Journal of Computer Science

Original Research Paper

Arabic Online Event-Based System for Monitoring and Extracting Infectious Disease-Related Information

Meshrif Alruily

Department of Computer Science, Jouf University, Sakaka, Saudi Arabia

Article history
Received: 29-11-2018
Revised: 10-01-2019
Accepted: 21-01-2019

Email: mfalruily@ju.edu.sa

Abstract: With the revolution of the internet, online data play a significant role in identifying disease outbreaks. This has led researchers, governments and organizations to pay close attention to such data in order to employ and exploit them in developing event-based systems. This research studies the infectious disease outbreaks domain in the Arabic language. In this paper, the Arabic Surveillance Infectious Disease Outbreak System (ASIDOS), which is able to extract infectious disease-related information from unstructured data published by newswires is developed. For identifying the features extraction and performing the data analysis, the word association methodology was adopted. The proposed system is validated through experiments using a corpus collated from different sources. Precision, recall and F-measure are used to evaluate the performance of the proposed information extraction method. The overall results achieved are: precision 94%, recall 74% and F-measure 83%.

Keywords: Arabic Health Domain, Event-Based System, Infectious Disease, Information Extraction, Pattern Recognition, Text Mining

Introduction

During the past few years, the spread of many different pandemic diseases has increased worldwide; for example, the disease caused by the Ebola virus was first reported by the World Health Organization (WHO) in Guinea in 2014, which then spread rapidly to many West African countries causing hundreds of deaths (Guinea 346, Liberia 181, Nigeria 1, Sierra Leone 37) (Washington, 2015; PAHO/WHO, 2014). It was also transmitted to other countries outside of the African continent, including Italy, the United Kingdom, Spain and United States of America. The latest information about Ebola can be accessed via the following link: http://www.who.int/csr/don/archive/disease/ebola/en/.

In addition to the outbreak of Ebola in Africa, respiratory syndrome coronavirus (SARSCoV) was identified in Asia (2002-2003), an outbreak of the pandemic disease H1N1 influenza virus occurred worldwide (2009) and the Middle East Respiratory Syndrome (MERS) was found in Saudi Arabia (2012-to date) (Velasco *et al*., 2014; CDC, 1996). Therefore, the threat of infectious disease outbreaks to public health has prompted countries and organizations to develop several early warning surveillance systems (Choi *et al*., 2016). However, the traditional surveillance systems or indicator-based surveillance systems require public health networks (Sentinel networks) to collect predefined structured data about diseases on a routine basis from indicator sources, such as over-the-counter and emergency department visits (Christaki, 2015; Collier and Doan, 2012). The World Health Organization (WHO) defines this type of system thus, “A passive surveillance system relies on the cooperation of health-care providers-laboratories, hospitals, health facilities and private practitioners-to report the occurrence of a vaccine-preventable disease to a higher administrative level” (WHO, 2014) Therefore, in the case of using passive surveillance systems, regular submission of monthly, weekly or daily reports of disease data by all health facilities is required. Although implementing this type of system has some advantages, such as its ability to cover all parts of a country and its statistical power, it takes a couple of weeks for disease patterns to be detected and the results regarding possible outbreaks to be disseminated; furthermore, not all countries have the required infrastructure to implement this system (Velasco *et al*., 2014; Collier and Doan, 2012; WHO, 2014; Ramalingam, 2016). On the other hand, as a result of the technological revolution of the internet, another type of surveillance system has emerged. This type is known as the event-based surveillance system. The WHO defines it thus, "Event-based surveillance is the organized and rapid capture of information about

events that are a potential risk to public health" (WHO, 2008). Generally, event-based surveillance systems can be described as real-time monitoring of diseases 24/7 through gathering information from informal sources, such as online news. According to Blench *et al*. (2009), the WHO's investigations into the majority of disease outbreaks are obtained through diverse informal online sources.

The remainder of the paper is organized as follows. In Section 2, a background to the topic and a review of related work are presented. Section 3 provides the built corpus that contains reports on various infectious disease events collected from different online sources. Data analysis is provided in Section 4. In Section 5, an overview of the Arabic Surveillance Infectious Disease Outbreak System (ASIDOS) architecture, together with an explanation of the methodology used for event information extraction is provided. Section 6 presents the experiments and the performance evaluation. Finally, the conclusion of the work is presented in Section 7.

Related Work

There are two types of surveillance systems: indicator-based surveillance systems (syndromic surveillance) and event-based surveillance systems (Agheneza, 2011). This section will examine event-based surveillance systems, which use informal data related to disease outbreaks collected from newswires to detect and extract infectious disease outbreak related-information, such as disease type, location name, date and number of victims in order to issue early warning. Greater concentration will be placed on event-based surveillance systems that support Arabic.

Indicator-based Surveillance Systems

This type of system relies on structured data collected from various official sources, such as emergency departments, telephone calls and over-the-counter drug sales to detect and track increases in disease incidence rates (Christaki, 2015). Many detection approaches are utilized for achieving this task. These approaches are classified into three types: temporal, spatial and spatiotemporal surveillance techniques (Tsui *et al*., 2010). Further information on these types of systems was examined in (Tsui *et al*., 2010; 2008), recent reviews of these systems can be found in (Azzedin *et al*., 2014; Abat *et al*., 2016). Moreover, the national communicable diseases surveillance systems developed between 2000 and 2016 in developed countries were reviewed in (Bagherian *et al*., 2017).

Event-Based Surveillance Systems

Many event/web-based surveillance systems have been developed for processing news reports relating to epidemic diseases from informal sources (online newspapers and social media platforms) (Christaki, 2015).

Table 1: The identified event-based systems

System	Year	Country	Homepage
ProMED-mail (Hugh-Jones, 2001) (Woodall, 1997)	1994	USA	www.promedmail.org
GPHIN (Mawudeku and Blench, 2006)	1997	Canada	https://gphin.canada.ca/cepr/listarticles.jsp?language=en_CA
EWRS (Guglielmetti *et al*., 2006)	1998	Europe	https://ewrs.ecdc.europa.eu
EpiSimS (Valle, 2000)	2000	USA	http://www.lanl.gov/projects/mathematical- computational-epidemiology/agent-based-modeling.php
GOARN (Heymann and Rodier, 2001)	2000	USA	https://extranet.who.int/goarn/
MiTAP (Damianos *et al*., 2002)	2001	USA	http://mitap.sdsu.edu
Proteus-BIO (Grishman *et al*., 2002)	2002	USA	Not available
Argus (Wilson, 2007)	2004	USA	www.biodefense.georgetown.edu
MedISys and PULS (EMM, 2002)	2004	Europe	http://medisys.newsbrief.eu
BioCaster (Collier *et al*., 2008)	2006	Japan	http://born.nii.ac.jp/ OR http://www.biocaster.org/
EpiSPIDER (Tolentino *et al*., 2007)	2006	USA	www.epispider.org
GODSN (Sharib *et al*., 2006)	2006	USA	Not available
HealthMap (Brownstein and Freifeld, 2007)	2006	USA	www.healthmap.org
InSTEDD (Taha and Tada, 2008)	2006	USA	http://instedd.org
Google Flu Trends (Jeremy *et al*., 2009)	2008	USA	www.google.org/flutrends
Influenzanet (Koppeschaar, 2011)	2008	Europe	www.influenzanet.eu
Animal disease-related event recognition system (Svitlana and William, 2010)	2010	USA	Not available
Automatic online news monitoring and classification (Zhang *et al*., 2009)	2010	USA	Not available
GET WELL (Hulth and Rydevik, 2011)	2010	Sweden	www.smittskyddsinstitutet.se
CIDARS (Yang *et al*., 2011)	2011	China	Not available
EpiCore (Lorthe *et al*., 2018)	2013	USA	https://epicore.org
Alshowaib's system (Alshowaib, 2014)	2014	KSA	Not available
Healthtweets (Dredze *et al*., 2014)	2014	USA	www.healthtweets.org
DESRM (Nguyen, 2015)	2015	Vietnam	Not available
Flutrack (Talvis *et al*., 2014)	2015	Greece	www.flutrack.org
Online Diagnostic System (Okokpujie *et al*., 2017)	2017	Nigeria	Not available
ARGO (Yang *et al*., 2017)	2017	5 countries	Not available
PADI-web (Goel *et al*., 2018)	2018	France	http://epia.clermont.inra.fr

The most comprehensive review for systems proposed between 1994 and 2006 was conducted by Velasco *et al*. (2014), who reviewed studies of infectious disease surveillance publications between 1990 and 2011 in detail, which yielded 13 event-based systems and used 15 review items including system name, year started, purpose, country, system category, language, disease type and public access. In this review, 28 systems were identified, which can be seen in Table 1. Some are multilingual and are able to handle Arabic news, such as Argus, GPHIN, HealthMap, MedISys and PULS and ProMED-mail utilizing translation tools to translate Arabic news reports into English. The following shows further investigation of these systems:

- A Global Detection and Tracking System for Biological Events (Argus) is based on a multilingual analytic team to detect and track global biological events. It uses a taxonomy of nearly 200 indicators for detecting 130 infectious disease outbreaks in 175 countries (Wilson, 2007). Data generated from WHO and ProMed is used to monitor the biological events (Chen *et al*., 2010).
- (GPHIN) is based on translation tools to support 8 languages including Arabic. It was developed by Health Canada in collaboration with the World Health Organization (WHO). For monitoring disease outbreaks, the GPHIN system compiles public health data from various online sources. Moreover, it is able to monitor other events, such as plant and animal diseases, contaminated food and water and chemical incidents. GPHIN is used by the Food and Agriculture Organization of the United Nations (FAO), WHO and Disease Control and Prevention (CDC, 1996). Therefore, the use of GPHIN by such official organizations indicates its power and accuracy in identifying global public health events (Agheneza, 2011). However, information generated by GPHIN is only presented in English and French languages and is not free (Mawudeku and Blench, 2006; WHO, 2015).
- The HealthMap consists of five parts: data gathering from different sources (newswires, Really Simple Syndication (RSS) feeds, ProMED-mail and WHO), classification, database, web backend and web frontend (Freifeld et al., 2008). HealthMap is similar to GPHIN, as it is a multilingual web-based system, but is free and publicly available, unlike, information produced by HealthMap is presented in seven languages: Arabic, French, Chinese, Spanish, Portuguese, English and Russian. With regard to non-English news reports, HealthMap uses translation to process them
- The Medical Information System (MedISys) and Pattern-based Understanding and Learning System (PULS) is also a multilingual early-warning surveillance system. It was developed at the Joint Research Centre of the European Commission (JRC) and is part of the Europe Media Monitor (EMM, 2002) software package. MedISys is based on predefined keywords for collecting reports from many online sources in different languages. However, information extraction is only performed on reports in English via the Plus system, i.e. it makes use of a translation tool to handle text written in other languages. Three types of access levels are provided by MedISys: free public access, restricted access outside the European Commission (EC) for public health professionals and full access inside the EC
- ProMED-mail (http://www.promedmail.org) is a program of the International Society for Infectious Diseases (ISID) (http://www.isid.org), developed in 1994 by Hugh-Jones (2002) to monitor emerging human, plant and animal disease outbreaks (Hugh-Jones, 2002; Woodall, 1997). It is a multilingual system with free public access and no subscription fees are required. In addition, it relies on public health reports obtained from its subscribers and ordinary users through submit ***Info from*** (http://www.promedmail.org/submitinfo). The information presented by the system is firstly examined by experts of ProMED-mail prior to publication (Woodall, 1997; ProMED, 2010). Therefore, many surveillance systems depend on health warning reports produced by ProMED-mail, such as Argus, BioCaster and HealthMap

As can be seen, most systems developed in the USA and Europe to process data are written in their own native language and are then enhanced to serve other languages by utilizing a translator engine, which helps monitor the spread of pandemic diseases worldwide.

On the other hand, few systems were found in the literature that are able to directly process Arabic health data, i.e., without using translation engines. For example, in (Samy *et al*., 2012) two approaches were proposed for recognizing and extracting medical terms from the Arabic medical dataset. The first approach is based on a gazetteer that contains 3473 Arabic medical terms translated from English medical terms resources (SNOMED and UMLS). The second approach is based on 410 Arabic terms that are equivalents of Latin prefixes and suffixes commonly used in the medical and health domain.

Furthermore, a named entity recognition system, NAMERAMA, has been proposed to identify cancer disease related-information, such as disease names, symptoms, treatment and diagnosis methods from Arabic texts in the medical domain (Alanazi, 2017). The system relies on the Bayesian Belief Networks (BBN) algorithm. However, both systems are not event-based surveillance systems and have not been developed for processing data related to infectious disease outbreaks.

Finally, a special type of system has been developed for tracking specific epidemic diseases, based on online social networks, such as Twitter, Facebook and Instagram. For

example; the flutrack system (http://flutrack.org) was developed for tracking the spread of influenza epidemics based on data published by Twitter users. In addition, Tracking Flu Infections on Twitter (Lamb *et al*., 2013), HealthTweets.org website, a platform using Twitter for public health surveillance (Dredze *et al*., 2014) and the ARGO system for monitoring dengue fever epidemics using internet-based sources (Yang *et al*., 2017) all belong to this type of system. Further information about systems that rely on online social networks can be found in (Al-garadi *et al*., 2016).

Datasets

To the author's knowledge, there is currently no available dataset for infectious disease outbreaks. As is well known, availability of a suitable corpus is the base of text mining research. In order to identify and understand the language's behaviour used in the health domain to describe events of infectious disease outbreaks a specialized dataset must be built. Therefore, five corpora containing only Arabic data on epidemic diseases was manually compiled from different sources to conduct this research and one of the contributions of this study was to create such a linguistic resource (e.g., corpus). These corpora contain many news reports written in Arabic that describe various types of events of infectious diseases. The five corpora contain a total of 317 files that comprise 76, 230 tokens. Each corpus represents different events related to a specific disease from the following list:

- Dengue حمى الضنك
- Ebola إيبولا
- H1N1 influenza virus أنفلونزا الخنازير
- SARSCoV سارس
- Middle East Respiratory Syndrome (MERS) كورونا

Data Preparation

The following disease names: Dengue (حمى الضنك), Ebola (إيبولا), H1N1 influenza virus (أنفلونزا الخنازير), SARSCoV (سارس), Middle East Respiratory Syndrome (MERS) (كورونا) are searched in each corpus to find their occurrences numbers and concordance as can be seen in Table 2.

In addition, n-gram is used in creating five independent sub-datasets by extracting sequences of ten words from both sides of the target word, i.e., the epidemic disease names mentioned in Table 2 from the five original datasets. It was found that these disease names can be written in different forms as in Table 3. In addition, it was found that الخنازير / H1N1 disease name is sometimes written as it is pronounced in the English language " اتش 1 إن 1". However, the sparseness problem in the disease names, such as typographic and spelling variants, can be reduced by performing normalization and stemming.

Table 2: The frequency distribution of the disease names in the five corpora

Arabic disease name	English name	Occurrences
إيبولا	Ebola	357
كورونا	MERS	256
سارس	SARSCoV	227
حمى الضنك	Dengue	214
أنفلونزا الخنازير	H1N1	161

Table 3: Disease names and their other forms

Disease	Arabic different forms
Ebola	أيبولا، الايبولا، ايبولا
MERS	الكورونا, كورونا
SARSCoV سارس	السارس, سارس
Dengue	الضنك ، بالضنك
H1N1	اتش1 إن1، انفلونزا الخنازير

Table 4: The epidemic diseases indicator words and their inflected forms

Indicators words	Inflected forms
Virus فيروس	بفايروس, بفيروس، فايروس
Fever حمى	بحمى
Flu انفلونزا	بأنفلونزا، بانفلونزا، أنفلونزا, انفلونزا,بإنفلونزا
Epidemic وباء	الوباء، وباء
Disease مرض	بمرض

Table 5: The victims indicators words and their inflected forms

Indicators words	Inflected forms
Case حالة	حالات، حالتين
Infected إصابة	إصابات، اصابات، مصابين، اصابتين
Death وفاة	وفيات، وفاتين، وفاتان

Moreover, any Arabic number was replaced with the word "number رقم" in all five corpora.

Word Frequency

Reducing the original corpora will help identify the greatest number of word associations within a specific extracted window to the target token. Moreover, it helps in extracting the most common words among and between these corpora. As illustrated below, Fig. 1 shows distribution of the top 30 words in أيبولا Ebola dataset.

Informative words appear in this analysis, some of which indicate types of epidemic diseases, such as فيروس virus, حمى fever, أنفلونزاflu, وباء epidemic and مرض disease. Table 4 shows their inflected forms.

In addition, some words indicate the number of victims, as in Table 5.

Further analysis will be conducted on these words in the next sections.

Common Words Analysis

Another analysis was performed to identify the 25 common words among the five corpora. For example, Fig. 2, 3, 4 and 5 show the comparison between the Ebola outbreak disease and other diseases that are listed in Table 2.

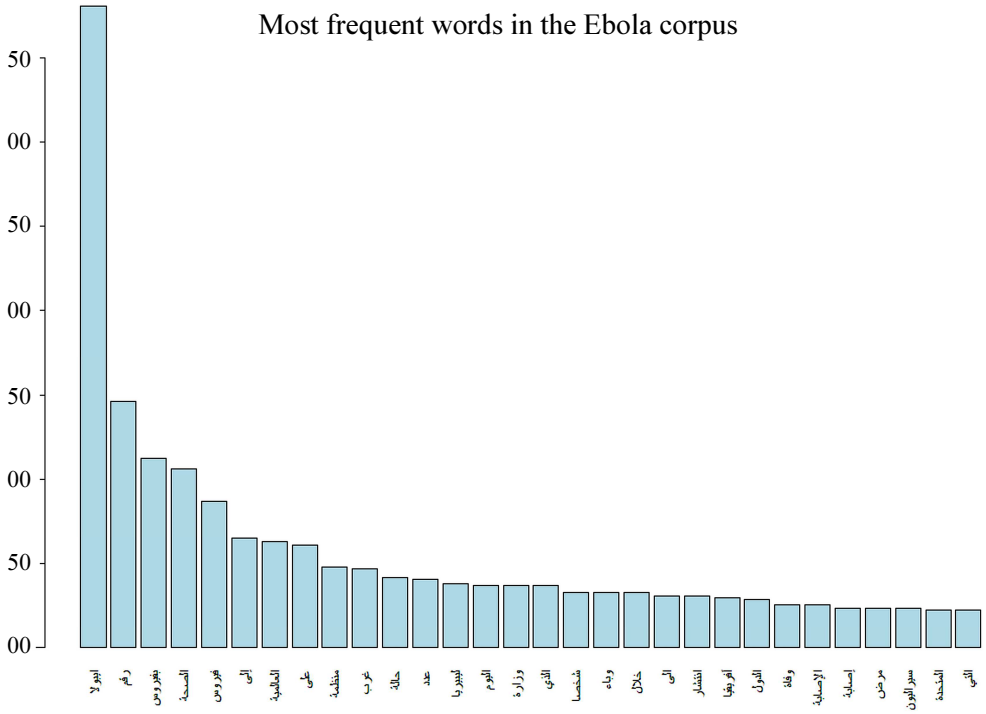

Fig. 1: The thirty most frequent words in the Ebola أيبولا corpus

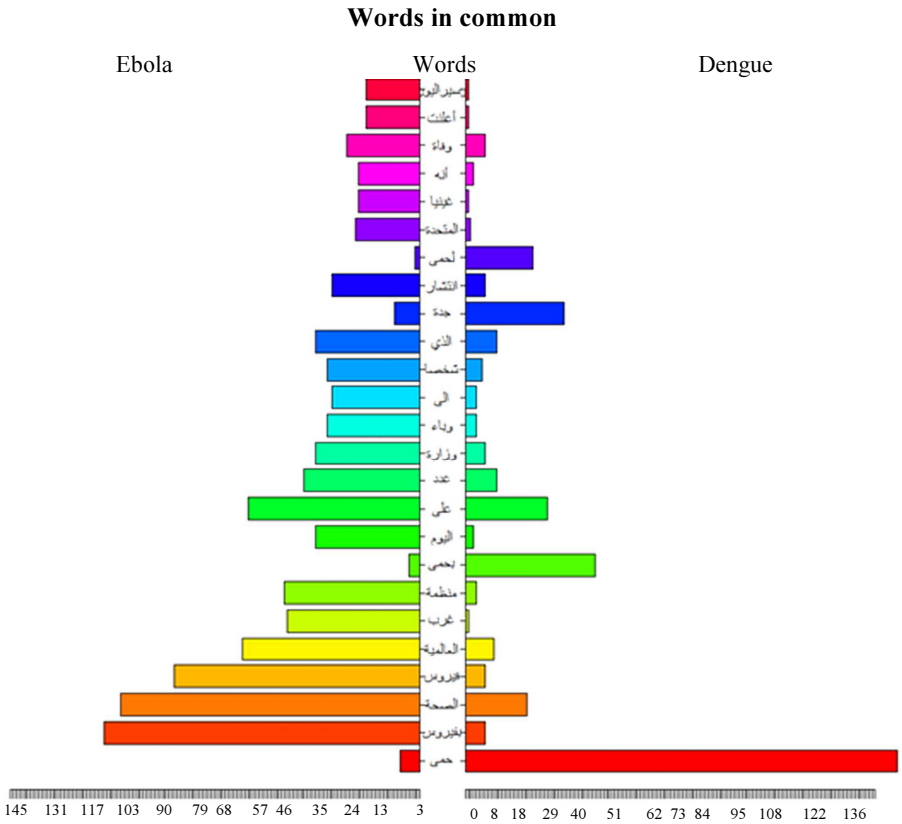

Fig. 2: Ebola vs Dengue

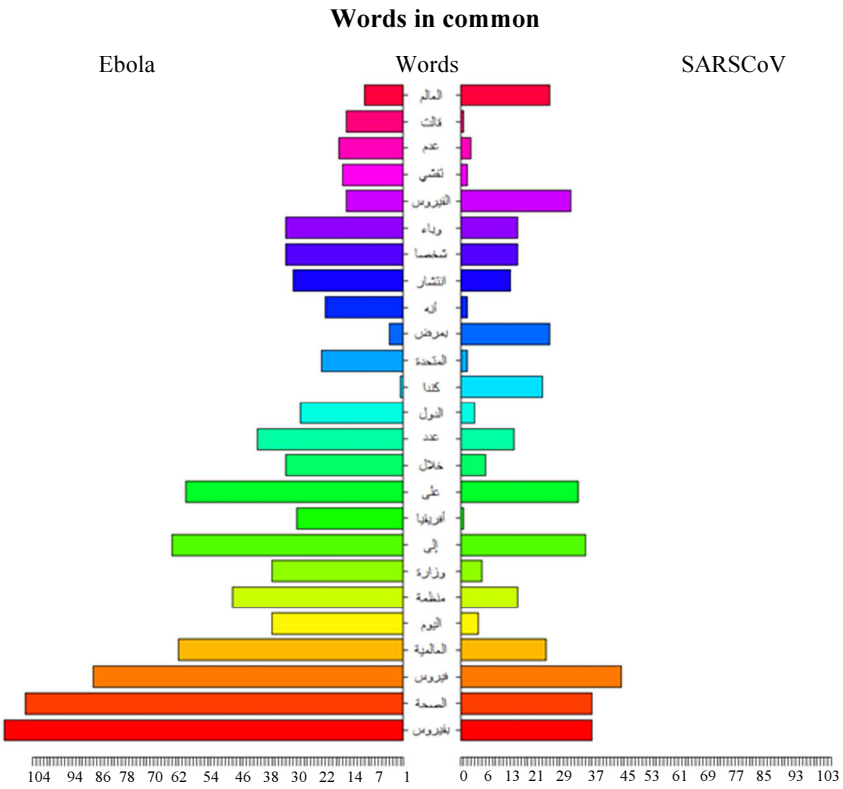

Fig. 3: Ebola vs SARS

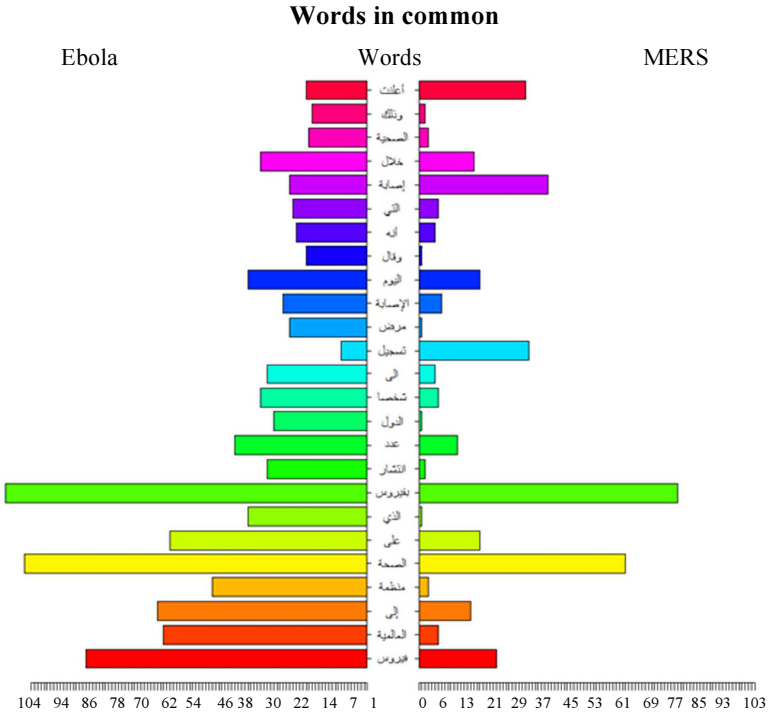

Fig. 4: Ebola vs MERS

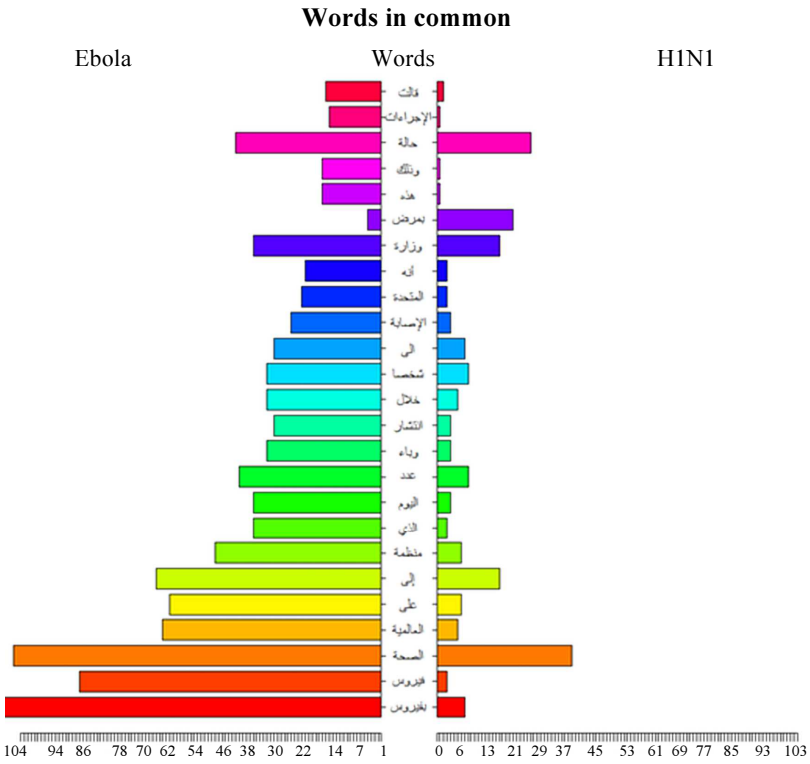

Fig. 5: Ebola vs H1N1

As can be seen, many words, such as "fever حمى", " flu انفلونزا" and "virus فيروس" are common between the five corpora. The derived common words have strong association with the infectious disease names as well be seen below.

Words Cluster Analysis

Examining words organization in sentences is an important step in identifying how the structure might affect identification of epidemic disease-related information. Therefore, the focus will be placed on identifying the relevant words (keywords) regarding disease names, event locations and number of victims.

The word association methodology was used to derive keywords from text data (disease name concordance words corpus), i.e. finding or extracting relations between units (n-gram span 10 and collocation span 5). In other words, the content of the five corpora was analysed in the search for relations between the words. Therefore, clustering analysis was used to model hierarchical relations between words that frequently occur. Word cluster analysis (text-based dendrogram) was performed to compute the differences between each row of the matrix using packages of R language. This analysis is based on the dissimilarities in the distance between the Term-Document Matrix (TDM). In order to avoid a clutter problem, the sparsity of TDM was adjusted to 0.999. This aims to limit the number of words in the TDM and therefore, make the clustering analysis easier to interpret. The following is the clustering analysis for the five corpora that represent the five diseases إيبولا Ebola, كورونا MERS, حمى الضنك Dengue, انفلونزا الخنازيرH1N1, سارس SARSCoV.

We use the hcluster package provided in R language for implementing the hierarchical clustering algorithm. Therefore, for performing word clustering, the following steps are applied:

- Loading the data
- Creating Corpus for storing text documents
- Applying the function (tm_map) to the corpus for data preprocessing, such as removing punctuation marks, extra white spaces and stopwords and replacing "tab" with a space.
- Creating "Term document matrix" by Term Document Matrix function.
- Applying removeSparseTerms function.
- Converting data in the form of a distance matrix by using dist function to compute the euclidean distance between the documents, i.e., calculating the differences between each row of the matrix.
- Applying hclust function to perform cluster analysis on the dissimilarities of the distance matrix.
- Generating dendogram graphs by running plot function.

As previously mentioned, in order to prevent a cluttered dendogram and to facilitate interpretation, the sparsity value was 0.999. This aims to limit the number of words. As can be seen, valuable clusters appear in the Figures.

The Figures show the hierarchical relations between words in each specific disease corpus. In Fig. 6, 7 and 8, the disease names "Ebola ايبولا", "MERS كورونا" and "SARSCoV سارس" always appear related to the word "virus فيروس" whereas in Fig. 9 and 10 the two disease names "Dengue الضنك" and "H1N1 الخنازير" always appear related to the words "fever حمى" and "انفلونزا", respectively. Moreover, the previous disease names have a co-occurrence with the words "disease مرض" and "epidemic وباء". Therefore, these words can be utilized to identify disease names from the news reports.

On the other hand, in the preprocessing phase, all Arabic numbers that indicate the number of victims in the dataset were changed to the word "number رقم". This word appears in the clustering analysis related to the words "cases حالات", "death وفيات" and "infected اصابات" as in "3 cases - حالات 3", "3 deaths - وفيات 3" and "3 infected-اصابات3". As is well known, Arabic nouns are inflected for gender (masculine and feminine) and number (singular, dual and plural). As a result, the number of victims are sometimes written in the form of dual nouns without writing an explicit number, such as "2 cases - حالتين", "2 deaths - وفاتين" and" 2 infected - اصابتين". Similarly, this may happen when the singular form is used, as in "one case - حالة", "one infected - اصابة" and" one death - وفاة".

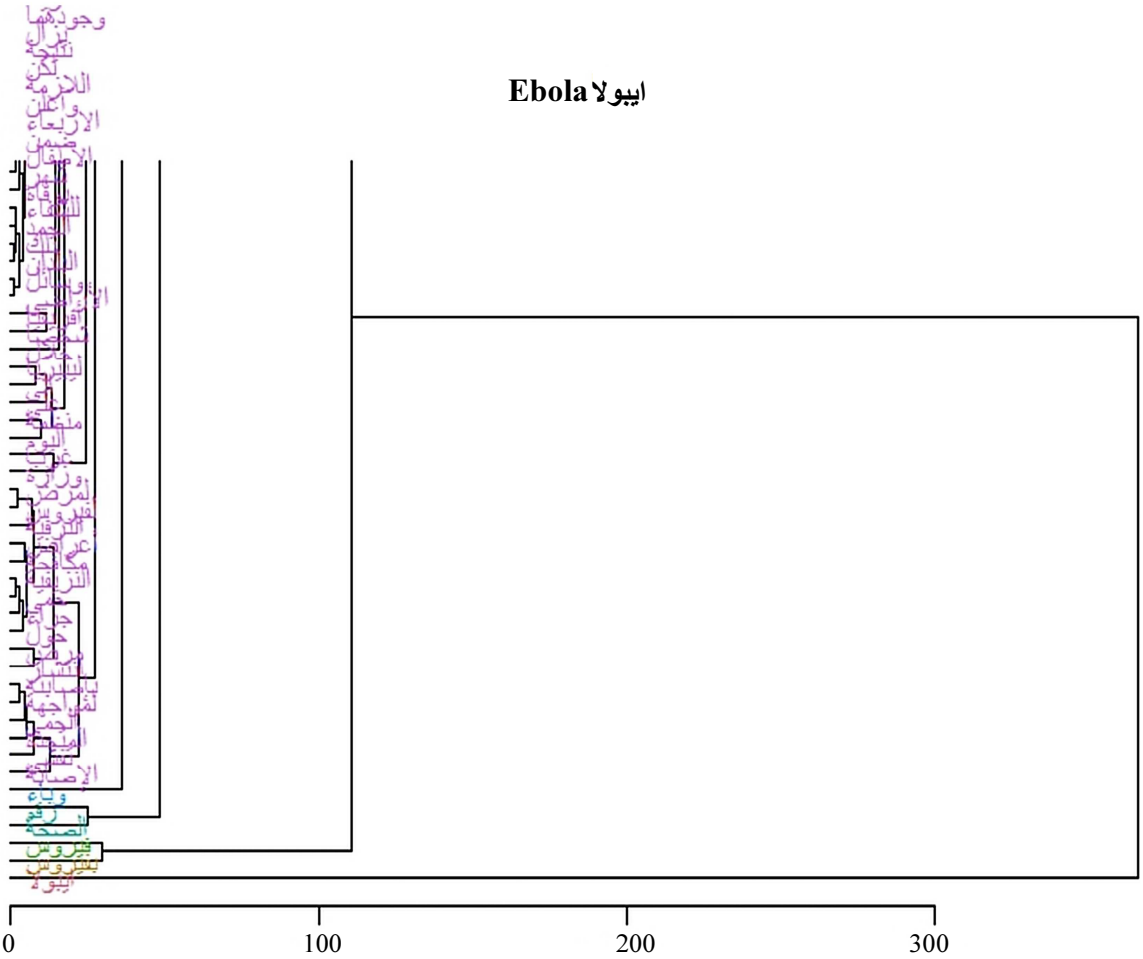

Fig. 6: The hierarchical clustering (dendogram graph) for the Ebola ايبولا corpus

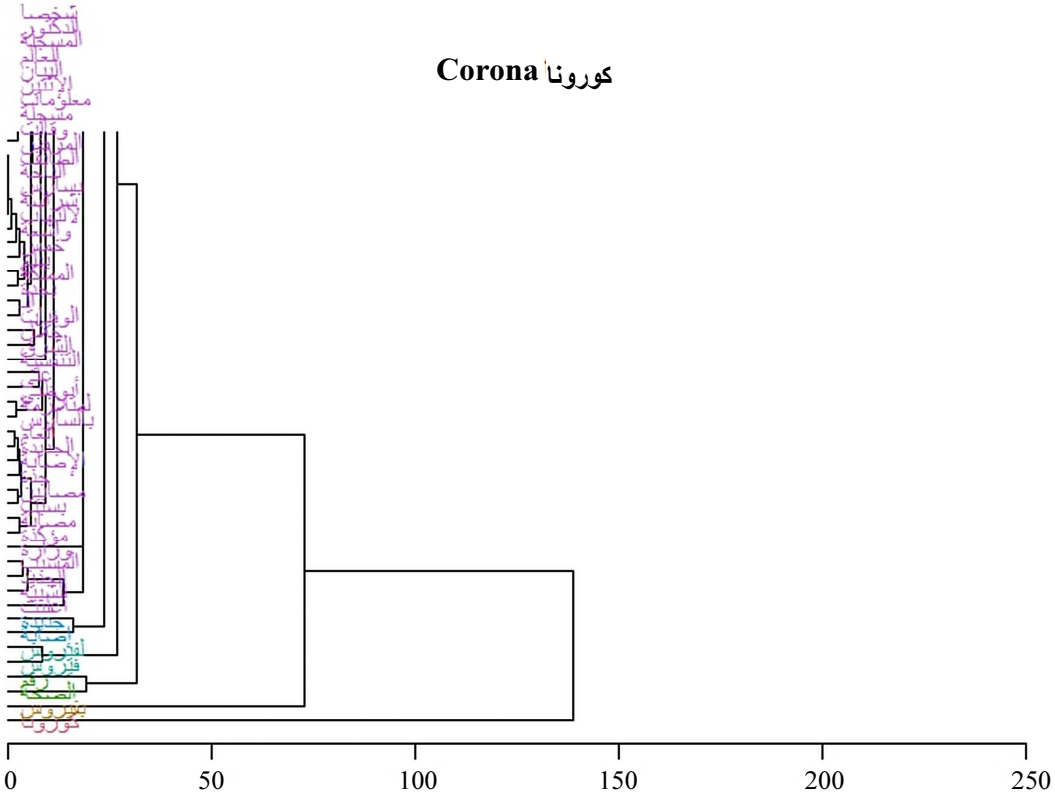

Fig. 7: The hierarchical clustering (dendogram graph) for the MERS/Corona كورونا corpus

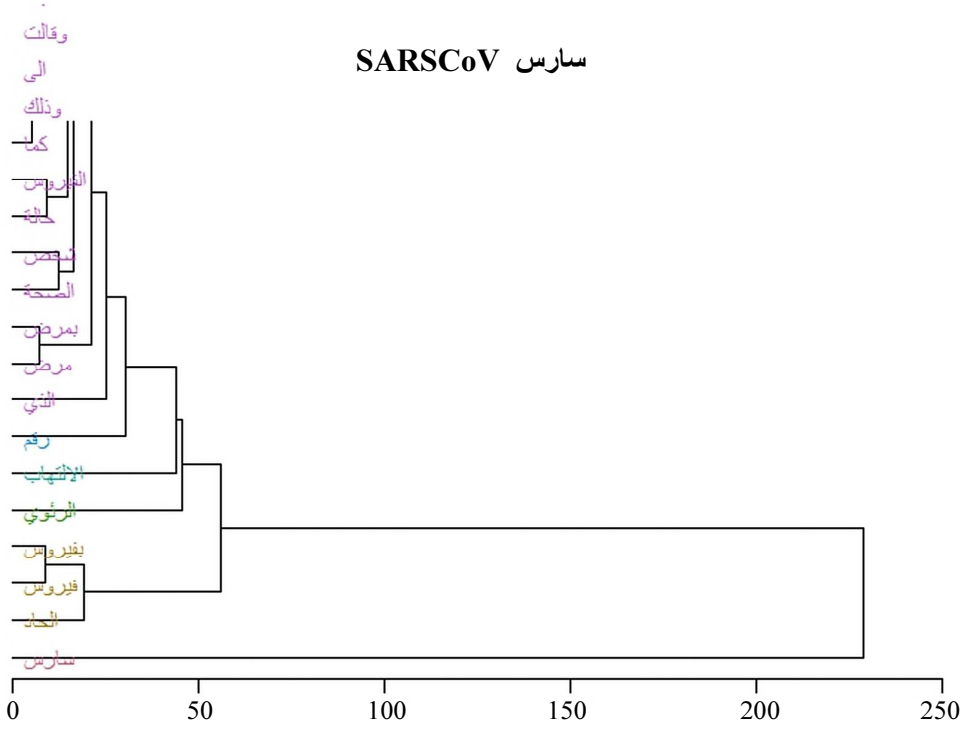

Fig. 8: The hierarchical clustering (dendogram graph) for the SARSCoVسارس corpus

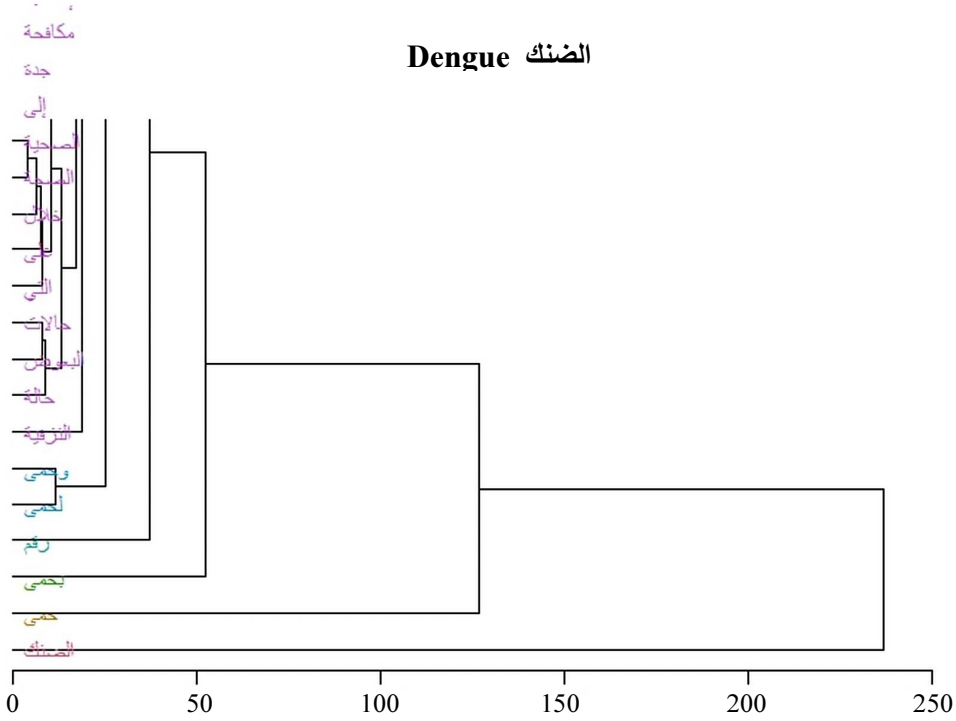

Fig. 9: The hierarchical clustering (dendogram graph) for the Dengue الضنك corpus

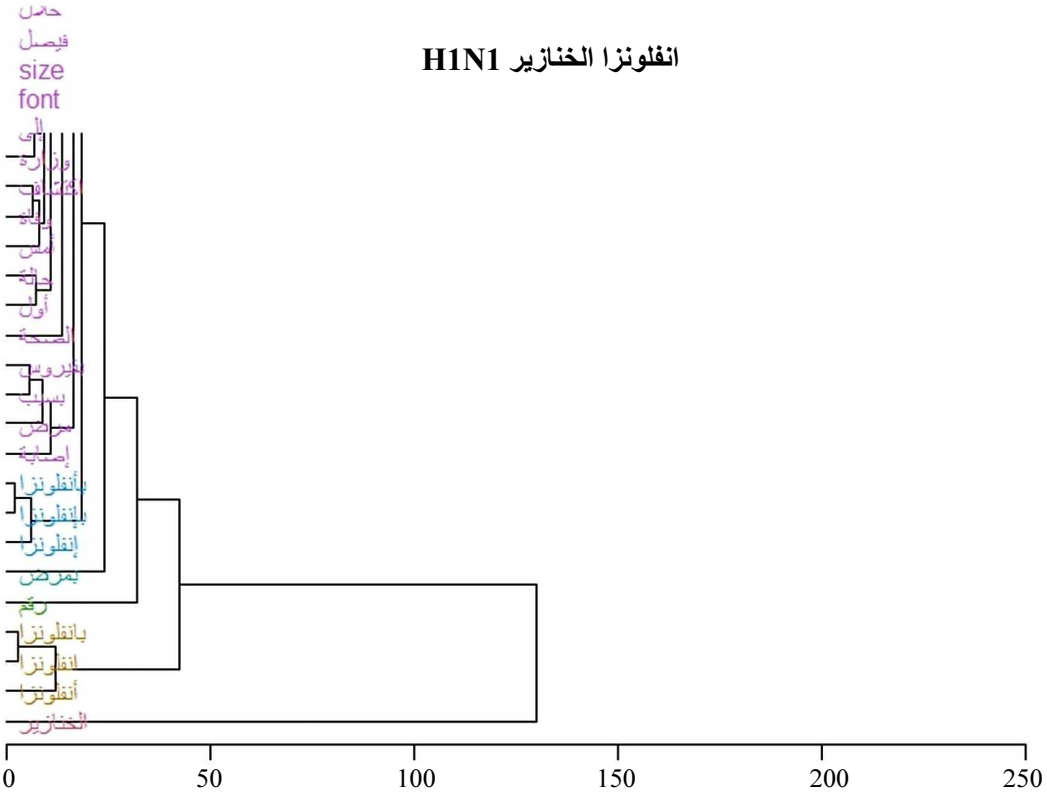

Fig. 10: The hierarchical clustering (dendogram graph) for the H1N1 انفلونزا الخنازير corpus

With regard to location and date of events, it is not necessary to perform an analysis to study the word relations. Disease outbreak location can be located through a pattern-matching method using predefined lists that contain names of countries and cities. Regarding the date, a report date is adopted, as the infectious disease date and <pubdate><\pubdate> tag is used for recognizing date, which will be presented later.

Arabic Surveillance Infectious Disease Outbreak System (ASIDOS)

In this section, the extraction methods for the four entities: disease name, location, number of victims and date are explained. The architecture of the Arabic Surveillance Infectious Disease Outbreak System (ASIDOS) is depicted in Fig. 11. The ASIDOS system is now online (Alruily, 2018).

ASIDOS is a real-time system for monitoring diseases 24/7 via information gathering from informal sources, such as online news, for detecting and tracking disease outbreaks. ASIDOS comprises several stages as follows:

- Data collection
  The process used is based on receiving Really Simple Syndication (RSS) feeds from predefined sources that normally publish reports on public health that include disease outbreaks.
- Information extraction engine
  In this stage, xml files are processed in order to extract information of outbreak incidents (disease name, location, number of victims, date). It was noticed that the disease-related information often occurs within the <title><\title> tag and if an entity or entities are missing, the system moves to <body><\body> tag to find the remaining entities.

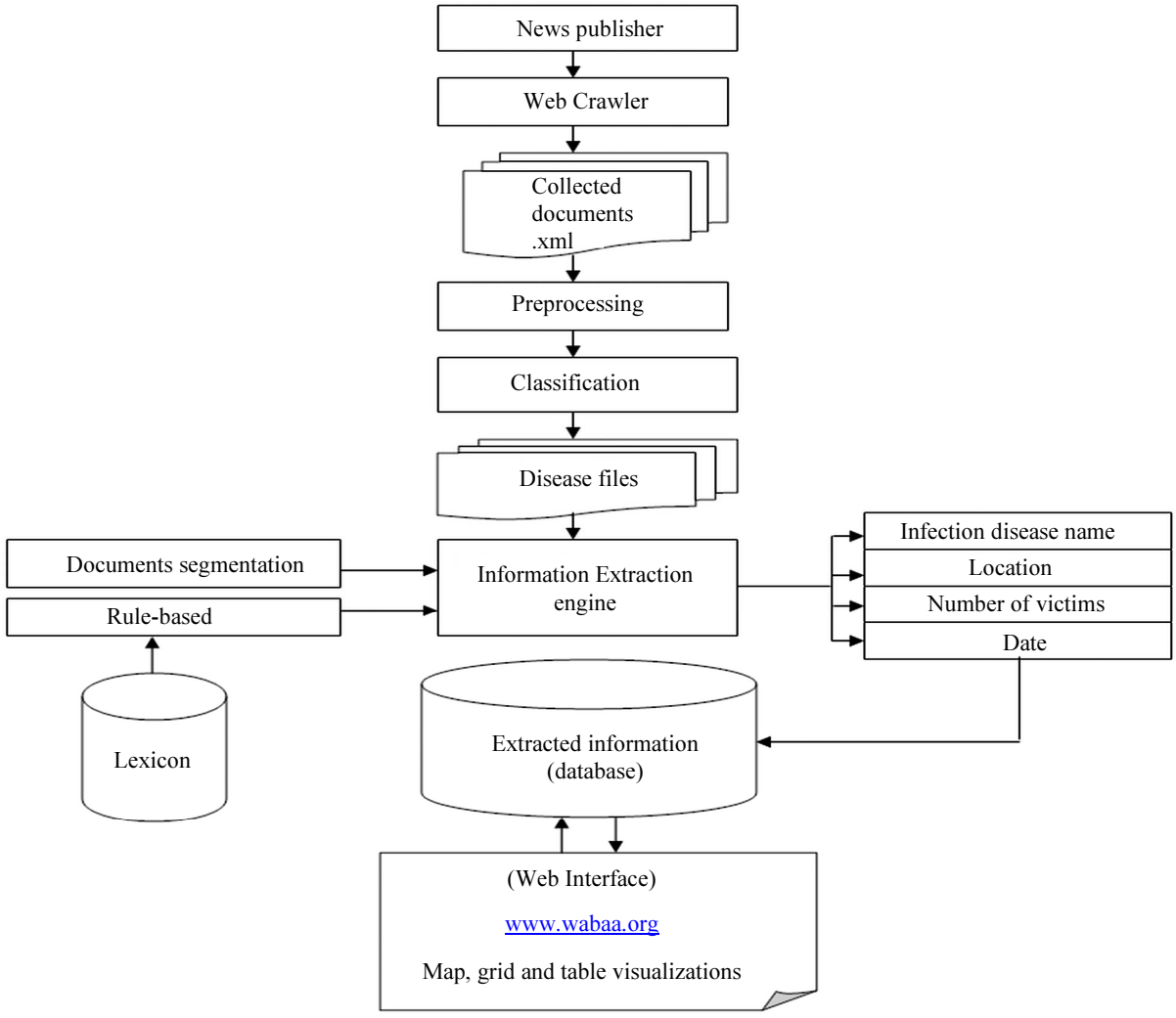

Fig. 11: ASIDOS architecture

The developed system is based on rules using regular expressions that were inferred, as seen in Section 4. In addition, a gazetteer containing keywords that indicate the position of patterns of interest within a body of text is used:

1. Extracting the infectious disease
   For extracting an infectious disease from incident narrative reports, regular expression consisting of a single keyword from the keywords list in Table 6 is used. Therefore, the word that follows the keywords is an in infectious disease.
   The following is the list of regular expressions for extracting infectious diseases within a body of text:

   (*Regex*) Keyword [\s]?[\ w]+

2. Extracting location of event
   Simply put, here a straightforward pattern-matching method is used for identifying the place of a disease outbreak. A location gazetteer was created for achieving this task.
3. Extracting number of victims
   For extracting the number of people affected in the incidence of a disease outbreak, the following regular expressions are utilized:

   (*Regex*) \d[\d] + [\s]?Keyword

   In the case of these regular expressions failing to extract the number of victims, the list of keywords in Table 7 is used for performing pattern matching. These keywords indicate that the number of victims is either one or two.
4. Extracting date of event
   Regarding the date, a report date embedded in the xml file is adopted as the outbreak of infectious disease date and <pubdate><\ pubdate> is used for recognizing date.

Table 6: The victims indicators words and their inflected forms

Keyword	English translation
فيروس	virus
حمى	fever
انفلونزا	flu
وباء	epidemic
مرض	disease

Table 7: Singular and dual nouns that implicitly represent number of effected victims

Type	keyword	English translation
Singular noun	حالة	one case
	اصابة	one infected
	وفاة	one dead
Dual noun	حالتين - حالتان	two cases
	اصابتين - اصابتان	two infected
	وفاتين - وفاتان	two dead

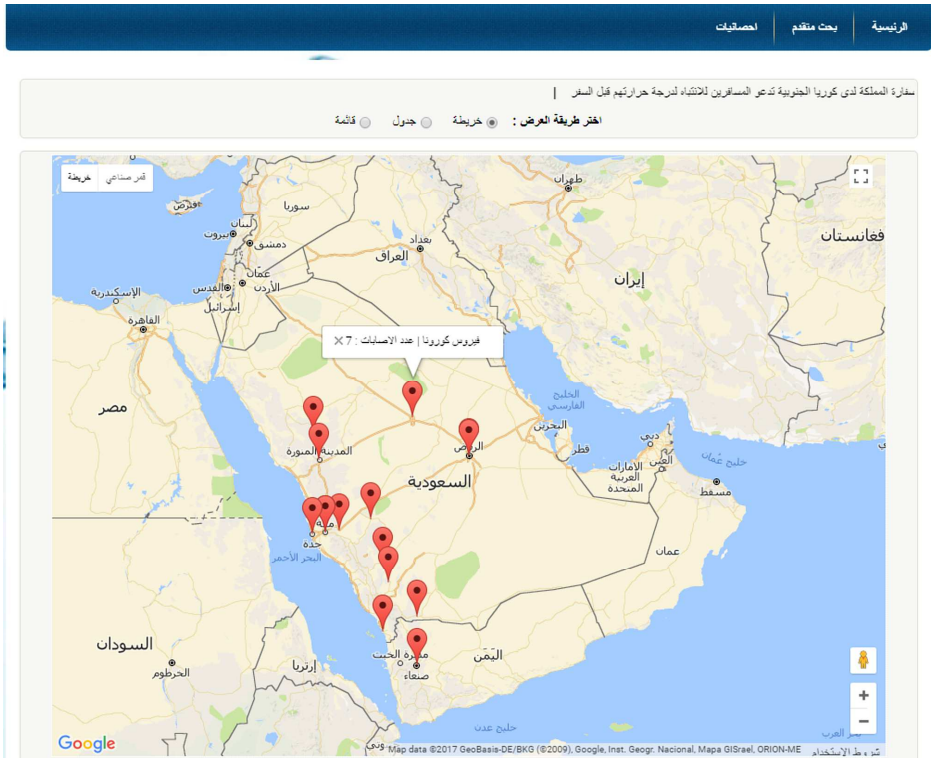

Fig. 12: Map used for plotting locations of infectious diseases

Fig. 13: The extracted information is presented in grid format

Fig. 14: The extracted information is presented in table format

Fig. 15: Searching window for searching by a disease name, location, source and period of time (last month, last three months and last six months)

Fig. 16: Statistical information can be generated for specific disease within a specific year

Interface

After detecting and extracting outbreaks of infectious disease-related information from the online textual news reports, they are mapped so as to be visualized in logical representation. Therefore, a relational database is created in order that extracted facts can be stored and then visualized in different forms, e.g. temporal and spatial visualizations. A web server is used in order to make the database available online and can be accessed through the website (www.wabaa.org). A web-based interface is provided to make the database

accessible. The system interface contains three types of visualizations: map, table and grid. Spatial plotting for locations of infectious diseases was achieved by using Google Maps technology. Figures 12, 13 and 14 are snapshots of ASIDOS illustrating the three forms of visualization.

Additionally, the searching ability in the database is available with many options to users. They are able to search by disease name, location, source and period of time (last month, last three months and last six months), as shown in Fig. 15. Moreover, advanced searching is also available, which comprises previous search options together. Furthermore, the system is able to generate statistical information for users about a specific disease within a specific year, as can be seen in Fig. 16.

Experiments and Evaluation

The experiments were performed on new and untouched corpus. This dataset contains 266 articles collected from different sources. Moreover, reports of new types of infectious disease outbreaks were added to this corpus, such as typhoid, cholera, malaria and tuberculosis, which will test the performance of the ASIDOS system more efficiently. To evaluate the performance of this system the Precision (P), Recall (R) and F-measure metrics are used.

Experiments

The following experiments test the performance of the ASIDOS system for extracting outbreaks of epidemic information, i.e., disease, location, number of victims and date.

- Disease
  In this experiment, the derived regular expressions and list of keywords for pattern matching were tested to extract infectious disease. The ASIDOS system was able to extract 266 disease names out of 287. The precision is high 100% as well as the recall 93%. It yielded an overall f-measure of 96%. The reason for not identifying some disease names was that a number of news reports did not have any keywords that are used in the regular expressions and, as a result, the extraction process was not implemented.
- Location
  The system was able to correctly identify 205 locations out of 328. The precision, recall and f-measure results are 98%, 63% and 76%, respectively. It can be observed that the recall value is low for various reasons. Arabic is a highly inflected language with a very complex morphology. For example, the word "Beijing" the capital of China, can be written in two forms in the Arabic language news "بكين" and "بيجين". Therefore, all the different forms of countries' or cities' names must be added to the location gazetteers in order that value of recall can be improved. Moreover, in any events, cities' names in the Arabic language are preceded by either the preposition "في" or "ب", both of which mean "in". The preposition "ب" is fused at the beginning of the word as a prefix. Hence, the locations cannot be extracted using straightforward pattern matching. In addition, sometimes, as in the sentence " سجلت عدة وفايات بانفلونزا الطيور في الساحل السوري", the name of the location "Syrian coast" is written in the form of an adjective not as a noun and therefore, cannot be recognized. In some cases, two or more locations are written in a report as places of disease outbreaks but the system is designed to identify only one location.
- Victims
  In this experiment, the system was evaluated and achieved 84% precision, 69% recall and 75% f-measure. Although the system failed to extract 104 entities and 44 were wrongly identified, the results seem relatively satisfactory. The reason for not extracting a number of entities was due to certain cases where Arabic numbers occur after the keywords, as in the sentence: "وفاة 7 بالكوليرا". Moreover, sometimes the regular expressions created for the extraction process is not implemented when the keywords are not found in the text, as in " تسجيل كورونا ل 4 أفراد من عائلة واحدة في الرياض". In consequence, the number of victims cannot be extracted.
- Date
  The publishing date of news reports is considered the date of the outbreak of an infectious disease. For extracting the date, the xml <pubdate><\pubdate> tag indicating the date is used. The system successfully extracted it with no errors.

To the author's knowledge, no event-based epidemic disease surveillance system has been developed for directly processing Arabic texts, i.e. without using translation engine. For example, the Global Public Health Intelligence Network (GPHIN) is a multilingual system supporting eight languages including the Arabic language but uses a translation machine to translate non-English reports into English in order to process them. It also presents information in English and French languages. HealthMap system is freely and publicly available not as GPHIN. HealthMap also uses a translation machine to support other languages including Arabic texts and presents information in Arabic. HealthMap's Arabic page (http://www.healthmap.org/ar/) was visited many times; the last visit was on 21 September 2017, in all visits it did not work. The system was evaluated using data written in English; the overall accuracy is 84% and the ASIDOS system outperforms it with 94%.

Table 8: The overall ASIDOS system evaluation results

	Precision	Recall	F-measure
ASIDOS	94%	74%	83%

Moreover, MedISys and PULS are able to process 60 languages. Although MedISys is able to retrieve data from many resources in different languages, Plus system is responsible for information extraction tasks and is only able to process reports in English language, i.e., it makes use of machine translation to perform text extraction processing for other languages. The system achieved 72% accuracy. The overall performance results for the ASIDOS system is listed in Table 8.

Conclusion

The main aim of this work was to develop an event-based surveillance system for extracting infectious disease, location of disease outbreak, number of affected victims and date of disease outbreaks from Arabic health news reports. An overview of the architecture of the proposed system was presented. Also, the performance of Arabic Surveillance Infectious Disease Outbreak System (ASIDOS) was evaluated through implementing experiments. The overall results achieved were: precision 94%, recall 74% and f-measure 83%. The system interface and its features were explained and can be accessed via the link www.wabaa.org. The extracted information is visualized in various ways and a user is able to search in the old data.

Many contributions were delivered from this research. The main implication was developing the ASIDOS system, which is the first system developed and available online in the Middle East and North Africa (MENA) region to track outbreaks of infectious disease. In addition, the first analysis on infectious diseases data written in Arabic was performed using R in this research. Moreover, building an infectious disease corpus is one of the contributions of this study, as currently, there is no available specific dataset containing reports about infectious diseases.

Finally, the ASIDOS system could be improved in the future through extending its work to cover other events, such as biological events and diseases that affect animals and plants. Also, disseminating warning messages to subscribers could be added to the system to alert them of any disease outbreaks.

Acknowledgement

The author would like to thank Jouf University for their support. Special thanks go to Abdulmajeed Alanzi for the programming support.

Funding Information

The financial support provided by Jouf University through Grant No. 34/243.

Ethics

This article is original and has not been published elswhere. The corresponding author confirms that there are no ethical issues involved.

References

Abat, C., H. Chaudet, J.M. Rolain, P. Colson and D. Raoult, 2016. Traditional and syndromic surveillance of infectious diseases and pathogens. Int. J. Infect. Dis., 48: 22-28. DOI: 10.1016/j.ijid.2016.04.021

Agheneza, T., 2011. A systematic literature review on event-based public health surveillance systems. Thesis of PHD, Hamburg University of Applied Sciences, Germany.

Alanazi, S., 2017. A named entity recognition system applied to Arabic text in the medical domain. PHD Thesis of the Philosophy, Staffordshire University, UK.

Al-Garadi, M.A., M.S. Khan, K.D. Varathan, G. Mujtaba and A.M. Al-Kabsi, 2016. Using online social networks to track a pandemic: A systematic review. J. Biomed. Informatics, 62: 1-11. DOI: 10.1016/j.jbi.2016.05.005

Alruily, M., 2018. Arabic Surveillance Infectious Disease Outbreak System (ASIDOS). http://www.wabaa.org/

Alshowaib, W.N., 2014. Rule-based information extraction from disease outbreak reports. Int. J. Computa. Linguistics, 5: 37-58.

Azzedin, F., J. Yazdani, S. Adam and M. Ghaleb, 2014. A generic model for disease outbreak notification systems. Int. J. Comput. Sci. Informat. Technol., 6: 137-154. DOI: 10.5121/ijcsit.2014.6409

Bagherian, H., M. Farahbakhsh, R. Rabiei, H. Moghaddasi and F. Asadi, 2017. National communicable disease surveillance system: A review on information and organizational structures in developed countries. Acta Informa. Med., 25: 271-276. DOI: 10.5455/aim.2017.25.271-276

Blench, M., H. Tolentino, C.C. Freifeld, K.D. Mandl and A. Mawudeku *et al*., 2009. Use of unstructured event-based reports for global infectious disease surveillance. Emerg Infect. Dis., 15: 689-95. DOI: 10.3201/eid1505.081114

Brownstein, J.S. and C.C. Freifeld, 2007. Healthmap: The development of automated real-time internet surveillance for epidemic intelligence. Eurosurveillance, 12: E071129.5. DOI: 10.2807/esw.12.48.03322-en

CDC, 1996. Middle East Respiratory Syndrome (MERS). https://www.cdc.gov/coronavirus/mers/

Chen, H., D. Zeng and P. Yan, 2010. Argus. In: Infectious Disease Informatics, 1st Edn., Springer US, Boston, MA, ISBN-10: 9781441912787, pp: 177-181.

Choi, J., Y. Cho, E. Shim and H. Woo, 2016. Web-based infectious disease surveillance systems and public health perspectives: A systematic review. BMC Public Health, 16: 12-38. DOI: 10.1186/s12889-016-3893-0

Christaki, E., 2015. New technologies in predicting, preventing and controlling emerging infectious diseases. Virulence, 6: 558-565. DOI: 10.1080/21505594.2015.1040975

Collier, N., S. Doan, A. Kawazoe, R.M. Goodwin and M. Conway *et al*., 2008. Biocaster: Detecting public health rumors with a web-based text mining system. Bioinformatics, 24: 2940-2941. DOI: 10.1093/bioinformatics/btn534

Collier, N. and S. Doan, 2012. Geni-db: A database of global events for epidemic intelligence. Bioinformatics, 28: 1186-1188. DOI: 10.1093/bioinformatics/bts099

Damianos, L., J. Ponte, S. Wohlever, F. Reeder and D. Day *et al*., 2002. Mitap, text and audio processing for bio-security: A case study. Proceedings of the 14th Conference on Innovative Applications of Artificial Intelligence, (AAI' 02), IEEE Xplore Press, Canada, pp: 807-814.

Dredze, M., R. Cheng, M.J. Paul and D. Broniatowski, 2014. Health tweets.org: A platform for public health surveillance using twitter. Proceedings of the AAAI Workshop on the World Wide Web and Public Health Intelligence. Johns Hopkins Universit, pp: 2-3.593-596.

EMM, 2002. Medical Information System (MedISys). http://emm.newsbrief.eu/overview.html

Freifeld, C.C., K.D. Mandl, B.Y. Reis and J.S. Brownstein, 2008. Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. J. Am. Medi. Informatics Association, 15: 150-157. DOI: 10.1197/jamia.M2544

Goel, R., S. Fadloun, S. Valentin, A. Sallaberry and M. Roche *et al*., 2018. Epidnews: An epidemiological news explorer for monitoring animal diseases. Proceedings of the 11th International Symposium on Visual Information Communication and Interaction, (INCI' 2018), IEEE Xplore Press, ACM, USA, pp: 1-8. DOI: 10.1145/3231622.3231624

Grishman, R., S. Huttunen and R. Yangarber, 2002. Real-time event extraction for infectious disease outbreaks. Proceedings of the 2nd International Conference on Human Language Technology Research, Mar. 24-27, IEEE Xplore Press, San Francisco, CA, USA, pp: 366-369. DOI: 10.3115/1289189.1289229

Guglielmetti, P., D. Coulombier, G. Thinus, F.L. Van and S. Schreck, 2006. The early warning and reponse system for communicable diseases in the Eu: An overview from 1999 to 2005. Euro. Surveill, 11: 215-220 7-8. DOI: 10.2807/esm.11.12.00666-en

Heymann, D.L. and G.R. Rodier, 2001. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. Lancet Infect. Dis., 1: 345-353. DOI: 10.1016/S1473-3099(01)00148-7

Hugh-Jones, M., 2002. Global awareness of disease outbreaks: The experience of promed-mail. Public Health Reports, 116: 27-31. DOI: 10.1093/phr/116.S2.27

Hulth, A. and G. Rydevik, 2011. Get well: An automated surveillance system for gaining new epidemiological knowledge. BMC Public Health, 11: 252. DOI: 10.1186/1471-2458-11-252

Jeremy, G., H.M. Matthew, S.P. Rajan, B. Lynnette and S.S. Mark *et al*., 2009. Detecting influenza epidemics using search engine query data. Nature. DOI: 10.1186/1471-2458-11-252

Koppeschaar, C., 2011. Influenzanet. www.influenzanet.eu

Lamb, A., M.J. Paul and M. Dredze, 2013. Separating fact from fear: Tracking flu infections on twitter. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, pp: 789-795.

Lorthe, T.S., M.P. Pollack, B. Lassmann, J.S. Brownstein and E. Cohn *et al*., 2018. Evaluation of the epicore outbreak verification system. Bulletin World Health Organizat., 96: 327-343. DOI: 10.2471/BLT.17.207225

Mawudeku, A. and M. Blench, 2006. Global Public Health Intelligence Network (GPHIN). Proceedings of the 8th Conference of the Association for Machine Translation in the Americas (MTS'06), EEE Xplore Press, Hawaii, USA, pp: 299-303.

Nguyen, M.T., 2015. Desrm: A disease extraction system for real-time monitoring. Int. J. Comput. Vision Robot., 5: 282-301. DOI: 10.1504/IJCVR.2015.071341

Okokpujie, K., A. Orimogunje, E. Noma-Osaghae and O. Alashiri, 2017. An intelligent online diagnostic system with epidemic alert. Int. J. Innovative Sci. Res. Technol., 2: 327-331.

PAHO/WHO, 2014. Ebola Virus Disease (EVD), implications of introduction in the Americas, https://www.paho.org/hq/dmdocuments/2014/06-aug-2014-cha-evd-preparedness-response-americas.pdf

ProMED, 2010. About promed-mail.

Ramalingam, B., 2016. Real-time monitoring in disease outbreaks: Strengths, weaknesses and future potential, IDS Evidence Report.

Samy, D., A.M. Sandoval, C.B. Diaz, M.G. Salazar and J.M. Guirao, 2012. Medical term extraction in an arabic medical corpus. Proceedings of the 8th Language Resources and Evaluation Conference, (REC' 12), IEEE Xplore Press, Istanbul, Turkey.

Sharib, A.K., C.O. Patel and R. Kukafka, 2006. Godsn: Global news driven disease outbreak and surveillance. AMIA Annu. Symp. Proc., 2006: 983.

Svitlana, V. and H.H. William, 2010. Computational knowledge and information management in veterinary epidemiology. Proceedings of the IEEE International Conference on Intelligence and Security Informatics, May 23-26, IEEE Xplore Press, Vancouver, Canada, pp: 120-125. DOI: 10.1109/ISI.2010.5484764

Taha, K.H. and N.D. Tada, 2008. International System for Total Early Disease Detection (INSTEDD) platform. Advances Dis. Surveillance, 5: 108.

Talvis, K., K. Chorianopoulos and K.L. Kermanidis, 2014. Real-time monitoring of flu epidemics through linguistic and statistical analysis of twitter messages. Proceedings of the 9th International Workshop on Semantic and Social Media Adaptation and Personalization, Nov. 6-7, IEEE Xplore Press, Corfu, Greece, pp: 83-87. DOI: 10.1109/SMAP.2014.38

Tolentino, H., R. Kamadjeu, P. Fontelo, F. Liu and M. Pollack *et al*., 2007. Scanning the emerging infectious diseases horizon-visualizing promed emails using epispider. Advances Disease Surveillance, 2: 169.

Tsui, K., L. Tsui, D. Goldsman, W. Jiang and S.Y. Wong, 2010. Recent research in public health surveillance and health management. Proceedings of the Prognostics and System Health Management Conference, Jan. 12-14, IEEE Xplore Press, Macao, China, pp: 1-7. DOI: 10.1109/PHM.2010.5413455

Tsui, K.L., W. Chiu, P. Gierlich, D. Goldsman and X. Liu *et al*., 2008. A review of healthcare, public health and syndromic surveillance. Quality Engineering, 20: 435-450. DOI: 10.1080/08982110802334138

Valle, S.D., 2000. Agent-based modeling.

Velasco, E., T. Agheneza, K. Denecke, G.R. Kirchner and T. Eckmanns, 2014. Social media and internet-based data in global systems for public health surveillance: A systematic review. Milbank Quart, 92: 7-33. DOI: 10.1111/1468-0009.12038

Washington, 2015. Modeling the spread of ebola.

WHO, 2008. A guide to establishing event-based surveillance,

WHO, 2014. National passive surveillance.

WHO, 2015. Epidemic intelligence-systematic event detection.

Wilson, J.M., 2007. Argus: A global detection and tracking system for biological events. Advances Disease Surveillance.

Woodall, J., 1997. Stalking the next epidemic: Promed tracks emerging diseases. Public Health Rep., 112: 78-82.

Yang, S., S.C. Kou, F. Lu, J.S. Brownstein and N. Brooke *et al*., 2017. Advances in using internet searches to track dengue. PLoS comput. Biolo., 13: e1005607. DOI: 10.1371/journal.pcbi.1005607

Yang, W., Z. Li, Y. Lan and J. Wang, 2011. A nationwide web-based automated system for outbreak early detection and rapid response in china. Western Pacific Surveillance Response, 2: 10-5. DOI: 10.5365/WPSAR.2010.1.1.009

Zhang, Y., Y. Dang, H. Chen, M. Thurmond and C. Larson, 2009. Automatic online news monitoring and classification for syndromic surveillance. Decision Support Syst., 47: 508-517. DOI: 10.1016/j.dss.2009.04.016