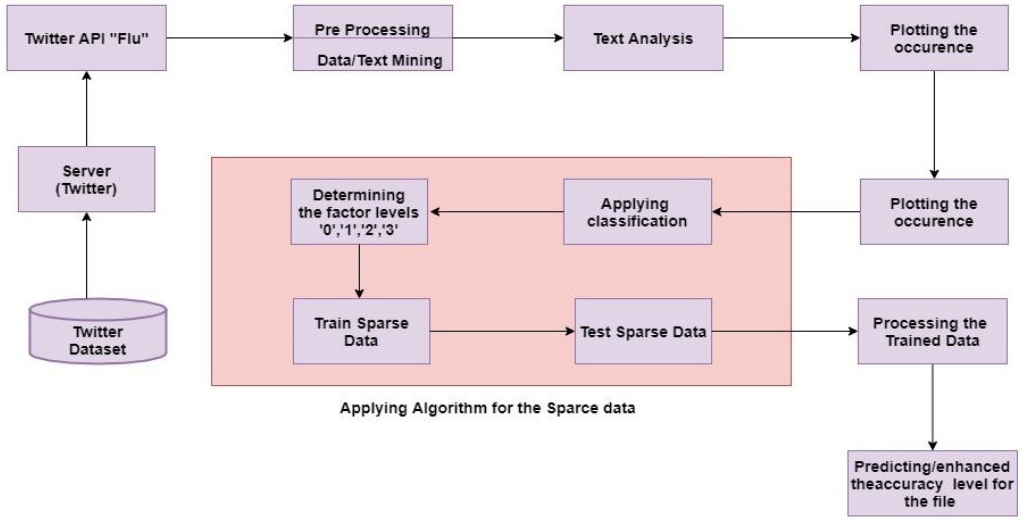



**Fig. 1.** The architecture of data preprocessing.

## 3.2 Unsupervised Learning

This method uses a large number of x-vectors of similar estimation without using any class names factors. A different figure of authenticity has been enforced to revamp, leaving the parameter release to defective closures. The goal is to generate data to keep track of whether it fails under different loads [38, 39]. In discretionary technique, the methodology is to think of the main data class 1 and construct a beneath normal of a comparable size distinguished as class 2. The manufactured beneath normal is created by randomly exploring the univariate transports of the primary data. Here is how a singular individual from class 2 is made—the chief organizes assessed from the N regards $\{x(1,n)\}$. The consequent encourage is inspected separately from the N regards $\{x(2,n)\}$, and so on. Along the way, class 2 has a flow of independent discretionary components, and everyone in the main data has the same univariate allocation as the relating variable [40, 41]. The dependency organization of the primary data is now in class 2. In any case, there are two classes, and this fictitious two-class question can be felt in the self-assured boondocks[42, 43]. This allows the sum of the self-assertive boondocks choices to be applied to the principal unlabeled instructive assortment[44]. Fig. 2 demonstrates the architecture of unsupervised learning.

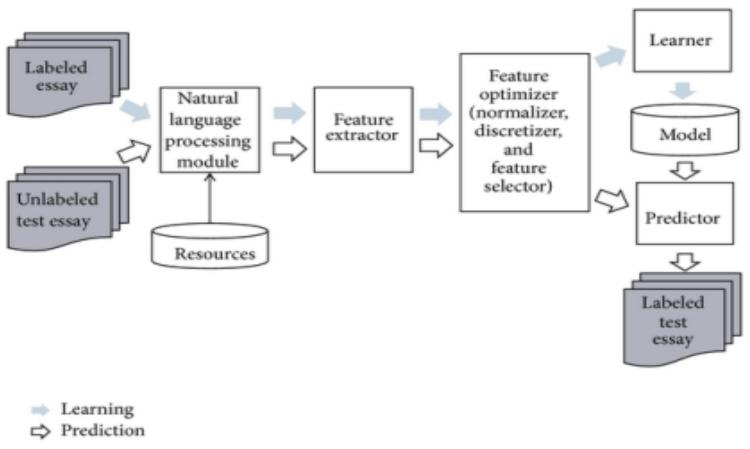


**Fig. 2.** The architecture of unsupervised learning.

## 3.3 Algorithm

**Input:** Collect the information from Twitter API

1. Tokenize the words and perform stemming for the preparation set and speak to them as a corpus
2. Using the corpus train the dataset first, to apply the random forest calculation
3. Given a training group $\alpha = \alpha_1, \ldots, \alpha_n$ with responses $\beta = \beta_1, \ldots, \beta_n$, stowing on and on picks an arbitrary value: For $x = 1, \ldots, Bm$: Test, with substitution, n preparing from $\alpha, \beta$; call these. $\alpha_x, \beta_x$
4. Train a characterization or deterioration tree $f_b$ on $\alpha_x, \beta_x$.
5. Test the dataset by setting the seeds for predicting the similarity matrix
6. By using the similarity matrix to get the probability value(accuracy)

**Output:** Calculating the matrix to predict the accuracy.

Fig. 3 illustrates the applied algorithms for the dataset that predict RF value from 0 to 3.

| | predictRF | | | |
|---|---|---|---|---|
| | 0 | 1 | 2 | 3 |
| 0 | 103 | 93 | 0 | 0 |
| 1 | 22 | 632 | 0 | 0 |
| 2 | 4 | 37 | 5 | 0 |
| 3 | 0 | 4 | 0 | 1 |

**Fig. 3.** Applied algorithms for the dataset.

## 3.4 Bootstrap Clustering

The preparation calculation for random forests relates to the overall method of bootstrap aggregating. Given a training group $\alpha = \alpha_1, \ldots, \alpha_n$ with responses $\beta = \beta_1, \ldots, \beta_n$, implemented separately and picks a random illustration with substitution of the preparation set to these examples as shown in Equation (1):

$$For\ x = 1, \ldots, Bm \quad (1)$$

Test, with substitution, n preparing models from $\alpha, \beta$; like $\alpha_x, \beta_x$. Train an arrangement or relapse tree $f_b$ on $\alpha_x, \beta_x$. Prospects for hidden parameters $x'$ are usually made in the midst of organizing by discovering predictions from all of the individual relapse trees on $x'$, as shown in Equation (2):

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b\ (x') \quad (2)$$

This bootstrapping system improves model execution by lowering the model's distinction without increasing the trend. This means that while a single tree's estimations are extremely sensitive to the commotion in its preparation group, the average of many trees isnot. The length of the trees is notrelated.Preparing various trees on a singular preparing group might give unequivocally associated trees; bootstrap analyzing is a strategy for de-associating the trees by demonstrating them unmistakable preparing sets. In addition, the difference of the projections from the entire character deterioration trees

on $x'$ as shown in Equation (3) is commonly used as a gauge of the expectation's vulnerability.

$$\delta = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x'-\hat{f})^2}{B-1}} (3)$$

The numberof tests, *B*, might be a complementary parameter. Usually, some trees are used, dependent upon the preparation set. A perfect amount of trees *B* would be established with cross-approval or by watching the out-of-pack blunder. The average forecast botch on every preparation test $x_i$, with only the trees that did not have $x_i$ in their bootstrap test. The preparation and test blunder will increase by and large level off the valuable trees.

## 3.5 Outliers Prediction

The random forest algorithm is a current figuring that is unexact in precision. As the timberland building progresses, it establishes an internal fair-minded check of the theory flaw. It gives assessments of what components are noteworthy in the plan. Moreover, it can influence an excessive number of information factors without variable deletion. It also has a ground-breaking system for surveying missing data and maintains exactness when a colossal degree is missing. The created timber grounds can be set aside for later use on other data. The random forest algorithm has procedures for modifying screw-ups in class masses of unbalanced instructive lists. Models that give information on the association between the components and the course of action are evaluated. It plots regions between sets of cases produced during data collection, uncovering anomalies, and providing interesting perspectives on the data. Furthermore, it offers a preliminary method for perceiving variable co-tasks. This tree is created using another bootstrap test from the primary information. Around 33% of the cases are kept separate from the bootstrap test and not used in the *k*-th tree's advancement. A significant adjustment can be done by describing special cases near their gathering. This way, a special case in class *j* is a case with regions to unique class *j* cases is nearly nothing. Portray the ordinary proximity for the remaining planning class is generated as shown in Equation(4):

$$\bar{P}(n) = \sum_{d(k)=j} prox^2(n,k) (4)$$

The crude exception proportion for case *n* is characterized as shown in Equation (5):

$$\frac{nsample}{p(n)} (5)$$

The regular proximity is so close to nothing that the center of the unrefined measures has been discovered within each class, without divergence from the center. Remove the center from all rough measurements and hole by the greatest divergence to arrive at the final abnormality proportion.

## 3.6 Interactions

If the data is divided on a single variable, state *m*, in a tree, factors *m* and *k* impart whether the data is divided on k either purposefully or gradually possible. The information gathered depends on the *g*(*m*) for each tree in the forest. The situated formation for every tree and the all-out qualification of their positions are discovered in the center estimation in the general trees. This process is also used to hypothesize that the two elements are liberated from the last eliminated from the past. The results of this test procedure should be scrutinized. Only two or three instructive lists have been attempted. If one component is linked to another, dividing one by the other reduces the likelihood of a neighboring value affecting another value. The distance between any two factors within a split is compared with their

hypothetical difference if the values are free. The last is subtracted from the first, implying that a revolting collaboration has occurred. The result comprises a code list, which shows the amount of the genes relating to id within 1 to 10. The associations are rounded to the nearest whole number, and the matrix is generated using a two-column list that specifies which gene number will be number 1 in the table, and so on. Fig. 4 presents the result given by the random forest algorithm.

The supportive devices are established in arbitrary random forests. The regions at first shaped an $N\times N$ position. After a tree is created, aggregate the information, both getting ready and OOB rate of error is estimated. If cases $k$ and $n$ have a similar incurable center point, increase their distance by one. Close to the end, institutionalize the zones by isolating them according to the number of trees they contain. Clients discovered that they could not fit a $N\times N$arrangement into their short-term memory even with extensive enlightening records. A change minimized the vital memory size to$N\times N$, where $T$ is the number of trees in the timberland. To speed up the calculation of the increased value replacement, the customer is given the option of only keeping the largest areas in each case. When a test group is available, the areas of each case in the test group and each case in the planning set can be enrolled in the same way. Extra enlistment is a rational percentage.

| **1** | **0** | **13** | **2** | **4** | **8** | **-7** | **3** | **-1** | **-7** | **-2** |
|---|---|---|---|---|---|---|---|---|---|---|
| **2** | 13 | 0 | 11 | 14 | 11 | 6 | 3 | -1 | 6 | 1 |
| **3** | 2 | 11 | 0 | 6 | 7 | -4 | 3 | 1 | 1 | -2 |
| **4** | 4 | 14 | 6 | 0 | 11 | -2 | 1 | -2 | 2 | -4 |
| **5** | 8 | 11 | 7 | 11 | 0 | -1 | 3 | 1 | -8 | 1 |
| **6** | -7 | 6 | -4 | -2 | -1 | 0 | 7 | 6 | -6 | -1 |
| **7** | 3 | 3 | 3 | 1 | 3 | 7 | 0 | 24 | -1 | -1 |
| **8** | -1 | -1 | 1 | -2 | 1 | 6 | 24 | 0 | -2 | -3 |
| **9** | -7 | 6 | 1 | 2 | -8 | -6 | -1 | -2 | 0 | -5 |
| **10** | -2 | 1 | -2 | -4 | 1 | -1 | -1 | -3 | -5 | 0 |

**Fig. 4.** The Random forest algorithm gives the result.

## 3.7 Geospatial Data Analytics

Geospatial information refers to data that has a geographical perspective. This type of data has features or a location associated with it, indicating its position in space. They are created with this type of information, and all of us are aware of it on some level. Geospatial information has two fundamental structures, namely, the vector-based and raster-based structures. Seeing information using a visual guide makes it easier to understand how circumstances are evolving and how to respond to them. Moreover, seeing how spatial conditions change over time helps an organization plan for a change and decide on future activities. Also, seeing area-based information helps associations understand why a few areas and nations, for example, the United States, are more fruitful for business than others [45, 46]. The spatial estimation scale is a dynamic issue in the spatial examination. More detail is accessible at the modifiable areal unit problem (MAUP) point passage. Scene biologists induce the progression of scale-invariant measurements for parts of the environment that are fractal. No scale-free technique for investigation is generally settled upon for spatial insights [47].

Spatial sampling includes deciding a predetermined number of geographical space areas for estimating marvels that are dependent on reliance and heterogeneity. Dependency means that since one region can predict another's estimation, we do not need to worry about perceptions in both places. However, heterogeneity implies that this relationship shifts over time, and as a result, we cannot put our trust in a monitored level of dependence outside a small region. Essential spatial testing plans should be incorporated arbitrarily and orderly [48]. These essential plans can be applied at numerous levels in an assigned spatial progressive system (e.g., urban zone, city, and neighborhood). It is also possible to

misuse subordinate information, such as property estimations, as a guide in a spatial testing plan to quantify instructive accomplishment and salary. Spatial models, such as autocorrelation insights, relapse, and insertion, can likewise direct example plans.

The spatial interaction of gravity models gauges individuals' progression, material, or data between areas in a geographical space [49]. Propulsive factors, such as the number of suburbanites in communities; goal appeal factors, such as the measure of office space in business zones; and proximity connections between the areas estimated in wording, such as driving separation or travel time, are examples of components [50]. Furthermore, topological connections between regions must be distinguished, particularly in light of the often-conflicting relationship between separation and topology; for example, two spatially close neighborhoods may not demonstrate any notable cooperation if separated by an interstate. After indicating the valuable types of these connections, the investigator can assess model parameters using watched stream information and standard estimation strategies, such as the conventional least squares method or most extreme probability [51]. Contending goal adaptations of spatial association models incorporate the nearness among the goals (or beginnings) notwithstanding the source goal vicinity. This catches the impacts of goal (starting point) bunching on streams. Computational strategies, such as fake neural systems, can likewise assess spatial association connections among areas and deal with loud and subjective information and it is presented in Fig. 5. Further geospatial patients healthcare data available on cloud and representing variety [52–54] requires secure applications to keep intact the data.

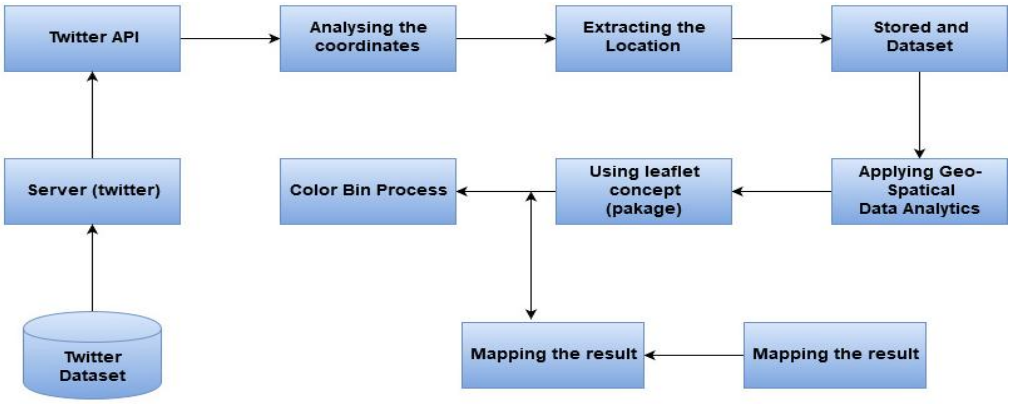



**Fig. 5.** The architecture of geospatial data analytics.

# 4. Experimental Analysis

The process of this work was done by collecting Twitter tweets for a specific (#hash tag) (flu) from the Twitter API using the Twitter development account. We choose R-programming for data analytics and pre-processing of text data. In R, we use data mining, text mining, and natural language processing (NLP) libraries to pre-process and calculate the TF-IDF (term frequency-inverse document frequency) and corpus for getting the correct featured text in the document. Besides, we use the unsupervised machine learning algorithm to train and test our text data in the dataset. The procedure is demonstrated as below:

1. First, set seeds to our text data.
2. Train our text data based on our seeds.
3. Trained our text data using machine learning.
4. Test our trained data.
5. Apply the random forest algorithm for classification.

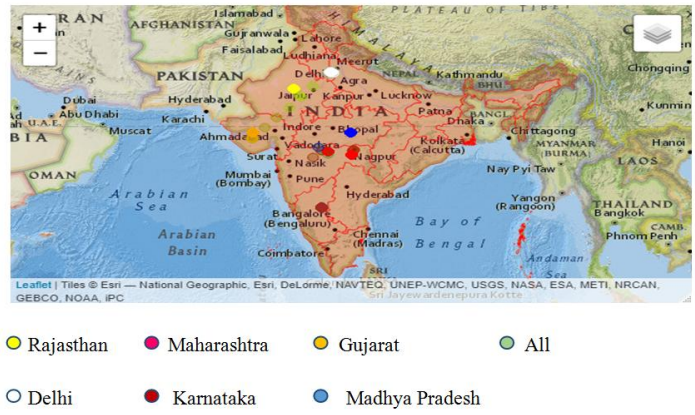



**Fig. 6.** Geospatial data prediction using leaflet.

Fig. 6 presents the geospatial data prediction using a leaflet. After applying the algorithm, we collected few tweets with geo-coordinate directions of Asia, specifically India. With the longitude and latitude coordinates that we gathered from various tweets, we created this map. In the R software, there is a library known as a leaflet, especially for geospatial data analytics. In this study, we use the methodology of vector-based geospatial data analytics. This vector-based analytics comprises of data files to points, lines, or polygon data and shapefiles. The shapefile design stores the data as a crude geometric structure like focuses, lines, and polygons. These structures, combined with data attributes associated with each shape, depict the geographical data. Fig. 7 presents the choropleth mapping using a leaflet.

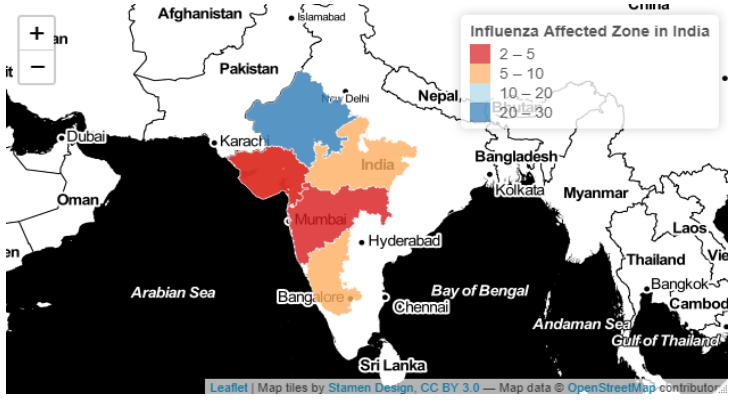
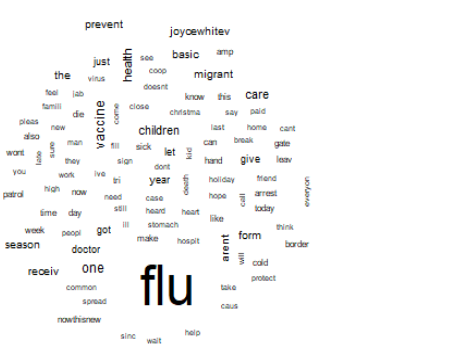



**Fig. 7.** Choropleth mapping usinga leaflet.



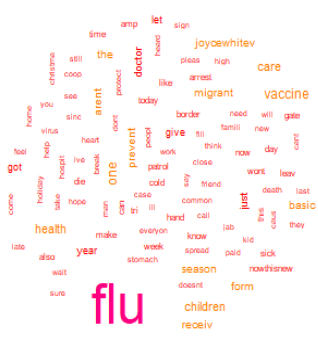



**Fig. 8.** Examples of word cloud: corporate office area (left) and urban area (right).

Cell automata demonstrate a fixed spatial system, such as matrix cells, and choose how to direct the condition of a phone based on its neighbors' needs. As time passes, spatial examples emerge as the state of cells change according to their surroundings, thus altering the conditions for future timeframes. For instance, cells can deliver to areas in an urban zone, and their states can be various kinds of land use. Examples that can arise out of the straightforward cooperation of neighborhood land utilize incorporate office areas and urban spread, which is demonstrated in Fig. 8. The accuracy determines how capable of recognizing anomalies in the methods described in this paper (Fig. 9). Recall detects all anomalies that must be addressed before the technique can be implemented, and the curve depicts the tradeoff between recall and accuracy using the threshold value parameter.

Fig. 10 demonstrates the accuracy performance for the proposed technique compared with the exiting technique and it shows that the proposed technique has the increased accuracy. The accuracy of the existing was 78% of the prediction. We found an optimized solution for thiscurrent problem with a different approach and improved results as 82% of the accuracy prediction model.

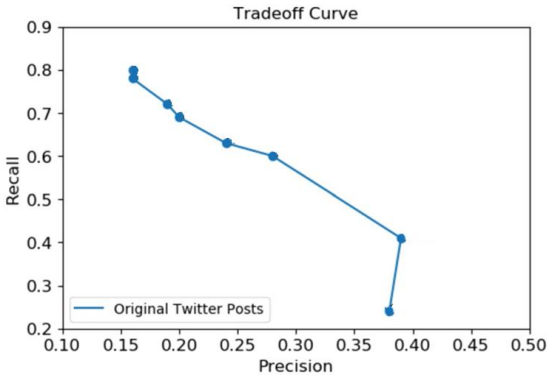



**Fig. 9.** Precision versusrecall.

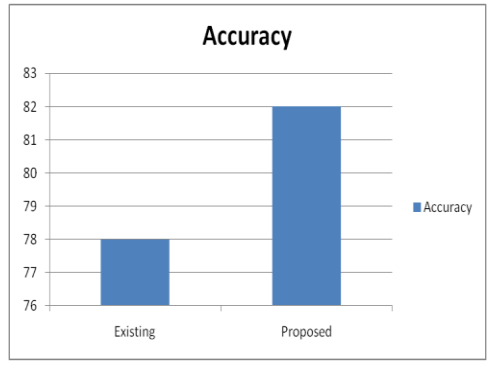



**Fig. 10.** Accuracy.

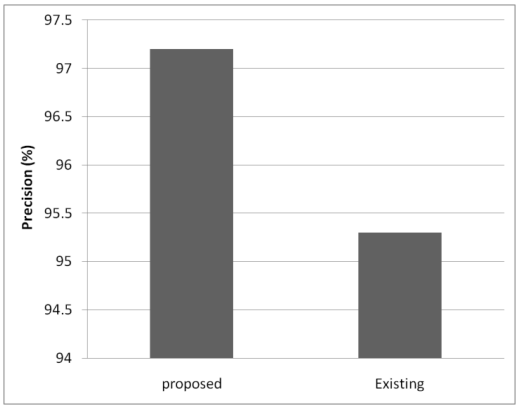



**Fig. 11.** Result of precision.

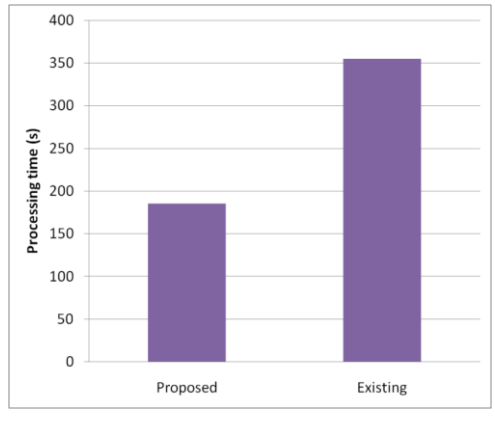



**Fig. 12.** Processing time.

Fig. 11 demonstrates the precision comparison for the proposed technique with the existing technology that ensures the proposed technique's efficiency, which is highly precise than the existing technique. Fig. 12 illustrates the processing time comparison of the proposed technique with the

existing technique and the result shows that the proposed technique has minimized the amount of processing time.

# 5. Conclusion and Future Work

This methodology helps recognize different manifestations and preventive measures for influenza-like illnesses. For the most part, social media platforms are not exact and do not provide legitimate outcomes consistently. Thus, we proposed a way to isolate accommodating data from a huge amount of information. We grouped the tweets depending on their comparability. This approach can also be employed in different frameworks. Isolating valuable information from a massive amount of data, especially those with long content, is challenging. This design has its limitations, but it is thought to be sufficient and tested explicitly for short contents and he task can be linked to web scratching.

Furthermore, sack of words can be discovered in a far more thoughtful way to achieve logically precise results. Similarly, to include all countries in map plotting using spatial analytics and perceive tweets based on a broad range of geo-coordinates, the zones must be categorized as Europe, America, North America, and other components. The results obtained predicted the “flu” from the tweets with an accuracy of 98.3%.

## Acknowledgements

We acknowledge support provided by the Center for Smart Society 5.0 (CSS5) at School of Computer Science and Engineering, Taylor’s University, Malaysia and Department of CSE, Ramco Institute of Technology, Tamilnadu, India and VIT University Vellore.

## Author’s Contributions

Conceptualization, YHR, MK, AAA, AKL, NZJ. Supervision, AAA, AKL, NZJ. Writing of the original draft, SV, YHR. Writing of the review and editing, SV.Validation, YHR, MK. Formal analysis, AAA, AKL, NZJ. Data curation, YHR, MK. All authors have checked and agreed the submission.

## Funding

We are thankful to Shaqra University for their support.

## Competing Interests

The authors declare that they do not have any conflict of interests. This research does not involve any human or animal participation.

# References

[1] E. K. Kim, J. H. Seok, J. S. Oh, H. W. Lee, and K. H. Kim, “Use of Hangeul Twitter to track and predict human Influenza infection,” *PLoS One*, vol. 8, no. 7, article no. e69305, 2013. https://doi.org/10.1371/journal.pone.0069305

[2] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu, and B. Y. Liu, "Predicting flu trends using Twitter data,” in *Proceedings of 2011 IEEE Conference on Computer Communications Workshops*, Shanghai, China, 2011, pp. 702-707.

[3] K. Lee, A. Agrawal, and A. Choudhary, “Forecasting influenza levels using real-time social media streams,” in *Proceedings of 2017 IEEE International Conference on Healthcare Informatics (ICHI)*, Park City, UT, 2017, pp. 409-414.

[4] E. D’Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, “Real-time detection of traffic from Twitter

stream analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2269- 2283, 2015.

[5] J. Capdevila, J. Cerquides, and J. Torres, "Recognizing warblers: a probabilistic model for event detection in Twitter," in *Proceedings of 2016 Anomaly Detection Workshop in theInternational Conference on Machine Learning (ICML)*, New York, NY, 2016.

[6] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using Twitter," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, 2011, pp. 1568-1576.

[7] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using Twitter data: demonstration on flu and cancer," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, IL, 2013, pp. 1474-1477.

[8] F. Zhang, J. Luo, C. Li, X. Wang, and Z. Zhao, "Detecting and analyzing influenza epidemics with social media in China," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Cham, Switzerland: Springer, 2014, pp. 90-101.

[9] S. Grover and G. S. Aujla, "Prediction model for Influenza epidemic based on Twitter data," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 7, pp. 7541-7545, 2014.

[10] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the US during the Influenza A H1N1 pandemic," *PLoS One*, vol. 6, no. 5, article no. e19467, 2011. https://doi.org/10.1371/journal.pone.0019467

[11] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu, and B. Y. Liu, "Online social networks flu trend tracker: a novel sensory approach to predict flu trends," in *Biomedical Engineering Systems and Technologies*. Berlin, Germany: Springer, 2012, pp. 353-368.

[12] B. Alkouz and Z. Al Aghbari, "Analysis and prediction of influenza in the UAE based on Arabic tweets," in Proceedings of *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*, Shanghai, China, 2018, pp. 61-66.

[13] N. Rokbani, R. Kumar, A. Abraham, A. M. Alimi, H. V. Long, and I. Priyadarshini, "Bi-heuristic ant colony optimization-based approaches for traveling salesman problem," *Soft Computing*, vol. 25, pp. 3775-3794, 2020. https://doi.org/10.1007/s00500-020-05406-5

[14] S. Jha, D. Prashar, H. V. Long, and D. Taniar, "Recurrent neural network for detecting malware," *Computers & Security*, vol. 99, article no. 102037, 2020. https://doi.org/10.1016/j.cose.2020.102037

[15] M. Musleh, "Spatio-temporal visual analysis for event-specific tweets," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, Snowbird, UT, 2014, pp. 1611-1612.

[16] S. P. Brennan, A. Sadilek, and H. Kautz, "Towards understanding global spread of disease from everyday interpersonal interactions," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, Beijing, China, 2013, pp. 2783-2789.

[17] A. Guille and C. Favre, "Event detection, tracking, and visualization in Twitter: a mention-anomaly-based approach," *Social Network Analysis and Mining*, vol. 5, article no. 18, 2015. https://doi.org/10.1007/s13278-015-0258-0

[18] M. A. Al-Garadi, M. S. Khan, K. D. Varathan, G. Mujtaba, and A. M. Al-Kabsi, "Using online social networks to track a pandemic: a systematic review," *Journal of Biomedical Informatics*, vol. 62, pp. 1-11, 2016.

[19] A. A. Aslam, M. H. Tsou, B. H. Spitzberg, L. An, J. M. Gawron, D. K. Gupta, K. M. Peddecord, A. C. Nagel, C. Allen, J. A. Yang, et al., "The reliability of tweets as a supplementary method of seasonal Influenza surveillance," J. Med. Internet Res., vol. 16, no. 11, p. e250, 2014. https://doi.org/10.2196/jmir.3532

[20] H. Achrekar, A. Gandhe, R. Lazarus, S. H. Yu, and B. Y. Liu, "Twitter improves seasonal influenza prediction," in *Proceedings of the International Conference on Health Informatics*, Vilamoura, Portugal, 2012, pp. 61-70.

[21] H. F. Huo and X. M. Zhang, "Modeling the influence of Twitter in reducing and increasing the spread of influenza epidemics," *SpringerPlus*, vol. 5, article no. 88, 2016. https://doi.org/10.1186/s40064-016-1689-4

[22] I. Kagashe, Z. Yan, and I. Suheryani, "Enhancing seasonal Influenza surveillance: topic analysis of widely used medicinal drugs using Twitter data," *Journal of Medical Internet Research*, vol. 19, no. 9, article no. e315, 2017. https://doi.org/10.2196/jmir.7393

[23] V. K. Jain and S. Kumar, "An effective approach to track levels of influenza-A (H1N1) pandemic in India using twitter," *Procedia Computer Science*, vol. 70, pp. 801-807, 2015.

[24] N. Ahmed, S. C. Quinn, G. R. Hancock, V. S. Freimuth, and A. Jamison, "Social media use and influenza vaccine uptake among White and African American adults," *Vaccine*, vol. 36, no. 49, pp. 7556-7561, 2018.

[25] M. J. Cumbraos-Sanchez, R. Hermoso, D. Iniguez, J. R. Pano-Pardo, M. A. A. Bandres, and M. P. L. Martinez, "Qualitative and quantitative evaluation of the use of Twitter as a tool of antimicrobial stewardship," *International Journal of Medical Informatics*, vol. 131, article no. 103955, 2019. https://doi.org/10.1016/j.ijmedinf.2019.103955

[26] F. Wang, H. Wang, K. Xu, R. Raymond, J. Chon, S. Fuller, and A. Debruyn, "Regional level influenza study with geo-tagged Twitter data," *Journal of Medical Systems*, vol. 40, article no. 189, 2016. https://doi.org/10.1007/s10916-016-0545-y

[27] P. Grover, A. K. Kar, and G. Davies, "'Technology enabled Health': insights from twitter analytics with a socio-technical perspective," *International Journal of Information Management*, vol. 43, pp. 85-97, 2018.

[28] I. Hellsten, S. Jacobs, and A. Wonneberger, "Active and passive stakeholders in issue arenas: a communication network approach to the bird flu debate on Twitter," *Public Relations Review*, vol. 45, no. 1, pp. 35-48, 2019.

[29] K. Gunaratne, E. A. Coomes, and H. Haghbayan, "Temporal trends in anti-vaccine discourse on Twitter," *Vaccine*, vol. 37, no. 35, pp. 4867-4871, 2019.

[30] A. Khatua, A. Khatua, and E. Cambria, "A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks," *Information Processing & Management*, vol. 56, no. 1, pp. 247-257, 2019.

[31] N. Levy, M. Iv, and E. Yom-Tov, "Modeling influenza-like illnesses through composite compartmental models," *Physica A: Statistical Mechanics and its Applications*, vol. 494, pp. 288-293, 2018.

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

[33] S. G. K. Patro, B. K. Mishra, S. K. Panda, R. Kumar, and H. V. Long, "Knowledge-based preference learning model for recommender system using adaptive neuro-fuzzy inference system," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 3, pp. 4651-4665, 2020.

[34] S. G. K. Patro, B. K. Mishra, S. K. Panda, R. Kumar, H. V. Long, D. Taniar, and I. Priyadarshini, "A hybrid action-related k-nearest neighbour (HAR-KNN) approach for recommendation systems," *IEEE Access*, vol. 8, pp. 90978-90991, 2020.

[35] V. Puri, S. Jha, R. Kumar, I. Priyadarshini, M. Abdel-Basset, M. Elhoseny, and H. V. Long, "A hybrid artificial intelligence and Internet of Things model for generation of renewable resource of energy," *IEEE Access*, vol. 7, pp. 111181-111191, 2019.

[36] H. V. Long, L. H. Son, M. Khari, K. Arora, S. Chopra, R. Kumar, T. Le, and S. W. Baik, "A new approach for construction of geodemographic segmentation model and prediction analysis," *Computational Intelligence and Neuroscience*, vol. 2019, article no. 9252837, 2019. https://doi.org/10.1155/2019/9252837

[37] S. Jha, R. Kumar, M. Abdel-Basset, I. Priyadarshini, R. Sharma, and H. V. Long, "Deep learning approach for software maintainability metrics prediction," *IEEE Access*, vol. 7, pp. 61840-61855, 2019.

[38] Z. Li, J. Wu, S. Mumtaz, A. M. Taha, S. Al-Rubaye, and A. Tsourdos, "Machine learning and multi-dimension features based adaptive intrusion detection in ICN," in *Proceedings of 2020 IEEE International Conference on Communications (ICC)*, Dublin, Ireland, 2020, pp. 1-5.

[39] H. Yang, J. Wen, X. Wu, L. He, and S. Mumtaz, "An efficient edge artificial intelligence multipedestrian tracking method with rank constraint," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4178-4188, 2019.

[40] L. Zhang, J. Wu, S. Mumtaz, J. Li, H. Gacanin, and J. J. Rodrigues, "Edge-to-edge cooperative artificial intelligence in smart cities with on-demand learning offloading," in *Proceedings of 2019 IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, 2019, pp. 1-6.

[41] K. M. S. Huq, S. Mumtaz, J. Rodriguez, P. Marques, B. Okyere, and V. Frascolla, "Enhanced c-ran using D2D network," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 100-107, 2017.

[42] H. Liang, J. Wu, S. Mumtaz, J. Li, X. Lin, and M. Wen, "MBID: micro-blockchain-based geographical dynamic intrusion detection for V2X," *IEEE Communications Magazine*, vol. 57, no. 10, pp. 77-83, 2019.

[43] Z. Zhou, H. Yu, C. Xu, Z. Chang, S. Mumtaz, and J. Rodriguez, "BEGIN: big data enabled energy-efficient vehicular edge computing," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 82-89, 2018.

[44] Y. Liu, X. Fang, M. Xiao and S. Mumtaz, "Decentralized beam pair selection in multi-beam millimeter-wave networks," *IEEE Transactions on Communications*, vol. 66, no. 6, pp. 2722-2737, 2018.

[45] S. Annamalai, R. Udendhran, and S. Vimal, "An intelligent grid network based on cloud computing infrastructures," in *Novel Practices and Trends in Grid and Cloud Computing*. Hershey, PA: IGI Global, 2019, pp. 59-73.

[46] S. Annamalai, R. Udendhran, and S. Vimal, "Cloud-based predictive maintenance and machine monitoring for intelligent manufacturing for automobile industry," in *Novel Practices and Trends in Grid and Cloud Computing*. Hershey, PA: IGI Global, 2019, pp. 74-89.

[47] S. Pradeepa, K. R. Manjula, S. Vimal, M. S. Khan, N. Chilamkurti, and A. K. Luhach, "DRFS: detecting risk factor of stroke disease from social media using machine learning techniques," *Neural Processing Letters*, 2020. https://doi.org/10.1007/s11063-020-10279-8

[48] M. Ramamurthy, I. Krishnamurthi, S. Vimal, and Y. H. Robinson, "Deep learning based genome analysis and NGS-RNA LL identification with a novel hybrid model," *Biosystems*, vol. 197, article no. 104211, 2020.https://doi.org/10.1016/j.biosystems.2020.104211

[49] P. Sampath, G. Packiriswamy, N. Pradeep Kumar, V. Shanmuganathan, O. Y. Song, U. Tariq, and R. Nawaz, "IoT Based health-related topic recognition from emerging online health community (med help) using machine learning technique," *Electronics*, vol. 9, no. 9, article no. 1469. 2020. https://doi.org/10.3390/electronics9091469

[50] S. Vimal and P. Subbulakshmi, "Secure data packet transmission in MANET using enhanced identity-based cryptography," *International Journal of New Technologies in Science and Engineering*, vol. 3, no. 12, pp. 35-42, 2016.

[51] G. Thomas, A. Sampaul, Y. H. Robinson, E. G. Julie, V. Shanmuganathan, S. Rho, and Y. Nam, "Intelligent prediction approach for diabetic retinopathy using deep learning based convolutional neural networks algorithm by means of retina photographs," *CMC-Computers Materials & Continua*, vol. 66, no. 2, pp. 1613-1629, 2021.

[52] A. P. Singh, N. R. Pradhan, S. Agnihotri, N. Jhanjhi, S. Verma, U. Ghosh, and D. Roy, "A novel patient-centric architectural framework for blockchain-enabled healthcare applications," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5779-5789, 2021.

[53] A. Ullah, M. Azeem, H. Ashraf, A. A. Alaboudi, M. Humayun, and N. Z. Jhanjhi, "Secure healthcare data aggregation and transmission in IoT: a survey," *IEEE Access*, vol. 9, pp. 16849-16865, 2021.

[54] S. Ali, Y. Hafeez, N. Z. Jhanjhi, M. Humayun, M. Imran, A. Nayyar, S. Singh, and I. H. Ra "Towards pattern-based change verification framework for cloud-enabled healthcare component-based," *IEEE Access*, vol. 8, pp. 148007-148020, 2020.