Research on Prediction of Infectious Diseases, their spread via Social Media and their link to Education

Olanrewaju T. Aduragba
Department of Computer Science
Durham University
Lower Mountjoy, South Road, Durham, DH1 3LE UK
olanrewaju.m.aduragba@durham.ac.uk

Alexandra I. Cristea
Department of Computer Science
Durham University
Lower Mountjoy, South Road, Durham, DH1 3LE UK
(+44)1913342761
alexandra.i.cristea@durham.ac.uk

ABSTRACT

Infectious diseases are a great plague, especially in low and middle income countries. Beyond the actual treatment, an important role is played by early prevention mechanisms, and education of society at large, about existing risks. This paper tackles these two important challenges, describing the current state of the art in this area, and pointing towards the need for both further, more inclusive research, as well as better education in affected countries on infectious diseases.

CCS Concepts

• Applied computing → Health informatics

• Computing methodologies → Artificial intelligence

Keywords

Social media mining, disease outbreak prediction, public health education, deep learning

1. INTRODUCTION

In the last couple of years, public health has become much more of a concern, possibly due to the expanding awareness of the prevailing risks, as well as lack of knowledge in terms of where such diseases start, as well as educative prevention - with the risk of serious infectious diseases and human mortality attributed to infection estimated to increase to up to 15 million deaths annually by 2030 [1]. Particularly, in low and middle-income countries, the spread of infectious diseases is a major concern and remains a significant threat to their economic development [2].

To avoid the consequences of these epidemics, early detection and prediction of disease outbreaks on one hand, and education on the other, are emphasised to mitigate infectious disease outbreaks [3]. With the increasing volume of information and new media types available via the Internet, digital surveillance has grown to include social media. Social media platforms such as Twitter have recently become sources of most up-to-date information and commentary on current and significant events taking place in people's lives and during various natural disasters. For instance, Twitter has over 500 million users that send more than 500 million messages on a daily basis and 4.3 billion Facebook messages are posted everyday [4] [5]. Recently, social media contents have been increasingly sent from mobile phones and devices which strengthens the chance that they will contain geographic information [6]. Such 'rich' contents can be leveraged to discover local trends in health updates, making digital surveillance of infectious diseases plausible [7].

Overall, social media can be seen a collector of real-time information that could be used by public health institutions as an additional information source for acquiring early warnings - thereby assisting them to mitigate the public health threats [8].

At the same time, the recent development of deep learning has allowed researchers to achieve remarkable success in various research areas in machine learning, by detecting interesting patterns and structures in high-dimensional data [9]. The success of deep learning and automatic data processing present the possibility of utilising the large amount of data generated from social media as data source for tracking and predicting disease spread. The main contributions of this paper are as follows:

- A review of why social media is suitable for predicting infectious diseases spread.
- A review of infectious disease tracking and prediction through social media.
- A review of public health education related to disease spread

2. RELATED WORK

Several researches have been carried out, focussing specifically on the task of reviewing prior contributions on prediction and detection of disease-spread through social media. These include reviews on influenza-related diseases, such as a review on existing alternative solutions that track flu outbreak in real-time, using both social media and the web [10], forcasting the dynamics of influenza outbreaks [11], actionable disease surveillance and outbreak management using social media [12]. They review existing alternative techniques, including machine learning models, mathematical/computational models, topic models, graph data mining that track flu outbreak in real time using social media, web blogs, internet search data and traditional data sources. On the other hand, other review work analyses existing research that centres around the Ebola virus disease and its visibility on social media [13], or consider studies that use consumer-generated data, such as social media and restaurant reviews, to track and monitor foodborne illness [14].

Unlike the existing literature review, our approach focuses on the review of studies that predict the spread of a greater variety of infectious diseases, including but not limited to Cholera, Ebola, flu and Zika, through a more comprehensive variety of social media sources, such as Twitter, Facebook, Instagram and LinkedIn. In addition, our study will discuss the social media data collection process, features and classifier performances of the deep learning techniques used in the reviewed studies, while linking it to the different types of education on disease detection and prediction.

3. METHODOLOGY

This paper aims to research published papers and articles in recent years that used social media to track and predict the spread of infectious diseases. The paper collection was limited to papers or articles published within the last 10 years, because this period corresponds with the growth of social media popularity. Research articles pertinent to infectious diseases prediction on social media that were published in English between years 2009 to 2019 were searched for on Google Scholar and PubMed. To identify papers and articles related to infectious disease prediction, a number of keywords including "infectious disease prediction using social media" or "epidemic forecasting using social media" or "infectious disease prediction using deep learning" were used. Additional keywords derived from the search phrases such as "Twitter", "Facebook", "Instagram", "ebola", "zika", "lassa fever", "cholera" were also added to the search arguments. These keywords comprises a list of specific form of social media based on their global popularity and the list of infectious diseases that have the potential to become international threats [15]. This review adopted the basic definition of social media as a "*group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content*" [16]. Epidemic forecasting was used in the search term because it is also commonly used to represent the prediction of disease outbreaks by epidemiologists and health-care providers [17]. After searching, the relevance filter tool in Google Scholar and PubMed was used to sort the results based on high relevance and only results where the titles are relevant to the search criteria were selected. After the paper selection and initial filtering of titles, a further filtering of abstracts was performed to filter out the papers based on certain selection criteria. Since our focus was on prediction of infectious diseases from social media data, the paper selection criteria were limited to the following categories:

- Predicting the spread of infectious diseases using social media data.
- Predicting the spread of infectious diseases on social media using deep learning techniques.
- Epidemiological predictions using machine learning and deep learning techniques.

Subsequently, a full-text screening was performed and the final number of papers were selected for the review. Figure 1 summarizes the process of paper selection.

4. DATA

Initially, the search performed with the combination of the search terms returned 17,900 entries on Google Scholar and 307 entries on PubMed. By default, the search results on Google Scholar are already sorted by relevance while on PubMed the sort by Best Match needed to be applied. After this process, a topic screening of the first 500 results on Google Scholar and all results on PubMed was performed. Based on the title, studies that were evidently not related to the keywords were eliminated. Across the two search databases, after removing duplicates, a total of 807 papers were retrieved between the years 2009 – 2019. The next step was filtering based on the selection criteria. In this step, the abstracts were screened if they directly address either of the following categories mentioned above. 55 papers were eliminated after the abstract screening so the remaining number of papers decreased to 19. From the removed papers, 42% (n=23) were study reviews that survey exisiting literature or discuss why social media is suitable for outbreak surveillance, 31% (n=17) did not describe the methods and techniques used for prediction and 27% (n=15) failed to describe the data used, or they did not discuss prediction results. Subsequently, the remaining papers were analysed by reading the full text. The full texts of 11% (n=2) of the screened papers were not available, 5% (n=1) did not appear in a peer-reviewed journal or conference proceeding with a good ranking and 5% (n=1) had low citation count with respect to the publication year, thus they were excluded from our studies. The final number of selected articles that were considered for this research was 15 articles.

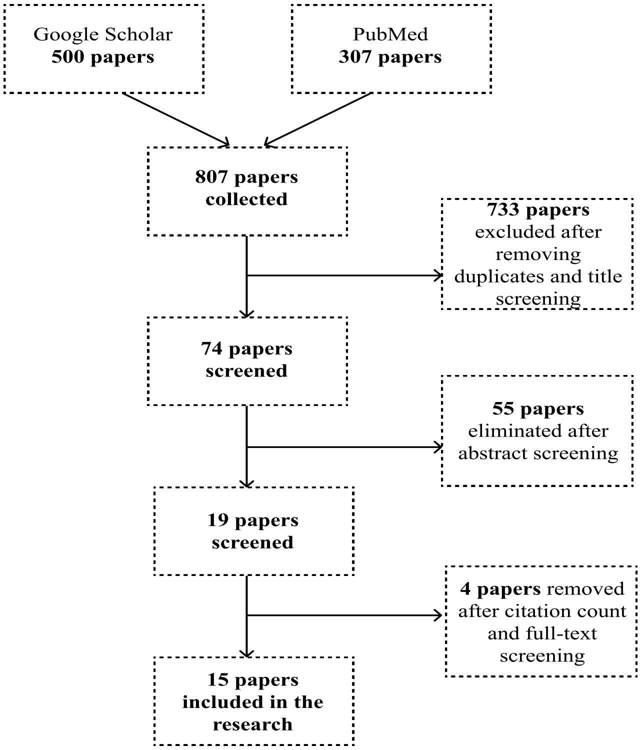

Google Scholar
500 papers
PubMed
307 papers
807 papers collected
733 papers excluded after removing duplicates and title screening
74 papers screened
55 papers eliminated after abstract screening
19 papers screened
4 papers removed after citation count and full-text screening
15 papers included in the research

Figure 1. Paper Selection Process.

The majority of the papers were published in conferences with core A and A* rankings such as the IEEE International Conference on Data Mining and Proceedings of the conference on empirical methods in natural language processing and journals with high impact factor of (up to 3.434) such as Preventive Medicine. The most-cited paper is [18] with up to 540 citations. The range of the citation count is between 2 and 540, depending on the publication year. The low visibility of some papers could be as a result of their recent publication year (e.g. 2019). Generally, the research area covered is not considered a mainstream research area hence this could also affect the visibility and influence.

5. RESULTS

5.1 Automatic Processing Results

Titles and abstracts of the research papers contain free text that sum up the major aspects of the research. Employing a text analysis technique, this study extracted free text from the research paper's title and abstract to explore what are the word phrases that were frequently used in the selected studies. As a preprocessing step, stop words from titles and abstracts were removed. Stop words (e.g., the, of, or...) are certain parts of English text that are meaningless or non-informative to our analysis. Another very important preprocessing step that was done is stemming. This is a common Natural Language Processing (NLP) technique that is used to transform topically similar words to their root. For example, "predicting", "predicted" all have similar meanings, by stemming them they are reduced to a common base form i.e.

"predict". This prevents similar words to be treated as separate entities with different frequencies and importance to the text.

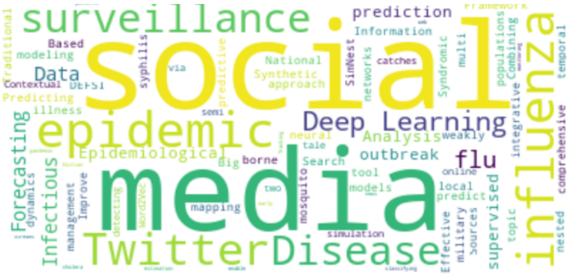


Figure 2. Word cloud of words in Titles.

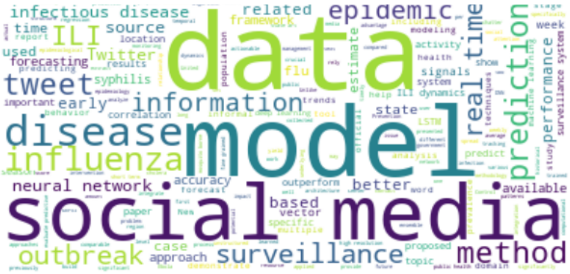


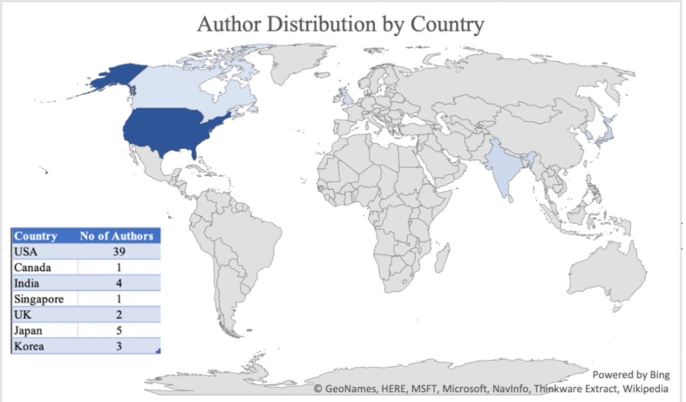
Figure 3. Word cloud of words in Abstracts.

Figures 2 and 3 show, respectively, the result of the frequency analysis of the words in the titles and abstracts of our selected studies for reviewing. In Figure 2, the most common word mentioned along with social media in the titles is Twitter. This highlights the popularity of Twitter as the most used social media data source for prediction in our studies. The rise of deep learning work is also obvious from the image. From Figure 3, we can identify the studied diseases in our studies. The words in the abstract highlight the prevalence of influenza and similar illnesses with infectious disease prediction.

5.2 Manual Processing Results

The majority of the included papers utilised several data sources for the prediction of disease. Out of the 15 papers included in the study, 60% used data from social media and other sources including hospital visit records, weather data, news feeds, search query data, Scholarly abstracts, temperature and humidity data in their prediction, 33% used only social media data and the remaining 7% used other sources apart from social media, such as Google Correlate Terms, Google Health Trends and medical reports. By far the most used primary sources of data on social media for predicting infectious diseases in the selected studies was Twitter, with up to 87% of the works using it. This popularity might be as a result of Twitter containing a large volume of data which are readily available in real time without technical challenges common with other data sources. Notably, no other social media, such as Facebook and Instagram, were used in the screened papers – possibly due to the restrictions inserted relatively recently in Facebook in terms of data usage outside the primary scope of the platform.

In terms of location where the research was conducted, 60% were conducted in North America (mainly the United States of America (USA)), 20% in Asia (including Korea, Japan and India), 1% on a global scale, and less than 1% in Europe and South America. Although diseases, e.g. Ebola and Malaria that are more common in Africa than other parts of the world, were included, the low internet penetration rate in African countries presented a challenge in choosing them as research context [19]. Similarly, USA is the country that is most published with respect to institutional affiliation of the authors, with 71% of the publications, followed by Japan, India and Korea with 9%, 7% and 5% respectively. Authors from UK, Canada and Singapore represent less than 5% of the publications (as shown in Figure 4).


Author Distribution by Country
Powered by Bing
© GeoNames, HERE, MSFT, Microsoft, NavInfo, Thinkware Extract, Wikipedia

Country	No of Authors
USA	39
Canada	1
India	4
Singapore	1
UK	2
Japan	5
Korea	3

Figure 4. Author Distribution by Country.

Out of the articles examined the prediction of various diseases, 60% of them concentrated on forecasting Influenza-like Illness. The other diseases that are investigated include malaria, syphilis, cholera and campylobacter. Some articles analysed more than one disease, such as [19], which predicted the spread of Zika and Ebola, or [20] who studied the spread of chickenpox, scarlet fever and malaria. Overall, there is similarity in the trend between the locations that have been surveyed and the disease types predicted. This could also be due to high social media penetration rate in North America and widespread of influenza, that keeps re-emerging every year, with an average of about 8% of the U.S. population catching flu each season [21].

Statistics including mean, standard deviation and one sample t-test were conducted on the number of words and number of authors. Table 1 provides the results of the statistical analysis, the average number of words in the studies is 3426.6 while the average number of authors is 3.6. All of the p-values obtained for the number of words and number of authors show that they are statistically significant (at 1% level). This indicates that for articles in this study, there is no relationship between articles in terms of the number of words and number of authors.

6. DISCUSSION

6.1 Interpretation of Results

This study started with a large collection of papers; however, the finally analysed papers which fulfilled all conditions was relatively limited (only 15 papers). Thus, conclusions drawn from these can only be interpreted as early pointers towards the statements made, and are not conclusive. Further research is necessary to establish their generalisabity for the whole sector and their validity across the whole research area.

Table 1. Statistical analysis of No. of words and No. of authors

	mean	std. dev.	t-statistic	p-value
No of words	7630.3	3426.6	-22901.09	< 0.01
No of authors	3.6	1.18	-4.58	<0.01

Our word clouds are very useful for quickly visualising the main topics in this area. Whilst analysing titles and abstracts renders somewhat different results, there are interesting similarities – such as the usage of deep learning models in the prediction, and the application field of influenza. Social media is clearly the most prominent topic, and, in hindsight, may have been best removed together with stop words etc. Differences are also interesting to analyse. Real time appears as a very important topic within the abstracts. This means that researchers are more interested now-a-days in the timely analysis of the data, to have real-time responses, and can thus intervene based on the data processing. It is clear that post-factum processing, whilst interesting, is less effectual, and being able to quickly bring help to the affected areas is of extremely high relevance for this line of research. The data usage is another topic which clearly stands out in the abstract. This conforms to the extremely popular area, which is data analytics and data processing, in the current research landscape across subjects. 'Big data' has been innundating us in the recent years, and the researcher communities find themselves compelled to make sense of it for a great variety of areas. It is a wealth of the current time, which however comes with new responsibilites of making sure it is used in the best way possible to support, enhance and augument human life. In terms of this usage, data analytics is only the first step. Importantly, the transition from 'simple' descriptive analysis, where the data is collected and explained, usually for human consumption, in more comprehensive and compact ways, needs to make way for more advanced data processing, such as diagnostic, predictive and ultimately, prescriptive data analytics. Diagnostic data analytics is to answer classically difficult AI questions, such as 'why did it happen?', explaining thus why and how the reasoning has happened. Predictive analytics answers 'what will happen?', thus using usually past data to predict the future. Finally, prescriptive analytics, the most challenging one, is answering questions such as 'what should I do?', which means to inform and guide people towards new actions, based on information from the data. Usually this happens based on past data, where the user of an information system is guided towards the most productive action, based on data and a form of user modelling (i.e., information about the current user). This bridges the way between the data analytics area and the user modelling and personalisation area, and thus is very important in creating future-proof human-centric systems.

Twitter is clearly emerging as a leading venue for prediction of infectious diseases and their spread. This is due to its availability and the fact that the information can be freely processed by reseachers, as well as its wide use across countries, its fast response time in terms of events occuring world-wide, amongst others. It is thus important to continue this line of research in further studies, to see at what level Twitter can be an accurate predictor, and, possibly most importantly, an early intervention mechanism for infectious diseases. It is interesting however that other social media are not used as much in these predictions. It would be interesting to further analyse if this is an omission, or just a convenience-driven approach, or actually if Twitter is fundamentally different to other social media, and is more appropriate for prediction in this area.

The fact that the analysed papers are very dissimilar in terms of number of words, may be due to the different venues this type of research is published. This is a good thing, potentially showing the wide spread interest in this area, and the fact that very different venues publish such material. Clearly more research is necessary to establish this matter at a more relevant level in terms of the mass of papers published on this area. This can be followed up by an analysis of our original body of papers extracted.

The fact that the number of authors is different also seems to show that such research is produced in a variety of groups, or smaller and larger size, pointing to the potential interest of different stakeholders in the area of infections diseases and its spread. This matter can also be further analysed in a similar manner.

In terms of country distribution, it is worrying that countries with the greatest spread of infectious diseases, such as many of the African countries, are the least involved in the research on these matters. Here, it is clear that education is an important factor, which connects to our next subsection.

6.2 Education versus Spread and Research

Recent annual epidemics of the Ebola virus disease and the Meningococcal disease have resulted in more than 20,000 deaths since 2017 across West Africa [22][23]. Although there has been increased availability of vaccines and developments made in the medical field, middle and low-income countries remain vulnerable to emerging and re-emerging epidemics that threaten their communities.

Detection and surveillance of infectious diseases provide epidemiological intelligence to assist health practitioners in managing disease outbreaks [24]. Although digital surveillance cannot replace traditional surveillance of infectious diseases, they are useful in filling the critical gaps. To support disease warnings and early detection, health information dissemination is very important. The conveyance of valuable and effective information is the basis of disease outbreak surveillance [25]. In our research, we found that social media can support and contribute to early warning systems in outbreak surveillance.

On the other hand, public health education on infectious diseases is critical to manage disease outbreaks. A notable example of the importance of education on disease outbreaks is the Senegal national response to the Ebola outbreak in 2014 [26]. Long before the first and only case was reported, there was extensive health education on Ebola around the country. Unlike other countries in the region, Senegal created a high level of alertness by providing the necessary health education that helped reduce the threat of Ebola to the barest minimum during the Ebola epidemic.

Several studies have revealed that the severity of an epidemic is strongly linked to the education and social behaviour in a population [27][28]. Some of the issues emerging from our findings on country distribution of research on disease prediction relate directly to poor education. Traditional and cultural beliefs also play an important role in the transmission of these diseases. Despite the knowledge on the spread of infectious diseases, victims recieve treaments from relatives or traditional health practitioners who have little or no experience about the treatment. This practice increases the chances of family members getting sick when they come in contact with infected relatives [29].

7. CONCLUSIONS

This paper has analysed a collection of papers in the area of prediction of infectious diseases, especially based on data from social media. Twitter has been shown to be the main prediction source of data. Interestingly, if possibly unsurprising, most research in this area comes from the high income countries, although the affected targets are mostly low and middle income countries. It is imperative to involve the targets in such research in the future, starting with a better support for education, as discussed here.

8. REFERENCES

[1] H. Heesterbeek *et al.*, “Modeling infectious disease dynamics in the complex landscape of global health,” *Science*, vol. 347, no. 6227. 2015.

[2] D. E. Bloom, D. Cadarette, and J. Sevilla, “The Economic Risks and Impacts of Epidemics,” 2018.

[3] 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, pp. 1–10, Dec. 2016.

[4] J. Deng, F. Qiao, H. Li, X. Zhang, and H. Wang, “An overview of event extraction from twitter,” *Proc. - 2015 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov. CyberC 2015*, pp. 251–256, 2015.

[5] Jeff Schultz, “How Much Data is Created on the Internet Each Day? | Micro Focus Blog,” 2017. [Online]. Available: https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/. [Accessed: 11-Jun-2019].

[6] M. Salathé, “Digital pharmacovigilance and disease surveillance: Combining traditional and big-data systems for better public health,” *J. Infect. Dis.*, vol. 214, no. Suppl 4, pp. S399–S403, Dec. 2016.

[7] A. Jimeno Yepes, A. MacKinlay, and B. Han, “Investigating Public Health Surveillance using Twitter,” *Acl-Ijcnlp 2015*, no. BioNLP, pp. 164–170, 2015.

[8] S. Romano, S. Di Martino, N. Kanhabua, A. Mazzeo, and W. Nejdl, “Challenges in detecting epidemic outbreaks from social networks,” in *Proceedings - IEEE 30th International Conference on Advanced Information Networking and Applications Workshops, WAINA 2016*, 2016, pp. 69–74.

[9] N. Rusk, “Deep learning,” *Nature Methods*, vol. 13, no. 1. Nature Publishing Group, p. 35, 30-Dec-2015.

[10] A. Alessa and M. Faezipour, “A review of influenza detection and prediction through social networking sites,” *Theor. Biol. Med. Model.*, vol. 15, no. 1, pp. 1–27, 2018.

[11] E. O. Nsoesie, J. S. Brownstein, N. Ramakrishnan, and M. V. Marathe, “A systematic review of studies on forecasting the dynamics of influenza outbreaks,” *Influenza Other Respi. Viruses*, vol. 8, no. 3, pp. 309–316, May 2014.

[12] L. E. Charles-Smith *et al.*, “Using social media for actionable disease surveillance and outbreak management: A systematic literature review,” *PLoS One*, vol. 10, no. 10, p. e0139701, Oct. 2015.

[13] I. C. H. Fung *et al.*, “Ebola virus disease and social media: A systematic review,” *Am. J. Infect. Control*, vol. 44, no. 12, pp. 1660–1671, Dec. 2016.

[14] R. A. Oldroyd, M. A. Morris, and M. Birkin, “Identifying methods for monitoring foodborne illness: Review of existing public health surveillance techniques,” *J. Med. Internet Res.*, vol. 20, no. 6, p. e57, Jun. 2018.

[15] W. H. Organization, “Managing Epidemics: Key Facts About Major Deadly Diseases,” 2018.

[16] A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media,” *Bus. Horiz.*, vol. 53, no. 1, pp. 59–68, Jan. 2010.

[17] F. S. Tabataba *et al.*, “A framework for evaluating epidemic forecasts,” *BMC Infect. Dis.*, vol. 17, no. 1, p. 345, Dec. 2017.

[18] E. Aramaki, S. Maskawa, and M. Morita, “Twitter catches the flu: detecting influenza epidemics using Twitter,” *Emnlp*, pp. 1568–1576, 2011.

[19] A. Khatua, A. Khatua, and E. Cambria, “A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks,” *Inf. Process. Manag.*, vol. 56, no. 1, pp. 247–257, Jan. 2019.

[20] S. Chae, S. Kwon, and D. Lee, “Predicting Infectious Disease Using Deep Learning and Big Data,” *Int. J. Environ. Res. Public Health*, vol. 15, no. 8, p. 1596, Jul. 2018.

[21] Centers for Disease Control and Prevention [CDC], “Key Facts about Influenza (Flu) & Flu Vaccine | Seasonal Influenza (Flu) | CDC,” *Centers for Disease Control and Prevention (CDC)*, 2015. [Online]. Available: http://www.cdc.gov/flu/keyfacts.htm. [Accessed: 24-Jun-2019].

[22] World Health Organization, “Ebola virus disease fact sheet 103: August 2015,” 2015. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs103/en/#. [Accessed: 07-Jun-2019].

[23] WHO, “Meningococcal meningitis Fact Sheet,” 2015. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/meningococcal-meningitis. [Accessed: 07-Jun-2019].

[24] H. Hu, H. Wang, F. Wang, D. Langley, A. Avram, and M. Liu, “Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network,” *Sci. Rep.*, vol. 8, no. 1, p. 4895, Dec. 2018.

[25] M. Odlum and S. Yoon, “What can we learn about the Ebola outbreak from tweets?,” *Am. J. Infect. Control*, vol. 43, no. 6, pp. 563–571, Jun. 2015.

[26] O. O. Oleribe *et al.*, “Ebola virus disease epidemic in West Africa: Lessons learned and issues arising from West African countries,” *Clin. Med. J. R. Coll. Physicians London*, vol. 15, no. 1, pp. 54–57, Feb. 2015.

[27] H. Xiang, N. N. Song, and H. F. Huo, “Modelling effects of public health educational campaigns on drinking dynamics,” *J. Biol. Dyn.*, vol. 10, no. 1, pp. 164–178, Jan. 2016.

[28] B. Levy *et al.*, “Modeling the role of public health education in Ebola virus disease outbreaks in Sudan,” *Infect. Dis. Model.*, vol. 2, no. 3, pp. 323–340, Aug. 2017.

[29] G. Fitzpatrick *et al.*, “Describing readmissions to an ebola case management centre (CMC), Sierra Leone, 2014,” 2014.