A Review on Event-Based Epidemic Surveillance Systems that Support the Arabic Language

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

***Abstract*—With the revolution of the internet, many event-based systems have been developed for monitoring epidemic threats. These systems rely on unstructured data gathered from various online sources. Moreover, some systems are able to handle more than one language to cover all news reports related to disease outbreaks worldwide. The aim of this paper is to examine existing systems in terms of supporting the Arabic language. The 28 identified systems were evaluated based on different criteria. The results of this evaluation show that only 5 systems support the Arabic language using translation tools; hence, disease outbreaks in news reports written in Arabic are not directly processed. In other words, no existing event-based system in the literature has yet been developed specifically for Arabic health news reports to monitor epidemic diseases.**

Keywords*—Public health; infectious disease; event extraction; disease surveillance system; arabic language

I. 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) [1], [2]. 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 reached 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) [3], [4]. Therefore, the threat of infectious disease outbreaks to public health has prompted countries and organizations to develop several early warning surveillance systems [5]. However, the traditional surveillance systems or indicator-based surveillance systems need public health networks (Sentinel Networks) in order to collect predefined structured data about diseases on a routine basis from indicator sources, such as over-the-counter drugs and emergency department visits [6], [7]. 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 for the results regarding possible outbreaks to be disseminated; furthermore, not all countries have the required infrastructure to implement this system [3], [7], [8], [9]. On the other hand, as a result of the technological revolution of the internet, another type of surveillance system has emerged. This type is called the event-based surveillance system [10]. 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 Keller et al. [11], WHO's investigations into the majority of disease outbreaks are obtained through diverse online informal sources.

The remainder of the paper is organized as follows: in Section II, a background to the topic and a review of related work are presented. Section III describes methods used in performing this work. Section IV explores disease outbreak surveillance systems in Arabic. Section V discusses the state of the art of outbreak surveillance systems. Section VI presents the results and discussion. Finally, the conclusion of this work is presented in Section VII.

II. Related Work

According to Agheneza et al. [12], surveillance systems are classified into two types based on the type of data being processed. The first type is the indicator-based surveillance system (syndromic surveillance). The second type is the event-based surveillance system that collects and processes unstructured data from formal and informal sources, such as newspapers, reports, and medical websites. The current paper will review and focus solely on the event-based surveillance system, which utilizes text mining techniques to process media sources (news reports related to disease outbreaks) for text understanding, i.e. detecting and extracting infectious disease outbreak-related information, such as disease type, location name, date, and number of victims, if any. Also, more focus will be placed on systems that are able to process Arabic unstructured texts in the health domain.

A. Indicator-based Surveillance Systems

To date, many existing public indicator-based surveillance applications have been introduced to detect and track increases in disease incidence rates based on structured predefined information collected from different official sources, such as emergency room visits, telephone calls, and over-the-counter drug sales (syndromic/clinical surveillance data) [6]. Many detection methods are used to perform this task. Tsui et al. [13] categorized these methods into three types: temporal (SPC, regression, time series, and forecast-based methods), spatial (scan statistics), and spatiotemporal surveillance techniques.

Detailed information on these types of health surveillance systems can be seen in [13], [14]; the most recent review of these systems can be found in [15], [16]. In addition, in [17] the national communicable diseases surveillance systems proposed between 2000 and 2016 in developed countries were reviewed.

B. Event-based Surveillance Systems

On the other hand, many event-based surveillance systems have been developed for manipulating gathered unstructured data relating to infectious disease outbreaks from informal or non-traditional sources, such as online newspapers, news reports and social media [6]. Keller et al. [11] examined and compared the performance of three existing systems: EpiSPIDER, HealthMap and Global Public Health Intelligence Network (GPHIN), which have been developed to process event-based outbreak information. However, the most comprehensive review was conducted by Velasco et al. [3], who reviewed studies of infectious disease surveillance publications between 1990 and 2011 in detail, yielding 13 event-based systems. These systems were developed between 1994 and 2006, as listed in Table I, and used 15 review criteria: system name, system category, country, year started, coordinating organization, purpose, jurisdiction, supporting Arabic, disease type, public access, data processing, dissemination of data, most avid users, system evaluation and Homepage.

Choi et al. [5] performed a systematic review of web-based infectious disease surveillance systems, published in 2016. They identified 11 web-based surveillance systems, including GOARN, GPHIN, MedISys , BioCaster, HealthMap, ProMED, and EpiSPIDER, which had already been reviewed by Velasco et al. [3]. However, they also reviewed new systems: EpiSimS (now known as Object-oriented Platform for People in Infectious Epidemic OPPIE) [18], Google Flu Trends [19], GET WELL [20] and Influenzanet [21], which were not mentioned in [3], as can be seen in Table I.

To the best of the authors' knowledge, the recent review article was carried out by Yan et al. [22] and was published in October 2017. Developed systems in articles published between 2006 and 2016 were evaluated in terms of their methods, timeliness and accuracy outcomes.

III. Methods

Web-based infectious disease surveillance systems were systematically reviewed by focusing on multilingual systems, and systems dedicated to the Arabic language. Many electronic databases, such as Google Scholar, PubMed, Web of Science, IEEE Xplore Digital Library, and CiteSeerx were visited for reviewing the English literature published between 1994 and 2018. Moreover, many different terms or keywords were used for achieving the search; these include "surveillance systems", "infectious disease systems", "event/internet-based surveillance systems", "Arabic surveillance systems", "syndromic surveillance", "biosurveillance", and "Arabic text mining systems".

IV. Disease Outbreak Surveillance Systems in Arabic

Some of the developed systems, such as Argus, BioCaster, GOARN, GPHIN, HealthMap, MedISys and PULS, MiTAP and ProMED are able to process texts written in languages other than English. As previously mentioned, the aim of this study is to investigate event-based surveillance systems that can process Arabic texts in the domain of disease outbreaks. Al-Mahmoud and Al-Razgan [23] conducted a systematic review of published works on Arabic text mining between 2002 and 2014. The review showed that the topics of the articles were limited to a few different domains, such as opinion mining, crime domain, social networks, Arabic Wikipedia, and Islamic studies. Therefore, it is expected that disease outbreaks in news reports written in Arabic have not been directly processed; i.e. the 5 developed systems that support the Arabic language in Table I use translation tools to translate Arabic online news reports of disease outbreaks into English in order to process them for identifying the desired information. Further investigation on these systems can be seen below.

1) ProMED-mail
   Monitoring Emerging Diseases (ProMED-mail) is a multilingual early warning system of emerging disease outbreaks developed in 1994 by Hugh-Jones [24], [25]. It monitors human, plant and animal diseases worldwide. Moreover, some surveillance systems depend on health warning reports produced by ProMED-mail, such as Argus, BioCaster and HealthMap. This system is available to the public and no subscription fees are required. The ProMED-mail's source of data depends on reports obtained from its subscribers. Currently, there are more than 70,000 subscribers from over 185 countries. With regard to the reports produced by the system, these are reviewed by a number of experts before dissemination [25], [51]. The system relies on its subscribers for performing the translation task. Although ProMED-mail utilizes translated Arabic news reports, disseminating information in Arabic is not provided.
2) The Global Public Health Intelligence Network (GPHIN)
   GPHIN is a multilingual internet-based system developed by Health Canada in collaboration with the World Health Organization (WHO). The system is able to collect public health reports from global media sources, such as newswires and websites on a real-time basis in order to monitor infectious disease outbreaks. As previously mentioned, GPHIN is a multilingual system supporting eight languages (English, Chinese, Spanish, Portuguese, Russian, Arabic, French and Farsi) to monitor disease outbreaks by using machine translation to translate non-English reports into English, and vice versa. According to Wang and Barry [52], the GPHIN team supports the Indonesian language. Moreover, not only can GPHIN track events related to disease outbreaks or infectious diseases but it can also track other events, such as animal diseases, chemical incidents, plant diseases, and contaminated food and water. Most of the WHO information is provided by GPHIN. Furthermore, the Centers for Disease Control and Prevention (CDC), and the Food and Agriculture Organization of the United Nations (FAO) use GPHIN on a daily basis [12]. However, GPHIN is not free and official organizations must pay to subscribe. It also presents

TABLE I. DIFFERENT DEVELOPED TYPES OF EVENT-BASED SYSTEMS

No	System	Country	Year	Supporting Arabic	Homepage	Victim type	Working
1	ProMED-mail [24], [25]	USA	1994	Yes	www.promedmail.org	Human, animal, plant	Yes 28th Sep 18
2	GPHIN [26]	Canada	1997	Yes	https://gphin.canada.ca/cepr/listarticles.jsp?language=en_CA	Human, animal and plant	Yes 28th Sep 18
3	EWRS [27]	Europe	1998	No	https://ewrs.ecdc.europa.eu	Human	Yes 28th Sep 18
4	EpiSimS [18]	USA	2000	No	http://www.lanl.gov/projects/mathematical-computational-epidemiology/agent-based-modeling.php	Human	Yes 28th Sep 18
5	GOARN [28]	USA	2000	No	https://extranet.who.int/goarn/	Human	Yes 28th Sep 18
6	MiTAP [29]	USA	2001	No	http://mitap.sdsu.edu	Human	No 28th Sep 18
7	Proteus-BIO [30], [31]	USA	2002	No	Not available	Human, animal and plant	————
8	Argus [32]	USA	2004	Yes	www.biodefense.georgetown.edu	Human, animal, plant	No 28th Sep 18
9	MedISys and PULS [33]	Europe	2004	Yes	http://medisys.newsbrief.eu	Human, animal and plant	Yes 28th Sep 18
10	BioCaster [34]	Japan	2006	No	http://born.nii.ac.jp/ OR http://www.biocaster.org/	Human, animal, plant	No 28th Sep 18
11	EpiSPIDER [35]	USA	2006	No	www.epispider.org	Human, animal, plant	No 28th Sep 18
12	GODSN [36]	USA	2006	No	Not available	Human	————
13	HealthMap [37]	USA	2006	Yes	www.healthmap.org	Human, animal, plant	Yes 28th Sep 18
14	InSTEDD [38]	USA	2006	No	http://instedd.org	Human	Yes 28th Sep 18
15	Google Flu Trends [19]	USA	2008	No	www.google.org/flutrends "No longer publishing current estimates of Flu and Dengue fever based on search patterns." [39]	Human	Yes 6th Oct 17
16	Influenzanet [21]	Europe	2008	No	www.influenzanet.eu	Human	Yes 28th Sep 18
17	Animal disease-related event recognition system [40]	USA	2010	No	Not available	Animal	————
18	Automatic online news monitoring and classification [41]	USA	2010	No	Not available	Human and animal	————
19	GET WELL [20]	Sweden	2010	No	www.smittskyddsinstitutet.se	Human	Yes 28th Oct 18
20	CIDARS [42]	China	2011	No	Not available	Human, animal, food and waterborne	————
21	EpiCore [43]	USA	2013	No	https://epicore.org	Human and animal	28th Sep 18
22	Alshowaib's system [44]	KSA	2014	No	Not available	Human	————
23	Healthtweets [45]	USA	2014	No	www.healthtweets.org	Human	Yes 28th Sep 18
24	DESRM [46]	Vietnam	2015	No	Not available	Human	————
25	Flutrack [47]	Greece	2015	No	www.flutrack.org	Human	Yes 28th Sep 18
26	Online Diagnostic System [48]	Nigeria	2017	No	Not available	Human	————
27	ARGO [49]	Mexico, Brazil, Thailand, Singapore and Taiwan	2017	No	Not available	Human	————
28	PADI-web[50]	France	2018	No	http://epia.clermont.inra.fr/	Animal	Yes 28th Sep 18

information in English and French languages [26], [53].

3) A Global Detection and Tracking System for Biological Events (Argus)
Argus is an event-based system developed at the Georgetown University Medical Center; its aim is to provide early warning alerts through detecting and tracking global biological events that might affect human, plant, and animal health. It is able to cover multilingual data (40 languages) through a multilingual analytic team. For biological event detection, the system relies on a taxonomy of nearly 200 indicators. For assessing biological event evolution, a heuristic staging model called the Wilson-Collmann Scale is proposed. Moreover, Argus can monitor the spread of 130 infectious disease outbreaks covering 175 countries [32]. Argus depends on official and unofficial disease reports generated from WHO and ProMed, respectively, as indicators of possible biological events [54].

4) MedISys and PULS
The Medical Information System (MedISys) is part of the Europe Media Monitor (EMM) software family developed at the Joint Research Centre of the European Commission (JRC). It is an automatic multilingual early warning system used to identify potential public health threats, such as communicable diseases, bioterrorism, and radiological, nuclear and chemical incidents. In the medical domain, the system relies on predefined keywords for gathering textual reports from several web-based sources in various languages. It monitors around 1400 medical websites, 3750 generic news portals, 20 commercial newswires, and more than 10000 Really Simple Syndication (RSS) feeds in 60 languages [33]. MedISys issues automatic notification by sending SMS and emails to the end users. Moreover, the latest trends about diseases can be seen through MedISys's web interface. With regard to extracting disease-related information, the Pattern-based Understanding and Learning System (PULS) is utilized [55]. PULS extracts the disease name, number of victims, and their conditions, location and date. According to Agheneza [12], PULS is only able to process reports in the English language. MedISys provides three types of access levels [55]:
   - Free access for public
   - The access level is restricted for public health professionals outside the European Commission (EC)
   - Full access inside the EC

   MedISys and PULS website interface supports Arabic via the link http://medisys.newsbrief.eu/medisys/homeedition/ar/home.html. However, not all information provided is in Arabic and sometimes Google translate is used for translation some none-Arabic news reports.

5) HealthMap
HealthMap is a multilingual automated real-time web-based surveillance system. This system is free and is publicly available [37]. It can also be browsed in 7 languages: English, French, Portuguese, Russian, Chinese, Arabic and Spanish. Similarly, HealthMap uses a translation engine to handle non-English articles. HealthMap comprises five components as follows: data gathering from diverse online sources: newswires, Really Simple Syndication (RSS) feeds, ProMED Mail, and WHO Classification, Database, Web Backend and Web Frontend [56]. The system's tools are Linux, Apache, MySQL and PHP. It also utilizes free services provided by other developers, such as Google Translate API, Google Maps, GoogleMap API for PHP and xajax PHP AJAX library.

In this review, few systems were found in the literature that are able to directly process Arabic health data, i.e. without using translation engines. One such system, the named entity recognition system, NAMERAMA, has been developed to identify disease related-information such as diagnosis methods, symptoms, disease names and treatment methods from textual reports in the Arabic medical domain [57]. This system uses Bayesian Belief Networks (BBN) to extract aforementioned entities and is comprised of two stages: the first is the processing stage, which includes preprocessing, data analysis and feature extraction; the second stage is based on BBN for performing the classification task. The AMIRA tool is used for applying Part of Speech (POS) to the data, and an annotated corpus is used to evaluate the proposed system. However, only 27 articles were used for evaluating the performance of the NAMERAMA system, which is considered a very small dataset. In addition, this system only focused on identifying cancer disease-related information, i.e. other types of diseases were not covered. Table II lists the evaluation results.

In [58], two methods for identifying and extracting medical terms from the Arabic medical corpus were proposed. Their

TABLE II. THE SYSTEM EVALUATION RESULTS

Measure	Disease entity	Treatment method entity	Diagnosis method	Symptom categories
Precision	96.60%	69.33%	84.91%	71.34%
Recall	90.79%	70.99%	53.36%	49.34%
F-measure	93.60%	70.15%	65.53%	58.33%

work forms part of the Multimedica project, funded by the Spanish Ministry of Science and Innovation. The aim of the project is the development of multilingual resources and tools that include the Spanish, Arabic, and Japanese languages to process published reports by news agencies in the health domain. The first proposed approach is based on a gazetteer that contains 3473 Arabic medical terms. The terms used are translated from English medical terms resources (SNOMED and UMLS) using Google translator. In contrast, the second approach uses 410 Arabic terms that are the equivalents of Latin prefixes and suffixes commonly used in the medical and health domain. The evaluation results show that the first approach achieved 100% accuracy and outperformed the second approach; however, it only achieved 54% recall, which is relatively low.

With regard to infectious disease outbreaks, Alruily et al. [59] presented a preliminary work on developing a web-based surveillance system to track infectious diseases by extracting disease-related information from Arabic news textual reports. However, no results were reported because it was a foundational study.

V. LATEST DISEASE OUTBREAK SURVEILLANCE SYSTEMS

As can be seen in Table I, the most recent system to be developed is that of Alshowaib [44], who used a rule-based approach to extract disease outbreak-related information, i.e. named entities, such as disease name, date and location, location of the reporting authority, and outbreak incident. The rules were created based on analysis of textual disease outbreak reports. This system is solely dedicated to the English language. The performance of this system can be seen in the following Table III.

TABLE III. THE SYSTEM EVALUATION RESULTS

Measure	Entities	Relations	Events
Precision	1.00	0.70	0.90
Recall	0.75	0.70	0.80
F-measure	0.88	0.70	0.85

Nguyen and Nguyen [46], [60] developed the Disease Extraction System for Real-time Monitoring (DESRM). This system is used for Vietnamese online news. The approach used for performing this task depends on semantic rules and machine learning to extract infectious disease events. DESRM consists of two components: disease event identification from textual data, and disease event information extraction. For identifying phrases and detecting disease events, semantic rules and machine learning (maximum entropy model) are used.
For extracting related information of disease events: time, disease name, and place in the second component, Name Entity Recognition (NER) rules and dictionary are utilized. Table IV and Table V present the performance of the system.

TABLE IV. DESRM CLASSIFICATION EVALUATION

Approach	Precision	Recall	F-measure
Semantic rules & machine learning	75.07	79.76	77.33
Baseline (using only machine learning)	72.35	77.84	74.97

TABLE V. DESRM EVENT EXTRACTION EVALUATION

Approach	Precision	Recall	F-measure
Rules & NER	89.47	94.44	91.89
Baseline (using Rules)	83.55	92.02	87.58

The China Infectious Diseases Automated-Alert and Response System (CIDARS) was developed in 2008 by the Chinese Center for Disease Control and Prevention [42]. CIDARS uses three early warning methods: spatial-temporal model, temporal model and fixed threshold detection method [61]. Although CIDARS is able to detect signals of infectious disease, many false positive signals are produced [62]. Investigations performed in 2017 on surveillance and early warning systems of infectious disease developed between 2012 and 2016 in China can be seen in [63].

The EpiCore global surveillance project was established by the International Society for Infectious Diseases, the Skoll Global Threats Fund, HealthMap, the Program for Monitoring Emerging Diseases (ProMED-mail) and the Public Health Interventions Network (TEPHINET) in 2013 [43], [64]. The EpiCore system is an online platform used to verify informal health alert reports related to potential disease outbreaks by health experts. Moreover, a web-based diagnostic with epidemic alerts was proposed by Okokpujie et al. [48]. It is also able to prescribe medications based on symptoms and is used for issuing alerts about the outbreak of epidemic diseases. This system relies on data provided by the users. Hyper Text Mark-up Language, Cascading Style Sheets, Javascript, Ajax, PHP, MySQL were utilized for developing this system. For analyzing the collected data, a medical diagnostic engine called Infermedica was used.

Moreover, a web-based diagnostic with epidemic alert was proposed by Okokpujie et al. [48]. Also, it is able to prescribe medication based on the symptoms and issuing alert about outbreak of epidemic diseases. This system relies on data provided by the users. Hyper Text Mark-up Language, Cascading Style Sheets, Javascript, Ajax , PHP , MySQL were utilized for developing this system. For analyzing the collected data a medical diagnostic engine called Infermedica was used.

Arsevska et al. [50] developed the Platform for Automated Extraction of Disease Information from the web (PADI-web) to monitor infectious animal diseases. It uses data collected from Google News to extract epidemic disease related-information, such as number of victims, dates and locations. The information extraction process relies on rule-based techniques of data mining. For evaluating the performance of PADI-web, 352 news reports were used, achieving F-scores 95% of the diseases, 85% of the number of cases, 83% of dates and 80% of locations, respectively.

Osaghae et al. [65] proposed a web-based grassroots epidemic alert system. However, this type of system is a passive surveillance system, as defined by WHO, because it relies on official data collected from official places, such as health centers, hospitals and registered laboratories. Therefore, it is not covered in this review.

Several existing systems have been developed for specific disease outbreaks, such as influenza epidemics, but these are limited to a specific data source, e.g. Online Social Networks (OSN), such as Twitter. For example, Elhadad et al. [66] investigated social media in order to extract information about food-borne disease outbreaks by monitoring restaurants in New York City. A prototype was developed based on supervised machine learning to detect the review comments or discussions about a food poisoning incident, or the people affected by the incident. The system yielded high results; however, no specific numbers relating to the results were reported. Furthermore, the system works only on Yelp data collected from the Yelp website. Additionally, Talvis et al. [47], 2014, proposed the flutrack system (http://flutrack.org) for monitoring the spread of influenza epidemics. Every 20 minutes, the system gathers tweets written in English using the Twitter API to detect potential flu outbreaks. In other words, the aim of this system is to track and visualize influenza epidemics in real time. The list of searching tags, namely, influenza, flu, chills, headache, sore throat, runny nose, sneezing, fever, and dry cough were used to extract and track flu-related tweets. The system was evaluated and achieved an accuracy of 92%. Related works of systems proposed for epidemics of seasonal influenza can be seen in the Google Flu Trends system [19], MappyHealth application [67], Tracking Flu Infections on Twitter [68], detecting influenza epidemics by analyzing Twitter messages [69], HealthTweets.org: a Platform for Public Health Surveillance Using Twitter [45], and the ARGO system for monitoring dengue fever epidemics [49]. For further reading on these types of systems, a systematic review of the literature conducted on proposed systems published between 2004 and 2015 that detect and track a pandemic using online social networks was published in 2016 by [70]. In addition, the most recent review was presented by Pollett et al. [71] in 2017 for evaluating the internet-based biosurveillance performance of diseases caused by bacteria, parasites and viruses. Furthermore, other important disease outbreak surveillance systems are listed below.

1) Proteus-BIO
   Grishman et al. [30], [31] developed the Proteus-BIO system for creating and automatically updating a database with information on infectious disease outbreaks. Proteus-BIO system consists of five phases:
   - Data gathering of online daily news and from medical sources, such as WHO, ProMed and a medical forum
   - Text zoner, based on rule-based techniques
   - Information extraction engine (multi-phase finite-state transducer) to extract this information: disease name, date, location and number, type, and status of victims
   - Interface to a relational database
   - Web-based database browser connected to the original texts

   The technique used for the information extraction task is based on a multiphase finite-state transducer that includes tokenization, lexical lookup, finite-state pattern matching, and noun phrase and clause. The performance of this system achieved P= 0.79, recall 0.41, and F-measure 0.54.
2) BioCaster
   BioCaster is an automated web service that monitors global online media for detecting infectious disease outbreaks [34], [72]. The system is able to process 1700 Really Simple Syndication (RSS) feeds from different sources: the World Health Organization (WHO) outbreak reports, ProMED-mail, Google News, and the European Media Monitor. The system is limited to seven languages: English, Japanese, Chinese, Spanish, Thai, Vietnamese and French; it comprises four components: event recognition, topic classification, disease/location detection and named entity recognition. In addition, the visualization is provided in this system to the user by plotting extracted information on a Google map. BioCaster has been evaluated on a gold standard corpus of annotated news articles; for all named entity classes the system achieved an F-measure of 76.97. With regard to topic classification performance, the system was able to achieve 0.89 precision, 0.97 recall, and F-measure 0.93. However, BioCaster depends on ontology, which is a limitation in the system, as it is unable to identify new diseases or locations.
3) Automatic online news monitoring and classification for syndromic surveillance
   Zhang et al. [41] developed an automatic online news monitoring and classification system for syndromic surveillance on infectious disease. The system consists of three components:
   - Data collection
     Programs of web crawler are used to collate online news of infectious disease.
   - Data representation and feature selection
     Three methods of document representation: Bag of Words, Noun Phrases, and Named Entities, are used for transforming the full text to document vectors. Regarding the feature selection implementation, Correlation-based Feature Selection (CFS) was chosen.
   - Classification and evaluation
     For performing classification tasks, learn Bayes net (LBN), K-nearest neighbour (KNN), Support Vector Machine (SVM) and Naïve Bayesian (NB) are used. SVM achieved the best performance in the evaluation results.

VI. Results and Discussion

In general, the aim of this review was to examine existing event-based surveillance systems focusing on finding systems developed for the Arabic language. It found that 5 systems supported Arabic in different ways but were originally created for the English language. These systems, such as MedISys and PULS, GPHIN and HealthMap use translator engines to translate from non-English to English. However, a small number of systems have been developed for several specific languages other than English, such as Swedish and Vietnamese. With regard to the Arabic language, however, this review shows that no existing surveillance system for monitoring infectious disease outbreaks has yet been created to directly process Arabic texts. Processing texts to identify certain entities depending on translation is not sufficient; indeed, translation may in fact lead to identifying incorrect information. As Agheneza claimed [12], most event-based surveillance systems are based

in the USA and Europe, and a few systems are based in Asia. However, only one system was found in the area of the Middle East and North Africa (MENA); this was developed by Alshowaib [44] but to date is not available online and developed to handle English health news reports.

Arabic is a Semitic language with a very complex morphology, as it is highly inflectional, and therefore, dealing with texts written in Arabic is highly complicated. Arabic is comprised of 28 letters (3 vowels and 25 consonants) that are used to form words. For correct pronunciation, the diacritical marks are used and are placed around the letters. Arabic can appear in different forms: Modern Standard Arabic (MSA), dialectal Arabic, and classical Arabic [73]. Ibrahim et al. [74] claimed that the main dialects in the Arabic word are: Gulf, Iraqi, Moroccan, Levantine, Yemeni, and Egyptian.

Furthermore, the Arabic medical domain faces a number of challenges. For instance, diglossia is common in the health community in Arabic countries [58]. According to Samy et al. [58], the English language is the primary teaching language in Egypt, Iraq, Jordan, and the Arabic Gulf countries for teaching all health courses at universities whereas French is used at universities in the North African Arab countries. Moreover, the communication languages used by health workers, either verbal or written, are English and French, even for patients prescriptions or their health reports.

VII. Conclusion

The aim of this review is to examine existing event-based surveillance systems focusing on finding systems developed for the Arabic language. This review shows that no existing early warning surveillance system for monitoring outbreaks of infectious diseases has yet been created to directly process Arabic texts, i.e, without using translator tools. This result might be due to some difficulties in the Arabic language itself and in the medical domain in particular. However, other event-based surveillance systems developed for detecting and tracking disease outbreaks are presented, and the components and performance of these systems are discussed. 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.

References

[1] University of Washington. (2015) Modeling the spread of ebola. https://sites.math.washington.edu/~morrow/mcm/mcm15/38725paper.pdf.

[2] Pan American Health Organization / World Health Organization (PAHO/WHO). (2014) Ebola virus disease (evd), implications of introduction in the americas. http://www.paho.org/hq/index.php.

[3] E. Velasco, T. Agheneza, K. Denecke, G. R. Kirchner, and T. Eckmanns, "Social media and internet-based data in global systems for public health surveillance: A systematic review," *Milbank Quart*, vol. 92, pp. 7–33, 2014.

[4] Centers of Disease Control and Prevention (CDC). (2018) Middle east respiratory syndrome (mers). https://www.cdc.gov/coronavirus/mers/.

[5] J. Choi, Y. Cho, E. Shim, and H. Woo, "Web-based infectious disease surveillance systems and public health perspectives: a systematic review," *BMC Public Health*, vol. 16, no. 1, p. 1238, 2016.

[6] E. Christaki, "New technologies in predicting, preventing and controlling emerging infectious diseases," *Virulence*, vol. 6, no. 6, pp. 558–565, 2015, pMID: 26068569.

[7] N. Collier and S. Doan, "Geni-db: a database of global events for epidemic intelligence," *Bioinformatics*, vol. 28, no. 8, p. 1186, 2012.

[8] World Health Organization. (2018, Aug.) National passive surveillance. http://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/passive/en/.

[9] B. Ramalingam. (2016) Real-time monitoring in disease outbreaks: Strengths, weaknesses and future potential. http://opendocs.ids.ac.uk/opendocs/handle/123456789/9940.

[10] World Health Organization. (2008) A guide to establishing event-based surveillance. http://www.wpro.who.int/emergingdiseases/documents/eventbasedsurv/en/.

[11] M. Blench, H. Tolentino, C. C. Freifeld, K. D. Mandl, A. Mawudeku, G. Eysenbach, and J. S. Brownstein, "Use of unstructured event-based reports for global infectious disease surveillance," *Emerg Infect Dis*, vol. 15, no. 5, May 2009.

[12] T. Agheneza, "A systematic literature review on event-based public health surveillance systems," Master's thesis, Hamburg University of Applied Sciences, 2011.

[13] K. L. Tsui, D. Goldsman, W. Jiang, and S. Y. Wong, "Recent research in public health surveillance and health management," in *2010 Prognostics and System Health Management Conference*, 2010, pp. 1–7.

[14] K.-L. Tsui, W. Chiu, P. Gierlich, D. Goldsman, X. Liu, and T. Maschek, "A review of healthcare, public health, and syndromic surveillance," *Quality Engineering*, vol. 20, no. 4, pp. 435–450, 2008.

[15] F. Azzedin, J. Yazdani, S. Adam, and M. Ghaleb, "A generic model for disease outbreak notification systems," *International Journal of Computer Science & Information Technology*, vol. 6, no. 4, p. 137, 2014.

[16] C. Abat, H. Chaudet, J.-M. Rolain, P. Colson, and D. Raoult, "Traditional and syndromic surveillance of infectious diseases and pathogens," *International Journal of Infectious Diseases*, vol. 48, pp. 22 – 28, 2016.

[17] H. Bagherian, M. Farahbakhsh, R. Rabiei, H. Moghaddasi, and F. Asadi, "National communicable disease surveillance system: A review on information and organizational structures in developed countries," *Acta Informatica Medica*, vol. 25, no. 4, p. 271, 2017.

[18] S. D. Valle. (2018, Aug.) Agent-based modeling. http://www.lanl.gov/projects/mathematical-computational-epidemiology/agent-based-modeling.php.

[19] G. Jeremy, H. M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, p. 1012, 2009.

[20] A. Hulth and G. Rydevik, "Get well: an automated surveillance system for gaining new epidemiological knowledge," *BMC Public Health*, vol. 11, no. 1, p. 252, 2011.

[21] C. E. Koppeschaar, V. Colizza, C. Guerrisi, C. Turbelin, J. Duggan, W. J. Edmunds, C. Kjelsø, R. Mexia, Y. Moreno, S. Meloni *et al.*, "Influenzanet: Citizens among 10 countries collaborating to monitor influenza in europe," *JMIR public health and surveillance*, vol. 3, no. 3, 2017.

[22] S. Yan, A. Chughtai, and C. Macintyre, "Utility and potential of rapid epidemic intelligence from internet-based sources," *International Journal of Infectious Diseases*, vol. 63, pp. 77–87, 2017.

[23] H. Al-Mahmoud and M. Al-Razgan, "Arabic text mining a systematic review of the published literature 2002-2014," in *Cloud Computing (ICCC), 2015 International Conference on*, April 2015, pp. 1–7.

[24] M. Hugh-Jones, "Global awareness of disease outbreaks: The experience of promed-mail," *Public Health Reports*, vol. 116, no. 2, pp. 27–31, 2001.

[25] J. Woodall, "Stalking the next epidemic: Promed tracks emerging diseases," *Public Health Rep*, vol. 112, no. 1, pp. 78–82., 1997.

[26] A. Mawudeku and M. Blench, "Global public health intelligence network (gphin)," in *7th Conference of the Association for Machine Translation in the Americas*, 2006.

[27] P. Guglielmetti, D. Coulombier, G. Thinus, V. F. Loock, and S. Schreck, "The early warning and reponse system for communicable diseases in the eu: an overview from 1999 to 2005," *Euro Surveill*, vol. 11, no. 12, pp. 7–8, 2006.

[28] D. L. Heymann and G. R. Rodier, "Hot spots in a wired world: Who surveillance of emerging and re-emerging infectious diseases," *Lancet Infect Dis*, vol. 1, no. 5, pp. 345–353, 2001.

[29] L. Damianos, J. Ponte, S. Wohlever, F. Reeder, D. Day, G. Wilson, and L. Hirschman, "Mitap text and audio processing for bio-security: A case study," in *Proceedings of the 14th Conference on Innovative Applications of Artificial Intelligence - Volume 1*, ser. IAAI'02. AAAI Press, 2002, pp. 807–814.

[30] R. Grishman, S. Huttunen, and R. Yangarber, "Real-time event extraction for infectious disease outbreaks," in *Proceedings of the Second International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 366–369.

[31] ——, "Information extraction for enhanced access to disease outbreak reports," *J. of Biomedical Informatics*, vol. 35, no. 4, pp. 236–246, 2002.

[32] J. M. Wilson, "Argus: a global detection and tracking system for biological events," *Advances in Disease Surveillance*, vol. 4, no. 21, 2007.

[33] European Commission's Joint Research Centre. (2018, Sep.) Medical information system (medisys). http://emm.newsbrief.eu/overview.html.

[34] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi, "Biocaster: detecting public health rumors with a web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, 2008.

[35] H. Tolentino, R. Kamadjeu, P. Fontelo, F. Liu, M. Pollack, and L. Madoff, "Scanning the emerging infectious diseases horizon - visualizing promed emails using epispider," in *Advances in Disease Surveillance*, 2007.

[36] S. A. Khan, C. O. Patel, and R. Kukafka, "Godsn: Global news driven disease outbreak and surveillance," in *AMIA Annual Symposium Proceedings*. American Medical Informatics Association, 2006, p. 983.

[37] J. S. Brownstein and C. C. Freifeld, "Healthmap: the development of automated real-time internet surveillance for epidemic intelligence," *Eurosurveillance*, vol. 12, no. 48, 2007.

[38] T. A. Kass-Hout and di Tada Nicolas, "International system for total early disease detection (instedd) platform," *Advances in Disease Surveillance*, vol. 5, no. 2, p. 108, 2008.

[39] Google. (2018, Sep.) Google flu trends. https://www.google.org/flutrends/about/.

[40] V. Svitlana and H. H. William, "Computational knowledge and information management in veterinary epidemiology," in *IEEE International Conference on Intelligence and Security Informatics, ISI 2010, Vancouver, BC, Canada*, 2010, pp. 120–125.

[41] Y. Zhang, Y. Dang, H. Chen, M. Thurmond, and C. Larson, "Automatic online news monitoring and classification for syndromic surveillance," *Decision Support Systems*, vol. 47, no. 4, pp. 508–517, 2009.

[42] W. Yang, Z. Li, Y. Lan, J. Wang, J. Ma, L. Jin, Q. Sun, W. Lv, S. Lai, Y. Liao *et al.*, "A nationwide web-based automated system for early outbreak detection and rapid response in china," *Western Pacific Surveillance and Response*, vol. 2, no. 1, 2011.

[43] T. S. Lorthe, M. P. Pollack, B. Lassmann, J. S. Brownstein, E. Cohn, N. Divi, D. J. Herrera-Guibert, J. Olsen, M. S. Smolinski, and L. C. Madoff, "Evaluation of the epicore outbreak verification system," *Bulletin of the World Health Organization*, vol. 96, no. 5, p. 327, 2018.

[44] W. N. Alshowaib, "Rule-based information extraction from disease outbreak reports," *International Journal of Computational Linguistics (IJCL)*, vol. 5, pp. 37–58, 2014.

[45] M. Dredze, R. Cheng, M. J. Paul, and D. Broniatowski, "Healthtweets.org: A platform for public health surveillance using twitter," in *AAAI Workshop on the World Wide Web and Public Health Intelligence*, 2014, pp. 593–596.

[46] M.-T. Nguyen and T.-T. Nguyen, "Desrm: A disease extraction system for real-time monitoring," *Int. J. Comput. Vision Robot.*, vol. 5, no. 3, pp. 282–301, 2015.

[47] K. Talvis, K. Chorianopoulos, and K. L. Kermanidis, "Real-time monitoring of flu epidemics through linguistic and statistical analysis of twitter messages," in *2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization*, 2014, pp. 83–87.

[48] K. Okokpujie, A. Orimogunje, E. Noma-Osaghae, and O. Alashiri, "An intelligent online diagnostic system with epidemic alert," *International Journal of Innovative Science and Research Technology*, vol. 2, no. 9, pp. 327–331, October 2017.

[49] S. Yang, S. C. Kou, F. Lu, J. S. Brownstein, N. Brooke, and M. Santillana, "Advances in using internet searches to track dengue," *PLoS computational biology*, vol. 13, no. 7, p. e1005607, 2017.

[50] R. Goel, S. Fadloun, S. Valentin, A. Sallaberry, M. Roche, and P. Poncelet, "Epidnews: An epidemiological news explorer for monitoring animal diseases," in *Proceedings of the 11th International Symposium on Visual Information Communication and Interaction*, ser. VINCI '18. New York, NY, USA: ACM, 2018, pp. 1–8.

[51] ProMED. (2018, Sep.) About promed-mail. http://www.promedmail.org/aboutus/.

[52] Y. AndreaWang and M. Barry, "Making online outbreak surveillance work for all," *Annals of global health*, vol. 83, no. 3-4, 2017.

[53] World Health Organization. (2018, Sep.) Epidemic intelligence - systematic event detection. http://www.who.int/csr/alertresponse/epidemicintelligence/en/.

[54] H. Chen, D. Zeng, and P. Yan, *Argus*. Boston, MA: Springer US, 2010, pp. 177–181.

[55] S. RALF, F. FLAVIO, V. D. G. ERIK, B. Clive, V. E. Peter, and Y. Roman, *Text Mining from the Web for Medical Intelligence*. IOS Press, 2008.

[56] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports," *Journal of the American Medical Informatics Association*, vol. 15, no. 2, pp. 150–157, 2008.

[57] S. Alanazi, "A named entity recognition system applied to arabic text in the medical domain," Ph.D. dissertation, Staffordshire University, 2017.

[58] D. Samy, A. M. Sandoval, C. B. Diaz, M. G. Salazar, and J. M. Guirao, "Medical term extraction in an arabic medical corpus," in *Proceedings of the 8th Language Resources and Evaluation Conference*, 2012.

[59] M. Alruily, N. Hammami, and M. Goudjil, "Arabic web-based surveillance system for monitoring infectious disease outbreaks," *2015 Internet Technologies and Applications (ITA)*, pp. 358–360, 2015.

[60] M.-T. Nguyen and T.-T. Nguyen, "Extraction of disease events for a real-time monitoring system," in *Proceedings of the Fourth Symposium on Information and Communication Technology*, ser. SoICT '13. New York, NY, USA: ACM, 2013, pp. 139–147.

[61] W. Yang, Z. Li, Y. Lan, J. Ma, L. Jin, S. Lai, Y. Liao, W. Lv, Q. Sun, and J. Wang, "Chapter 7 - china infectious diseases automated-alert and response system (cidars)," in *Early Warning for Infectious Disease Outbreak*, W. Yang, Ed. Academic Press, 2017, pp. 133 – 161.

[62] R. Wang, Y. Jiang, X. Guo, Y. Wu, and G. Zhao, "Influence of infectious disease seasonality on the performance of the outbreak detection algorithm in the china infectious disease automated-alert and response system," *Journal of International Medical Research*, vol. 46, no. 1, pp. 98–106, 2018, pMID: 28728470.

[63] H. Zhang, L. Wang, S. Lai, Z. Li, Q. Sun, and P. Zhang, "Surveillance and early warning systems of infectious disease in china: From 2012 to 2014," *The International journal of health planning and management*, vol. 32, no. 3, pp. 329–338, 2017.

[64] Z. Haddad, L. Madoff, E. Cohn, J. Olsen, A. Crawley, J. Brownstein, M. Smolinski, J. Shao, M. Pollack, and D. Herrera-Guibert, "The epicore project: Using innovative surveillance methods to verify outbreaks of emerging infectious diseases," *International Journal of Infectious Diseases*, vol. 45, p. 19, 2016.

[65] E. N. Osaghae, K. Okokpujie, C. Ndujiuba, O. Okesola, and I. P. Okokpujie, "Epidemic alert system: A web-based grassroots model," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 3809–3828, 2018.

[66] N. Elhadad, L. Gravano, D. Hsu, S. Balter, V. Reddy, and H. Waechter, "Information extraction from social media for public health," in *KDD at Bloomberg Workshop, Data Frameworks Track (KDD 2014)*, 2014.

[67] B. Norris, C. Boicey, and M. Silverberg. (2012) Mappyhealth application. http://mappyhealth.com.

[68] A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: Tracking flu infections on twitter," in *NAACL*, 2013.

[69] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA 10. New York, NY, USA: ACM, 2010, pp. 115–122.

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

[71] S. Pollett, B. M. Althouse, B. Forshey, G. W. Rutherford, and R. G. Jarman, "Internet-based biosurveillance methods for vector-borne diseases: Are they novel public health tools or just novelties?" *PLoS neglected tropical diseases*, vol. 11, no. 11, p. e0005871, 2017.

[72] N. Collier, "What's unusual in online disease outbreak news?" *Journal of Biomedical Semantics*, vol. 1, no. 1, pp. 1–18, 2010.

[73] N. Boudad, R. Faizi, O. h. t. Rachid, and R. Chiheb, "Sentiment analysis in arabic: A review of the literature," *Ain Shams Engineering Journal*, 2017.

[74] H. S. Ibrahim, S. M. Abdou, and M. Gheith, "Idioms-proverbs lexicon for modern standard arabic and colloquial sentiment analysis," *International Journal on Natural Language Computing*, vol. 4, 2015.