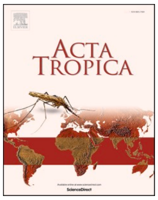
Contents lists available at ScienceDirect

Acta Tropica

journal homepage: www.elsevier.com/locate/actatropica



Text mining in mosquito-borne disease: A systematic review

Song-Quan Ong [a,*], Maisarah Binti Mohamed Pauzi [b], Keng Hoon Gan [b]

[a] *Institute for Tropical Biology and Conservation, Universiti Malaysia Sabah, Jalan UMS, Kota Kinabalu, Sabah 88400, Malaysia*
[b] *School of Computer Sciences, Universiti Sains Malaysia, Penang 11800, Malaysia*

ARTICLE INFO

Keywords:
Mosquito-borne
Vector-borne
Text analysis
Dengue
Malaria
Zika

ABSTRACT

Mosquito-borne diseases are emerging and re-emerging across the globe, especially after the COVID19 pandemic. The recent advances in text mining in infectious diseases hold the potential of providing timely access to explicit and implicit associations among information in the text. In the past few years, the availability of online text data in the form of unstructured or semi-structured text with rich content of information from this domain enables many studies to provide solutions in this area, e.g., disease-related knowledge discovery, disease surveillance, early detection system, etc. However, a recent review of text mining in the domain of mosquito-borne disease was not available to the best of our knowledge. In this review, we survey the recent works in the text mining techniques used in combating mosquito-borne diseases. We highlight the corpus sources, technologies, applications, and the challenges faced by the studies, followed by the possible future directions that can be taken further in this domain. We present a bibliometric analysis of the 294 scientific articles that have been published in Scopus and PubMed in the domain of text mining in mosquito-borne diseases, from the year 2016 to 2021. The papers were further filtered and reviewed based on the techniques used to analyze the text related to mosquito-borne diseases. Based on the corpus of 158 selected articles, we found 27 of the articles were relevant and used text mining in mosquito-borne diseases. These articles covered the majority of Zika (38.70%), Dengue (32.26%), and Malaria (29.03%), with extremely low numbers or none of the other crucial mosquito-borne diseases like chikungunya, yellow fever, West Nile fever. Twitter was the dominant corpus resource to perform text mining in mosquito-borne diseases, followed by PubMed and LexisNexis databases. Sentiment analysis was the most popular technique of text mining to understand the discourse of the disease and followed by information extraction, which dependency relation and co-occurrence-based approach to extract relations and events. Surveillance was the main usage of most of the reviewed studies and followed by treatment, which focused on the drug-disease or symptom-disease association. The advance in text mining could improve the management of mosquito-borne diseases. However, the technique and application posed many limitations and challenges, including biases like user authentication and language, real-world implementation, etc. We discussed the future direction which can be useful to expand this area and domain. This review paper contributes mainly as a library for text mining in mosquito-borne diseases and could further explore the system for other neglected diseases.

1. Introduction

The mosquito-borne disease is one of the central public health issues globally, which causing more than 700,000 deaths annually (WHO--World Health Organization, 2020). The recent advances in text mining technologies that have been used to prevent and predict the course of many public health issues are undoubtedly valuable (Simpson and Demner-Fushman, 2012) and they hold the promise of managing diseases in the coming future (Safdari et al., 2021). The rapid progress and growth in text mining are attributable to improvements in machine learning algorithms and the surge in the volume of text data available on the internet (Tran, 2019). Text mining (TM) is “the discovery and extraction of interesting, nontrivial knowledge from free or unstructured text” (Kao and Poteet, 2007). The state of text mining in public health is reviewed relatively regularly. However, numerous studies and surveys (Simpson and Demner-Fushman, 2012; Safdari et al., 2021; Kao and Poteet, 2007) strongly indicate that general-purpose text mining tools are not well suited for the mosquito-borne disease domain because it is highly specialized. In this context, Zweigenbaum et al. (2007) explained the importance of specific subject knowledge for annotating the correct

[*] Corresponding author.
E-mail address: songquan.ong@ums.edu.my (S.-Q. Ong).

https://doi.org/10.1016/j.actatropica.2022.106447
Received 18 February 2022; Received in revised form 31 March 2022; Accepted 1 April 2022
Available online 14 April 2022

entity in biomedical text. Cowell and Smith (2010) explored the role of named entities in the identification and classification of the epidemic of infectious diseases. They emphasized the role of ontology-based text mining system in the domain, such as the name for the mosquito that carried the virus. To tackle the issue of inconsistent entities in text mining related to mosquito-borne diseases, Rajapakse et al. (2008) developed an ontology-centric navigation infrastructure in an acquisition engine that standardizes multiple text resources on dengue and enables data integration and knowledge sharing.

Despite the restricted nature of the domain, text mining is of interest to both researchers and the public community. Some web-based tools, such as ProMed-mail (http://www.promedmail.org), facilitate open data resources access to the community. Such tools are manually curated systems that provide access to reports by public health experts. BioCaster is a public health surveillance system that detects biomedical rumors (Collier et al., 2008). Nevertheless, due to the specific goals of text mining in mosquito-borne diseases, public health personnel, epidemiologists, and entomologists are better positioned to define useful tasks. This is also supported by the work of Cohen and Hunter (2008), who note that the most fruitful method of biomedical text mining will require the efforts and the abilities of both epidemiologists and computational linguist.

These text mining techniques hold the potential to combat complicated mosquito-borne diseases such as malaria and dengue. For instance, Villanes et al. (2018) discovered a new cluster of dengue fever in India by using news media. On the other hand, Carlos et al. (2017) analyzed and predicted the outbreak of dengue in a shorter time by using a large volume of data from social media. In contrast, traditional syndromic-based surveillance of mosquito-borne disease approaches is mostly manual and involves multiple stages. This usually causes a delay in providing the data to public health personnel for decision-making (Lee et al., 2021). Recently, real-time or near real-time web-based data sources have allowed rapid responses in contrast to traditional surveillance systems (Newkirk et al., 2012; Neill, 2012) and have successfully contributed to infectious disease management and control. The incidences are detected rapidly by using web-based data form social media platforms. In such situations, the users' postings have led to a tremendous increase in public engagement. For mosquito-borne diseases, previous studies (Safdari et al., 2021; Ghani et al., 2022; García-Díaz et al., 2020; Saire, 2019a, 2019b) are focused on the early detection of diseases and mostly covered the technologies, applications, corpus sources, and their limitations in the general biomedical and public health domain. However, to the best of our knowledge, a recent review of text mining in the mosquito-borne disease domain is not available.

In this paper, we aim to review the available text mining techniques published in the domain of mosquito-borne disease, mainly in the past few years when the data mining technologies has flourished. We also discuss the future challenges and directions. Specifically, we evaluate a set of research questions that allows us to obtain the latest trends of text mining in mosquito-borne disease. Furthermore, we aim to obtain an overview of the recent technologies used for analyzing the data used by these systems. These research questions (RQs) are stated as follows:

RQ1: What are the most common sources of the corpus used for text mining in mosquito-borne disease?
RQ2: Which text mining techniques are commonly used among authors of research papers in text mining of mosquito-borne diseases?
RQ3: What are the usages of text mining in mosquito-borne disease?
RQ4: What are the limitations and challenges of text mining in mosquito-borne diseases?

The methodology for the selection of these articles is discussed in Section 2, and the results are reported in Section 3. We address RQ1 by discussing a few corpora that involve data collection and mining techniques in Section 3.1. In Section 3.2, we answer RQ2 by further reviewing a few main text mining techniques by using both literature review and articles selection by subject experts, although the actual work of text mining is much broader and diverse. Subsequently, in Section 3.3, we continue with the techniques applied in mosquito-borne disease management, which answers RQ3. To answer RQ4, Section 4 mainly discusses the limitations and challenges. Section 4.2 highlights the original contribution of this writing. It consists of a visualization of the trend and statistical analysis of the articles for the four questions, followed by a conclusion and future work in Section 5.

2. Article selection and methodology

We applied bibliometric methods and leveraged the Boolean logic search technique to the databases. We expand the keyword of "mosquito-borne disease" according to the World Health Organization (2020), which are "Malaria" (caused the most fatal diseases), "Dengue" (the most prevalent disease), "Zika" (recently emerged and outbreak), and four of the re-emerging diseases – "Chikungunya", "Yellow fever", "West Nile fever" and "Japanese encephalitis". The selection criteria for the research articles were set to incorporate papers published from 2016 to 2021 taking into account the Zika outbreak in 2015, the years with the most related publications. The scientific databases of PubMed and Scopus were explored to provide a comprehensive bibliography of research papers on text mining in mosquito-borne disease literature. The selection of databases was based on two main rationales: (1) the peer-reviewed process of the articles that required indexing in the database, therefore ensuring the quality and reliability of the selected articles. (2) PubMed was the main database for most of the publications of infectious diseases, including mosquito-borne diseases.

For Scopus, we advanced searched database for query: ((TITLE-ABS-KEY (text AND mining) AND TITLE-ABS-KEY (malaria)) AND PUBYEAR > 2015) OR ((TITLE-ABS-KEY (text AND mining) AND TITLE-ABS-KEY (dengue)) AND PUBYEAR > 2015) OR ((TITLE-ABS-KEY (text AND mining) AND TITLE-ABS-KEY (zika)) AND PUBYEAR > 2015) OR ((TITLE-ABS-KEY (text AND mining) AND TITLE-ABS-KEY (chikungunya)) AND PUBYEAR > 2015). Ultimately, 46 articles were extracted based on the search terms. PubMed, which accesses the MEDLINE database, was also searched for publications by searching this resource with the following query: ((text mining) AND (malaria)) OR ((text mining) AND (dengue)) OR ((text mining) AND (zika)). From this search we obtained 112 articles.

A total of 294 articles were thus identified in the initial query of the databases. We further filtered the articles by removing articles in conference proceedings from 2016 to 2021 and obtained a total of 158 articles. Each of the abstracts and titles of the articles was assessed independently by each author of this paper. If the abstract, title or both explained the text mining technique and the mosquito-borne disease, then we considered them for further research; otherwise, they were filtered out. According to this procedure, an article was selected only when all the authors agreed on its relevance to the text mining in mosquito-borne disease. The next step was to consider the articles that had utilized text mining in their methodologies, such as relation/event extraction, named entities recognition (NER), part of speech (POS), linguistic-based, or statistics-based. Fig. 2.1 describes the procedures followed in selecting the relevant articles for this study.

3. Results

3.1. Sources of corpus

3.1.1. Twitter

Twitter is a microblog text-based social media platform that is excellent for exploring public concerns and interests about mosquito-borne diseases. The platform is an effective source of rich information to conduct text mining to obtain valuable insights. In this review, Twitter is the most common corpus that is used to conduct sentiment analysis to understand the discourse and event on the topic of malaria (Lee et al., 2021; Boit and El-Gayar, 2020; Oyelade et al., 2021), dengue (Ghani et al., 2022; García-Díaz et al., 2020; Saire, 2019a, 2019b; Abulaish et al., 2019; Parwez et al., 2018), and Zika. (García-Díaz et al., 2020; Safarnejad et al., 2019; Jordan et al., 2018; Li et al., 2017; Li, 2020; Tripathy et al., 2017; Khatua and Khatua, 2016). It occupied 13/27 (48%) of the total selected publications. Most of the studies examine the public discourse and their association with the incidences of the disease or information extraction of drug-disease association. Boit and El-Gayar (2020) studied the public discourse with a supervised learning method, namely, the Crimson Hexagon's ReadME algorithm, for classification of the keywords extracted by SQL. They determined the categorized polarity of tweets and were able to lead to the recommendation of treatment of malaria. García-Díaz et al. (2020) and Saire (2019a) focused on the development of an intelligent indicator system from Twitter to detect Dengue incidence, which greatly assists in disease surveillance in India. Khatua and Khatua (2016) investigated the immediate and long-term effects of Zika outbreaks by both volumetric and word co-occurrence analysis and examined the trend of mild symptoms such as fever as the immediate effect and birth defects as the long-term effect of Zika. Other social media platforms, such as Facebook, are not common within our scope of the study and they occupy just 1% of the reviewed studies.

3.1.2. PubMed

PubMed accesses the MEDLINE database, which has historically been the gold standard of biomedical research for the past 127 years. This corpus has served as the common resource of text mining in the biomedical domain. It consists mainly of scientific articles from the research community in the form of text documents, in contrast to Twitter, which consists of text from the lay public. The task of accessing PubMed by using text mining is mainly focused on the association of the text, which can generate insights such as diagnosis (by the relation of symptoms) (Abulaish et al., 2019; Parwez et al., 2018) and treatment (by the relation of active ingredients of drugs and disease) (Zhao et al., 2016; Selvaraj and Periyasamy, 2016). Abulaish et al. (2019) used text processing such as part-of-speech (POS) tagging and grammatical dependency relations to build a system with disease symptom extraction and characterization from biomedical documents. Parwez et al. (2018) studied the dependency relations of the text to determine the disease-symptoms relationship for malaria and dengue and, subsequently, the relationship of a newly emerged symptom to the most relevant infectious disease. Li (2020) applied the literature-based discovery approach and supervised machine learning methods to predict previously unknown research links between medical subject headings (MeSH) terms in Zika and CRISPR research. Other platforms for scientific article databases used for text mining in mosquito-borne diseases are ProMED-mail (4%), University e-repository (Muchene and Safari, 2021) (4%), and Google/Bing search engine (4%). Some articles from the selected publication mine the information from news documents from the database of LexisNexis (11%) and Reuters Media (4%).

3.1.3. Other sources

Other corpora, such as news reported by a journalist, could be an excellent alternative to retrieve mosquito-borne disease related text. Zhang et al. (2020) used topic clustering on the text documents in the LexisNexis news database to obtain the topic model that can monitor the disease through time and across the different regions. LexisNexis also allows a user to conduct a Boolean logic search to investigate the frequency of news of dengue in the area, which indicates a high infection rate (Zhang et al., 2019). Other news corpora, such as Reuters media, may not be accessible publicly and therefore relatively uncommon to be used as a news text corpus.

3.2. Text mining techniques

Text mining is a multidisciplinary field based on information extraction, machine learning, linguistics-based, and statistics-based data mining (Tran, 2019). In more detail, the processes involved are tokenization, normalization, lemmatization, stemming, named entity recognition (NER), Part of Speech (POS) and then relation and event extraction, sentiment analysis, and usually followed by but not restricted to machine learning model build-up for classification or clustering. In this section, we review two techniques that are most interested in the studies of text mining in mosquito-borne diseases: sentiment analysis, which usually extracts and classifies the comments' polarity on the disease in social media, and information extraction, which creates an automatic system that is able to retrieve unstructured text and transform it to structural information.

3.2.1. Sentiment analysis

Sentiment analysis is an excellent tool for harvesting opinions from reviews or expressions of different users on a particular subject matter. The polarity of the opinion, namely, negative, positive, or neutral assists greatly in understanding the discourse from the community and making a decision concerning public health (Yusof et al., 2015; Aggarwal and Aggarwal, 2017). Therefore, sentiment analysis can be a classification process that preprocesses the text through operations such as tokenization, part-of-speech (PoS), and standardization, followed by classification algorithms such as machine learning.

Sentiment analysis is probably the single major focus of research in mining opinion from the Twitter platform, which is rich in information and perspectives of the public. The motivation here is mainly on the surveillance of mosquito-borne diseases (Ghani et al., 2022; García-Díaz et al., 2020; Saire, 2019a, 2019b), which involves disease classification with demographic variables (Quwaider and Alfaqeeh, 2016), detection (Ghani et al., 2022; García-Díaz et al., 2020; Saire, 2019a), prediction (Carlos et al., 2017; Jordan et al., 2018) and public awareness (Tripathy et al., 2017; Oza et al., 2016). Boit and El-Gayar (2020), Oyelade et al. (2021), Jordan et al. (2018), and Villanes et al. (2018) classified the topics, sentiments, and polarity to monitor and predict the public response to mosquito-borne diseases. You et al. (2020) used keyword extraction and the TextRank algorithm to study keyword co-occurrence to track the trends of Zika outbreaks. Huang et al. (2018) used topic modeling and sentiment analysis and determined the disease-medical conditions using the public opinion polarities.

3.2.2. Information extraction

Information extraction refers to the process of automatically retrieving structured facts from unstructured and semi structured text. In the mosquito-borne disease domain, most of the unstructured text includes scientific articles appearing in the literature and clinical information systems. Information extraction is often performed as an initial processing step for other text mining applications, such as sentiment classification, topic modeling or clustering. Based on our review, literature-based discovery is the second most performed text mining technique after sentiment analysis in the mosquito-borne disease domain. The technique is focused on the extraction of explicit relations between entities and events using the co-occurrence-based method (Safarnejad et al., 2019; Khatua and Khatua, 2016) and dependency relations (Abulaish et al., 2019; Parwez et al., 2018; Zhao et al., 2016; Selvaraj and Periyasamy, 2016). The techniques can be used to generate rich document summaries that are able to develop question answering systems (an organized way to link the answer to certain research questions from other literature). An entity can be any word or series of words that consistently refers to the same thing; for instance, the entity of Zika, Dengue and Malaria are classified as "mosquito-borne disease"; fever, rashes and vomit are classified as "symptoms". These proposed systems aim to discover the relationships that are absent in the text but that can be inferred from other information, for instance, where the location that reported the greatest number of dengue cases could be considered a cluster. Literature-based discovery mainly seeks new knowledge from existing publications/literature in an automated/semiautomated way and finds the relationship between two target events. For instance, we may discover the rationale of the surging of Zika by looking at the literature and linking it with co-occurring events. The main task of literature-based discovery is to utilize the scientific literature to uncover "hidden," previously unknown, or neglected relationships between existing knowledge.

To study the climate-sensitive diseases Malaria and Dengue, which follow seasonal patterns where the mosquitoes were influenced by precipitation and temperature, Parwez et al. (2018) used part-of-speech (POS) tagging and grammatical dependency relations to determine the disease symptoms and their relation from the PubMed database. To investigate the role of herba, Artemisia carvifolia, as a potential treatment for malaria, Zhao et al. (2016) aimed to examine the grammatical dependency relations of the documents in PubMed to determine the treatment and drug-disease association. Khatua and Khatua (2016) used the co-occurrence of words and conducted hierarchical clustering to investigate the immediate and long-term effects in the Twitter community after the outbreak of Zika.

3.2.3. Other techniques

Keyword extraction and topic modeling accounted for 15% and 18%, respectively, in our reviewed papers. This ranked as third and fourth in the reviewed papers. A common practice of keyword extraction studies (Oza et al., 2016; You et al., 2020) was extracting the name of mosquito-borne diseases such as 'dengue' and 'Zika'', and further with co-occurrence study with a ranking algorithm. For instance, the frequency of the disease occurring in a corpus could be linked to certain keyword (such as vector or parasite) frequencies. This ultimately aims to extract valuable information from the corpora, such as search engines, for disease surveillance. Topic modeling was conducted by using a supervised (Villanes et al., 2018; Oyelade et al., 2021; Jordan et al., 2018; Li, 2020; Muchene and Safari, 2021) or unsupervised (Zhang et al., 2020) machine learning algorithm. Supervised machine learning, such as logistic regression and support vector machine (SVM), was commonly applied to classify the semantic polarity of the topic that is available in the corpora. Other traditional techniques, such as bibliometric and Boolean logic (by using "AND" and "OR"), are still very effective in retrieving important information, such as the frequency and co-occurrence of the keywords related to mosquito-borne diseases.

3.3. Usage in mosquito-borne disease

3.3.1. Surveillance

Fig. 5.1 shows the distribution of the application of text mining in mosquito-borne disease texts; surveillance was the majority, followed by treatment. Mosquito surveillance programs are probably the most important component to prevent the disease and control the pest mosquito population (Lu et al., 2021). The target platform for text mining depends on the research question and objective; for instance, if the demographic is more of a public community, then social media such as Twitter is appropriate to understand the discourse of opinion and assist in making a decision. Ghani et al. (2022) tracked the keywords of dengue on Twitter by introducing a novel processing pipeline that combines filtration polarity and Apache Flume. Saire (2019a, 2019b) focused on the public opinion or sentiments reflected on Twitter about Dengue transmission, which proposed a detection method that enables surveillance of the infected person based on the symptoms. Carlos et al. (2017) identified the text pattern in Twitter tweets and constructed a prediction model to predict dengue occurrence. For relatively more formal corpora such as news or email, Zhang et al. (2019) and Villanes et al. (2018) focused on topic clustering on dengue and Zika in tropical countries from the LexisNexis database. You et al. (2020) retrieved Zika information from ProMED-mail by using keyword extraction and the TextRank algorithm and later obtained the time/geographical information for surveillance by using co-occurrence networks. Nonetheless, most of the studies focused on the technical parts—text analytics— but some key elements such as mosquito populations, were not covered at all.

3.3.2. Treatment

There are no licensed vaccines or specific treatments, for many of the mosquito-borne diseases. Hence the control of these diseases depends greatly on controlling of the mosquitos' populations (WHO, 2020). However, efforts to develop drugs and vaccines are in progress at an exciting and encouraging pace. The search for suitable drugs or active ingredients could be supported by the text mining techniques applied to biomedical databases. Ellis et al. (2020) aim to summarize the treatment in parasitology research from scientific journal databases such as MEDLINE, Web of Science, and Scopus by using bibliometric methods. Selvaraj and Periyasamy (2016) developed a framework of drug-protein association for malaria treatment, in which the relation between the drug and the response of the patient was obtained. Huang et al. (2018) aim to determine the disease-medical condition association by using topic modeling and sentiment analysis on the polarity of the discourse of Reuter media. The effort is to recommend the correct treatment to the user if they have encountered the medical condition.

3.3.3. Other applications

Discourse understanding is also a popular application of text mining in mosquito-borne diseases. Such studies aim to obtain public concern via polarity (García-Díaz et al., 2020; Oyelade et al., 2021; Safarnejad et al., 2019), emotional analysis (Boit and El-Gayar, 2020), and public awareness (Li, 2020) of the disease. In addition to domain problems, some of the studies aim to improve the technical perspective of text mining in mosquito-borne diseases, such as proposing a new algorithm in topic modeling (Yusof et al., 2015).

4. Discussion

4.1. Limitation and challenges

4.1.1. Bias

For social media platforms such as Twitter, the major bias will be the authentication of the user profile. However, information such as location, username, number of followers/friends, user background, etc., can be obtained along with the social media content. However, it is very difficult to determine if it is the same person who has multiple accounts posting about the outbreak and influencing the mining of results.

Another bias is that certain platforms of databases, such as social media, consist of a certain range of user ages. For example, Facebook and Twitter are more common for users aged more than 25 years old, but the younger generation prefers to use Instagram and Snapchat as their social platforms (US Generation, 2022). This is strongly supported by the observation that social media data are skewed toward active users who are often teenagers or young adults (Nsoesie et al., 2016). In addition, based on two databases, the search of other important yet neglected diseases, such as Chikungunya and West Nile fever, is extremely low in number.

Language could be another major bias. The collection and analysis of most of the databases used by 27 articles were restricted to the English language, which has an allusion to the generalizability of the study. This is due to the fact that a significant proportion of the total corpus, such as social media posts, news, and scientific papers, are reported in the English language. Future studies can explore different popular languages, such as Mandarin, Malay, or Hindi, to analyze the corpus that has mentioned mosquito-borne disease.

4.1.2. Real-world inferences

The levels of reflectivity of the topics in corpora could be questionable. Most of the authors from the articles that have been reviewed in this study agreed that it is unclear how an actual situation of a mosquito-borne disease influences the documents (Abulaish et al., 2019; Parwez et al., 2018; Muchene and Safari, 2021; Zhang et al., 2020) and perception and discourse on social media (Boit and El-Gayar, 2020; Ghani et al., 2022; Saire, 2019b; Zhang et al., 2019; Quwaider and Alfaqeeh, 2016). Carlos et al. (Carlos et al., 2017) emphasize that the drawback of the correlation between real-world events and their impact on web-based discussions of mosquito-borne disease is not well investigated and understood. Therefore, most of the text mining techniques involved in the analytics of mosquito-borne disease required event extraction or semantic analysis with NLP, which provides much more granular control and refinement for additional improvement of topic and theme development (Boit and El-Gayar, 2020).

Nevertheless, if we focus the text mining on mosquito-borne diseases for social media such as Twitter, the users themselves may not encounter, represent, or relate to the diseases. As suggested by Liang and Fu, most tweets are posted by a few extremely active users (Liang and Fu, 2015), which may not be the patient or community that has an issue of the mosquito or mosquito-borne disease. Furthermore, the occurrence of many tweets containing the words “sick” and “fever” may be ambiguous and not imply that there are many sick people.

4.2. Contribution of this review

This paper aims to study the latest trends in text mining in the mosquito-borne domain. We prefer to use Scopus and PubMed databases as our main resource due a robust review process is required for the

Table 1
Summarization of the reviewed articles based on their disease, corpora, TM technique and application.

Author	Mosquito borne disease	Corpora	Techniques	Application
Muchene and Safari (2021)	Malaria	University of Nairobi e-repository	• Topic modeling • Latent Dirichlet Allocation LDA	Text mining technique improvement
Ellis et al. (2020)	Malaria	Scientific journal databases	• Bibliometric methods	Treatment
Boit and El-Gayar (2020)	Malaria	Twitter	• Sentiment analysis and emotion analysis	Discourse understanding
Abulaish et al. (2019)	Dengue, Malaria	PubMed	• Information extraction • Part-Of-Speech (POS) tagging • Dependency relations	Text mining technique improvement
Parwez et al. (2018)	Malaria, Dengue	PubMed	• Information extraction • Part-Of-Speech (POS) tagging • Dependency relations	Treatment, Disease’s symptoms and their relation
Oyelade et al. (2021)	Malaria	Twitter	• Information extraction • Part-Of-Speech (POS) tagging • Sentiment Analysis	Discourse understanding and classification
Quwaider and Alfaqeeh (2016)	Malaria	Facebook	• Disease Classification	Disease Classification
Zhao et al. (2016)	Malaria	PubMed	• Dependency relations	Treatment and drugs-disease association
Selvaraj and Periyasamy (2016)	Malaria	PubMed, UniprotKB and MeSH	• Dependency relations	Treatment and drugs-disease association
Ghani et al. (2022)	Dengue	Twitter	• Text analysis • Sentiment analysis	Surveillance
García-Díaz et al., (2020)	Zika, Dengue, Chikungunya	Twitter	• Sentiment analysis	Surveillance, Detection, Discourse understanding
Zhang et al. (2019)	Dengue, Zika	LexisNexis	• Topic clustering	Surveillance
Saire (2019a)	Dengue	Twitter	• Sentiment analysis	Surveillance, Detection
Saire (2019b)	Dengue	Twitter	• Sentiment analysis	Surveillance, Detection
Zhang et al. (2020)	Dengue	LexisNexis	• Boolean logic search	Surveillance
Villanes et al. (2018 Jan)	Dengue	LexisNexis	• Topic clustering	Surveillance
Carlos et al. (2017)	Dengue	Twitter	• Text pattern identification	Surveillance, Prediction
You et al. (2020)	Zika	ProMED-mail	• Information extraction • Keyword extraction • TextRank algorithm • Keywords co-occurrence networks	Surveillance
Safarnejad et al. (2019)	Zika	Twitter	• Co-occurring event/keywords	Discourse understanding
Li (2020)	Zika	PubMed	• Information extraction • Literature-based Discovery approach • Link prediction • Supervised classification	Medical subject headings (MeSH) prediction in Zika and CRISPR research
Jordan et al. (2018)	Zika	Twitter	• Topics classification	Surveillance, Prediction
Li et al. (2017)	Zika	Twitter	• Textual analysis of comment texts	Topic Discover
Tripathy et al. (2017)	Zika	Twitter	• Text classification	Community support, public awareness
Khatua and Khatua (2016)	Zika	Twitter	• Information extraction • Co-occurrence of words • Hierarchical clustering	Post disease management
Oza et al. (2016)	Zika	Google, Bing, and YouTube	• Keywords extraction with search engine	Public awareness
Safarnejad et al. (2019)	Zika	Twitter	• Information extraction • Co-occurring trending events	Discourse understanding
Huang et al. (2018)	Zika	Reuters media	• Topic modeling • Sentiment analysis • Polarity analysis	Treatment, Disease-medical conditions association

specialized area. Our paper has the following contributions that have not been considered carefully in the previous papers:

- Provides the article selection queries from different digital libraries databases for selecting the relevant articles (Section 2).
- Present an overview of the corpora used for text mining in the mosquito-borne disease domain studied by the selected articles (Section 3.1).
- Presents an overview of text mining techniques used for text mining in the mosquito-borne disease domain (Section 3.2).
- Presents an overview of text mining applications in the mosquito-borne disease domain
- Summarization of 27 articles that studied text mining in mosquito-borne disease (Table 1).
- Comparison between the trend and relevant probability (= relevant articles/ total filtered queries) of queries from different digital libraries by their mosquito-borne disease (Fig. 2).

As can be seen from Fig. 2, the average number of articles retrieved from PubMed is higher than Scopus, this is due to a lower number of publishers available in the Scopus which required indexation. However, when we compare the article relevant percentage (A.R.P.), the Scopus database can provide higher relevancy.

5. Conclusion and future direction

Literature on text mining for mosquito-borne diseases was summarized, and the techniques and corpora were discussed. Twitter was the major source of text data for mosquito-borne disease. Twitter data are collected in real time and are available from many users across different geographic regions via API. However, since tweets are microblog-based and posted with different purposes, clear targets and objectives with correct methods must be used to extract the desired information. For techniques and applications, each article used a different analysis, and due to the lack of standardization, direct comparisons were difficult to perform. Nevertheless, the two main interests of research on text mining in mosquito-borne diseases were semantic analysis and information extraction. This is understandable since decision-making in public health requires more practical information and insights to manage the disease effectively. Most of the applications were concentrated on surveillance, which is the most important component in mosquito-borne disease. Surveillance includes incidence detection and prediction. However, almost none of the articles related the vector mosquito to the studies, which makes the studies less comprehensive (Fig. 1).

This review provides an overview of the text mining in mosquito-borne disease. Referring to the quantities of articles from the databases, which indicate the domain still has high potential to explore. While there remain some challenges and offer great opportunities for improvement, and this review sets the groundwork for the development of text mining in utilizing the untapped wealth of data to improve the management of a mosquito-borne disease. We expect this review could lead to the discovery of new gaps and development of new technologies, new capabilities for research of text mining in mosquito-borne disease. Some of the future works that we have observed can be:

(1) Integration of more online data like weather conditions, demographic information, vector population to obtain better generalization.
(2) Specify the research question of the domain. With a narrower research question, like what are the vectors that caused Zika in the sub-urban area? What are the complications of co-current diseases such as Zika and Dengue?
(3) Integrating text and image analysis.
(4) Incorporating the input features such as locations
(5) Classification of users' posts into different categories such as health, news, ads.

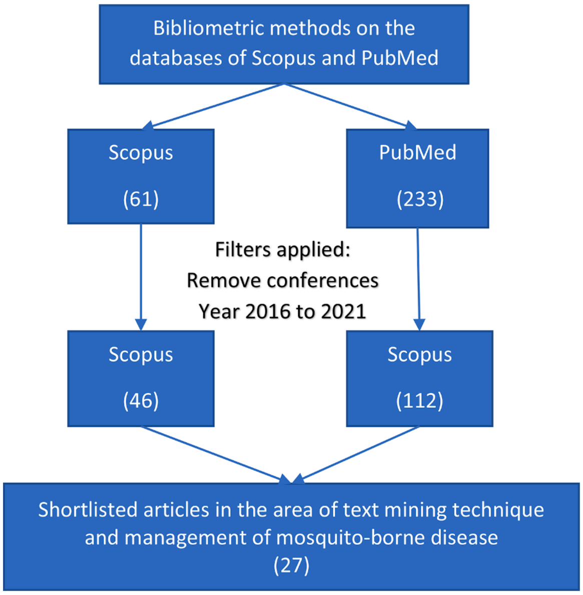

Bibliometric methods on the databases of Scopus and PubMed
Scopus (61)
PubMed (233)
Filters applied: Remove conferences Year 2016 to 2021
Scopus (46)
Scopus (112)
Shortlisted articles in the area of text mining technique and management of mosquito-borne disease (27)

Fig. 1. Search methodology for the selection of relevant articles.

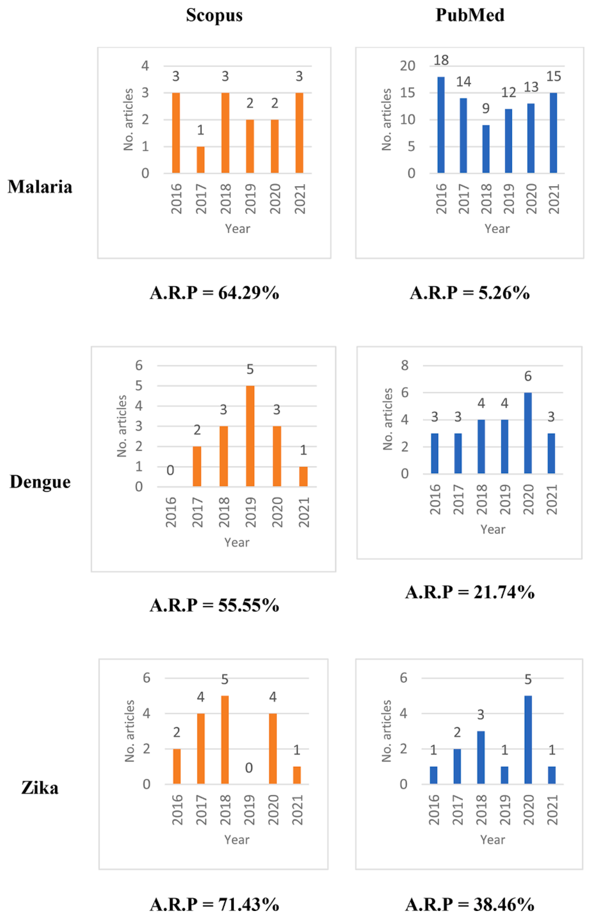


Fig. 2. Trends and articles relevant percentage for the databases used in this review for the articles that using text mining in mosquito-borne disease from 2016 to 2021.

(6) Attempt in the performance of topic modeling algorithm.
(7) Implementation of predictive models in the real-world mosquito-borne disease

Additional information

Correspondence and requests for materials should be addressed to SQO.

CRediT authorship contribution statement

Song-Quan Ong: Conceptualization, Visualization, Validation, Writing – review & editing. **Maisarah Binti Mohamed Pauzi:** Conceptualization, Validation, Writing – review & editing. **Keng Hoon Gan:** Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgment

We thank all the anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript.

References

Abulaish, M., Parwez, M.D.A., Jahiruddin, 2019. DiseaSE: a biomedical text analytics system for disease symptom extraction and characterization. J. Biomed. Inform. 100, 103324 https://doi.org/10.1016/j.jbi.2019.103324. Dec.

Aggarwal, U., Aggarwal, G., 2017. Sentiment analysis: a survey. Int. J. Comput. Sci. Eng. 5 (5). Open Access Surv. Pap.

Boit, J., El-Gayar, O., 2020. Topical mining of malaria using social media. A text mining approach. Fac. Res. Publ. Jan.Accessed: Dec. 10, 2021. [Online]. Available: https://scholar.dsu.edu/bispapers/222/.

Carlos, M.A., Nogueira, M., Machado, RJ., 2017. Analysis of dengue outbreaks using big data analytics and social networks. In: Proceedings of the 4th International Conference on Systems and Informatics (ICSAI). IEEE, pp. 1592–1597. Nov 11.

Cohen, K.B., Hunter, L., 2008. Getting started in text mining. PLoS Comput. Biol. 4 (1), e20. https://doi.org/10.1371/journal.pcbi.0040020.

Collier, N., et al., 2008. BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics 24 (24), 2940–2941. https://doi.org/10.1093/bioinformatics/btn534. Dec.

Cowell, L.G., Smith, B., 2010. Infectious disease ontology. Infect. Dis. Inform. 373–395. https://doi.org/10.1007/978-1-4419-1327-2_19.

Ellis, J.T., Ellis, B., Velez-Estevez, A., Reichel, M.P., Cobo, M.J., 2020. 30 years of parasitology research analysed by text mining. Parasitology 147 (14), 1643–1657. https://doi.org/10.1017/S0031182020001596. Dec.

García-Díaz, J.A., Cánovas-García, M., Valencia-García, R., 2020. Ontology-driven aspect-based sentiment analysis classification: an infodemiological case study regarding infectious diseases in Latin America. Future Gener. Comput. Syst. 112, 641–657. https://doi.org/10.1016/j.future.2020.06.019. Nov.

Ghani, N.A., et al., 2022. Tracking dengue on twitter using hybrid filtration-polarity and apache flume. Comput. Syst. Sci. Eng. 40 (3), 913–926. https://doi.org/10.32604/CSSE.2022.018467. Jan.

Huang, M., ElTayeby, O., Zolnoori, M., Yao, L., 2018. Public opinions toward diseases: infodemiological study on news media data. J. Med. Internet Res. 20 (5), e10047. https://doi.org/10.2196/10047. May.

Jordan, S., Hovet, S., Fung, I., Liang, H., Fu, K.-W., Tse, Z., 2018. Using twitter for public health surveillance from monitoring and prediction to public response. Data 4 (1), 6. Dec.

Kao, A., Poteet, S.R., 2007. Natural Language Processing and Text Mining. Springer Science & Business Media. Mar 6.

Khatua, A., Khatua, A., 2016. Immediate and long-term effects of 2016 Zika outbreak: a Twitter-based study. In: Proceedings of the IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 1–6. https://doi.org/10.1109/HealthCom.2016.7749496.

K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: demonstration on flu and cancer," CiteSeer. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.414.1896 (accessed Dec. 10, 2021).

Li, J., Shin, S.Y., Lee, H.C., 2017. Text mining and visualization of papers reviews using R language. J. Inf. Commun. Converg. Eng. 15 (3), 170–174. https://doi.org/10.6109/jicce.2017.15.3.170.

Li, M.H., 2020. Using link prediction methods to examine networks of Co-occurring MeSH terms in Zika and CRISPR research. In: Proceedings of the International Conference on Information. Springer, Cham, pp. 782–789. Mar 23.

Liang, H., Fu, K., 2015. Testing propositions derived from twitter studies: generalization and replication in computational social science. PLoS One 10 (8), e0134270. https://doi.org/10.1371/journal.pone.0134270. Aug.

Lu, X., Bambrick, H., Pongsumpun, P., Dhewantara, P.W., Toan, D.T.T., Hu, W., 2021. Dengue outbreaks in the COVID-19 era: alarm raised for Asia. PLoS Negl. Trop. Dis. 15 (10), e0009778 https://doi.org/10.1371/journal.pntd.0009778. Oct.

Muchene, L., Safari, W., 2021. Two-stage topic modelling of scientific publications: a case study of University of Nairobi, Kenya. PLoS One 16 (1), e0243208. https://doi.org/10.1371/journal.pone.0243208. Jan.

Neill, D.B., 2012. New directions in artificial intelligence for public health surveillance. IEEE Intell. Syst. 27, 56–59. https://doi.org/10.1109/MIS.2012.18.

Newkirk, R.W., Bender, J.B., Hedberg, C.W., 2012. The potential capability of social media as a component of food safety and food terrorism surveillance systems. Foodborne Pathog. Dis. 9, 120–124. https://doi.org/10.1089/fpd.2011.0990.

Nsoesie, E.O., Flor, L., Hawkins, J., Maharana, A., Skotnes, T., Marinho, F., Brownstein, J.S., 2016. Social media as a sentinel for disease surveillance: what does sociodemographic status have to do with it? PLoS Curr. https://doi.org/10.1371/currents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6.

J. Oyelade, E. Uwoghiren, I. Isewon, O. Oladipupo, O. Aromolaran, and K. Michael, "Machine learning and sentiment analysis: examining the contextual polarity of public sentiment on malaria disease in social networks." Accessed: Dec. 10, 2021. [Online]. Available: http://eprints.covenantuniversity.edu.ng/14812/1/1037.pdf.

Oza, K.S., Kumbhar, V.S., Kamat, R.K., 2016. Zika virus awareness: a text mining approach. Res. J. Pharm. Biol. Chem. Sci. 7 (6), 1942–1947.

Parwez, M.D.A., Abulaish, M., Jahiruddin, 2018. Biomedical text analytics for characterizing climate-sensitive disease. Procedia Comput. Sci. 132, 1002–1011. https://doi.org/10.1016/j.procs.2018.05.016.

Quwaider, M., Alfaqeeh, M., 2016. Social networks benchmark dataset for diseases classification. In: Proceedings of the IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), Vienna, Austria, pp. 234–239. https://doi.org/10.1109/W-FiCloud.2016.56.

Rajapakse, M., Kanagasabai, R., Ang, W.T., Veeramani, A., Schreiber, M.J., Baker, C.J.O., 2008. Ontology-centric integration and navigation of the dengue literature. J. Biomed. Inform. 41 (5), 806–815. https://doi.org/10.1016/j.jbi.2008.04.004. Oct.

Safarnejad, L., Xu, Q., Ge, Y., Bagavathi, A., Krishnan, S., Chen, S., 2019. Identifying influential factors on discussion dynamics of emerging health issues on social media: a computational study (Preprint)," JMIR Public Health Surveill. https://doi.org/10.2196/17175. Nov.

Safdari, R., Rezayi, S., Saeedi, S., Tanhapour, M., Gholamzadeh, M., 2021. Using data mining techniques to fight and control epidemics: a scoping review. Health Technol. 11 (4), 759–771. Jul.

Saire, J.E.C., 2019a. Building intelligent indicators to detect dengue epidemics in Brazil using social networks. In: Proceedings of the IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI), pp. 1–5. https://doi.org/10.1109/ColCACI.2019.8781976.

Saire, J.E.C., 2019b. Building dengue sensors for Brazil using a social network and text mining. Commun. Comput. Inf. Sci. 1096, 69–77. Accessed: Dec. 10, 2021. [Online]. Available: https://repositorio.usp.br/item/002986470.

Selvaraj, B., Periyasamy, &S., 2016. A framework of protein-drug association for malaria by text data mining of biomedical literature. Res. J. Pharm. Biol. Chem. Sci. 7 (4), 1493–1499.

Simpson, M.S., Demner-Fushman, D., 2012. Biomedical text mining: a survey of recent progress. Min. Text Data, 465–517. https://doi.org/10.1007/978-1-4614-3223-4_14.

H. Tran, "A survey of machine learning and data mining techniques used in multimedia system a preprint," 2019. Accessed: Dec. 10, 2021. [Online]. Available: https://personal.utdallas.edu/~trunghieu.tran/publications/ResearchGate19_Survey_ML_DM.pdf.

Tripathy, B.K., Thakur, S., Chowdhury, R., 2017. A classification model to analyze the spread and emerging trends of the Zika virus in Twitter. Computational Intelligence in Data Mining. Springer, Singapore, pp. 643–650.

V. Petrock. US Generation Z Technology and Media Use. What Usage Looks Like for the First Generation with 24/7 Access to Connected Devices. 2022 https://www.emarketer.com/content/us-generation-z-technology-and-media-use.

Villanes, A., Griffiths, E., Rappa, M., Healey, CG., 2018. Dengue fever surveillance in India using text mining in public media. Am. J. Trop. Med. Hyg. 98 (1), 181. Jan.

WHO, "Dengue and severe dengue," Who.int, Jun. 23, 2020. https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue.

WHO-World Health Organization (2020). "Vector-borne diseases," WHO. int, Mar. 02, 2020. https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases.

World Health Organization, "Vector-borne diseases," Who.int, Mar. 02, 2020. https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases.

You, J., Expert, P., Costelloe, C., 2020. Using text mining to track outbreak trends in global surveillance of emerging diseases: ProMED-mail. Semant. Sch. https://doi.org/10.1101/2020.01.10.20017145.

Yusof, N.N., Mohamed, A., Abdul-Rahman, S., 2015. Reviewing classification approaches in sentiment analysis. Communications in Computer and Information Science. Springer, Singapore, pp. 43–53 vol. 545.

Zhang, Y., Ibaraki, M., Schwartz, F.W., 2019. Disease surveillance using online news: an extended study of dengue fever in India. Trop. Med. Health 47 (1). https://doi.org/10.1186/s41182-019-0189-y. Dec.

Zhang, Y., Ibaraki, M., Schwartz, F.W., 2020. Disease surveillance using online news: dengue and Zika in tropical countries. J. Biomed. Inform. 102, 103374 https://doi.org/10.1016/j.jbi.2020.103374. Feb.

Zhao, Y.P., Wang, H., Yang, G., Qiu, Z.D., Qu, X.B., Zhang, XB., 2016. Exploring pharmacological principle of Artemisia carvifolia with textmining technology. China J. Chin. Mater. Med. https://doi.org/10.4268/cjcmm20161622. Aug.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K.B., 2007. Frontiers of biomedical text mining: current progress. Briefings Bioinf. 8 (5), 358–375. https://doi.org/10.1093/bib/bbm045. Jun.