OPEN ACCESS

Research Article

Application of Twitter and web news mining in infectious disease surveillance systems and prospects for public health

Anwendung von Twitter und Web News Mining in Überwachungssystemen für Infektionskrankheiten und Perspektiven der öffentlichen Gesundheit

Kia Jahanbin[1]
Fereshte Rahmanian[1]
Vahid Rahmanian[2]
Abdolreza Sotoodeh Jahromi[2]

1 Research Center for social determinants of health, Jahrom University of Medical Sciences, Jahrom, Iran

2 Zoonoses Research Center, Jahrom University of Medical Sciences, Jahrom, Iran

Abstract

Aims: With the advancements of communication technology and growing access to social networks, these networks now play an important role in the dissemination of information and news without going through the time-consuming channels of official news networks. Analysis of social networking data is a new, interesting branch of text mining science. This study aimed to develop a text mining technique for extracting information about infectious diseases from tweets and news on social media.

Methods: A method called "Fuzzy Algorithm for Extraction, Monitoring, and Classification of Infectious Diseases" (FAEMC-ID) was developed by the use of fuzzy modeling of the Takagi-Sugeno-Kang type. In addition to the real-time classification, the method is able to update its vocabulary for new keywords and visualize the classified data on the world map to mark the high risk areas.

Results: As an example, the monitoring was performed for measles-related news items over a 183-hour period from 01/03/2019 (01:00 am) to 08/03/2019 (12:00 pm), which were related to 2,870 tweets from 2,556 users. This monitoring showed that the number of tweets posted from each region ranged from 1 to 47, with the highest number, 47 tweets, belonging to Canada. The origins of most measles-related news were in the Americas and Europe, and they were mostly from the United States and Canada.

Conclusion: The performance analysis of the developed method in comparison with other algorithms in the literature demonstrated the excellent precision of the method with a recall ratio of 88.41% and the high inter-correlation of data in each class. The proposed algorithm can also be used in the development of more effective monitoring and tracking systems for other human and even animal health hazards.

Keywords: fuzzy classification, surveillance system, Twitter, text mining, infectious disease

Zusammenfassung

Zielsetzung: Mit der Weiterentwicklung der Kommunikationstechnologie und dem wachsenden Zugang zu sozialen Netzwerken spielen diese Netzwerke eine wichtige Rolle zur Verbreitung von Informationen und Nachrichten, ohne dass die zeitaufwendigen Kanäle offizieller Nachrichtennetzwerke durchlaufen werden müssen. Die Analyse sozialer Netzwerkdaten ist ein neuer, interessanter Zweig der Text-Mining-Wissenschaft. Diese Studie zielt darauf ab, eine Text-Mining-Technik zu entwickeln, um Informationen über Infektionskrankheiten aus Tweets und Nachrichten in sozialen Medien zu extrahieren.

Methode: Als Analysemethode wurde der sog. „Fuzzy-Algorithmus zur Extraktion, Überwachung und Klassifizierung von Infektionskrankheiten"

(FAEMC-ID) unter Verwendung des Fuzzy-Modells des Takagi-Sugeno-Kang-Typs entwickelt. Zusätzlich zur Echtzeitklassifizierung kann die Methode neue Schlüsselwörter aktualisieren und die klassifizierten Daten auf der Weltkarte visualisieren, um Hochrisikobereiche zu markieren.

Ergebnisse: Als Beispiel wurde das Monitoring für Nachrichten mit Bezug zu Masern über einen Zeitraum von 183 Stunden vom 01.03.2014 (01:00 Uhr) bis 08.03.2014 (12:00 Uhr) durchgeführt, das 2.870 Tweets von 2.556 Benutzern umfasste. Das Monitoring ergab als Anzahl der von jeder Region geposteten Tweets 1 und 47 mit der höchsten Anzahl von 47 Tweets aus Kanada. Der Ursprung der meisten Nachrichten über Masern war in Amerika und Europa; die Tweets stammten größtenteils aus den Vereinigten Staaten und Kanada.

Schlussfolgerung: Die Analyse der entwickelten Methode liefert im Vergleich zu anderen Algorithmen in der Literatur eine ausgezeichnete Präzision mit einer Rückrufquote von 88,41% und einer hohen Interkorrelation der Daten in jeder Klasse. Der vorgeschlagene Algorithmus kann auch zur Entwicklung wirksamer Überwachungs- und Nachverfolgungssysteme für andere Gesundheitsgefahren für Mensch und Tier verwendet werden.

Schlüsselwörter: Fuzzy Klassifikation, Surveillance System, Twitter, Text Mining, Infektionskrankheit

Introduction

Today, social media generate vast amounts of data on a daily basis in a wide variety of areas including technology, medicine, history, political and social news, sports, and many other fields. These data can be refined and analyzed to extract economically and scientifically valuable knowledge and have therefore piqued the interest of researchers in many areas [1], [2], [3].

In recent years, big data science has emerged as a powerful tool for collecting, storing, managing, and analyzing data on a large scale [4]. Big data can be characterized by five features: volume, variety, velocity, variability, and veracity. Among these features, the most important is the volume or size, according to which data can be classified into three categories [5]:

1. Structured: Data that is organized in a predefined schema.
2. Semi-structured: Data that does not require a predefined schema.
3. Unstructured: Data that is stored without any defined structure or schema.

A great portion of all data produced and consumed across the world is in textual form. The science of text mining is focused on the extraction of high-quality information from textual data [6]. The major applications of text mining include texts categorization, concept/entity extraction, text clustering, text summarization, sentiment analysis, and entity relation modeling [7].

Web-news mining from media and social networks is one of the major applications of text mining in social sciences. An automated news-mining-based system can monitor, analyze, and classify news according to its contents, which is useful not only for managing news articles but also for developing recommenders and security systems [1].

Twitter is one of the world's most popular social networks. The highly interesting applications of this micro-blogging platform have attracted the attention of researchers. At present, Twitter has over 11 million active users, who post about 6 million tweets every day, including instant messages and comments. Given the easily accessible and extremely rich information contained in tweets, they can be used in a wide range of applications, including the analysis of political trends, product performance, and the monitoring of health-related events [8], [9] .

In the model proposed in this paper, the unstructured data about infectious diseases like influenza, HIV/AIDS, malaria, measles, poliomyelitis, tuberculosis, plague, Ebola and cholera are extracted from Twitter and then subjected to text cleanup, term filtering, and finally categorization operations. Since the focus of the work is on real-time application, the model is implemented with the help of a fuzzy rule-based evolutionary algorithm called Eclass1-MIMO.

Literature review

In 2014, the term "social big data" was used for the first time to refer to the data generated by social networks [4], [10]. This includes, for example, the 30 million tweets posted every day, the 3,000 photos uploaded to Flickr every minute, and the 15 million blog posts written on a daily basis. These social networking data can have scientifically and economically significant uses in many fields including sociology, psychology, politics, commerce, and healthcare [8], [11], [12], [13].

Text mining can be discussed from two perspectives – the type of knowledge extracted and applications. Applications of text mining can be categorized as follows:

- **Security applications:** Text mining packages have extensive use in security software, especially for analyzing online plain texts such as websites and weblogs for national security protection purposes [8].
- **Biomedical applications:** A wide range of text mining tools and software has been developed for biomedical applications [10]. For example, PubGene is a well-known Internet service that combines biomedical text mining with network visualization [14].
- **Online media applications:** Media corporations such as the Tribune Company have utilized text mining to achieve enhanced data clarity and create more interesting contents for readers. This science has also been used in the public sector to develop software for the monitoring and tracking of terrorist activities [15].
- **Business and marketing applications:** Text mining is finding extensive use in business and marketing intelligence and particularly in customer relations management [16], [17].
- **Sentiment analysis:** Sentiment analysis can be discussed from the perspective of the type of information extracted and its application. For example, sentiment analysis has been used for the analysis of movie reviews [18] and also for comment recognition in the field of artificial emotional intelligence [2], [19].
- **Academic applications:** Text mining is one of the major tools that large publishers use for data categorization and retrieval from large databases [8].
- **Text categorization:** Text categorization is an automatic process whereby text data are organized into multiple predefined categories or classes. One of the applications of text categorization is the opinion categorization, which gives an insight into the opinion of users of social networks like Facebook or Twitter about a certain topic (e.g. a law, a treatment, a political view, etc.) [20].
- **Text clustering:** Unlike text categorization, text clustering is focused on the unsupervised management of text documents [21].
- **Text summarization:** Automatic text summarization algorithms are language-independent (multilingual) tools for generating a summary of a text [5], [22], [23].

This paper presents a method based on a Takagi-Sugeno-Kang (TSK) fuzzy system called the Eclass1-MIMO model for the categorization of news on Twitter about infectious diseases with epidemic potential. In developing the method, the authors aim to create an accurate text categorization system with real-time applicability for marking high risk areas based on tweets for improved monitoring and timely control of growing epidemics and related damage.

Methods

One of the most effective ways to prevent and control epidemics is to monitor and track the news about the spread of contagious diseases. This section explains the general frame and main structure of the proposed model for the collection of raw data about a select group of contagious diseases from related news and tweets and the analysis of these data.
The proposed method consists of 4 phases:

1. Data cleanup and integration and term extraction
2. Web and tweet crawling
3. Applying fuzzy rules and fuzzy classifier
4. Visualization

The first phase consists of data cleanup, data integration, and term extraction steps. The term extraction step consists of letter case homogenization (transforming all words to lowercase), tokenization, stemming, filtering stop words (removing pronouns, auxiliary verbs, and so on), and term filtering with the TF-IDF method.
In the proposed method, classification and evolving fuzzy rules are developed with the help of fuzzy rule-based classification package (FRBS) [24], [25]. The evolving fuzzy system plays a fundamental role in the text analysis, i.e. updating the terms being extracted from the database [26]. This is important because, considering the large volume and unpredictable nature of the news and tweets related to infectious diseases and the likely emergence of new terms over time, the terms used in classification must be regularly updated. To resolve this issue, the proposed method makes use of evolving fuzzy rules and implements the text classification scheme with the Eclass1-MIMO method based on TSK rules [27], [28], [29], [30].
The visualization component of the proposed method aims to assist real-time monitoring and tracking of the onset and spread of epidemics, which can greatly contribute to the efficacy of active health and research systems in this area. Details of the proposed method are illustrated in Figure 1.

Data cleanup, data integration, and term extraction

As shown in Figure 1, the first phase of the proposed method consists of three steps:

1. Text cleanup: This step involves processing the tweet and news contents to remove redundant characters such as @ (“at” sign), # (hashtag), rt (for retweets), emotions, metadata, links, etc., which should be cleaned before classification [31]
2. Data integration: After text cleanup, tweets and news are integrated into related classifications.
3. Term extraction: This step consists of the following processes:
    - **Tokenization:** In this step, the streams of textual data are decomposed into words, symbols, phrases,

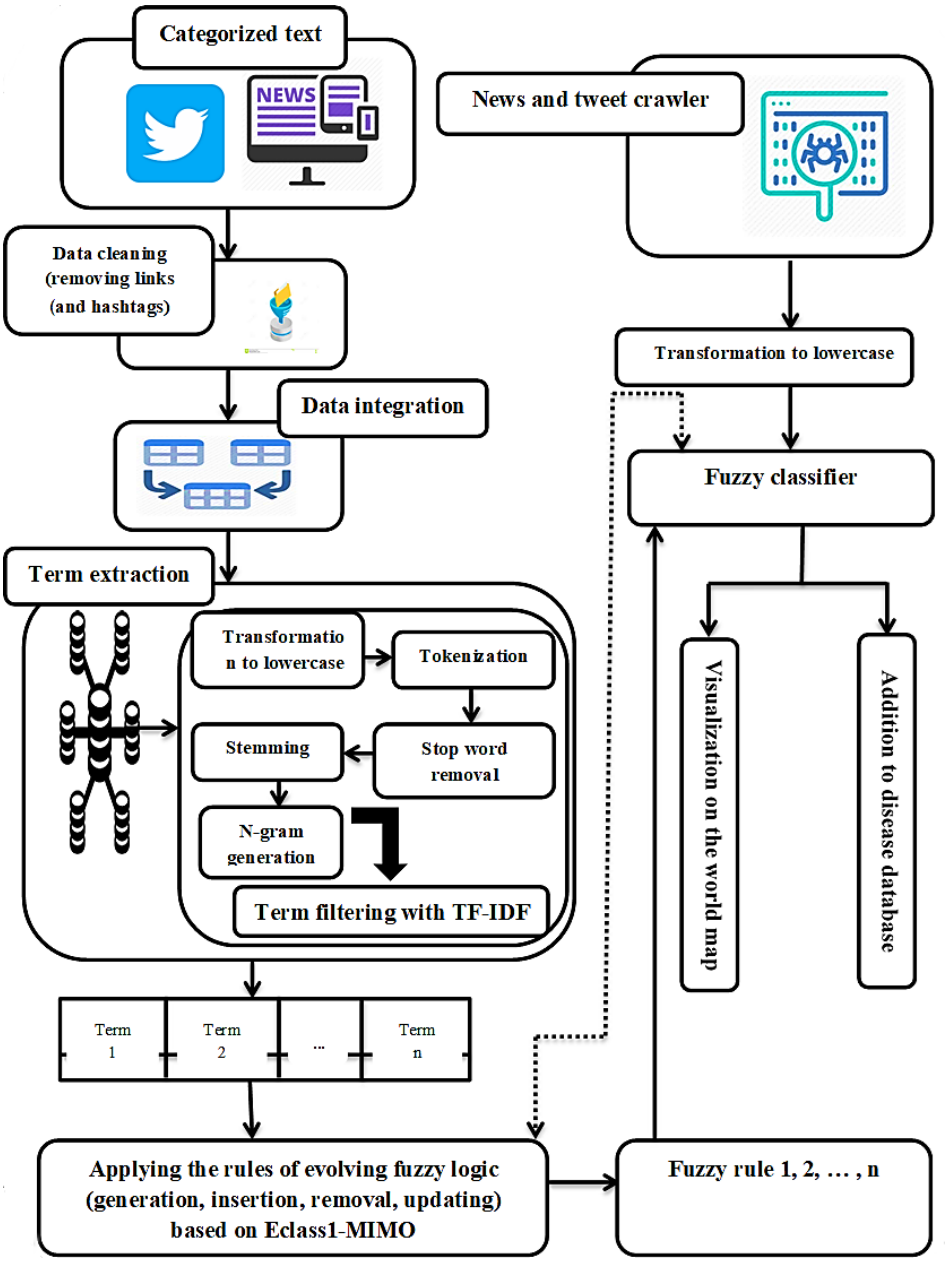

Figure 1: Framework of data collection and monitoring for infectious diseases

and other meaningful elements as well as keywords that are valuable for classification, clustering and analysis of texts [1], [31].

- **Homogenization:** In this step, all words in the database are transformed to lowercase in order to prevent redundant terms [31].
- **Stopword filtering:** This step involves finding and removing pronouns, prepositions, and "to be" verbs from the text [32].
- **Stemming:** In this step, the inflected and derived words (with prefixes, suffixes, etc.) are converted to their base form in order to reduce the number of redundant terms [1], [31]. In this work, stemming is done with the help of the Snowball algorithm [32].
- **n-gram generation:** n-gram is an alternating sequence of n items (characters, letters, etc.); an n-gram is said to be a unigram if n=1, bigram if n=2, and trigram if n=3. n-gram generation has extensive use in language identification [20] and speech recognition, and contributes to the identification of keywords that are not valuable by themselves. In this study, the learning accuracy of the model is improved by the use of bigrams [33], [34].
- **Term filtering:** The tokenization step extracts all terms of each tweet without considering the frequency of each term, which can reflect its importance. The term filtering step involves removing the terms that rarely appear in the text, the terms that have a constant distribution, and the terms that appear too frequently in the text in order to prevent the redundant growth of the term set [1].

Database collection method

In the proposed method, the news about smallpox, influenza, malaria, measles, poliomyelitis, tuberculosis, plague, Ebola and cholera in various news sites is collected by a powerful API called Newsapi. This API collects the news of 54 countries from 134 major news organizations including CNN, BBC, CBC, Washington Post, etc. The code written in Ruby for extracting news about measles, for example from Twitter between 01/03/2019 and 08/03/2019 is presented below:

```
Require 'open-url'
url='https:/newsapi.org/v2/everything'
    ‘language=en&’
    ‘q=measles disease&’
    ‘from=2019-03-01&’
    ‘to=2019-03-08&’
    ‘sortBy=relevancy&’
    ‘apikey=[write your api]’
Req=open(url)
Response_body-req.read
Put response_body
```

Tweets crawler was coded with the R language. For example, the following code was used to crawl the HIV-related tweets from 01/03/2019 to 08/03/2019:

```
Library(twitterR)
Consumer_key=“[your consumer_key]”
Consumer_secret=“[your consumer_key]”
Access_token=“[your access_token]”
Access-secret=“[your access_secret]”
Setup_twitter_oauth(Consumer_key,  Consumer_secret, Access_token, Access-secret)
Tw=SearchTwitter(“#HIV”,n=1e4,scince='2019-03-01')
```

Application of fuzzy rules and fuzzy classifier

The next step after extracting the terms related to each class involves the application of fuzzy rules and fuzzy classifier. The system developed for this phase consists of two steps:

1. Generation and updating of fuzzy rules
2. Classification of news/tweet related data

In the proposed system, fuzzy rules are generated and updated by a fuzzy model called Eclass1-MIMO, which is a multi-input-multi-output framework based on the rules of the TSK fuzzy system [26]. In addition to using the TSK fuzzy system, the Eclass1-MIMO model can remove useless potential terms with the help of an “aging” mechanism. Using this mechanism, the potential terms that have not been recently used to classify any text are removed from the list of keywords.
The rules of the TSK-based fuzzy model are defined as follows:

1. $Rule_i$ = IF (A_1 is around $Port_1$) AND ...
   AND (A_n is around $Prot_n$) Then = $J_i = A^{-t} * \Theta$

Where i is the rule number, n is the number of input variables (or terms) in $Rule_i$, $Port_i$ is the value of variables at A_i (obtained using tf-idf), $\bar{A}$ is the vector of input features, i.e. $\bar{A} = [1, x_1, x_2, ..., x_n]$, and y_i is the resulting output. The normalized output is obtained using the following equation [2]:

2. $$y_i = \frac{y_i}{\sum_{i=1}^{N} y_i}$$

The y_i values should sum up to 1: $\sum_{i=1}^{N} y_i = 1$

The normalized output can be interpreted to find a match with the existing classes. If classes are binary, “1” means the output is a member of the class, and “0” means it is not. If the objective function has more than two classes or multiple inputs and multiple outputs with (n+1)*k members (where k is the number of classes or classifications, and n is the number of terms), then:

3. $$\Theta = \begin{pmatrix} \Theta_{01}^{i} & \cdots & \Theta_{0k}^{i} \\ \vdots & \ddots & \vdots \\ \Theta_{n1}^{i} & \cdots & \Theta_{nk}^{i} \end{pmatrix}$$

The output of the fuzzy rules related to the k^{th} row of this vector is the normalized output for the class:

4. $$y^{-} = [y_1^{-}, y_2^{-}, \ldots, y_n^{-}]$$

In this study, the choice of using Eclass1-MIMO in the classification algorithm is made because of the dynamic adaptability of fuzzy rules to the changes in the input data stream.

Fuzzy rule generation, removal, and updating

Provided that the aging condition is met, the fuzzy rules for assigning the news or tweet A_z to classes or categories C_j are updated in the following steps:

1. Compute the potential terms of news or tweet A_z
2. Update the patterns (list of all existing terms) according to the potential terms of news or tweet A_z
3. Insert A_z as a new pattern (new pattern of the class C_j) if necessary.
4. Remove duplicate patterns if necessary

Comparison of the proposed method with other algorithms

For performance evaluation, the proposed method was compared with the conventional algorithms listed below. This comparison was made in terms of accuracy, misclassification, Kappa statistic, and absolute error.

1. Naïve Bayes algorithm: This is a simple classifier based on the Bayes theory, which has no configurable parameters [35].

2. Bayesian network algorithm: Bayesian networks (BNs) are a family of probabilistic graphic models (GMs) developed by the combination of graph theory, probability theory, and statistics. In this algorithm, each vertex of the graph represents a random variable and the edges between vertices represent the probabilistic dependence between the corresponding random variables [36].
3. Deep learning: Deep learning is a multi-layer feed-forward artificial neural network that is trained by a stochastic gradient descent scheme using back-propagation [37].
4. K-nearest neighbor's algorithm (KNN): In this algorithm, the parameter K is the number of closest training examples or the number of nearest neighbors in the feature space. After receiving K as an input, this puts the K nearest neighbors of an object to the same class. In this algorithm, distance is measured based on a distance criterion like Euclidian distance [38].
5. Learning Vector Quantization (LVQ) neural network: LVQ neural network is an artificial neural network based on local competitive learning. In this network, neurons are called codebooks or prototypes [39].
6. Support Vector Machine (SVM): SVM algorithms are supervised/unsupervised learning models developed for classification and regression analysis [40].

The comparison between the classification methods was performed by the use two of metrics, accuracy and confusion matrix, which represent, respectively, the degree and extent of text classification precision.

Results

The collected database consisted of 10,000 news items and tweets (selected), which, after data cleanup and integration, yielded 1,100 keywords in 9 classes. After applying a pruning technique (in the 10%–30% range) to remove the terms with low tf-idf index, the results provided in Table 1 were obtained. It should be noted that to improve the accuracy and speed of the process in real-time applications, pruning was performed with p<20%.

Table 1: The number of keywords after the application of pruning technique

Database	Without pruning	p<10%	p<20%	p<30%
Number of keywords	8,179	5,600	2,700	**1,100**

Figure 2 shows the accuracy and confusion matrix of the proposed algorithm, named Fuzzy Algorithm for Extraction, Monitoring and Classification of infectious Diseases or FAEMC-ID. While using FAEMC-ID, the highest and lowest precisions or recall ratios were obtained for cholera and plague.

As shown in Figure 2, unlike in other works [1], the precision of the method increases with the sampling volume. This reflects the applicability of the method to large-scale databases and hence in real-world applications.

In Table 2, FAEMC-ID is compared with the conventional algorithms commonly used previous works. This comparison is in terms of accuracy, misclassification, Kappa coefficient, and absolute error. As can be seen, the proposed method exhibits a higher accuracy in the classification of the test data. In addition, the high correlation of data in each class is reflected in the obtained Kappa coefficient.

With an automatic system for extraction of news and comments, one can rapidly build a large database of disease-related events. With the provided visualization process, it is also possible to track the geographical location of the sources of news or comments.

Figure 3 shows the results obtained by monitoring measles-related news in a continuous 183-hour period from 01/03/2019 (01:00 am) to 08/03/2019 (12:00 pm), which are related to 2,870 tweets from 2,556 users. The number of tweets posted from each region range from 1 to 47, with the highest number (47 tweets) from Canada. The origins of most measles-related news were in the Americas and Europe, and they were mostly from the United States and Canada.

This is consistent with the map illustrated in Figure 4, which was obtained from the United States Centers for Disease Control and Prevention (CDC).

Figure 5 displays the map of Ebola-related news and tweets obtained using the proposed method, and Figure 6 shows the map of Ebola epidemics according to the WHO. As can be seen, there is a high degree of consistency between these maps. The map of HIV/AIDS-related news and tweets for the study period is shown in Figure 7.

Discussion

In this study, we developed a new method based on the evolving fuzzy algorithm of TSK type for the extraction, monitoring, storage and visualization of news and tweets about various infectious diseases. To implement the method, more than 10,000 tweets and news were cleaned, integrated and classified with the help of the Eclass1-MIMO method, then visualized on the world map in real-time.

In recent years, many researchers have worked on classification, clustering, sentiment analysis, opinion mining and development of recommenders based on social data, but most of these works have focused either on news websites or Twitter [1], [41], [42], [43].

The findings of the present study are consistent with those of Angelov PP and Zhou X [26] and Bhattacharyya et al. [25], who reported the high efficacy of evolving fuzzy algorithms in real-time applications in terms of ensuring satisfactory precision, speed, and flexibility.

In the study by Iglesias et al., an evolving fuzzy algorithm with the Eclass1-MIMO method was used to classify the public news into 6 categories of science, health, technology, sports, arts, and commerce [1]. But unlike this

Table View Plot View

accuracy: 88.41%

	true Cholera	true ebola	true HIV	true influenza	true malaria	true measle	true plague	true poliomyel	true smallpox	class precisi...
pred. Cholera	145	0	0	8	6	0	0	20	0	81.00%
pred. ebola	0	269	10	0	0	0	11	5	0	91.18%
pred. HIV	0	1	400	5	0	3	0	0	0	97.79%
pred. influenza	0	1	0	60	8	0	12	0	0	74.07%
pred. malaria	0	0	6	0	130	8	0	16	0	81.25%
pred. measle	0	0	0	0	0	225	10	0	2	94.93%
pred. plague	0	0	0	0	4	4	120	0	9	89.55%
pred. poliom...	0	0	0	0	6	0	8	110	8	83.33%
pred. smallp...	0	0	0	0	0	10	6	10	120	82.19%
class recall	100.00%	99.00%	96.15%	82.19%	88.41%	90.00%	71.80%	68.32%	81.60%	

Figure 2: Accuracy and confusion matrix of the proposed algorithm

Table 2: Performance of the proposed algorithms in comparison with other algorithms

Algorithm name	Recall %	Misclassification %	Kappa coefficient	Absolute error
FAEMC-ID	88.41	11.59	0.873	**0.127**
Bayesian network	80.70	19.30	0.750	**0.250**
Naive Bayes	56.23	43.77	0.402	**0.598**
Deep learning	70.20	29.80	0.640	**0.360**
KNN (k=9)	29.60	70.40	0.117	**0.883**
LVQ	75.70	24.50	0.690	**0.310**
SVM	30.80	69.20	0.139	**0.861**

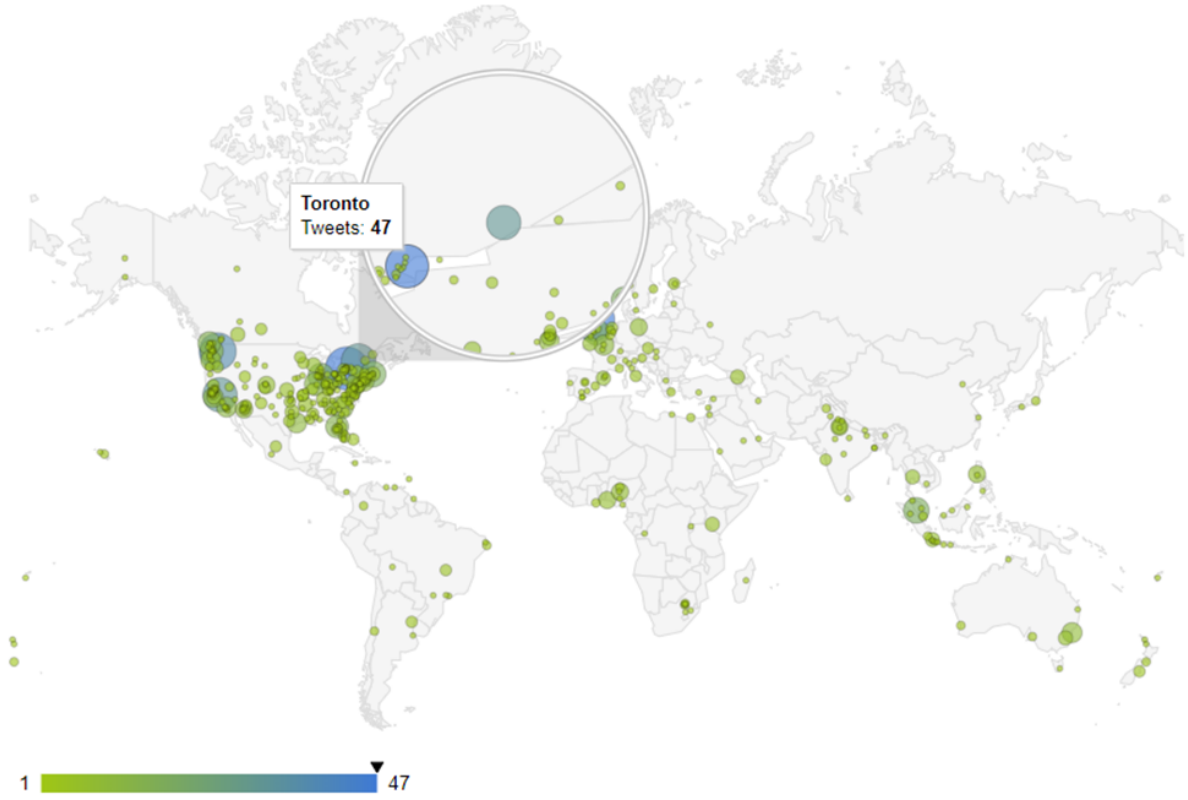

Figure 3: Monitoring of geographical distribution of the tweets about measles from 01:00 am 01/03/2019 to 12:00 pm 08/03/2019

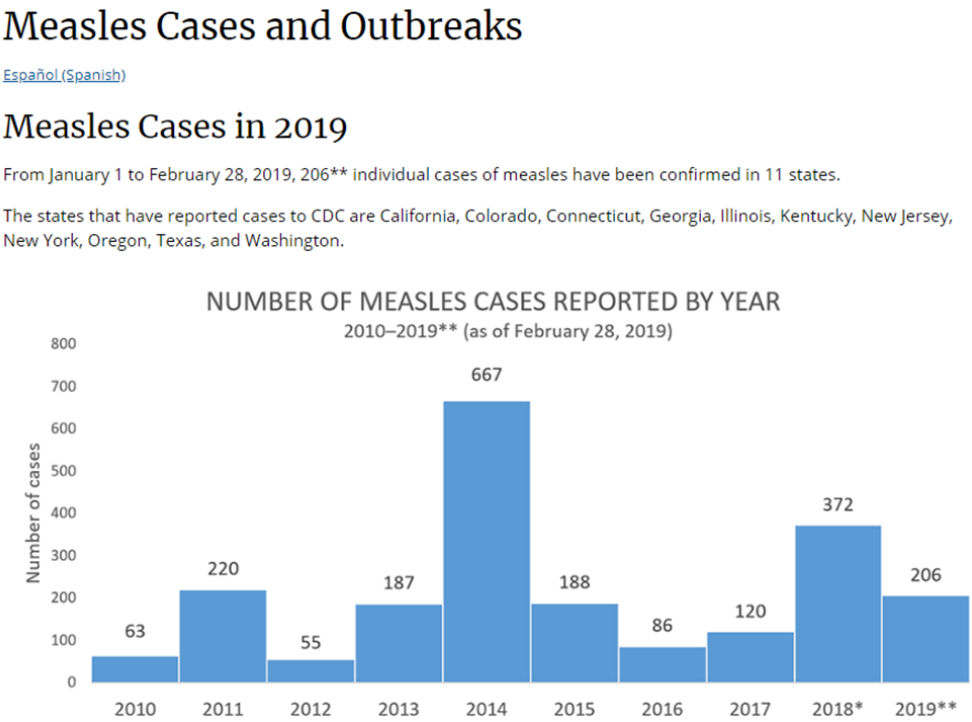

Figure 4: CDC report about the incident of measles in the United States [51]

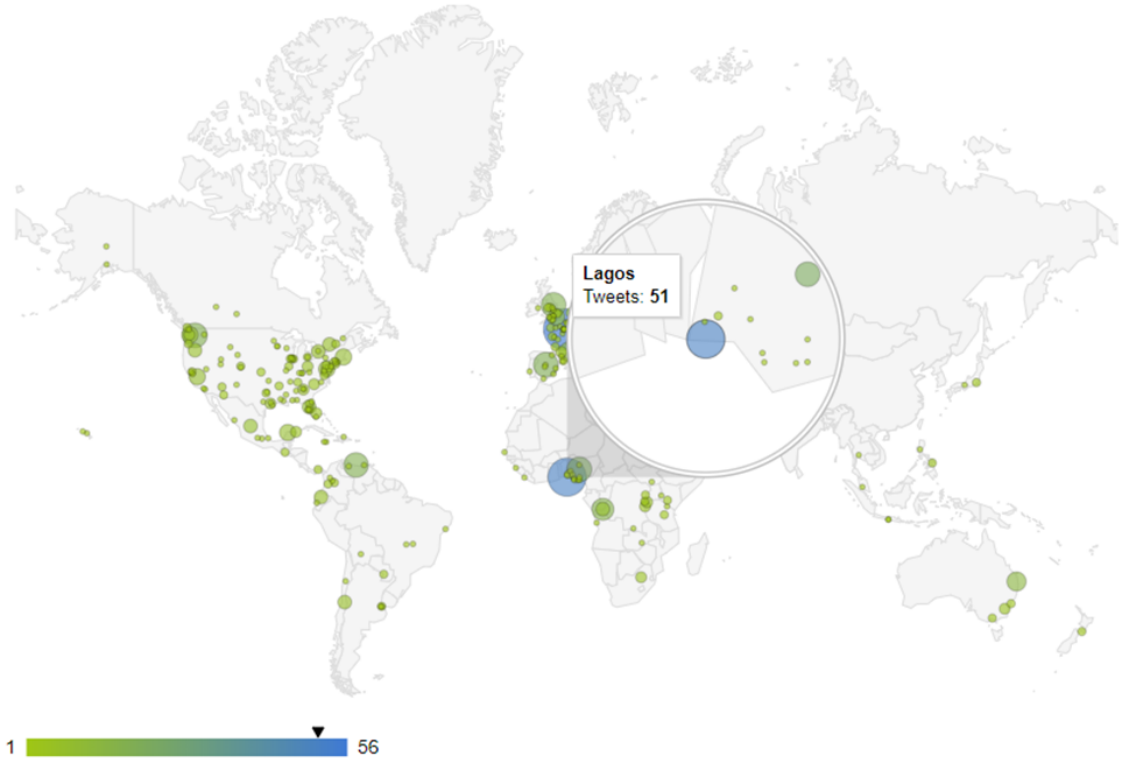

Figure 5: Monitoring of geographical distribution of the tweets about Ebola from 01:00 am 01/03/2019 to 12:00 pm 08/03/2019

model, in the proposed method, increasing the data size not only does not reduce the accuracy but actually improves it. Another advantage of the proposed method over similar works [44], [45], [46], [47] is the ability to visualize the results for improved monitoring and tracking of epidemics.

Also, the geographic origins of tweets posted about measles and Ebola were found to be consistent with official CDC and WHO reports about their incidence during the studied period. This reflects the efficacy of the proposed method in monitoring and tracking the targeted diseases.

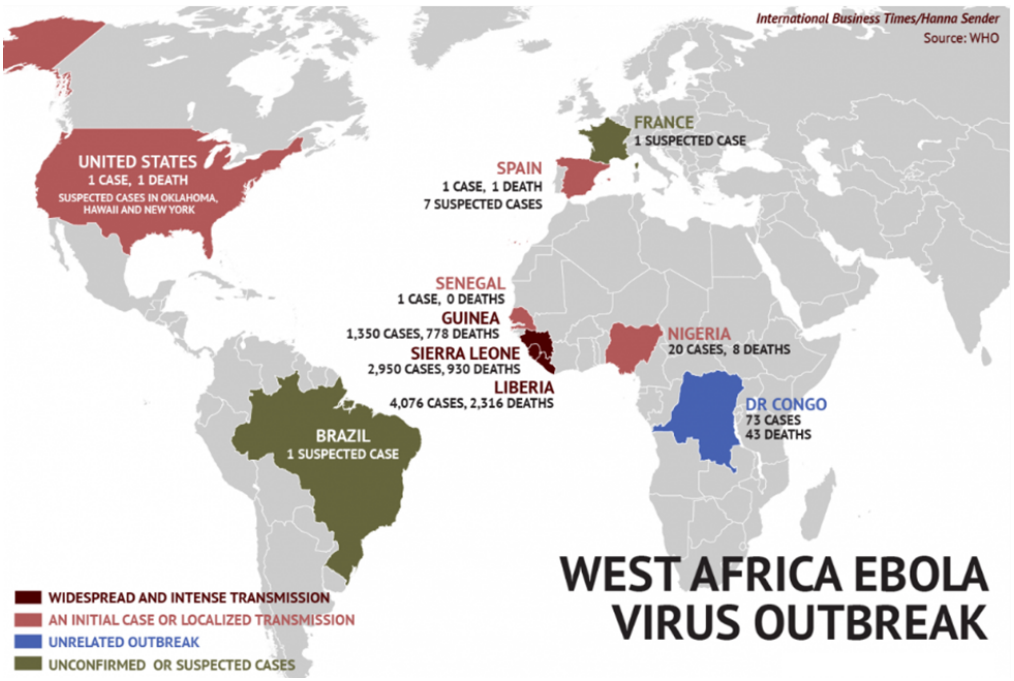

Figure 6: WHO report about the incident of Ebola [52]

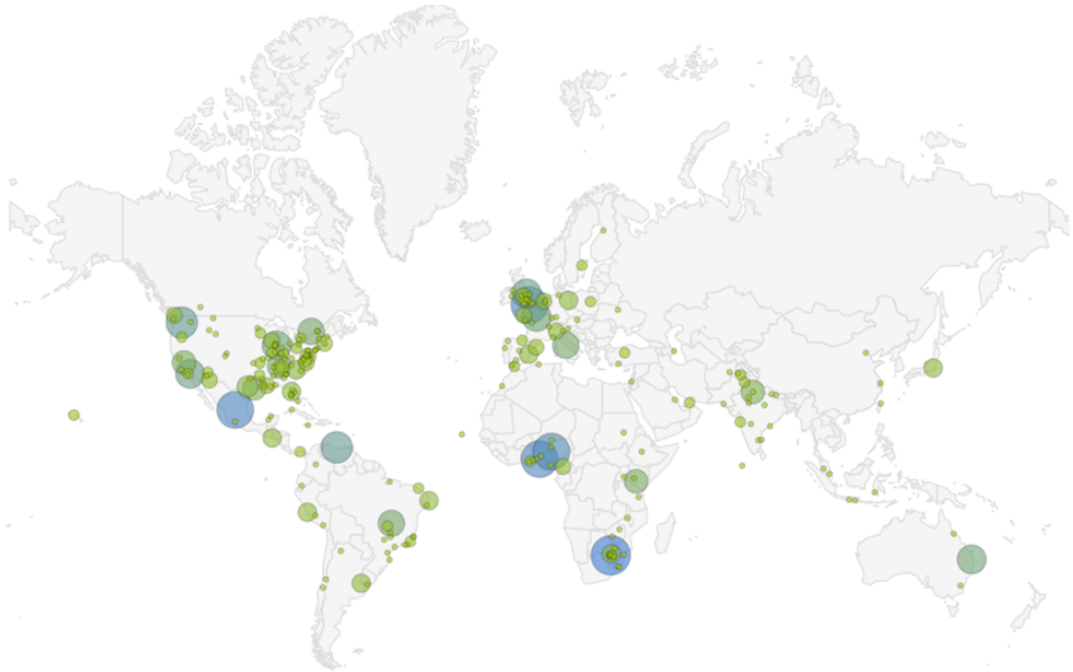

Figure 7: Monitoring of geographical distribution of the tweets about HIV from 01:00 am 01/03/2019 to 12:00 pm 08/03/2019

The evolving fuzzy method has also been used by Del Jesus [48] to enhance low-grade classification algorithms, by Lughofer [49] to solve the problems of online multi-class classification, and Lughofer [50] for online incremental feature dimension reduction. Our findings about the use of evolving fuzzy method are in agreement with the results of these studies in terms of high accuracy, high correlation of data in each class (kappa coefficient), and efficacy in online multiclass data analysis.

Study limitation

A limitation of the suggested method is that it cannot be used to monitor and track infectious diseases in areas with poor or no access to social networks such as Twitter and Facebook, and this includes poor countries, where morbidity and mortality due to infectious diseases are noticeably higher.

Conclusions

This paper presented a method for extraction, monitoring, storage, and visualization of data related to certain infectious diseases through news mining and tweet crawling. The proposed framework consists of four phases, including data collection with a code written in the R-programming language, text cleanup, classification with the evolving fuzzy model Eclass1-MIMO, and visualization. The fuzzy classification component was developed based on fuzzy TSK rules and evolving fuzzy model, and hence is able to update its vocabulary and remain efficient and accurate upon encountering new terms. Moreover, unlike previous methods, the proposed method exhibits satisfactory flexibility regarding the size of input data and can handle large datasets without a decline in classification accuracy. Other notable features of this method include the simultaneous extraction of news from tweets and websites, the real-time classification capability, data storage in one database, and visualization of data in real-time. The analysis of this proposed method in comparison with other algorithms in the literature showed its high accuracy (88.41%) and the high correlation of data within each class. The proposed algorithm can also be used in the development of more effective monitoring and tracking systems for other human and even animal health hazards.

Notes

Acknowledgments

We would like to express our gratitude to Dr. Antonio Iglesias at the University of Madrid for the helpful comments, the instructors of the online course “Machine Learning for Data Science and Analytics” provided by Columbia University for giving us better insight into the area of data and text mining, and also the members of the Iran Data Mining Group, who patiently answered our questions.

Competing interests

The authors declare that they have no competing interests.

References

1. Iglesias JA, Tiemblo A, Ledezma A, Sanchis A. Web news mining in an evolving framework. Information Fusion. 2016;28:90-8. DOI: 10.1016/j.inffus.2015.07.004
2. Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems. 2015;89:14-46. DOI: 10.1016/j.knosys.2015.06.015
3. Guellil I, Boukhalfa K. Social big data mining: A survey focused on opinion mining and sentiments analysis. In: 12th International Symposium on Programming and Systems (ISPS); 2015 Apr 28-30; Algiers, Algeria. IEEE; 2015. DOI: 10.1109/ISPS.2015.7244976
4. Mukkamala RR, Hussain A, Vatrapu R. Fuzzy-set based sentiment analysis of big social data. In: 18th International Enterprise Distributed Object Computing Conference (EDOC); 2014 1-5 Sept. 2014; Ulm, Germany. IEEE; 2014. DOI: 10.1109/EDOC.2014.19
5. Evans DK, Klavans JL, McKeown KR. Columbia newsblaster: Multilingual news summarization on the web. In: HLT-NAACL Demonstrations '04; 2004 May 2-7; Boston, USA. Stroudsburg, PA: Association for Computational Linguistics; 2004. DOI: 10.3115/1614025.1614026
6. Tan AH. Text mining: The state of the art and the challenges. In: Zhong N, Zhou L, editors. Methodologies for Knowledge Discovery and Data Mining. Third Pacific-Asia Conference PAKDD'99; 1999 Apr 26-28; Beijing, China. Springer; 1999. p. 65-70.
7. McCaig D, Bhatia S, Elliott MT, Walasek L, Meyer C. Text-mining as a methodology to assess eating disorder-relevant factors: Comparing mentions of fitness tracking technology across online communities. Int J Eat Disord. 2018 07;51(7):647-55. DOI: 10.1002/eat.22882
8. Russell MA. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. Sebastopol, CA: O'Reilly Media Inc; 2013.
9. Tang J, Chang Y, Liu H. Mining social media with social theories: a survey. ACM SIGKDD Explorations Newsletter. 2014;15(2):20-9. DOI: 10.1145/2641190.2641195
10. Ngoc PT, Yoo M. The lexicon-based sentiment analysis for fan page ranking in Facebook. In: International Conference on Information Networking 2014 (ICOIN 2014); 2014 Feb 10-12; Phuket, Thailand.
11. Ueda M, Mori K, Matsubayashi T, Sawada Y. Tweeting celebrity suicides: Users' reaction to prominent suicide deaths on Twitter and subsequent increases in actual suicides. Soc Sci Med. 2017 Sep;189:158-166. DOI: 10.1016/j.socscimed.2017.06.032
12. O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen HJII. Detecting suicidality on Twitter. Int Intervent. 2015;2(2):183-8. DOI: 10.1016/j.invent.2015.03.005
13. Colombo GB, Burnap P, Hodorog A, Scourfield J. Analysing the connectivity and communication of suicidal users on twitter. Comput Commun. 2016 Jan;73(Pt B):291-300. DOI: 10.1016/j.comcom.2015.07.018
14. Masys DR. Linking microarray data to the literature. Nat Genet. 2001 May;28(1):9-10. DOI: 10.1038/88324
15. Srivastava AN, Sahami M, editors. Text mining: Classification, clustering, and applications. New York: Chapman and Hall/CRC; 2009. DOI: 10.1201/9781420059458
16. Coussement K, Van den Poel D. Integrating the voice of customers through call center emails into a decision support system for churn prediction. Inform Manag. 2008;45(3):164-74. DOI: 10.1016/j.im.2008.01.005
17. Nassirtoussi AK, Aghabozorgi S, Wah TY, Ngo DCL. Text mining for market prediction: A systematic review. Expert Syst Applications. 2014;41(16):7653-70. DOI: 10.1016/j.eswa.2014.06.009
18. Gálvez RH, Gravano A. Assessing the usefulness of online message board mining in automatic stock prediction systems. J Comput Sci. 2017;19:43-56. DOI: 10.1016/j.jocs.2017.01.001
19. Valitutti A, Strapparava C, Stock O. Developing affective lexical resources. Psych J. 2004;2(1):61-83.

20. Zhang Z, Li X, Chen Y. Deciphering word-of-mouth in social media: Text-based metrics of consumer reviews. ACM Transactions on Management Information Systems (TMIS). 2012;3(1):5. DOI: 10.1145/2151163.2151168

21. Irfan R, King CK, Grages D, Ewen S, Khan SU, Madani SA. A survey on text mining in social networks. Knowl Eng Rev. 2015;30(2):157-70. DOI: 10.1017/S0269888914000277

22. Mani S. UGT1A1 polymorphism predicts irinotecan toxicity: evolving proof. AAPS PharmSci. 2001;3(3):2.

23. Litvak M, Last M, Friedman M. A new approach to improving multilingual summarization using a genetic algorithm. In: Proceedings of the 48th annual meeting of the association for computational linguistics; 2010 Jul 11-16; Uppsala, Sweden. Stroudsburg, PA: Association for Computational Linguistic; 2010.

24. Riza L, Bergmeir C, Herrera F, Benítez J. frbs: Fuzzy Rule-Based Systems for Classification and Regression in R. J Stat Softw. 2015;65(6):1-30. DOI: 10.18637/jss.v065.i06

25. Bhattacharyya S, Basu D, Konar A, Tibarewala D. Interval type-2 fuzzy logic based multiclass ANFIS algorithm for real-time EEG based movement control of a robot arm. Rob Auton Syst. 2015;68:104-15. DOI: 10.1016/j.robot.2015.01.007

26. Angelov PP, Zhou X. Evolving fuzzy-rule-based classifiers from data streams. IEEE Trans Fuzzy Syst. 2008;16(6):1462-75. DOI: 10.1109/TFUZZ.2008.925904

27. Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern. 1985(1):116-32. DOI: 10.1109/TSMC.1985.6313399

28. Ishibuchi H, Yamamoto T, Nakashima T. Hybridization of fuzzy GBML approaches for pattern classification problems. IEEE Trans Syst Man Cybern B Cybern. 2005;35(2):359-65. DOI: 10.1109/TSMCB.2004.842257

29. Boyacioglu MA, Avci D. An adaptive network-based fuzzy inference system (ANFIS) for the prediction of stock market return: the case of the Istanbul stock exchange. Expert Syst Appl. 2010;37(12):7908-12. DOI: 10.1016/j.eswa.2010.04.045

30. Bai Y, Zhuang H, Roth ZS. Fuzzy logic control to suppress noises and coupling effects in a laser tracking system. IEEE Trans Control Syst Technol. 2005;13(1):113-21. DOI: 10.1109/TCST.2004.833653

31. Basari ASH, Hussin B, Ananta IGP, Zeniarja J. Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Eng. 2013;53:453-62. DOI: 10.1016/j.proeng.2013.02.059

32. Gupta V, Lehal GS. A survey of text summarization extractive techniques. J Emerg Technol Innov Res Web Intell. 2010;2(3):258-68. DOI: 10.4304/jetwi.2.3.258-268

33. Broder AZ, Glassman SC, Manasse MS, Zweig G. Syntactic clustering of the web. Computer Networks and ISDN Systems. 1997;29(8):1157-66. DOI: 10.1016/S0169-7552(97)00031-7

34. Cavnar WB, Trenkle JM. N-gram-based text categorization. Ann arbor mi. 1994;48113(2):161-75.

35. Angelov P, Filev D. Simpl_eTS: A simplified method for learning evolving Takagi-Sugeno fuzzy models. In: The 14th IEEE International Conference on Fuzzy Systems FUZZ'05; 2005 May 25; Reno, USA. IEEE; 2015. DOI: 10.1109/FUZZY.2005.1452543

36. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Mach Learn. 1997;29(2-3):131-63. DOI: 10.1023/A:1007465528199

37. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May;521(7553):436-44. DOI: 10.1038/nature14539

38. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21-7. DOI: 10.1109/TIT.1967.1053964

39. Kohonen T. Learning vector quantization. In: Kohonen T, editor. Self-Organizing Maps. Berlin, Heidelberg: Springer; 1995. (SSINFL; Vol. 30). p. 175-89. DOI: 10.1007/978-3-642-97610-0_6

40. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst. 1998;13(4):18-28. DOI: 10.1109/5254.708428

41. Amato F, Moscato V, Picariello A, Piccialli F. SOS: A multimedia recommender System for Online Social networks. Future Gener Comput Syst. 2019;93: 914-923. DOI: 10.1016/j.future.2017.04.028

42. Li J, Li X, Zhu B. User opinion classification in social media: A global consistency maximization approach. Inform Manag. 2016;53(8):987-96. DOI: 10.1016/j.im.2016.06.004

43. Mathioudakis M, Koudas N. Twittermonitor: trend detection over the twitter stream. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data; 2010 Jun 6-10; Indianapolis, USA. New York: ACM; 2010. DOI: 10.1145/1807167.1807306

44. Al-Surimi K, Khalifa M, Bahkali S, El-Metwally A, Househ M. The potential of social media and internet-based data in preventing and fighting infectious diseases: from internet to twitter. Basel: Springer; 2016. p. 131-9. DOI: 10.1007/5584_2016_132

45. Ku LW, Chen HH. Mining opinions from the Web: Beyond relevance retrieval. J Am Soc Information Sci Technol. 2007;58(12):1838-50. DOI: 10.1002/asi.20630

46. Krishnalal G, Rengarajan SB, Srinivasagan K. A new text mining approach based on HMM-SVM for web news classification. Int J Comput Appl. 2010;1(19):98-104. DOI: 10.5120/395-589

47. Maghdid HS. Web News Mining Using New Features: A Comparative Study. IEEE Access. 2019;7:5626-41. DOI: 10.1109/ACCESS.2018.2890088

48. Del Jesus MJ, Hoffmann F, Navascués LJ, Sánchez L. Induction of fuzzy-rule-based classifiers with evolutionary boosting algorithms. IEEE Trans Fuzzy Syst. 2004;12(3):296-308. DOI: 10.1109/TFUZZ.2004.825972

49. Lughofer E, Buchtala O. Reliable all-pairs evolving fuzzy classifiers. IEEE Trans Fuzzy Syst. 2013;21(4):625-41. DOI: 10.1109/TFUZZ.2012.2226892

50. Lughofer E. On-line incremental feature weighting in evolving fuzzy classifiers. Fuzzy Sets Syst. 2011;163(1):1-23. DOI: 10.1016/j.fss.2010.08.012

51. Centers for Disease Control and Prevention. Measles Cases and Outbreaks 2019. [updated 2019 Jan 16; cited 2019 Mar 9]. Available from: https://www.cdc.gov/measles/cases-outbreaks.html

52. WHO. Ebola outbreak 2014. [updated 2015 Jul 23; cited 2019 Mar 8]. Available from: https://www.who.int/features/ebola/storymap/en/

Corresponding author:
Vahid Rahmanian
Zoonoses Research Center, Jahrom University of Medical Sciences, Jahrom, Iran, Phone: +98 9175985204
vahid.rahmani1392@gmail.com

Please cite as
Jahanbin K, Rahmanian F, Rahmanian V, Jahromi AS. Application of Twitter and web news mining in infectious disease surveillance systems and prospects for public health. GMS Hyg Infect Control. 2019;14:Doc19.
DOI: 10.3205/dgkh000334, URN: urn:nbn:de:0183-dgkh0003343

This article is freely available from
https://www.egms.de/en/journals/dgkh/2019-14/dgkh000334.shtml

Published: *2019-12-02*