Methods 129 (2017) 50–59

Contents lists available at ScienceDirect

Methods

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

Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks

Sungwoon Choi [a,b], Jangho Lee [a], Min-Gyu Kang [c], Hyeyoung Min [d], Yoon-Seok Chang [e,*], Sungroh Yoon [a,f,*]

[a] Electrical and Computer Engineering, Seoul National University, Seoul 08826, Republic of Korea
[b] IT Convergence, Korea University, Seoul 02841, Republic of Korea
[c] Internal Medicine, Chungbuk National University Hospital, Cheongju 28644, Republic of Korea
[d] RNA Biopharmacy Laboratory, College of Pharmacy, Chung-Ang University, Seoul 06974, Republic of Korea
[e] Department of Internal Medicine, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Gyeonggi-do 13620, Republic of Korea
[f] Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 15 January 2017
Received in revised form 29 July 2017
Accepted 31 July 2017
Available online 13 August 2017

Keywords:
Machine learning
Middle East respiratory syndrome (MERS)
Natural language processing
Sentiment analysis

ABSTRACT

From May to July 2015, there was a nation-wide outbreak of Middle East respiratory syndrome (MERS) in Korea. MERS is caused by MERS-CoV, an enveloped, positive-sense, single-stranded RNA virus belonging to the family Coronaviridae. Despite expert opinions that the danger of MERS might be exaggerated, there was an overreaction by the public according to the Korean mass media, which led to a noticeable reduction in social and economic activities during the outbreak. To explain this phenomenon, we presumed that machine learning-based analysis of media outlets would be helpful and collected a number of Korean mass media articles and short-text comments produced during the 10-week outbreak. To process and analyze the collected data (over 86 million words in total) effectively, we created a methodology composed of machine-learning and information-theoretic approaches. Our proposal included techniques for extracting emotions from emoticons and Internet slang, which allowed us to significantly (approximately 73%) increase the number of emotion-bearing texts needed for robust sentiment analysis of social media. As a result, we discovered a plausible explanation for the public overreaction to MERS in terms of the interplay between the disease, mass media, and public emotions.

1. Introduction

Middle East respiratory syndrome (MERS) is an infectious disease caused by the MERS-coronavirus (MERS-CoV) [1,2]. A large outbreak of MERS occurred in Korea from May to July 2015 [3,4]. Although the country had advanced medical systems with reliable public health monitoring capabilities, the outbreak, which started with a single case, caused massive public fear, affecting various aspects of civil life. Inappropriate initial responses and insufficient information were believed to cause unnecessary social chaos [5–7]. Because of worries about infection, the majority of the public refrained from performing normal social and economic activities [8–10]. For example, owing to a sharp decrease in Chinese tourists, the Korea Tourism Organization (KTO) reported that the number of tourists in June decreased by 41% compared with the previous year [11]. As a result, the country had to experience lower economic growth than originally estimated before the outbreak, even though the Korean government declared a *de facto* end to the MERS outbreak on July 28, 2015, approximately 10 weeks after the first confirmed case [12].

An overreaction to a moderate infectious disease by the public can cause various unnecessary complications. By contrast, negligence with regard to dangerous infections can result in a widespread pandemic that could have been controlled by sufficient public attention. For instance, although the incidence and mortality rate of tuberculosis in Korea are the highest among Organization for Economic Co-operation and Development (OECD) countries, public attention has not been drawn to this airborne disease as vividly as MERS.

Evidently, it would require significant time and resources to monitor public thoughts of and reactions to infectious diseases in a traditional way, which makes it inappropriate for the purpose

* Corresponding authors at: Department of Internal Medicine, Seoul National University Bundang Hospital, Seoul National University College of Medicine, Gyeonggi-do 13620, Republic of Korea (Y-S Chang), Electrical and Computer Engineering, Seoul National University, Seoul 08826, Republic of Korea (S. Yoon).
E-mail addresses: nebulach23@gmail.com (S. Choi), ubuntu@snu.ac.kr (J. Lee), irreversibly@gmail.com (M.-G. Kang), hymin@cau.ac.kr (H. Min), addchang@snu.ac.kr (Y.-S. Chang), sryoon@snu.ac.kr (S. Yoon).
URL: http://data.snu.ac.kr (S. Yoon).

http://dx.doi.org/10.1016/j.ymeth.2017.07.027
1046-2023/

of controlling infectious diseases requiring prompt public and government responses. Instead, utilizing social media can provide a rapid and effective means for monitoring public health on a large scale at low cost. A well-known example is Google Flu Trends, which provides query-based estimates of influenza activities for multiple countries [13].

To investigate what triggered public overreaction to MERS in Korea, we presumed that machine learning-based analysis of media outlets could provide a plausible explanation. From the Internet, we collected articles reported by 153 news media outlets in Korea and comments associated with these articles from day 1 (the first confirmed case on May 20, 2015) to day 70 (the *de facto* end declared by the government on July 28, 2015). In Korea, in addition to Twitter and Facebook (two widely used social networks world wide), short-text comments on news articles are extremely popular and often provide a common medium for expressing personal emotions and thoughts about social phenomena. The machine learning challenges in sentiment analysis using Twitter and Facebook data (such as short text lengths and semantic heterogeneity) would remain the same for mining the short-text comments we collected.

Based on the collected data (which consisted of 86,324,566 words from 490,749 articles and 3,901,985 comments), we performed thorough text mining and comparative analysis, focusing on the interplay between the disease, social/mass media, and public emotions. We developed a machine-learning engine for sentiment analysis of a large population. Our approach utilized information-theoretic and machine-learning techniques (such as topic modeling and word embedding) and included an effective method for extracting emotions from texts that hold sentiments (such as emoticons and so-called *Internet slang*). For comparative analysis, we additionally collected and analyzed articles and comments data for the H1N1 influenza epidemic in Korea in 2009 and the Ebola hemorrhagic fever reports in Korea in 2014.

Through our analysis results, we discovered a loop of information transfers [14] between the media and the public. We believe that this discovery may provide a reasonable explanation of the mechanism that triggered the overreaction to MERS in Korea. In addition, we report various analysis results that should be helpful for alleviating the excessive fear and overreaction of the public regarding nation-wide infectious diseases occurring in the future.

2. Background

2.1. Middle East respiratory syndrome (MERS)

MERS is caused by MERS-CoV, which is an enveloped, positive-sense, single stranded RNA virus belonging to the lineage C of the genus Betacoronavirus (βCoV) in the family Coronaviridae [15]. It was first isolated from the sputum of a 60-year-old man with pneumonia in Saudi Arabia in 2012, and has since spread to the Middle East, Africa, Europe, the United States, and Asia including Korea [1,16]. As of January 10, 2017, World Health Organization (WHO) has reported 1,879 laboratory-confirmed cases of MERS-CoV and 659 deaths associated with MERS-CoV [17]. 27 countries have been affected by an outbreak of MERS-CoV, but the majority of cases (>85%) have been reported from Saudi Arabia [17]. Dromedary camels (*Camelus dromedarius*) are known to be the natural source of infection in human, and consumption of contaminated milk, urine, or meat as well as direct contact with infected camels is the suspected transmission route [16]. In addition, human-to-human transmission through close contact of an infected individual with family members and health care workers was also confirmed [18,19]. For example, Chen X. et al. [20] reported that 94.1% of MERS-CoV cases in the 2015 outbreaks in Korea had a history of contact in hospital facilities, and six cases (3.2%) were infected with MERS-CoV through community contacts.

Clinical features of MERS-CoV infection in humans range from asymptomatic or mild infection to severe acute respiratory diseases, renal failure, and multi-organ failure leading to death [15]. A typical MERS symptom represents fever, shortness of breath, and cough commonly, but not always accompanied by pneumonia [17]. In addition, gastrointestinal symptoms, including diarrhea, nausea, and vomiting, have also been observed [21]. The global mortality rate was about 35.7% as of July 29, 2015, while it was 19.4% in Korea as of Aug 1, 2015 [3,4]. The high-risk group is males above the age of 60 with underlying conditions such as cancer, lung disease, and diabetes [17,22].

Currently, there is no specific therapeutic agent or approved vaccine against MERS-CoV. Broad-spectrum antiviral, ribavirin, in combination with interferon has been found to control MERS-CoV, but their clinical usage is limited due to toxicities [23–25]. Various attempts have been made to develop vaccines against MERS-CoV, and they are based on inactivated or attenuated viruses, viral vectors, virus-like particles, DNAs, or recombinant viral proteins [26]. In particular, subunit vaccines containing Receptor binding domain (RBD) of viral S protein has been shown to elicit strong neutralizing antibody responses in mice, representing a great potential for effective MERS-CoV vaccine development [27,28]. In addition, RBD is an attractive therapeutic target of anti-MERS-CoV drugs [26]. The RBD binds to CD26 or dipeptidyl peptidase 4 (DPP4) expressed on epithelial cells and initiates infection of the host cells [29]. Furthermore, viral proteases such as PLpro and 3CLpro, and viral accessory proteins are also potential targets for antiviral agents [30–32].

2.2. Review of the analysis techniques used in this study

In text mining, the latent Dirichlet allocation (LDA) is a generative, probabilistic model for discrete data [33] and widely used in natural language processing (NLP) for modeling corpora and discovering topics therein. LDA considers a document as a mixture of topics, whose distribution is assumed to have a Dirichlet prior. The applications of LDA include topic modeling, document classification, and collaborative filtering.

For representing words in a text for analysis, we utilize the Word2Vec method, an NLP algorithm that takes a corpus and returns vector representations of the words in the corpus [34]. Word2Vec builds a vocabulary from training data and then learns word representations by, for instance, either the continuous bag-of-words (CBOW) method or the continuous skip-gram method. These representations allow us to add and subtract concepts as if they were ordinary vectors. For instance, we can evaluate an interesting query "`queen − woman + man`" to the result '`king`.' According to Mikolov et al. [34], the CBOW model tends to be more efficient than the skip-gram in training time and has slightly better accuracy for handling frequent words. On the contrary, the skip-gram model is known to be better for limited training data with rare words of phrases.

In this study, we use the skip-gram model because of the need for handling infrequently occurring Internet slangs and their limited training data. In the skip-gram model, we associate each word $w \in W$ with a vector $\mathbf{v}_w \in \mathbb{R}^d$, where W is a vocabulary set, and d is the embedding dimensionality. Let us suppose that the training corpus contains a sequence of $2n+1$ words: $w_{i-n}, \ldots, w_{i-1}, w_i, w_{i+1}, \ldots, w_{i+n}$. The objective function of the skip-gram model is represented by the sum of the log probabilities of the n words surrounding the target word w_i [34]: