Mining Internet Media for Monitoring Changes of Public Emotions about Infectious Diseases

Sungwoon Choi[1,2], Jangho Lee[1], Sangheon Pack[2], Yoon-Seok Chang[3], and Sungroh Yoon[1]
[1]Data Science Laboratory, Electrical and Computer Engineering, Seoul National University, Seoul 08826, Korea
[2]Department of IT Convergence, Korea University, Seoul 2841, Korea
[3]Internal Medicine, Seoul National University Bundang Hospital, Seongnam 13620, Korea
Correspondence: sryoon@snu.ac.kr

***Abstract*—The Internet encompasses websites, email, social media, and Internet-based television. Given the widespread use of networked computers and mobile devices, it has become possible to monitor the behavior of Internet users by examining their access logs and queries. Based on large-scale web and text mining of Internet media articles and associated user comments, we propose a framework to rapidly monitor how the emotion of the public changes over time and apply the framework to a real case of an infectious disease. The proposed methodology will be helpful for performing cost-effective and time-efficient public health monitoring that otherwise would take orders-of-magnitude more time and resources if traditional epidemiology techniques were used.**

I. Introduction

Traditionally, it requires significant time and resources to monitor public emotions about infectious diseases. Since controlling infectious diseases requires prompt public and government responses, using traditional methods developed in epidemiology would be inappropriate in such cases. Instead, utilizing Internet media can provide a rapid and effective means for monitoring public health on a large scale at a lower cost. A well-known example is Google Flu Trends, which provides query-based estimates of influenza activities for multiple countries [1].

In this paper, we propose a framework that can collect Internet media articles reported by news outlets and user comments associated with these articles. Note that our work utilizes media articles and short-text comments on them, not social media (such as Twitter and Facebook) unlike most of the existing related methods.

The reasons are twofold: (1) most online news articles are professionally edited before unloads and do not usually contain inappropriate contents that are frequently found in social media; (2) in many countries, Twitter and Facebook, the two most widely used social networks, are not particularly popular, and collecting enough data for mining in such countries often poses a challenge. Note that the mining challenges in sentiment analysis using Twitter and Facebook data (such as short text lengths and semantic heterogeneity) remain the same for mining short-text comments to media articles.

Based on the collected data we perform thorough text mining and comparative analysis, focusing on the interplay between Internet media and public emotions. We developed a text-mining engine for sentiment analysis of a large population. Our approach utilizes information-theoretic and machine-learning techniques and includes an effective method for extracting emotions from texts that hold sentiments.

To validate our methodology, we applied it to a real infectious disease case that occurred in the Republic of Korea in 2015. To this end, we collected over 180,000 articles and two million comments on them from the Internet during an outbreak of the disease. Using our methodology, we could successfully discover how the public emotion about the diseases changed over time.

We anticipate that our approach will provide an efficient means for rapid monitoring how a large number of people emotionally react to infectious diseases, which then provides useful information for timely disease control purposes.

II. Related Work

There have been approaches that utilize social media for public health surveillance. Paul et al. [2] and Signorini et al. [3] traced the trends of the attack of diseases during outbreaks and analyzed the correlation between the trends and public responses. Corley et al. [4] reported that there was high correlation between the prevalence of influenza that occurred in autumn 2008 in the United States and the quantity of influenza-related personal blogs. Armaki et al. [5] proposed a support vector machine-based method to classify whether a Twitter user was infected by influenza or not based on the tweets of the user.

There are two relevant techniques that deserve explanation to facilitate further reading. Firstly, Word2Vec is a natural language processing (NLP) algorithm that takes a corpus and returns vector representations of the words in the corpus [6]. Word2Vec builds a vocabulary from training data and then learns word representations by continuous bag-of-words or continuous skip-gram. 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'`.

Secondly, transfer entropy (TE) is an information-theoretic measure to quantify directed transfer of information between two random processes [7], [8]. To define the TE from a random process to another, let X and Y denote two stationary Markov processes with order p and let t indicate their time indices. The TE from X to Y is defined as

$$T_{X\to Y} = H(Y[t]|Y[t-1:t-p]) - H(Y[t]|Y[t-1:t-p], X[t-1:t-p])$$

where $H(X)$ represents the Shannon entropy of process X.

III. Methods

To foster understanding of the reader, we start our explanation with a description of the experimental data we used. We then propose two types of analysis methods, public emotion analysis and TE analysis.

A. Data Preparation

We collected Internet media articles and associated short-text comments on Middle East Respiratory Syndrome (MERS), an infectious disease caused by the MERS-coronavirus (MERS-CoV) [9], [10]. Note that a large outbreak of MERS occurred in South Korea between May and July of 2015 [11], [12].

Specifically, we targeted the 153 media companies existing in Korea and extracted technicall all of their news articles and comments (written in Korean) containing the word "MERS" uploaded during the outbreak. The numbers of MERS-related articles and comments were 187,295 and 2,431,030, respectively. From each article and comment, we extracted the title, time, contents, and reply counts and stored them in the JavaScript Object Notation (JSON) format (http://json.org). We also collected MERS epidemic data from the Korean Ministry of Health and Welfare [12] and World Health Organization (WHO) websites [11].

B. Public Emotion Analysis

The input for this step was the set of 2,431,030 short-text comments. The output was the daily trend of each of the public fear emotion to MERS. In preprocessing step, we tokenized the text and removed unnecessary components while preserving emoticons and Internet slang. Then, we replaced Internet slang and emoticons with the words corresponding to their emotions. We used part-of-speech (POS) tagging to choose adjectives and nouns after that we created the vector representation of each word using the Word2Vec approach. To implement Word2Vec, we used the library provided by Mikolov et al. [6]. Using the resulting vector representations, we computed the proximity of fear emotion word to MERS appearing in the articles, producing daily time series of emotion trends.

C. TE Analysis

Based on the results from the article and comments analyses, we derived the time series for three types of variables for the MERS epidemic, mass media, and public emotion. To quantify the information transfers among them, we computed the TE values and generated the results shown in Fig. 1. We used the JIDT toolkit [13] for TE computation. The statistical significance of the TE values was tested using block bootstrapping [8].

IV. Results and Discussion

The objectives of our experiments were to reveal the interactions between Internet media, and public emotion. We thus collected the time-series of the numbers of the online articles (for the mass media variable), and the short-text comments on the articles (for the public emotion variable) during the MERS outbreak in South Korea.

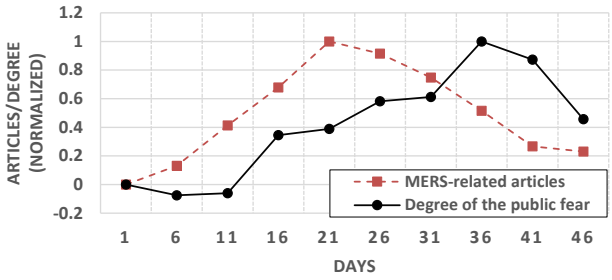

Fig. 1. Interactions between Internet media and public emotion revealed by the trends of the numbers of MERS-related articles and the degree of fear.

TABLE I. The Influence between Two Variables Measured by TE

Information Flow	Transfer Entropy
Mass Media to Public Fear	0.55
Public Fear to Mass Media	0.49

To understand the interplay between Internet media and public emotion, we showed the time series corresponding to each in Fig. 1. The two time series represent the normalized numbers of MERS-related articles and public fear. To quantify the flow of influence from one series to another, we measured the TE values between each pair of time series, as listed in Table I that shows the direction and magnitude of the TE values between variables.

We found a feedback loop of information transfers between the media and emotion variables, which represents an overreaction of the pubic to MERS. Understanding these interactions between the two variables analyzed may provide a helpful means to prevent similar reactions to infectious diseases from happening in the future.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science, ICT and Future Planning, MSIP) [No. 2011-0009963 and No. 2015M3A9A7029735], by the Future Flagship Program (10053249), by the Korean National Police Agency funded by MSIP (PA-C000001), by grant from the Seoul National University Bundang Hospital Research Fund (12-2013-009) and by a research grant from Samsung Advanced Institute of Technology.

References

[1] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[2] M. J. Paul and M. Dredze, in *ICWSM*, 2011, pp. 265–272.

[3] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic," *PloS one*, vol. 6, no. 5, p. e19467, 2011.

[4] C. Corley, A. R. Mikler, K. P. Singh, and D. J. Cook, in *BIOCOMP*, 2009, pp. 340–346.

[5] E. Aramaki, S. Maskawa, and M. Morita, in *Proceedings of the conference on empirical methods in NLP*. Association for Computational Linguistics, 2011, pp. 1568–1576.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[7] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[8] J. Kim, G. Kim, S. An, Y.-K. Kwon, and S. Yoon, "Entropy-based analysis and bioinformatics-inspired integration of global economic information transfer," *PloS one*, vol. 8, no. 1, 2013.

[9] A. M. Zaki, S. Van Boheemen, T. M. Bestebroer, A. D. Osterhaus, and R. A. Fouchier, "Isolation of a novel coronavirus from a man with pneumonia in saudi arabia," *New England Journal of Medicine*, vol. 367, no. 19, pp. 1814–1820, 2012.

[10] A. Zumla, D. S. Hui, and S. Perlman, "Middle east respiratory syndrome," *The Lancet*, 2015.

[11] WHO: Middle East respiratory syndrome coronavirus (MERS-CoV). (2015) http://www.who.int/csr/don/29-july-2015-mers-saudi-arabia/en/.

[12] Korean Ministry of Health & Welfare: MERS-CoV daily update. (2015) http://www.mw.go.kr/front_new/al/sal0301vw.jsp?par_menu_id=04&menu_id=0403&page=1&cont_seq=324570.

[13] J. T. Lizier, "Jidt: an information-theoretic toolkit for studying the dynamics of complex systems," *arXiv preprint arXiv:1408.3270*, 2014.