Using twitter and web news mining to predict the monkeypox outbreak

Jahanbin Kia, Jokar Mohammad, Rahmanian Vahid

Abstract

Monkeypox is a zoonotic disease caused by monkeypox virus (MPXV). MPXV is a double-stranded DNA virus from the genus Orthopoxvirus that was firstly detected in captive monkeys in 1958[1], with first reported MPXV infected case in humans in the Democratic Republic of the Congo in 1970. This zoonotic infection has since spread and become endemic in African countries, especially in West and Central Africa[2]. Human monkeypox cases were detected outside Africa in 2003, however, recent cases have not been reported in non-endemic countries until 2022[3].

Massive data is created daily in online social media and the internet in various fields such as medicine, technology, politics, history, arts, and social news, etc. Twitter is one of the most popular social networks, with about 35 million daily active users in the US alone. In previous studies, twitter data analysis has been proven useful in public health, especially in infectious diseases[4],[5],[6],[7]. Monitoring and analyzing social networks about the spread of infectious diseases is one of the ways to control and prevent epidemics. In this study, the Fuzzy Algorithm for Monitoring, Extraction and Classification (FAMEC) method[4],[6] was used to send an alert message to surveillance systems for timely detection of monkeypox outbreaks. The steps of the FAMEC method were performed as follows: 1) Collecting and clearing text and extracting words; 2) Storing data and applying the fuzzy classifier to classify words related to monkeypox; 3) Illustratiing on the world map (using heatmap layer on the image to show the intensity of transmitted tweets in different countries, calculating the percentage of transmitted tweets separately for each country); 4) Extracting words cloud to display the most used terms in tweets. Detailed descriptions of the study design, standardized data collection protocol, and methods of FAMEC have been provided before by the authors[4],[5].

For seven continuous days, from 11:00 am 16 May 2022 to 12:00 pm 22 May 2022, monkeypox tweets were investigated on the twitter social network. The collected database contained 384 560 tweets from 180 407 users; 1 874 861 users have re-tweeted or liked these posts; and these posts have been viewed 3 451 762 times by users. The main hashtags were #monkeypox #pox #vaccine. [Figure 1] illustrates the results obtained by monitoring monkeypox news for seven days, from 11:00 am 16 May 2022 to 12:00 pm 22 May 2022, related to 384 560 tweets from 180 407 users.
The most frequent tweets about monkeypox were in the Americas (44%), including the US and Canada, in Europe (31.9%), the UK, the Netherlands, and Italy. In addition, in Africa (15.4%), in two parts of West Africa, including Mauritania, Mali, and Senegal (West African species of

the virus), and Central Africa, the countries of Zambia and Zamia have the most monkeypox tweets. In addition, the percentage of tweets posts in Australia was 4%. This is in line with the data released by the World Health Organization (WHO)[3].

The number of researchers in health-related infodemiology has increased; moreover, most of them are working on trend analysis, categorizing, clustering, and sentiment analysis of social data on twitter, Google, and news websites[4],[6]. The application of tweeter mining in predicting and detecting various disease outbreaks is shown in different studies such as swine flu pandemic prediction, influenza epidemics detection, and utilized surveillance system for influenza and cancer detection[7]. Many studies have displayed the importance of tweets being compatible with the WHO and the Centers for Disease Control and Prevention reports and recognized that mining these tweets' data was practical to determine the geographic area of patients and monitor and predict the morbidity and mortality rates of COVID-19. Therefore, tweet mining helps in rapid assessment of treatment, better application of telemedicine, and sanitization of the area in epidemics[8]. A similar study used the FAMEC model to mine, text clean, and classify COVID-19 data from twitter, and accordingly, this highly predictable model was identical to the actual incidence of COVID-19 cases[6]. Therefore, promptly sending warning messages to policymakers and surveillance systems to detect quarantined medical centers is one of the FAMEC model applications. The incidence of monkeypox reported by the WHO is consistent with the geographical location of the tweets sent about monkeypox during the study period; thus, this result shows the potential of the FAMEC model to track and monitor this zoonotic disease.

The limitations of this study are as follows: this method cannot be used to track and monitor monkeypox in areas with poor or no access to social networks such as twitter and Facebook[5]; the language of posted tweets processing was English; hence, the result may be influenced by the language.

In conclusion, the analysis and processing of social media data have revolutionized infodemiology, which helps researchers investigate human-related events accurately. Furthermore, these social networks report various statistical data such as the most comments, photos, videos, etc. about social-trend diseases like monkeypox. Therefore, this allows for predicting monkeypox morbidity rates in each area and brings awareness to health policymakers to implement educational and preventional programs in the higher-risk regions. Finally, this may help decrease the incidence of monkeypox cases and even mortality in communities.