The Journal of Supercomputing (2021) 77:11738–11755
https://doi.org/10.1007/s11227-021-03726-3

Epidemic zone of COVID-19 from social media using hypergraph with weighting factor (HWF)

S. Pradeepa[1] · **K. R. Manjula**[1]

Accepted: 5 March 2021 / Published online: 29 March 2021

Abstract

Online social network is one of the most prominent media that holds information about society's epidemic problem. Due to privacy reasons, most of the users will not disclose their location. Detecting the location of the tweet users is required to track the geographic location of the spreading diseases. This work aims to detect the spreading location of the COVID-19 disease from the Twitter users and content discussed in the tweet. COVID-19 is a disease caused by the "novel coronavirus." About 80% of confirmed cases recover from the disease. However, one out of every six people who get COVID-19 can become seriously ill, stated by the World health organization. Inferring the user location for identifying the spreading location for the disease is a very challenging task. This paper proposes a new technique based on a hypergraph model to detect the Twitter user's locations based on the spreading disease. This model uses hypergraph with weighting factor technique to infer the spreading disease's spatial location. The accuracy of prediction can be improved when a massive volume of streaming data is analyzed. The Helly property of the hypergraph was applied to discard less potential words from the text analysis, which claims this work of unique nature. A weighting factor was introduced to calculate the score of each location for a particular user. The location of each user is predicted based on the one that possesses the highest weighting factor. The proposed framework has been evaluated and tested for various measures like precision, recall and F-measure. The promising results obtained have substantiated the claim for this work compared to the state-of-the-art methodologies.

Keywords Twitter data · Natural Language Processing · COVID-19 · Hypergraph · Helly property · Weighting factor

✉ K. R. Manjula
manjula@cse.sastra.edu

[1] School of Computing, Sastra Deemed University, Thanjavur 613401, India

1 Introduction

Online social networks, such as Facebook, Twitter, Four-square, LinkedIn, PInterest and Google Plus, employed as a monitoring tool to explore essential information about Event Detection [1], Election Prediction [2], Epidemic Forecasting [3, 4], human personality analysis and trend prediction. Twitter is a modern public platform where many users debate, discuss and share their views. It has become an arena to showcase world news and also proves to be one of the most quantitative and qualitative human information databases. Twitter users fill up the profile with their personal information. These data are not always available. 75% of the Twitter users did not disclose their location in his/her profile. The user's location is an essential piece of the attribute that could be exploited for various surveillance kinds.

Many researchers have focused on machine learning techniques to infer the user's location. Primary two ways to detect the Twitter user's location include (1) GPS-enabled geotagging and check-in, and (2) user-generated content. The first way is not always reliable and must be continuously updated [1, 5]. The second way is detecting the location based on their publicly available data [6]. The prediction performance of these models depends on the availability of local words in post contents. Figure 1 portrays the location prediction of the particular Twitter user using two different types of ways.

Detecting the particular Twitter user's location may help identify the epidemic zone for any type of disease. Here, Twitter API generates the streaming data based on the keywords given by the users. Twitter has been measured to be an essential resource for receiving health-related information discussed in society. This information has been shared by both official sources and citizens. Content discussed in the tweets related to diseases is beneficial for public health research [7].

Health-related topics are recognized from emerging online health communities (MedHelp) using unsupervised machine learning techniques [8]. Different

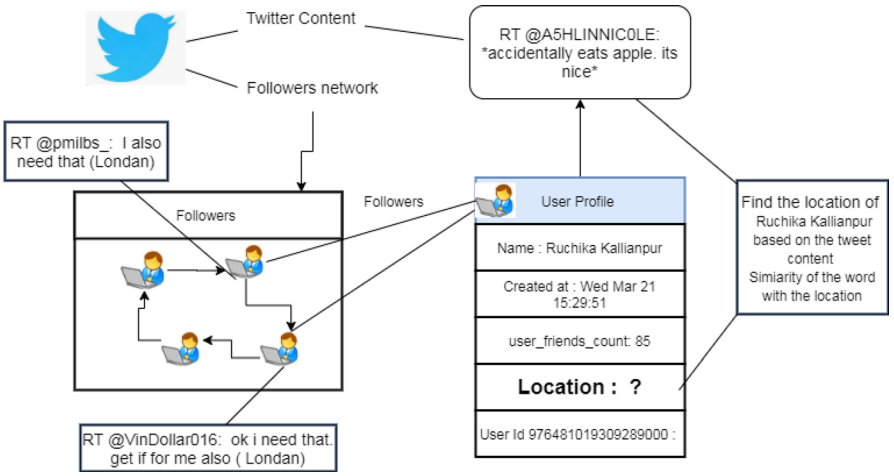


Fig. 1 Two different ways to predict the Twitter user's location

community centre dispatches their content through some online forum. Analysis of these types of content will help society to detect the new symptoms, prevention and causes for any type of disease.

Followers of Twitter users and self-reported contents play a significant role in the hybrid prediction model [9, 10]. If the particular user's followers declare some places in their tweets, the hybrid prediction model can use such data to recognize the user's location. Graph structure of the networks [11] is constructed from the follower's information. This technique explicit clustering features while inferring users' locations. Analyzing the social network graph is challenging when we worked with a high volume of data. Some researchers employed a graph-based machine learning technique to train a model between the users' known and unknown geolocation [10, 12].

Markov chain Monte Carlo simulation technique [9] is proposed, considering the Markov random field probability model to represent the network. This social network approximates the posterior probability distribution of the missing label. It combines the graph and content analyses for detecting the geolocation of the users. Pang et al. [5] introduced deep learning-based feature learning architecture, and it can detect the user's demographics location category.

The challengers are

- Due to the processing of unstructured data, it is difficult to identify the common framework for detecting Twitter users' location. It is serious to understand the performance of the existing models.
- Previous researchers have evaluated their methods using less volume of data and limited number metrics with their own data. It is essential to classify suitable metrics that can predict the user's location. Defining such a suite of metrics is a challenging task.

To respond to these challenges, this paper proposes the framework for detecting twitter's location by incorporating the following contributions.

- Other graphing technique used in the previous studies is produced single association between two entities. However, the hypergraph produces the multiway association between the nodes.
- Multiway association between the words and location can be established using hypergraph.
- Hypergraph measures the multiway association between the COVID-19 with the proper location.
- Applying the Helly property in the hypergraph reduces the noisy data or irrelevant data in the context.
- Weighting factor measure in the hypergraph produces the classifier result with good accuracy.
- We tested our methodology with ten trials. Finally, it presents new experiments with high accuracy.

The rest of the paper is organized as follows. Section 2 summarizes the current state of the art to learn the geographical location of social network agents. Section 3 shows the proposed framework (HWF) based on hypergraph with Helly property and weighting factor (IF), and Sect. 4 portrays the experiments, along with data collection and data analysis, followed by the experimental results and discussion. Section 5 concludes and shows future directions.

2 Related work

The microblogging service in Twitter has been used to gather information on events occurring in real time [13–15]. For example, when COVID-19 is spreading in society, people make many posts related to the novel coronavirus, enabling the detection of geographical locations for spreading disease.

An epidemic zone for any pandemic spreading disease can be detected based on the Twitter user's location. Most Twitter users do not disclose their location because of privacy reasons. Inferring the Twitter user location may help the government detect the epidemic zone for the spreading disease.

This section reviews current methodologies when predicting the location of the Twitter user [16–18]. Comparison of different algorithms is difficult because the evaluations are not performed on the same data set and standard configurations.

Name entity recognition (NER) [12] technique is used to extract the features from the text document, news and articles. It has attracted many researchers for text processing. Twitter's locations are geographically defined places like countries, rivers, mountains, highways, streets, factories, etc. NER is a part of location recognition in social media data. The results were evaluated with some components physically identified as to location by experts.

Agarwal et al. introduced an approach that combines the concept-based vocabulary and Stanford NER tool to identify the location information from tweets. To remove the noisy component from extracted location phrases, they used a naive Bayes classifier with the following features: the word's POS tags, three words after this word and three words before this word [19, 20].

Hence, instead of using social content, some prediction techniques rely on a social network's graph structure. These techniques exploit the network features while inferring user's locations using their social connections. Comparative analysis of nine network-based techniques is discussed in [21]. They used a bidirectional Twitter network dataset for predicting "tweet location" from a user's post. It shows the system performance when varying the gazetteer used to identify self-reported locations.

Some authors tried different machine learning technique algorithms validation [22, 23]: Naïve Bayes, support vector machine, decision tree and random forest using cross-validation. They experiment with different machine algorithms with an accuracy measure.

Markov chain Monte Carlo (MCMC) [9] algorithm and posterior probability estimator method are used to detecting the Twitter user's location. They compared the integrated data approach with Naive Bayes. MRW (multirank walk) and Predict the

country of origin of Twitter users with good accuracy. Home location prediction with similarities is posed by Potts model [16].

Form a network based on the followers and establish the relation between user living probability in a city. Using the probability distribution of the friendship between the friends, they infer the user's location [24, 25]. Prediction methodology assumes that users in the same social circle were likely to visit the same area [21]. New prevention, causes and fear of the PCOS problem have to detected from social media content analysis [26, 27].

Geographical proximity and structural proximity metric are used to outline communities in a user's ego-net [28]. Each metric on social media data can be evaluated and project the performance measures. In graph-based deep learning architecture based on user-generated content [10] Weighted most frequently visited (WMFV) cluster of the user.

In supervised keyword extraction methodology, the keyword extraction task is treated as a twofold order issue. A classifier decides if each word or phrase in the document is a keyword. Most of the research community uses different classifier algorithms, such as Naive Bayes classifier, support vector machine, maximum entropy modeling, decision tree and hidden Markov mode. The main problem of keyword extraction methodology is the requirement for a labeled corpus. The nature of the preparation corpus straightforwardly influences the model's presentation, consequently influencing the after effects of keyword extraction. Also, since there are not many marked corpuses accessible, the preparation set regularly should be labeled by clients themselves. Manual labeling of top-notch keywords from text prompts a lot of contrast in the experimental data, which are likewise costly, tedious and mistake inclined. Hence, how to get a top-notch preparing set is the bottleneck of these methodologies.

Here we developed a supervised keyword extraction tool using hypergraph-based Weighting factor with better 89% accuracy, which is better compared to the other existing supervised keyword extraction algorithms.

3 The methodology of the proposed HWF model

Online social networks (Twitter, Facebook, Flickr) are an essential source of information to supervise real-time events, such as epidemic surveillance, location-based election prediction, real-time events. But 30% of the substantial number of Twitter users choose not to disclose their geographical location. Detection of the user's location is an important task to produce a survey report related to any events. Here, we are focusing on the epidemic zone prediction of the COVID-19 using Twitter users' location and content discussed in the tweet. Figure 2 shows the overall architecture for the HWF used to infer the epidemic zone for the COVID-19.

The key contribution of the proposed framework is

1. Collect the tweet streaming data using Twitter API and preprocess the text data using NLTK.

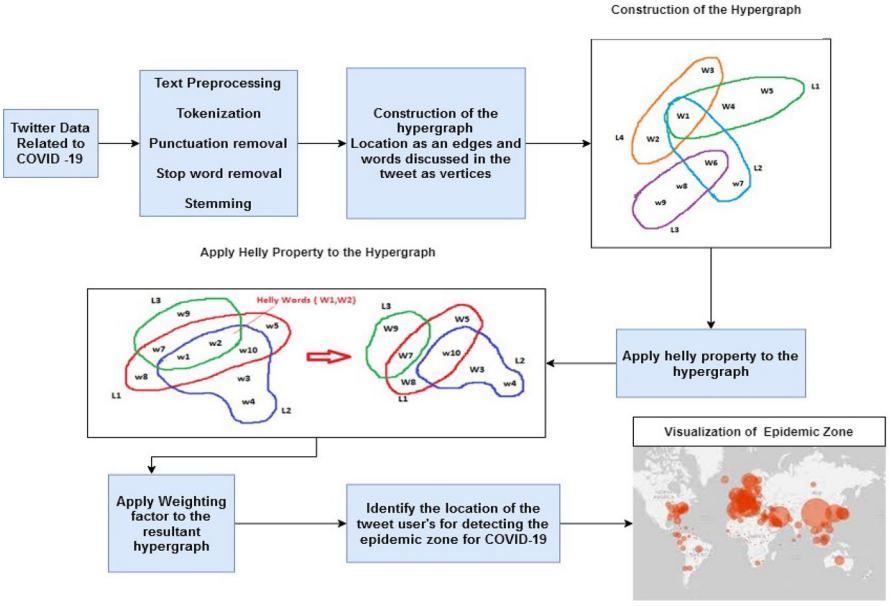


Fig. 2 HWF architecture

2. Construction of the hypergraph. Location as edge and words discussed in a tweet as vertex
3. Apply the Helly property to each word for excluding the repeated word in a different domain.
4. Apply the weighting factor for each word and classify the tweets by location.
5. Detect the spreading rate and geographical zone for the epidemic zone for the COVID-19.

3.1 Data gathering and preprocessing

Formulating effectual datasets that are utilized for exploratory studies and understanding the hidden patterns is necessary to preprocess the dataset. It is essential to do complete data analyses on tweet attributes. The prime idea is that spoken words from Twitter are a random combination of words related to COVID-19. This section is intended to discuss the approach and methodology for data collection and preprocessing. The data are accumulated from Twitter using keywords necessary for data collection and help recognize relevant tweets. To extract data utilizing keywords and medicinal terms, Twitter API [11] has been used. The important variables, also known as keywords, are collected through a methodology based on Jain and Kumar [5] that supplies vital keywords associated with public feelings and general trends and are widely trending on a particular day. Some of the keywords taken into consideration are: Keywords: {"COVID-19"," CORONA VIRUS"}.

Test preprocessing the tweets is an important task as these documents are filled with dialect, other language words and misspellings. Preprocessing of text is done to address the complications associated with the noise present in the document. Some of the key methods such as tokenization, stop words removal, stemming, lemmatization, feature weighting, dimensionality reduction and frequency-based methods [12] are applied. Due to privacy reasons, most Twitter users do not disclose their location. Finding the location of the Twitter user can help society to find the epidemic zone related to Coronavirus. For an effectual observation, it is imperative to identify the appropriate tweets that investigate coronavirus disease using surveillance and detection. Algorithm 1 given below is reposible for the data collection and the preprocessing steps are also explained.

Algorithm 1: Data collection and preprocessing

Input: Data collected from Twitter API related to the COVID-19
Output: List of words in a file after preprocessing the tweets
1. While all words in the document in a file are exhausted
    1.1. Tokenize the document.
    1.2. Remove the stopwords.
    1.3. Remove the punctuation.
2. Removing Tweets with less than three words
3. Removing duplicate tweets.

3.1.1 Removing stop words

Stop words removal is not a fast rule in natural language processing. Stopwords are excluded from the Twitter dataset, which concentrates the sentence with those words that are informative. This process will decrease the size of the dataset and improve performance evaluation. Google search engine removes the stopwords for fast and relevant text extraction. Example stop words are {'ourselves,' 'hers,' 'between,' 'yourself'}. Fifty stop words are removed from our document as they are irrelevant for our purpose.

3.1.2 Removing Tweets with less than three words

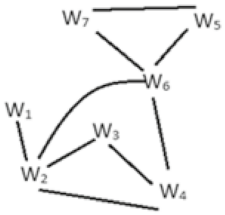
In our dataset, 20% of the tweet has less than three words. Those tweets are removed from the data set for improving the accuracy of an algorithm.

3.1.3 Removing duplicate tweet

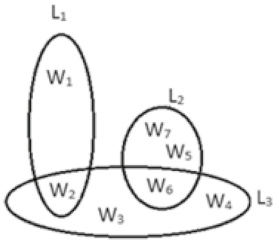
Here, if the tweets came from the same tweet id more than once, those tweets can be removed from the data set. Around 27% of the tweets are duplicated in our data set.

	L_1	L_2	L_3
W_1	1	0	0
W_2	1	0	1
W_3	0	0	1
W_4	0	0	1
W_5	0	1	0
W_6	0	1	0
W_7	0	1	1

(a)



(b)



(c)

Fig. 3 Different ways to represent the relationship between the location with the words discussed in the tweet

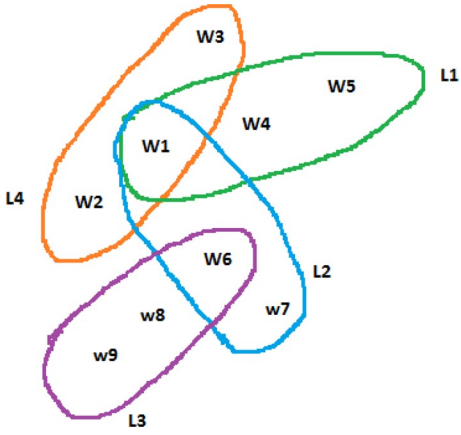


Fig. 4 Hypergraph representation

3.2 Construction of the hypergraph

In Fig. 3, an location set $V=\{ v_1,v_2,v_3,v_4,v_5,v_6,v_7\}$ and the word set $E=\{e_1,e_2,e_3\}$ is represented in three different ways. The element (V_i, e_j) is set to 1 if e_j is a word of location v_i, and 0 otherwise (a). It is a simple undirected graph in which two locations are joined together by an edge if there is at least one word in common. This graph cannot tell us whether the same word is the word of three or more locations or not (b). A hypergraph that could completely illustrate the high-order relationships among location and words (c).

Definition (hypergraph) A hypergraph is a graph in which an edge can connect more than two vertices as defined in [29–31]. A hypergraph H can be defined, $H = \{V, E\}$ where $V = \{v_1, v_2, v_3, \ldots, v_n\}$ is the set of vertices and $E = \{e_1, e_2, e_3, \ldots, e_m\}$ is the set of edges.

Figure 4 demonstrates the sample hypergraph structure for location prediction for the Twitter user. Here we represented the hypergraph by location in the tweets as

edges and words discussed in the location as vertices. Here, $V = \{L_1, L_2, L_3, L_4,\}$and $E = \{w_1, w_2, w_3, w_4, , w_5, w_6, w_7, w_8, , w_9,\} = \{\{w_1, w_2, w_3\}, \{w_1, w_4, w_5\}, \{w_6, w_8, w_9\}, \{w_1, w_6, w_7\}\}$. The uniformity of a hypergraph is judged by the number of vertices that each of its hyper-edges connects. We consider the words discussed in tweets to be the vertices $V = \{w_1, w_2, w_3, \ldots, w_n\}$ and the locations to be the edges $E = \{l_1, l_2, l_3 \ldots l_n\} = \{\{w_1, w_2, w_3\}, \{w_1, w_2\}, \{w_4, w_5\}\}$. Here, w indicates the words discussed in Twitter based on the location. The edge "l" indicates the locations available in the Twitter data set related to the COVID-19. Construction of the hypergraph is represented in Algorithm 2. The algorithm receives the tweet location with content. If the new location receives, then it will be considered as a new edge. If the location already exists in the hypergraph, it collects the words and adds them to the corresponding edges. This procedure is executed until the construction of the complete hypergraph from the tweet dataset. The 'p' variable is used to identify the number of edges in the hypergraph. If one word is available in all the locations, then the word's degree is equal to 'p' variable. Finally, we considered that word is a Helly word. These types of Helly words were removed from the hypergraph. We mentioned the usage of the "p" variable in the revised manuscript.

Algorithm 2: Construction of the Hypergraph

```
Input: Tweet ( location, content)
Output: Hypergraph H (E, V)

for each tweet ( Li , set of words in Li)
        If (Li == Ei∈ E)
            ++ f(Li)                ## number of each location in tweet dataset
        else
            Create edge(Ei)
        f(Ei)=1
        V(Ei) = set of words in Ei
         P++                        ## distinct location in dataset
end for
```

3.3 Helly Property in Hypergraph

Let H be a hypergraph. The sets that can be written and the union of different edges of H form a new hypergraph are denoted by H′. Let us suppose that H has the Helly property, and we want to state something similar for H′. We prove a conjecture of C, Berge and two negative results [32, 33]. Commonly discussed words in all the locations might not produce the correct location of the tweet. Here, commonly discussed words are removed from the hypergraph using the Helly property. In Fig. 5, w_1 and w_2 are the Helly words. Those words are removed from the hypergraph. Helly property in the hypergraph is used to identify the most discussed words in all the locations. If the word is present in all the locations, those words are considered less potential component words for our analysis. In our

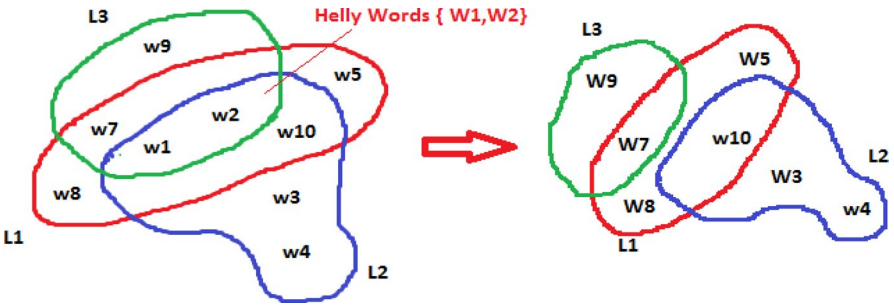


Fig. 5 Remove the Helly words from the hypergraph using the Helly property

work, 30% of the words are removed from the hypergraph based on the Helly property. The Helly property [34] identifies the most featured component of the entire hypergraph or set of vertices that are associated with every edge in the graph.

Without removal of Helly words from the hypergraph may lead to less accuracy. These types of analysis will improve the accuracy of an algorithm compare to the other existing algorithms. The demonstration of the removal of the Helly words is disrobed in algorithm 3. In this algorithm, it will traverse every edge in the hypergraph. If the vertex presents in all the edge in the hypergraph, then that vertex will be removed from the hypergraph.

Algorithm 3

Input: Hypergraph with Helly words, p (Total number of edges in the graph)
Output: Hypergraph without Helly words

for all edges e_j in H do
for each vetex v_i in e_i
$f(v_i) = f(E_j)$
if ($f(v_i) == p$)
Release the vertex v_i from all edges in H
end if
end for
end for

3.4 Weighting factor (WF)

The weighting factor for every word is calculated using the following procedure. The main objective of the work in this work is weighting factor calculation for the vertices in the hypergraph. The vertices are associated with more than one edge that has less weight. A word (vertex) uttered only in a particular edge is more important than that used by two or more edges. The weighting factor for the location prediction of the particular tweet is

$$\mathrm{WF(v)} \propto \left(N_{\mathrm{e}} - n_{\mathrm{ev}}\right) \tag{1}$$

Here, N_{e} represents the total number of edges in the hypergraph and n_{ev} denotes the number of edges (e)that uses a vertex v. A vertex used by many edges has a lower weight factor value than the one that is restricted to a few edges. If the vertex is unique in the hypergraph, then it gets the highest weight factor value.

$$\mathrm{WF(v)} = \left(N_{\mathrm{e}} - n_{\mathrm{ev}}\right) \tag{2}$$

The value of β can be chosen arbitrarily according to the dataset. It is not harmful to set its value to one. In our experiment, we set it to one. There is a possibility of a word being unique to a region, but only one user from a different region may have used it. The value of β will vary from word to word. So, we rewrite our equation of IF as

$$\mathrm{WF(v)} = \beta_{\mathrm{w}}\left(N_{\mathrm{e}} - n_{\mathrm{ev}}\right) \tag{3}$$

The value of β_{w} depends on the frequency of the words' usage in every region. Finally, we derived a generic formula for β_{w} is

$$\beta_{\mathrm{v}} = \prod\nolimits_{j=1\text{tom and}j\text{contains}w} \kappa_j * {}^{\mathrm{f}(v)}/_{\mathrm{f}(j)} \tag{4}$$

where f(j) represents the total number of vertices covered by the edge j. The fractional part represents the frequency of a particular component in a region. The Constant variable k is used to differentiate each edge. The number of words may be high in a particular region and low in another. It is a normalizing factor. It is unique for each edge or region. Algorithm 4 demonstrates the Weighting factor calculation of each vertex. This value predicts the location of the Twitter users based on the context.

Algorithm 4: Predict the location of the Twitter user using a Weighting factor

Input: Tweet content without location
Output: Tweet content with location

For all test tweets, $\mathrm{Te_i}$ do
  For all edges $\mathrm{e_j}$ in H do
    $\mathrm{Score(e_j)} = 0$
  End for
  For all components $\mathrm{c_k}$ in $\mathrm{Te_i}$do
    $\mathrm{Score(e_j)} += \mathrm{WF(c_k)} \forall \mathrm{e_j}$ with $\mathrm{c_k}$ contained in $\mathrm{e_j}$ // Calculate from formula 3.
  End For
  Find the edge $\mathrm{e_j}$ with the maximum score
  Declare $\mathrm{e_j}$ as the location of $\mathrm{Te_i}$
End For

4 Experimental result

4.1 Data preprocessing

Around 134,456 tweets on March 26, 2020, were collected from Twitter using Twitter API programs in Python. Here the tweets are collected by us based on the following keys given by the Twitter server, access_token, access_token_secret, consumer_key, consumer_secret. Sample tweets are shown in Table 1. Twitter words are tokenized using the functions, namely sent_tokenize and word_tokenize using the NLTK (Natural Language ToolKit) library in Python. The stopwords were removed by using the stopword package. We collected the data with Twitter id and removed duplicate data when the tweets are generated from the common tweet id. This type of preprocessing will improve the performance of the result. We found only 82,416 tweets after preprocessing the document.

Figure 6 shows the COVID-19 tracker from Twitter content and the location of the Twitter user. The geographical location of the spreading of COVID-19 is mapped with the report generated from the World Health Organization. Reports given by the world health organization are depicted in Fig. 7. These two figures are matched with the count. Word cloud of the Twitter content is shown in Fig. 8.

4.2 Construction of the hypergraph

In the hypergraph, the number of the vertex is 595,710, and the number of edges is 182. The number of Helly words is 178,713. 30% of the words are removed in each edge based on the Helly words. A weighting factor was calculated for the test dataset and predicted the Twitter user's location with the highest score. It observed that whenever a location beats another by a huge margin (of score), the test was successful. In samples represented in tweet dataset with weighting factor calculated, the Twitter users' locations are identified based on the highest score, represented in Table 2.

4.3 Accuracy

During the training phase, measuring accuracy plays a relevant role in model selection. Parameters were selected to maximize prediction accuracy on training samples.

Standard Accuracy measure for the classification:

The precision and recall values were calculated for each location based on the observed true positive, true negative, false positive and false negative values. Precision measures the ratio of the positive cases that were correct. Recall measures the ratio of positive cases that were correctly identified.

$$\text{Precision} = {}^{\text{TruePositive}}/_{(\text{Truepoitive} + \text{FalsePositive})}$$

Table 1 Sample tweet collected from Twitter using Twitter API

Here are Cuomo and de Blasio downplaying COVID-19 on March 26 "I want to make sure I tell the people of New York what I told my daughter â€ "in this situation, the facts defeat fear because the reality"	California USA
COVID-19 has engulfed the world with fear and anxiety. I highly recommend the daily practice of pranayama. This is an ancient practice of conscious breath control. It improves health, calms the mind, expands mental abilities and most importantly, it awakens you spiritually. https://t.co/TLNoASGzod	Chicagoillinois
At a times, when jails are being decongested and stringent PSA revoked, many Kashmiri families continue to fear for their captive kinâ€™s safety, as Covid-19 became all pervasive. @omer_farooq__ reports. https://t.co/BhvR4abO3j	Washington DC
We all are sitting in our homes in the fear of Covid-19. But in a way, its nature's way to retaliate. #nature #coronavirus #covid19 #humanity #environment #hiteshi #WFH #gocorona #ITcompany #values #inspiration #Covid_19 #CoronaWillEndSoon	Atlanta GA
I believe that Africans are wrong to think low Covid numbers mean immunity. I think it has to do with low mobility rates. Going into town is a budgeted journey for most. The real threat is overcrowding. If the virus hits a populated slum I fear the worst. I also hope I am wrong	Brisbane Australia
@srl>> Texas AG says… "criminal sanction" for those encouraging… mail-in ballot because of COVID-19… Is he going to "criminally sanction" the state judge who issued an injunction to allow it? I look forward to reading the bench warrant from t	Texas the USA

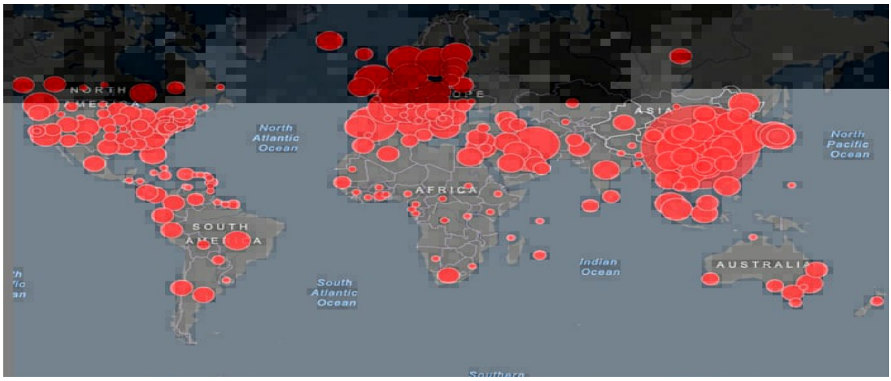


Fig. 6 COVID-19—Tracker from Twitter content and location of the Twitter user

$$\text{Recall} = {}^{\text{TruePositive}}/_{(\text{TruePositive} + \text{FalseNegative})}$$

When the reallocation is A, and it was correctly predicted, it contributes to the true positive of A. When the reallocation is A, and it was wrongly predicted as X, it contributes to the false negative of A. When the reallocation is X and was wrongly predicted to be A, it contributes to the false positive of A. F1–Measure is a harmonic mean of precision and recall

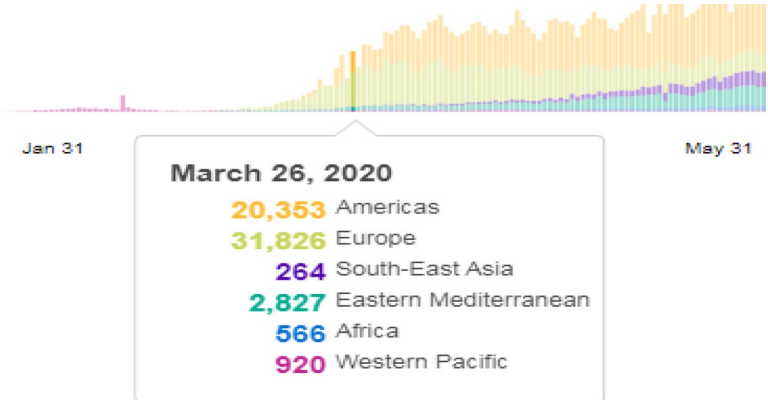


Fig. 7 Survey report released by the World Health Organization

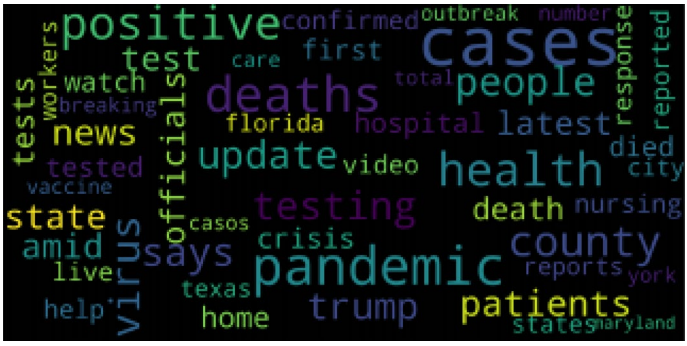


Fig. 8 Word cloud for Twitter content related to COVID-19

$$\text{F1 Measure} = \frac{2 \text{ X Precision X Recall}}{\text{Precision} + \text{Recall}}$$

The standard approach when evaluating the accuracy of a model is the holdout method. We divided the available data into two disjoint sets: the training set, which is fed into the learning algorithm, and the test set, used only to evaluate the model's performance. Commonly, two-thirds of the data items are destined for the training and the remainder is reserved for the testing, but other proportions like 70:30, 80:20 and 90:10.

The results obtained by the three methods are summarized in Table 2. The proposed method HWF and two baselines that consider the information about the two data sources independently in the 10 classes dataset are obtained. HWF was produced high accuracy compared to the other existing classifier algorithms (DTC—decision tree classifier, and MCM—Markov modeling, MEM—maximum entropy modeling, NB—Navie Bayes classifier) (Fig. 9).

Table 2 Infer the Twitter user location related to the COVID-19 based on the weighting factor

Sample tweet	Location 1	Location 2	Location 3	Location 4
$WF(w_1)$	0.16		0.16	
$WF(w_2)$		0.32	0.32	0.32
$WF(w_3)$				0.56
$WF(w_4)$		0.45	0.45	
$WF(w_5)$	0.32	0.32	0.32	
Max	0.32	0.45	0.45	0.56
Predicted location	Location 4			
Sample tweet	Location 1	Location 2	Location 3	Location 4
$WF(w_1)$	0.23	0.23		
$WF(w_2)$		0.16		0.16
$WF(w_3)$	0.24	0.24	0.24	
$WF(w_4)$		0.54		0.42
$WF(w_5)$			0.78	
Max	0.24	0.54	0.78	0.42
Predicted location	Location 3			

4.4 Time complexity

The time is taken for each of the algorithms compared and plotted as a graph, as shown in Fig. 10. To estimate the algorithm's time complexity, assume 'n' as the number of tweets and 'w' as the number of distinct vertices observed. For the process of creation of the hypergraph, it would require $1 + 2 + \cdots + w = O(w^2)$ time. For modeling the hypergraph, all vertices are checked for the Helly property. If the Helly property is observed, that particular vertex is removed, for which the time complexity is comparatively negligible. Hence, modeling takes $\theta(\sum_{edge=0}^{edge=e} w(edge))$, where w (edge) is the number of vertices covered by the edge' edge.' The testing process takes linear time if indexing is used. Hence, the total time complexity is $\omega(n^2)$, since there is no vertex common to all vertices. The following chart (Fig. 10) compares different classifier algorithms based on time complexity. It shows that the time taken for the HWF algorithm is less compared to others.

5 Conclusion and future work

This work proposed a new hypergraph-based weighting factor technique to detect Twitter user's location. The proposed model has been tested with a high volume of streaming data retrieved from Twitter API, and the same data set has been compared with other competent algorithms, namely Naïve Bayes classifier, Markov modeling, maximum entropy modeling and decision tree. The experimental result has been observed for 10 test cases, and it is evident that the location prediction accuracy

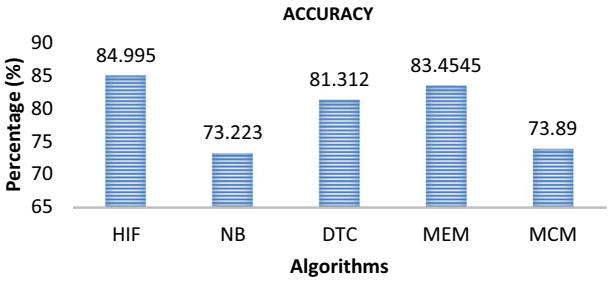


Fig. 9 Comparison of HWF accuracy with existing algorithms. *Note* NBC—Naïve Bayes classifier, MCM—Markov modeling, MEM—maximum entropy modeling, DTC—decision tree classifier

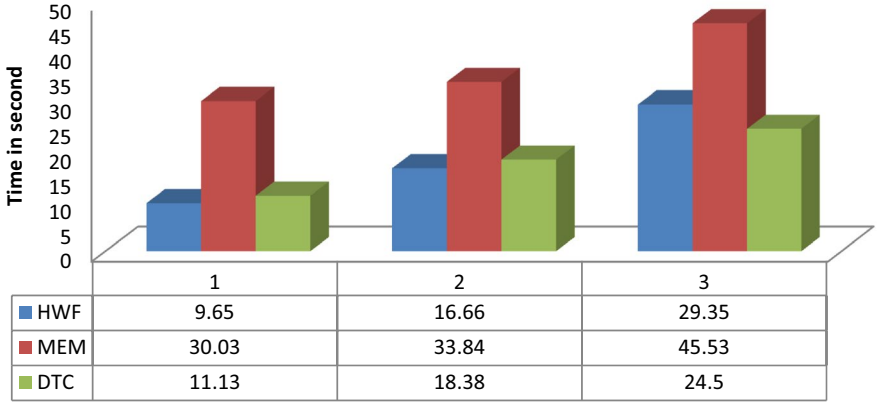


Fig. 10 Time complexity comparision of HWF, MEM, DTC

for the HWF supersedes the remaining for all the cases. Despite prediction accuracy, the time complexity involved in this process has also been found to be moderate. The complete framework has been tested by considering a single source of data from Twitter, and the same can be extended to incorporate multiple sources of data from other social networks. This work can suit most of the emerging applications like cyber investigation team, election polling and disaster management, predicting the users' location with higher accuracy in a short span of time. Geolocation is not just handled on Twitter, yet also numerous different stages like Facebook, Foursquare, Gowalla, etc. The forecast models proposed based on Twitter can likewise be adjusted to other online media locales, requiring a few changes. Be that as it may, before considering model transformations, we should be sure whether the three geolocation issues on Twitter, i.e., forecast of the home area, tweet area and referenced area, are pertinent to the objective stage or not.

References

1. Shahraki ZK, Fatemi A, Malazi HT (2019) Evidential fine-grained event localization using Twitter. Inf Process Manag 56(6):102045
2. Grover P, Kar AK, Dwivedi YK, Janssen M (2019) Polarization and acculturation in US Election 2016 outcomes–can twitter analytics predict changes in voting preferences. Technol Forecast Soc Chang 1(145):438–460
3. Jain VK, Kumar S (2018) Effective surveillance and predictive mapping of mosquito-borne diseases using social media. J Comput Sci 1(25):406–415
4. Pradeepa S, Manjula KR, Vimal S, Khan MS, Chilamkurti N, Luhach AK (2020) DRFS: detecting risk factor of stroke disease from social media using machine learning techniques. Neural Process Lett 9:1–9
5. Pang J, Zhang Y (2017) DeepCity: A feature learning framework for mining location check-ins. In: Eleventh International AAAI Conference on Web and Social Media 2017 May 3
6. Yamaguchi Y, Amagasa T, Kitagawa H, Ikawa Y (2014) Online user location inference exploiting spatiotemporal correlations in social streams. InProceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management 2014 November 3 (pp. 1139–1148)
7. Jordan SE, Hovet SE, Fung ICH, Liang H, Fu KW, Tse ZTH (2019) Using Twitter for public health surveillance from monitoring and prediction to public response. Data 4(1):6
8. Sampath P, Packiriswamy G, Pradeep Kumar N, Shanmuganathan V, Song OY, Tariq U, Nawaz R (2020) IoT Based health—related topic recognition from emerging online health community (med help) using machine learning technique. Electronics 9(9):1469
9. Rodrigues E, Assunção R, Pappa GL, Renno D, MeiraJr W (2016) Exploring multiple evidence to infer users' location in Twitter. Neurocomputing 1(171):30–38
10. Luceri L, Andreoletti D, Giordano S (2019) Infringement of tweets geolocation privacy: an approach based on graph convolutional neural networks. arXiv preprint arXiv:1903.11206. 2019 March 27
11. Rout D, Bontcheva K, Preoţiuc-Pietro D, Cohn T. Where's@ wally? a classification approach to geolocating users based on their social ties. In Proceedings of the 24th ACM Conference on Hypertext and Social Media 2013 May 1 (pp. 11–20)
12. Al-Nabki MW, Fidalgo E, Alegre E, Fernández-Robles L (2019) Improving named entity recognition in noisy user-generated text with local distance neighbor feature. Neurocomputing. 2019 December 5
13. Hasan M, Orgun MA, Schwitter R (2018) A survey on real-time event detection from the twitter data stream. J Inf Sci 44(4):443–463
14. Lakew AM, Tesema GA, Akalu TY (2019) Malaria Outbreak Investigation in Argoba District, South Wello Zone, Northeast Ethiopia, 2016: A case control study
15. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp 851–860
16. Haldar NA, Li J, Reynolds M, Sellis T, Yu JX (2019) Location prediction in large-scale social networks: an in-depth benchmarking study. VLDB J 28(5):623–648
17. Zheng X, Han J, Sun A (2018) A survey of location prediction on twitter. IEEE Trans Knowl Data Eng 30(9):1652–1671
18. Xin M, Wu L (2020) Using multi-features to partition users for friends recommendation in location based social network. Inf Process Manage 57(1):102125
19. Backstrom L, Sun E, Marlow C (2010) Find me if you can: improving geographical prediction with social and spatial proximity. In WWW'10: The 19th International World Wide Web Conference Raleigh North Carolina USA April, 2010. Association for Computing Machinery, New York, pp 61–70. https://doi.org/10.1145/1772690
20. Suresh A, Udendhran R, Balamurgan M (2019) Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers. Soft Comput. https://doi.org/10.1007/s00500-019-04066-4
21. Jurgens D, Finethy T, McCorriston J, Xu YT, Ruths D (2015) Geolocation prediction in twitter using social networks: a critical analysis and review of current practice. In Ninth International AAAI Conference on Web and Social Media 2015 April 21

22. Indira K, Brumancia E, Kumar PS, Reddy SP (2019) Location prediction on Twitter using machine learning Techniques. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) 2019 April 23 (pp. 700–703). IEEE
23. Balaji GN, Subashini TS, A Suresh (2014), An efficient view classification of echocardiogram using morphological operations. J Theor Appl Inf Technol, (JATIT) ISSN: 1992-8645, E-ISSN: 1817–3195, Vol. 67, No.3, September 2014, pp. 732–735
24. Bayat O, Ucan ON (2018) Estimation of Twitter user's nationality based on friends and followers information. Comput Electr Eng 66:517–530
25. Rahimi A, Vu D, Cohn T, Baldwin T (2015) Exploiting text and network context for geolocation of social media users. arXiv preprint arXiv:1506.04803. 2015 June 16
26. Pradeepa S, Geetha K, Kannan K, Manjula KR (2020) DEODORANT: a novel approach for early detection and prevention of polycystic ovary syndrome using association rule in hypergraph with the dominating set property. J Ambient Intell Humaniz Comput, 1–17
27. Suresh A, Kumar R, Varatharajan R (2018) Health care data analysis using evolutionary algorithm. J Supercomput 76:4262–4271. https://doi.org/10.1007/s11227-018-2302-0
28. Wagenseller P, Avram A, Jiang E, Wang F, Zhao Y (2019) Location prediction with communities in user ego-net in social media. In ICC 2019–2019 IEEE International Conference on Communications (ICC) 2019 May 20 (pp. 1–6). IEEE
29. Bretto A (2013) Hypergraphs: first properties hypergraph theory. Springer, Heidelberg, pp 23–42
30. Molnár B (2014) Applications of hypergraphs in informatics: a survey and opportunities for research. Ann Univ Sci Budapest Sect Comput 42:261–282
31. Bretto A (2013) Applications of hypergraph theory: a brief overview. In: Hypergraph Theory, Springer, Heidelberg, pp 111–116
32. Lin J, Cromley RG (2018) Inferring the home locations of Twitter users based on the spatiotemporal clustering of Twitter data. Trans GIS 22(1):82–97
33. Lu H, Niu W, Caverlee J (2018) Learning geo-social user topical profiles with bayesian hierarchical user factorization. In The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval 2018 June 27 (pp. 205–214). ACM
34. Mulder HM, Schrijver A (1979) Median graphs and Hellyhypergraphs. Discrete Math 25(1):41–50