Kuwait J. Sci. **45 (2)** pp 8-14, 2018

Rough set based intelligent approach for identification of H1N1 suspect using social media

Vinay K. Jain, Shishir Kumar *

Dept. of Computer Science & Engineering, Jaypee University of Engineering & Technology, Guna (M. P.), India

**Corresponding author: dr.shsihir@yahoo.com*

Abstract

Social media data offer unique challenges and opportunities for monitoring and surveillance of public health. The identification of epidemic suspect depends on doctor's experience, symptoms and laboratory tests. Delay in identifying the beginning of infectious epidemic results in a big damage to a society. To handle the cases of epidemic effectively, a low-cost, accurate and timely diagnosis system is needed. An intelligent technique based on Rough set theory for identifying suspect of H1N1 using social media, has been presented in this paper.

Classification of symptoms from the dataset has been performed using machine learning techniques. From the large number of symptom attributes mined from the dataset, H1N1 related symptom attributes, have been extracted. These extracted attributes contribute maximum to the decision-making process. Rough set theory has been used to evaluate significant attributes (symptoms) from symptom attribute set by generating reducts using indiscernibility relation. Identification of suspects is performed using significant conditional attributes and dependency rules generated from reducts. The utilization of presented social media based medical decision support system turn out to be an effective approach to assist government and health agencies in decision-making.

Keywords: H1N1; influenza; rough sets; swine flu; text classification.

1. Introduction

Latest Technologies are emerging to deal with the extraction of knowledge from these data sources in different domains (Hassanien & Ali, 2004; Tripathy *et al*., 2011; Moses & Deisy, 2015). Economic domain is the major sector, where extensive amount of work has been carried out in comparison to medical sciences (Tay & Shen, 2002). Internet is one of the most important resources in the field of surveillance systems for tracking disease outbreaks. It provides an opportunity for low-cost, fast computation of data in comparison to existing traditional surveillance systems (Jain & Kumar, 2015). Social media has been a primary focus in the domain of information retrieval and text mining (Pang & Lee, 2008). Social media users are conscious of their health and share their personal experience, with the causes or symptoms that are related to them (Ceron, 2013).

The medical diagnosis system varies from the level at which they attempt to deal with different complicated aspects of diagnosis like importance of symptoms, varied symptom patterns and the relation between diseases themselves (Nallamuth & Palanichamy, 2015). Delay in identifying the beginning of an infectious epidemic result in critical stage to a society (Glik, 2007). The work has been concerned in the domain of medical sciences for improving the facilities and enhancing in diagnosis of diseases using social media data and Rough set theory. Rough sets provide effective solution for managing the uncertainties and imprecision present in text data as compared to fuzzy sets, which deals with partial membership (Kumara & Inbaranib, 2015; Pandey *et al*., 2013).

With the popularity of social media platforms such as Twitter, an Intelligent Technique based on Rough sets theory has been proposed to identify the suspect of Swine Flu Virus (H1N1) by considering significant symptoms indicated in the tweets. The prominent two sub tasks of proposed work are:

Identification of important conditional attributes (symptoms) from tweets

Discovery of decision rules characterizing dependency between the values of conditional attributes and decision attributes

2. Public health emergencies

Social networking sites are one of the easiest and cheapest techniques for spreading information rapidly (Emarketer, 2015). Public terminology (keywords) used during the epidemic provides a measure to gain knowledge and help in information retrieval (Jain & Kumar, 2015). Impact of Influenza-A H1N1 in India during different time intervals is shown in Figure 1. Data has been collected for Indian region using Geo-tagging feature provided by the Twitter API (Twitter, 2015).

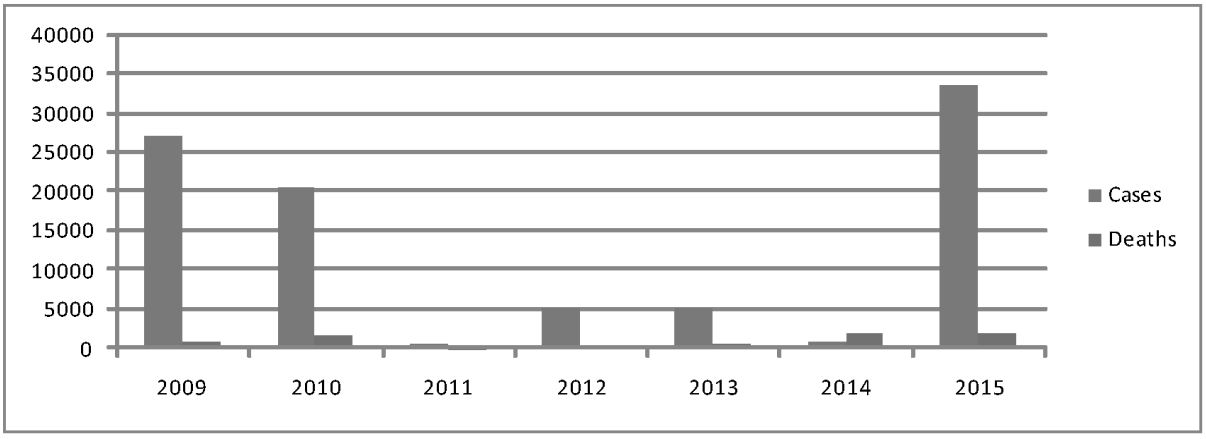

Fig. 1. Year wise cases of Influenza-A H1N1 (PIB, 2015)

3. Related work

The traditional method for identification of the suspect is based on clinical data sets or surveying results. Earlier work showed that social media data provide a rich source of information, which helps in identification of the disease outbreak in the world (Santos & Matos, 2014). Chew (2010) proposed a method based on precise keywords related to H1N1 during 2009 H1N1 pandemic, using Twitter. Hu *et al*. (2011) system is based on Google web search queries related to an influenza epidemic. Lampos & Cristianini (2010) used content analysis and regression models to measure and monitor levels of H1N1 pandemic in the United States. Aramaki *et al*. (2011) used support vector machine (SVM) for predicting influenza rates in Japan. Stewart & Diaz (2012) proposed a daily activity system related to corresponding disease and symptoms with an early-warning system. Bodnar & Salathe (2013) applied various classification techniques for detecting influenza. Parker *et al*. (2013) developed a low-cost framework for tracking public health condition trends via Twitter.

4. Rough sets theory

Rough sets theory is used to process uncertain and incomplete information (Pawlak, 1982). It is a mathematical tool used for approximate reasoning of decision making and helps in classification of objects. It deals with imprecision, vagueness and uncertainty in data analysis (Qian *et al*., 2008; Peng *et al*., 2008) and has successfully been applied to many practical problems (Pawlak, 1991; Hassanien & Ali, 2004). There are four main concepts of Rough set theory such as information systems, indiscernibility, reduction of attributes, and dependency (Pawlak, 1991; Jiye, 2007). Presently, research on Rough sets focus on many significant issues, like attribute reduction problems, approximation operator models axiomatic systems, generalizations (Jia *et al*., 2016). It helps in finding patterns in text data, dimensionality reduction, attribute dependency analysis, feature identification and classification, which make it suitable for text classification and rule generation (Miao *et al*., 2009). Reduct and Core, which play a vital role in finding significant rules, are important in the development of decision support system (Jia *et al*., 2016).

5. Proposed method

The proposed method is based on the identification of relevant tweets containing H1N1 related symptoms and analyzed using Rough set theory for identification of suspects. From a large number of attributes mined from the dataset, significant attributes that contribute maximum to the decision-making need to be extracted. The system is divided into five phases: data collection phase, preprocessing of tweets, symptom identification, information system and suspect identification phase. The system architecture is shown in Figure 2.

Step 1: Data collection

In this phase, identification of related tweets which indicates the presence of H1N1 symptoms are fetched using Twitter API (Twitter, 2015) with the help of significant keywords and medical science terms. Significant keywords collection methodology is based on Jain & Kumar (2015), which gives dynamic keywords. Some of the significant keywords are:

Keywords:{#SwineFlu,#flu,#H1N1,#Swine,#swinevirus,##h1n1,#influenza,#swinefluindia, #infuenzavirus, #delhiSwineflu ,#Fluvirus}

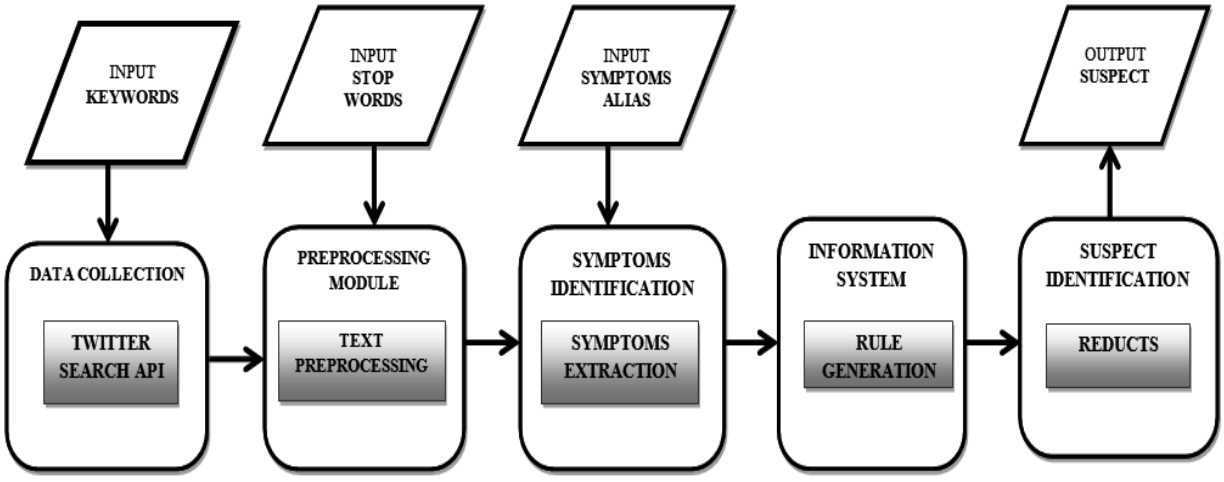

Fig. 2. Framework for identification of suspect

Step 2: Pre-processing of tweets

Every word present in a tweet is prominent in decision making. Thus, pre-processing is required to decrease the noise from tweets and filter out irrelevant tweets (Khan *et al*., 2016). Pre-processing steps such as tokenization, stop word removal, stemming, lemmatization, feature weighting, dimensionality reduction and frequency based methods were applied for normalization of tweets. Every relevant word related to H1N1 is a feature and classification techniques such as SVM, Naïve Bayes, Random Forest and Decision Tree were applied. The results are explained in terms of precision, recall, accuracy, and F-measure and shown in Table 1. The F-measure provides the overall performance of a classifier and is calculated using the following formula presented in Equation (1):

$$\text{F} - \text{measure} = \frac{2(\text{Precision})(\text{Recall})}{\text{Precision} + \text{Recall}} \tag{1}$$

Table 1. Comparison of classification techniques

Classifier	F-Measure	Precision	Recall
Naïve Bayes	0.77	0.70	0.86
SVM	0.77	0.70	0.84
Random Forest	0.70	0.66	0.75
Decision Tree	0.70	0.67	0.74

Step 3: Symptoms identification

Symptoms of H1N1 flu virus has been taken from Centers for Disease Control and Prevention (CDCP, 2015), National Health Portal of India (NHP, 2015) and Apollo Health services (Apollo health city, 2015). Every tweet is examined for match/similarity from the existing list of symptoms and if found, store them into the suspect information Table. To improve symptom analysis, alias corresponding to symptoms is also considered and presented in Table 2.

Table 2. Symptoms and corresponding alias and attributes.

Symptoms	Alias	Attribute
Fever	Mild fever, High temperature, feverishness	a_1
Cough & Cold	Chilly	a_2
Headache	Head pain	a_3
Running nose	Blocked nose	a_4
Sore throat	tonsillitis	a_5
Shortness of breath	breathing trouble, bronchitis	a_6
Loss of appetite	Fatigue, illness, dizziness	a_7
Diarrhoea	Loose motion	a_8
Vomiting	abdominal pain	a_9
High Risk Diseases	Sick	a_{10}

Significant symptom is considered as an attribute. New significant symptom is created, which represent High Risk Diseases(a10) = {Heart diseases, Kidney disorders, Pneumonia, Liver disorders, Metabolic disorders, Blood disorders, Endocrine disorders, Lung diseases, Asthma, Weak immune system}.

Step 4: Information system

Consider a suspect information system represented in Table 2 of 'n' Twitter users 'ti' where, i = (1, 2,…., n) as the set of objects of the universe with a set of attributes given in are found using Jhonson algorithm and Genetic algorithm using Rosetta tool (Hvidsten, 2013):

6. Results

Proposed approach provides following general rules for identification of H1N1 suspect. For suspect identification Rule 4 and Rule 5 represent the relevant symptoms from Jhonson algorithm and Rule 1 to Rule 5 represent relevant symptoms for the possible suspect using from Genetic Algorithm. Applying these rules on the test data, results in the outcomes presented in Table 4.

Table 3. Classical view of the Suspect information system.

Tweets	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	Decision
t_1	y	n	y	y	y	n	y	n	y	y	1
t_2	n	y	n	y	n	y	n	n	n	n	0
t_3	y	y	n	n	y	y	n	n	y	n	0
t_4	y	n	y	y	n	y	n	n	n	y	1
t_5	y	y	n	y	y	n	n	y	n	n	0
t_6	y	y	n	n	n	n	y	n	y	y	1
t_7	n	n	y	n	y	y	n	n	n	n	0
:	y	y	y	y	n	y	y	y	y	n	1
:	y	y	n	y	y	n	y	n	n	n	1
t_n	y	n	n	n	y	n	y	y	y	y	1

Table 2. The attribute Suspect is considered as the decision attribute with value=1, represent suspect and with value=0, represent not suspect. For particular, user 't1' is characterized in the table by the attribute value set (fever, yes), (cough & cold, no), (headache, yes), (running nose, yes), (sore throat, yes), (short of breath, yes), (loss of appetite, yes), (diarrhea, yes), (vomiting, yes), (High risk disease, yes) which form the information about the particular user. The reduced medical information system is presented in Table 3.

Step 5: Discovering important rules

In this phase, selection of important rules based on Rough sets theory is applied, which help in identifying the suspect from the data set. Consider a decision table L = (T, A, D), where T = {t0, t1, tm−1} is a set of records in the table. A = {a0, a1, ..., ap−1} is a set of the condition attributes and D is a set of the decision attributes. Let us consider decision tables with one decision attribute. A set of rules R is generated from this table L, where R = {Rule0, Rule1, ..., Rulen−1} is called as reduct (Jiye, 2007). Reducts contains a set of attributes that are sufficient to define all information in the data (Jiye, 2007). The new decision table is constructed as follows. A new decision table Sm×(n+1) has been prepared, where each record from the original decision table u0, u1, ..., um−1 is the row, and the columns of this new table consisting of Rule0, Rule1, ...,Rulen−1 and the decision attributes. Decision rules

Rules from Jhonson algorithm

Rule 1: a_3 (n) AND a_7 (n) => D (0)

Rule 2:a_1(y) AND a_7 (n) => D (0)

Rule 3:a_2 (n) AND a_7 (n) => D (0)

Rule 4:a_{10}(y) => D (1)

Rule 5: a_2(y) AND a_3(y) => D (1)

Rules from Genetic algorithm

Rule 1:a10(y) => D(1)

Rule 2: a_7(y) => D(1)

Rule 3:a_4 (n) AND a_6 (n) => D(1)

Rule 4:a_6 (n) AND a_9(y) => D(1)

Rule 5:a_6 (n) AND a_8 (n) => D (1)

Rule 6: a_7 (n) AND a_8 (n) => D(0)

Rule 7:a_7 (n) AND a_9 (n) => D(0)

Rule 8:a_3 (n) AND a_7 (n) => D(0)

Rule 9:a_6(y) AND a_8 (n) AND a10 (n) => D(0)

Rule 10: a_5(y) AND a_7 (n) => D(0)

Table 4. Identification of suspected Tweets.

Algorithm	No. of Tweets
Suspected tweets	28
Tweets with symptoms	702
Irrelevant tweets	1106
Total	1836

7. Conclusion

An effective social media based Medical Decision Support system has been developed for identification of suspect of H1N1 flu virus. The experimental results showed that Rough set theory helps in finding minimal set of attributes using attribute reduction techniques and hence, preserves the power of decision-making. Rough set theory find suitable in decision-making because it does not required preliminary information related to data, such as probability assignment in Bayesian data analysis and Dempster-Shafer theory, grade of membership in fuzzy set theory. Proposed approach is one of the novel attempts using social media data rather than clinical data sets for decision making in identification of suspects. This approach can be applied in real-world multi-attributes group decision making problems like in economic sector, election prediction, government policies, sports events etc. The experimental results strongly suggest that it can perform better, if data has been collected from different demographic regions with multiple time-intervals. This work is only for experimental study and generated rules may not applicable in same form in context to real life problems. Suitable amendments should be incorporated as per circumstances.

In future research work, incorporation of features and reduction algorithms along with other intelligent techniques may be considered to develop a social media based expert system.

References

Apollo Health Services (2015). Apollo Health Services, India (http://www.apollohealthcity.com /swine-flu/) [Accessed: 25-Sep-2015].

Aramaki E., Maskawa S. & Morita M. (2011). Twitter catches the Flu: detecting influenza epidemics using Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA.

Bodnar, T. & Salathé, M. (2013). Validating models for disease detection using Twitter. International World Wide Web Conference Companion. Rio de Janeiro, Brazil.

CDCP (2015). Centers for Disease Control and Prevention (http://www.cdc.gov /h1n1flu/qa.htm)[Accessed: 25-Sep-2015]

Ceron, A. C. (2013). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media & Society, **16**(2):340-358.

Chew, C. M. (2010). Pandemics in the age of twitter: A content analysis of the 2009 h1n1 outbreak. Master's thesis, University of Toronto.

Emarketer (2015). Emarketer (http://www.emarketer.com) [Accessed : 25-Sep-2015].

Glik, D. (2007). Risk communication for public health emergencies. Annual Reviews of Public Health, **28**:33-54.

Hassanien, A. E. & Ali, J.M.H. (2004). Rough set approach for generation of classification rules of breast cancer data. INFORMATICA, **15**(1):23–38.

Hu, X., Lei, T. & Liu, H. (2011). Enhancing accessibility of microblogging messages using semantic knowledge.20th ACM International Conference on Information and Knowledge Management, New York, NY, USA.

Hvidsten, T. R. (2013). A tutorial-based guide to the ROSETTA system: A Rough Set Toolkit for Analysis of Data.

Jain,V. K. & Kumar, S. (2015). An effective approach to track levels of influenza-A (H1N1) Pandemic in India Using Twitter. Procedia Computer Science, **70**(1):801–807.

Jia,X., Shang, L., Zhou, B. & Yao, Y. (2016). Generalized attribute reduction in Rough set theory. Knowledge-Based Systems, **91**(1):204–218.

Jiye, Li. (2007). Rough set based rule evaluations and their applications, Ph.D Thesis, University of Waterloo.

Khan, K., Ullah, A. & Baharudin, B. (2016). Pattern and semantic analysis to improve unsupervised techniques for opinion target identification. Kuwait Journal of Science, **43**(1):129-149.

Kumara, S. U. & Inbaranib, H. H. (2015). A novel neighborhood rough set based classification approach for medical diagnosis. Procedia Computer Science, **47**(1):351–359.

Lampos, V. & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web.In 2nd IAPR Workshop on Cognitive Information Processing, IEEE Press.

Miao, D., Duan, Q., Zhang, H. & Jiao, N. (2009). Rough set based hybrid algorithm for text classification. Expert Systems with Applications, **36**(5):9168–9174.

Moses, D. & Deisy, C. (2015). A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data. Kuwait Journal of Science, **42**(2):206-235.

Nallamuth, R. & Palanichamy, J. (2015). Optimized construction of various classification models for the diagnosis of thyroid problems in human beings. Kuwait Journal of Science, **42**(2):189-205.

NHP(2015). National Health Portal of India,(http://www.nhp.gov.in/diseaseaz/s/swineflu) [Accessed : 25-Sep-2015]

Pandey, P., Kumar, S. & Srivastava, S. (2013). Forecasting using Fuzzy time series for diffusion of innovation: Case of Tata Nano car in India. National Academy Science Letters, **36**(3):299-309.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, **2**(1):1–135.

Parker, P., Wei, Y., Yates, A., Frieder , O. & Goharian, N. (2013). A framework for detecting public health trends with Twitter.IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, ON, Canada.

Pawlak, Z. (1982). Rough sets. International Journal of Information and Computer Science. **11**(5):341-356.

Pawlak, Z. (1991). Rough sets: theoretical aspects of reasoning about data. Theory and decision library.Kluwer Academic Publishers. Norwell, MA, USA.

Peng, Y., Kou, G., Shi, Y. & Chen, Z.X. (2008). A descriptive framework for the of data mining and knowledge discovery. International Journal of Information Technology & Decision Making, **7**(4):639-682.

PIB(2015). Preventive measures for Swine flu(http://pib.nic.in/newsite/PrintRelease.aspx? relid=115710) [Accessed: 15-Aug-2015].

Qian, Y. H., Liang, J.Y., Li, D.Y. et al. (2008). Measures for evaluating the decision performance of a decision table in rough set theory. Information Sciences, **178**(1):181-202.

Santos, J.C. & Matos, S. (2014). Analysing Twitter and web queries for flu trend prediction. Theoretical Biology and Medical Modelling, **11**(1):1-11.

Stewart, A. & Diaz, E. (2012). Epidemic intelligence: for the crowd, by the crowd. In Proceedings of the 12th international conference on Web Engineering, Berlin.

Tay, F.E.H. & Shen, L. (2002). Economic and financial prediction using Rough sets model.European Journal of Operational Research, **141**(3):641–659.

Tripathy, B. K., Acharjya, D. P. & Cynthya, V. (2011). A framework for intelligent medical diagnosis using rough set with formal concept analysis. International Journal of Artificial Intelligence & Applications, **2**(2):1-14.

Twitter (2015). Twitter Developer Page (https://dev.twitter.com/docs/api/1/get/search) [Accessed: 25-June-2015].

Submitted: 11/01/2016
Revised : 03/05/2016
Accepted : 17/05/2016

نهج ذكي يعتمد على مجموعة الاستقراب لتحديد هوية المشتبه به المصاب بمرض H1N1 باستخدام وسائل التواصل الاجتماعي

فيناي كومار جاين١ ، شيشير كومار٢

[1]طالب دكتوراه بقسم علوم الحاسوب والهندسة، جامعة جايبي للهندسة والتكنولوجيا، غونا (M.P.)، الهند

[2،×] أستاذ بقسم علوم الحاسوب والهندسة، جامعة جايبي للهندسة والتكنولوجيا، غونا (M.P.)، الهند

*dr.shsihir@yahoo.com

الملخص

تقدم بيانات وسائل التواصل الاجتماعي تحديات وفرص فريدة لمراقبة الصحة العامة. ويعتمد تحديد هوية المشتبه فيه المصاب بوباء ما على خبرة الطبيب والأعراض والفحوصات المخبرية. التأخير في تحديد بداية الوباء المُعدي يؤدي إلى أضرار كبيرة تجاه المجتمع . وللتعامل مع حالات الوباء بشكل فعال، يتطلب الأمر توفير نظام تشخيص منخفض التكلفة ودقيق ومناسب. وفي هذا البحث، تم عرض تقنية ذكية تقوم على نظرية مجموعة الاستقراب لتحديد هوية المشتبه به المصاب بمرض H1N1 باستخدام وسائل التواصل الاجتماعي.

وتم تصنيف الأعراض من مجموعة البيانات باستخدام تقنيات التعلم الآلي. من عدد كبير من سمات الأعراض المستخرجة من مجموعة البيانات، تم استخراج سمات أعراض H1N1 ذات الصلة. هذه السمات المستخرجة تساهم إلى أقصى حد في عملية صنع القرار. تم استخدام نظرية مجموعة الاستقراب لتقييم سمات (أعراض) هامة من فئة سمات (أعراض) تم تعيينها عن طريق توليد الاختزال باستخدام علاقة غير قابلة للتعريف. يتم تحديد هوية المشتبه بهم باستخدام سمات شرطية كبيرة وقواعد التبعية الناتجة عن الاختزال. إن استخدام نظام دعم اتخاذ القرار الطبي القائم على وسائل التواصل الاجتماعي الحالية قد تحول إلى نهج فعال لمساعدة الحكومة والوكالات الصحية في اتخاذ القرارات.