보건학 박사학위논문

Development of Influenza Surveillance Model

Based on Internet Search Query and Social Media Data

인터넷검색쿼리와 소셜미디어 데이터를 활용한

사회인구학적 독감 감시모형개발

2015년 8월

서울대학교 대학원

보건학과 보건학 전공

우 혜 경

Development of Influenza Surveillance Model

Based on Internet Search Query and Social Media Data

by

Hyekyung Woo

A dissertation submitted to the faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of Doctor of Philosophy

Graduate School of Public Health

Seoul National University

Development of Influenza Surveillance Model Based on Internet Search Queries and Social Media Data

A dissertation submitted in partial fulfillment of the requirement for the degree of

Doctor of philosophy in Public Health

To the Faculty of the Graduate school of public health

at

Seoul National University

by

Hyekyung Woo

Data approved: June, 2015

Sung-Il	**Cho**
Ho	**Kim**
Chang-Gun	**Lee**
Taemin	**Song**
Youngtae	**Cho**

인터넷 검색쿼리와 소셜미디어 데이터를 활용한 사회인구학적 독감 감시모형개발

Development of Influenza Surveillance Model Based on Internet Search Queries and Social Media Data

지도교수 조 영 태

이 논문을 보건학박사 학위논문으로 제출함

2015년 4월

서울대학교 보건대학원

보건학과 보건인구학 전공

우 혜 경

우혜경의 박사학위논문을 인준함

2015 년 6 월

위 원 장 조 성 일 (인)

부 위 원 장 김 호 (인)

위 원 이 창 건 (인)

위 원 송 태 민 (인)

위 원 조 영 태 (인)

Abstract

Development of Influenza Surveillance Model Based on Internet Search Query and Social Media Data

Hyekyung Woo

Department of Public Health Science

Graduate School of Public Health

Seoul National University

Seasonal influenza epidemics present a significant public health challenge, and early detection is crucial for disease control. In the last few years, the availability of big data from novel sources has contributed substantially to influenza surveillance. The purpose of this study is to investigate, with an application to seasonal influenza epidemic, whether Internet-based online surveillance could be helpful to complement and intensify the traditional surveillance system in South Korea. In addition, I propose a pragmatic method for detecting the influenza epidemic in Korea using Internet based big data, especially social media data and web search engine query data. The concerns and specific approach of this study are summarized as follows: (1) The first study is to identify keywords as a predictor for detecting influenza epidemic using social media data, especially twitter and web blog. (2) The second study is to construct a forecast model for detecting influenza epidemic using search engine query data based on the keywords identified from social media data.

In the 1st study, I identified keywords predicting influenza epidemics from social

media data. I included data from Twitter and online blog posts to obtain a sufficient number of candidate predictors and to represent a larger proportion of the Korean population. The methods used this study include (a) initial keyword selection, (b) generation of the keyword time series, and (c) selection of optimal features for model building. I built the candidate models using the least absolute shrinkage and selection operator (Lasso), support vector machine for regression (SVR), and random forest regression (RFR) using the training set based on the features we selected. To find the model having the best performance, I evaluated the root mean square error (RMSE) of the predicted values and ILI incidence using the validation set. A total of 15 keywords optimally predicted influenza epidemic, evenly distributed across Twitter and blog data source. Predictions generated from using SVR model were highly correlated with the recent influenza incidence data (SVR model correlation: r=0.92, p<.001; RMSE=0.55).

In the 2nd study, I described a methodological extension for detecting influenza outbreaks using Internet search query; I provided a new approach for query selection through the exploration of contextual information gleaned from social media data. Additionally, I evaluated whether it is possible to use these queries for monitoring and predicting influenza epidemics in South Korea. My study was based on freely available weekly influenza incidence data and query data originating from the search engine on the Korean web site Daum between April 3, 2011, and April 5, 2014. In order to select queries related to influenza epidemics, several approaches were applied: (a) exploring influenza-related words in social media data (b) identifying the chief complaints related to influenza, and (c) using web query recommendations. Optimal feature selection by Lasso and SVR were used to construct a model for predicting influenza epidemics. A considerable proportion of optimal features for final models were derived from queries with reference to the social media data. The SVR model performed well: the prediction values were

highly correlated with the recent observed ILI (SVR model의 correlation: r= 0.956, p<.001; RMSE=0.39) and the virological incidence rate (SVR model의 correlation: r= 0.963, p<.001; RMSE=7.24).

My models for detecting national influenza incidence have the power to predict. These results demonstrate the feasibility of search queries and social media data in enhancing influenza surveillance in South Korea. The current study provides further evidence, based on a new approach, for linkages between the use of Internet-based data and the surveillance of emerging influenza incidence in South Korea. I found that internet-based influenza surveillance that combines search engine query data with social media data has the power to predict influenza outbreaks, exhibiting strong congruence with traditional surveillance data. Furthermore, in an attempt to exploit the complementary nature of the two types of data sources in this study, I fused information drawn from social media with the methodology for query-based influenza surveillance. As seen through my results, these new data sources may be compatible and complementary in predicting influenza incidence. In addition, the basic principles underpinning my approach could be applied to other countries, languages, infectious diseases and data sources.

Keywords: influenza, surveillance, population surveillance, infodemiology, infoveillance, Internet search query, social media, big data, forecasting, epidemiology, early response

Student number: 2011-30701

Contents

List of Tables

List of Figures

Chapter 1 General introduction

1.1 Public health and Health information

Obtaining useful and accurate health information is a major public health concern. According to Winslow (1920), public health is defined as ‘the science and art of preventing disease, prolonging life and promoting health through the organized efforts and informed choices of society, organizations, public and private, communities and individuals’ [1]. Although specific challenges related to public health have evolved over time, the significance of health information has been constantly emphasized. Public health decision-making is critically dependent on accurate information.

The advent of the Web 2.0 paradigm has had significant implications for the development of health-related information. As Internet availability and use have extensively and rapidly increased in the past 15 years, many people have shifted from a mode of passive consumption of health information to one of more active creation [2]. More widespread access to the Internet and the recent appearance of mobile devices and platforms such as smartphones and tablets make it possible for people to easily generate and share health information anytime and anywhere [2, 3]. In this setting, keywords such as *big data*, *social networks*, *crowdsourcing*, *user-generated content*, *folksonomy*, and *micro-blog* have emerged as playing a prominent role. Internet users can communicate via online platforms that allow social

interaction [4]. These circumstances have also changed how people seek information about health [5]. Ongoing technological changes may eventually provide a new means of obtaining contextualized health information that can be used for public health objectives.

Surveys represent a popular method for obtaining information related to public health but they can be costly and time consuming. In more current research pertaining to public health surveillance, multiple sources of information have been utilized to identify the various phases and manifestations of disease. Novel sources of data that are useful for health surveillance include Internet-based big data such as news reports, search engine query data, and social media data. Behavioral interventions have become increasingly important for achieving good public health outcomes, and there is tremendous potential for Internet-based big data to contribute to a better understanding of health-related phenomena and the behavior of individuals within a large population. The new technology of online health surveillance is garnering interest as a low-cost and effective approach for detecting, tracking, reporting, and managing public health in real time [6]. Although the technology is hampered by several limitations such as noise from irrelevant information, the lack of Internet accessibility in some countries, and the lack of a well-defined study population, most researchers agree that the new data paradigm has significant possibilities and implications with respect to health surveillance through the development of sophisticated methodologies.

1.2 Online surveillance for detection and monitoring infectious diseases

A main public health concern is the challenge of controlling emerging infectious diseases. Disease control requires an effective and rapid national surveillance system. Traditional surveillance systems rely on formal medical and/or public health networks involving professionals such as physicians, laboratory staff, epidemiologists, and other healthcare workers [5]. Such systems typically exhibit substantial time lags between the outbreak of disease and its detection, thereby resulting in notification delays. Late reporting by data providers and the hierarchical structure of national disease surveillance systems contribute to these time lags. Moreover, maintaining and operating surveillance systems is costly. Such limitations of traditional surveillance systems represent shared concerns around the world.

In South Korea, the Korea Centers for Disease Control and Prevention (KCDC) operates infectious disease surveillance systems. The national incidence of notifiable infectious diseases is calculated using data reported by medical providers, including Western and Asian medical doctors, to health institutions. Physicians working at clinics or hospitals report cases of infectious disease to their regional health centers. These data are reviewed by local health authorities and transferred to the KCDC. After final confirmation from the KCDC, a report of these national statistics is published [7] (Figure 1.1).

The KCDC publishes national data on infectious disease gathered by these surveillance systems on a weekly basis using a Web-based statistical system (http://stat.cdc.go.kr), typically with a 1-week reporting lag. Their national data regarding notifiable infectious diseases relies on the passive reporting of physicians, which results in a low reporting rate [7]. Although an Internet-based reporting system, developed for use in clinics and hospitals in 2009, facilitates more efficient and convenient data collection than paper-based reporting, the recent increase in emerging infectious diseases has led to a call for methodological improvements to ensure effective and rapid surveillance.

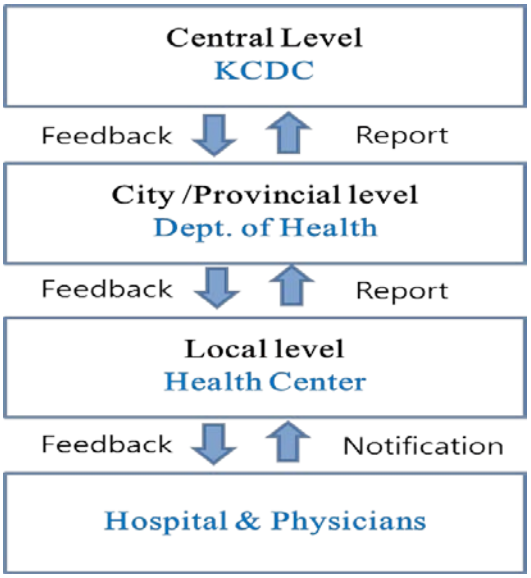


Figure 1.1 Hierarchical structure of the national infectious disease surveillance system in South Korea

Source: Korea Centers for Disease Control and Prevention (KCDC)

To date, the number of Internet users has reached almost 3 billion globally and Internet use penetration has reached 40%, with a penetration rate of 78% in developed countries and 32% in developing countries [8]. As a result of the development of Internet infrastructure around the world, Internet availability and use have increased in the past decade (Figure 1.2). South Korea was ranked first among 157 countries with respect to access to broadband Internet by the ICT Development Index, issued by the International Telecommunication Union (ITU) in 2014 [8, 9]; according to the 2013 survey, the Internet usage rate among Koreans aged 3 or older was 82.1% [9]. With the ongoing shift from a wired Internet based on desktop computers to a mobile Internet based on smartphones and tablets, the number of individuals using the Internet is likely to increase even more dramatically.

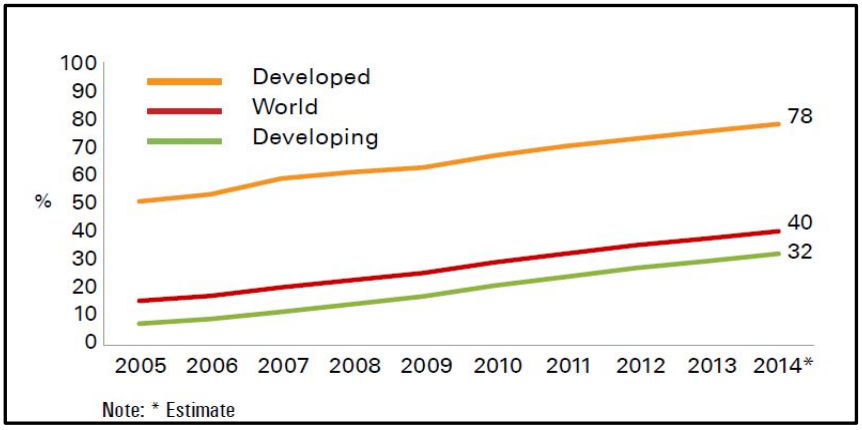


Figure 1.2 Percentage of individuals using the Internet, 2005–2014

Source: ITU World Telecommunication/ ICT Indicators database,

The World in 2014: ICT Facts and Figures, 2014 (http://www.itu.int/en/ITU-D/Statistics/)

These circumstances make it possible to develop online surveillance methods for detecting and monitoring infectious diseases. Internet usage is associated with the seeking and sharing of health information. Some users write expositions about their health through various social media channels such as blogs and Twitter. In addition, users leave logs and queries pertaining to health-related questions on the Internet search engines of websites. The access and availability of the Internet have made it much easier than before for people to seek and/or share health information themselves, leading to changes in how information about health is used [5, 10]. These transitions pave the way for a new paradigm in which health information is both contributed and led by users.

To counter emerging infectious diseases, novel technologies for surveillance should be designed in such a way as to complement the traditional surveillance system. The initial days of an epidemic represent a critical period for health authorities in terms of initiating appropriate interventions. An online surveillance system provides a novel approach, congruent with a traditional surveillance system, for monitoring public health [5]. Such a system allows for near real-time and cost-effective monitoring of the outbreaks of infectious diseases through rapid data collection. Therefore, it is increasingly desirable for public health authorities to seriously consider adopting and applying these Internet technologies for the purposes of assessing, protecting, and promoting public health. There have been attempts to monitor disease through online data sources, and although these efforts are still very much in their initial stages, they offer significant implications for public health, suggesting that, overall, it may be possible to bolster the

capacity of a traditional surveillance system through an online surveillance system.

1.3 Novel data Sources for Online Surveillance

The Internet has become an important tool for seeking health information, among the general public as well public health practitioners, physicians, epidemiologists, researchers, and other healthcare workers. Recent attempts at public health surveillance have relied on multiple data sources, from formal channels to informal channels. Sources of data on disease surveillance are disseminated on the Internet in various forms and languages. The specific approaches to online surveillance have depended on the nature of the data sources. Several studies have attempted online surveillance utilizing informal channels such as online news reports [11–13], social media data [14–20], and search engine query data [6, 21–24]. These sources provide information that may serve as a valuable complement to the data gathered by traditional infectious disease surveillance systems.

1.3.1 Online News Reports

In April 2000, the World Health Organization (WHO) launched a special network infrastructure, the Global Outbreak Alert and Response Network, to promote early awareness of disease outbreaks and enhance response preparedness. The network is based on a computer-driven tool for

the real-time gathering of information pertaining to outbreaks of disease [25]. This information, investigated by a network of WHO, is initially obtained from informal online sources such as news reports. However, these sources are not organized or integrated in such a way as to facilitate knowledge management and early detection of outbreaks [13].

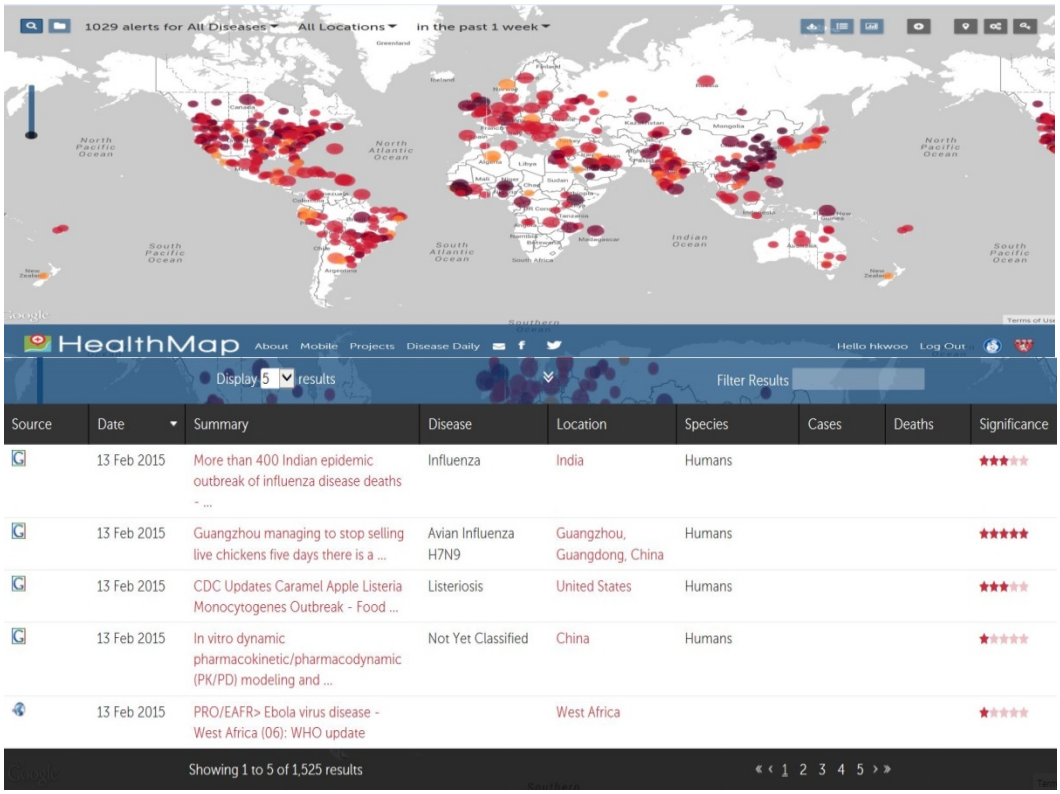


Figure 1.3 User interface of HealthMap. HealthMap monitors Web-based information related to outbreaks of infectious disease through a variety of media sources globally and displays results in near real time on a world map. (http://www.healthmap.org/en/index.php)

Several initiatives have been introduced to address this challenge, including the Program for Monitoring Emerging Diseases (ProMED-mail) [26], the Global Public Health Intelligence Network (GPHIN) [27], HealthMap [13], BioCaster [28], and EpiSPIDER [29]. These systems represent novel approaches to online health surveillance based on the event-based monitoring of infectious disease outbreaks by means of organized and integrated Internet-based news reports, online discussion sites, and other electronic media [12, 13, 27–29]. The promising potential of these systems has been demonstrated during recent outbreaks. For example, GPHIN played a significant role in detecting SARS more than 2 months prior to the first reports by the WHO [5].

Current surveillance systems overwhelmingly rely on media sources for data input, including news broadcasts, websites, news wires, local and national news reports, and even short message service (SMS) components of mobile communication systems [30]. This novel approach has resulted in innovative transitions within the field of global health surveillance. The great advantage of these systems is that they have access to detailed local and near real-time information pertaining to disease outbreaks. In addition, because they rely on open sources of information that are available to users free of charge, they substantially reduce the cost of monitoring outbreaks, a factor that is particularly beneficial for countries that lack the more costly traditional systems for public health surveillance [11]. Unstructured information on outbreaks exists in various forms on the Internet. Overall, these automated Web-crawling surveillance systems can be very useful for gathering and analyzing this information.

1.3.2 Social Media Data

Social media such as Twitter, Facebook, and blogs provide venues in which individuals express their emotions, thoughts, and desires. Thus, these media have the power to act both as a means of disseminating information and as sources of highly contextual information [4, 5]. Communication strategies related to public health include engagement with social media, thereby allowing health authorities to become aware of and respond to real or perceived concerns that are raised by the public [18]. During the 2009 *H1N1* epidemic, the US Centers for Disease Control and Prevention (CDC) posted information on Facebook to educate the public about the disease and the importance of vaccination. Moreover, the CDC responded immediately on Facebook and Twitter to health concerns regarding the contraction of the *H1N1* virus from eating pork [4].

Social media are also potentially useful with respect to assessing the incidence and distribution of disease. Several studies have investigated methods that use social media data to predict disease incidence, and have suggested that the analysis of such data may be useful for identifying emerging disease epidemics. For example, Corley et al. [31] proposed a method that evaluates blog posts discussing influenza. They analyzed the volume of posts containing the keywords *influenza* and *flu*, and found a high correlation between the volume of posts containing these keywords and CDC surveillance data reporting influenza-like illness (ILI) [31]. In a Chinese investigation, daily posts submitted to the Sina microblog and the Baidu website were analyzed [32]. The number of posts involving the keyword

H7N9 increased rapidly during the first 3 days of outbreaks of this avian influenza and remained at a high level for several subsequent days. The authors suggested that the first 3 days of an epidemic represent a critical period for an appropriate response from authorities and that Internet surveillance using data from social media can be used efficiently and economically for the control and prevention of public health emergencies.

Most studies that have used social media data have used data from microblogs, especially Twitter [15–20], which allows users to communicate through status updates limited to 140 characters. Chew and Eysenbach [18] demonstrated that tweets can be linked with influenza data reflecting *H1N1* incidence rates in the US by monitoring the use of the terms *H1N1* and *swine flu* over a period of time. Between May 1 and December 31, 2009, the relative proportion of tweets containing the word *H1N1* increased in an almost linear fashion ($R^2 = 0.788$, $p < .001$). Broniatowski et al. [20] recently developed a system for estimating influenza prevalence based on data from Twitter. They used an algorithm to detect actual influenza infections based on tweets, and showed that the rate of tweets indicative of influenza infection was correlated with surveillance data gathered by the US CDC ($r = 0.93$, $p<.001$). In addition, because some social media, including Twitter, are tagged for geographic location, location information can be used to identify the distribution of a disease. Furthermore, social media data can be used for near real-time content analysis.

1.3.3 Internet Search Engine Queries

One way to obtain insights that are relevant for public health monitoring and intervention is by detecting health-seeking behavior on the Internet. Recently, there has been increasing interest in the use of logs of queries submitted to search engines as novel and potential sources of data relevant for public health. Numerous studies have explored the usefulness of online search behavior for the purpose of monitoring disease outbreaks. The authors of these studies have noted that individuals faced with disease or ill health typically seek information on the Internet regarding their health state and possible remedies or countermeasures; logs of queries submitted to search engines for the purpose of seeking this information may be useful for the detection of emerging epidemics, which can be achieved by tracking changes in the volumes of relevant search queries.

There are several advantages to using search query data in health studies such as epidemiological analyses [33]: the data are almost in real time, it is possible to obtain information with regard to groups of individuals other than those who consult a doctor, a system using these data can easily be adapted to various diseases, and the data reflect a point in time close to onset of illness (provided that the people seeking information are in fact ill).

Ginsberg et al. [23] discussed the possibility of search engine queries being used to detect influenza epidemics, based on the concept that monitoring such queries can facilitate the early detection of infectious diseases. Studies have used a variety of different data sources, including data gleaned from various search engines such as Google, Yahoo!, and Baidu, and

have found that the data collected are highly correlated with disease incidence. Polgreen et al. [34] examined the relationship between searches for the keyword *influenza* and rates of influenza occurrence using queries submitted to the Yahoo! search engine. A Chinese study [6] presented a method of using Internet search query data from Baidu to detect influenza activity, motivated by the fact that Baidu is more widely used in China than Yahoo! or Google. Swedish studies [33, 35] examined influenza-related search activity by monitoring anonymous logs of search queries from the national medical website Vårdguiden.se (www.vardguiden.se). Google Trends, a publicly accessible online portal of Google Inc., has been used in many studies with a wider range of applications. Several studies have estimated a correlation between Google Flu Trends (GFT) and national data on ILI or laboratory-confirmed influenza.

Although Google is the most widely used search engine in the world, it is not dominant in South Korea. Local search engines based on the Korean language, such as Daum and Naver, are used more widely than Google. In addition, despite the fact that GFT is available in many countries, this is not the case in South Korea. Because of these limitations inherent to South Korean data, few studies have evaluated whether search query data have the potential to be of value for the detection or monitoring of influenza in this country [36, 37]. One recent study [36] suggested that GFT is insufficient for use in a model of influenza prediction in South Korea.

1.4 Further Challenges for Online Surveillance

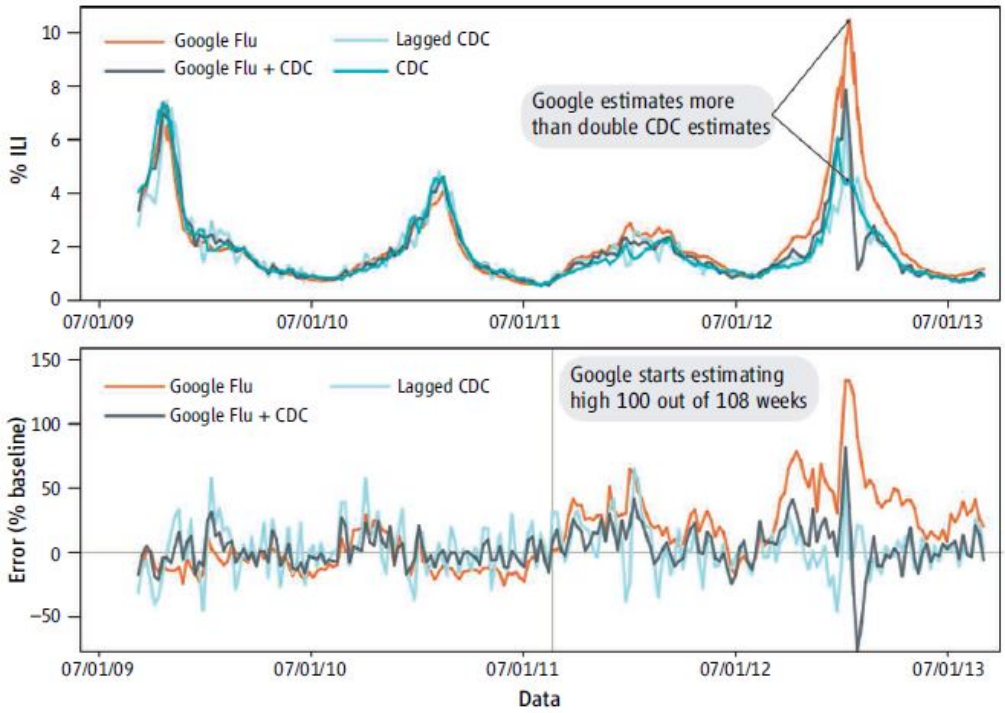
There is no disagreement regarding the potential of using Internet-based data as a highly informative source of data for real-time surveillance. However, this novel and developing approach has led to calls for more sophisticated methodologies that will increase the reliability of surveillance due to several limitations to the data that have been used. The first and most significant limitation is that it is unclear whether these data can be used to represent the entire population, given the lack of a well-defined study population. Second, noise from irrelevant information may decrease the accuracy of the surveillance data. Third, there are difficulties with respect to the classification and interpretation of textual data. Fourth, because individual behavior constantly changes, keywords submitted by individuals may be influenced by various factors over time. Such changes may alter or degrade the performance of an online surveillance system. Finally, emerging infectious diseases are a global public health concern because new diseases emerge all over the world. However, the lack of easy Internet accessibility in some countries is the biggest barrier to the development of an international approach for bolstering an online disease surveillance system worldwide. For these reasons, Internet-based surveillance systems should be viewed as extensions that reinforce the capacity of traditional surveillance systems rather than as alternatives [5].

1.4.1 Recent Errors of GFT

The problem of over-fitting is well known in the domain of big data analysis. Recent prediction errors arising from GFT have led to debates regarding limitations in obtaining reliable predictions from sources of big data [14, 38]. Lazer et al. [38] reported that GFT overestimated the prevalence of flu during the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%; according to the authors, these errors reflected the general limitations of big data analysis (Figure 1.4). This suggests that researchers need a better understanding of new data sources and better tools for extracting considerably more information from such sources.

In addition, during the 2009 pandemic of the influenza virus A (H1N1), the original GFT model did not correlate with data from the outpatient Influenza-like Illness Surveillance Network (ILINet) ($r = 0.290$), demonstrating another example of GFT failure [39]. In practice, GFT has also been unable to provide accurate estimates of non-seasonal influenza outbreaks. However, Cook et al. [39] updated the model based on GFT. This updated model, which included more search query terms than the original model, correlated more closely with data from the ILINet ($r = 0.95$) during the H1N1 pandemic. That study indicated that Internet search behavior changed during this pandemic, particularly with respect to the categories “influenza complications” and “term for influenza”. Therefore, it is apparent that changes in Internet search behaviors can affect the performance of online surveillance systems that are based on search queries. Any prediction

model based on Internet data requires a continuous evaluation of the performance of the system.



GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. **(Top)** Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. **(Bottom)** Error [as a percentage {[Non-CDC estmate)−(CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232 for combined GFT and CDC. All of these differences are statistically significant at $P < 0.05$. See SM.

Figure 1.4 GFT overestimation (Source : Lazer et al (2014) [38])

1.4.2 Further Challenges for Online Surveillance in South Korea

On May 16, 2014, the National Health Insurance Service (NHIS) of South Korea launched the National Health Forecast Alarm System on its website (http://hi.nhis.or.kr/). The NHIS analyzed medical data accumulated between 2008 and 2015 and selected four diseases (influenza, eye infections, food poisoning, and allergic contact dermatitis), all of which are good candidates for predictions based on monthly fluctuation rates of related keywords used on Twitter. The alarm system distinguishes diseases by region and age, and classifies the diseases with respect to four levels of risk: attention, caution, alert, and critical. The system informs users of daily precautions to be taken at each level (Figure1. 5).

Although the alarm system provides an information service using online surveillance, it is still very much in its infancy. It relies on a small number of related keywords, despite the rich source of available information associated with the various phases and manifestations of disease epidemics. Furthermore, Twitter use is not widespread in South Korea, although a significant number of people use social network services (SNSs); in Korea, only 19.4% of SNS users use Twitter. To represent a larger proportion of the Korean population, a study that focuses on multiple data sources is necessary.

To this end, it is important to assess the predictive utility of data from Twitter. To do so, it needs to be determined whether a prediction model based on data from social media has the capacity to predict the occurrence of diseases and their proliferation in South Korea. Thus, there is a need to develop a maximally reliable prediction model for the risk of disease,

classifying a relevant medical lexicon in accordance with the targeted disease's symptoms, causes, and time of occurrence.

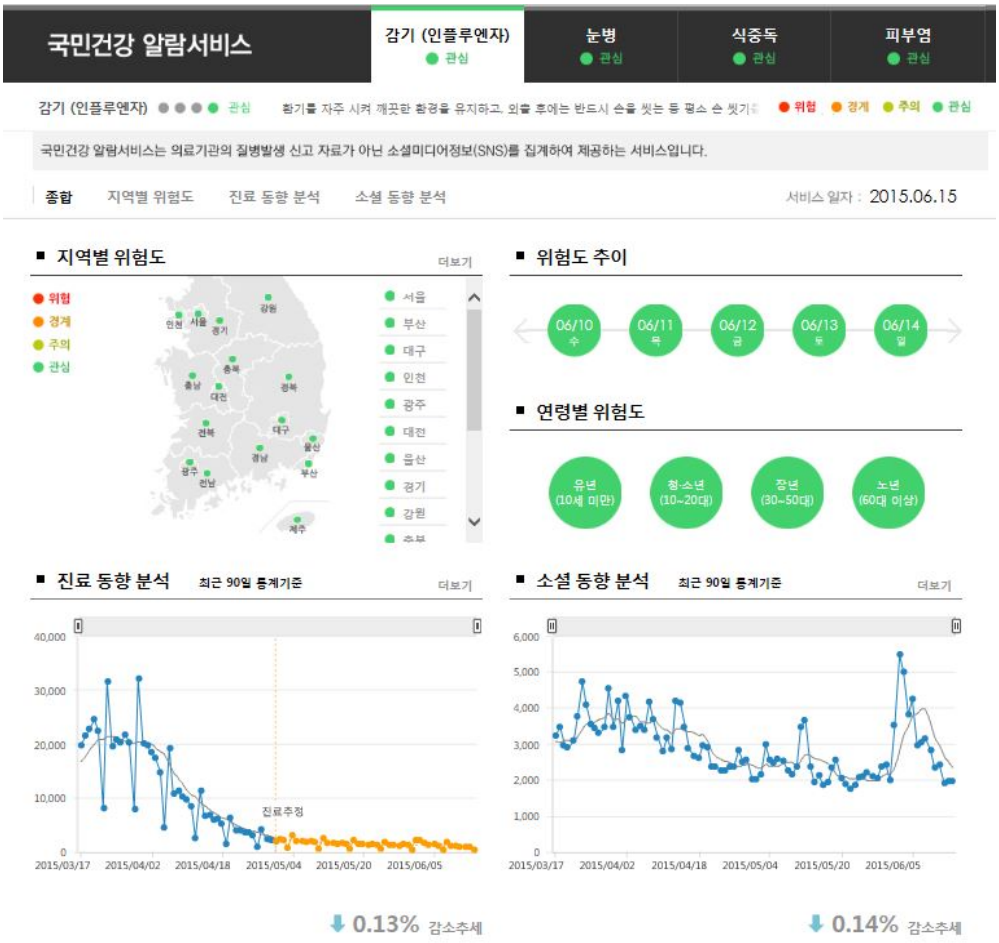


Figure 1.5 User interface of the National Health Forecast Alarm System. The system provides online surveillance for four diseases (influenza, eye infections, food poisoning, and allergic contact dermatitis) based on Twitter and medical data. (http://forecast.nhis.or.kr)

Nevertheless, most researchers agree that new data paradigms offer promising potential for the surveillance of infectious disease, provided that sophisticated methodologies can be developed [5]. Future research regarding online surveillance needs to devote careful attention to the previously discussed limitations. As the usage of Internet-based data for monitoring infectious disease epidemics increases over time, validation will be an important issue. To demonstrate the value of online surveillance as a complement to existing systems, it is crucial to provide empirical justification for the knowledge that can be derived from novel data sources and to develop advanced methodologies for analyzing the data.

1.5 Background and Objectives

1.5.1 Research Background

The recent increase in emerging infectious diseases has renewed public concerns regarding the global and national transmission of disease. Advanced models for controlling and responding to emerging epidemics are required; health organizations need accurate and timely disease surveillance systems. In the case of the KCDC, the surveillance of diseases nationwide has been based on data reported by medical providers to health institutions; national statistics from the previous week are summarized and published only after a process of data collection and aggregation, which in turn relies on confirmation from within the hierarchical structure of the surveillance network [7]. Accurate and timely disease surveillance systems require substantial financial resources and organized surveillance networks. These challenges have led to calls for new approaches and technologies to reinforce the capacity of traditional surveillance systems for detecting emerging infectious diseases, as discussed above.

1.5.2 Objectives

The purpose of this study is to investigate, with an application to seasonal influenza epidemic, whether Internet-based online surveillance could be helpful to complement and intensify the traditional surveillance

system in South Korea. In addition, I propose a pragmatic method for detecting the influenza epidemic in Korea using Internet based big data, in particular data from social media and search engine query.

1.5.3 Dissertation Design

The concerns and specific approach of this study can be summarized as follow:

⦿ The first study is to identify keywords as a predictor for detecting influenza epidemic using social media data, especially twitter and web blog.

1. Identifying the specific keywords or terms associated with influenza
2. Keyword filtering through processing keywords by excluding those unrelated to influenza epidemics
3. Listing complex keyword and simple keyword
4. Generating the keyword time series with application of several extract conditions

⦿ The second study is to construct a forecast model for detecting influenza epidemic using search engine query data based on the keywords identified from social media data.

1. Keywords re-indexing using traditional information source and the keywords identified in the social media data.
2. Fit several machine learning models
3. Model selection and validation

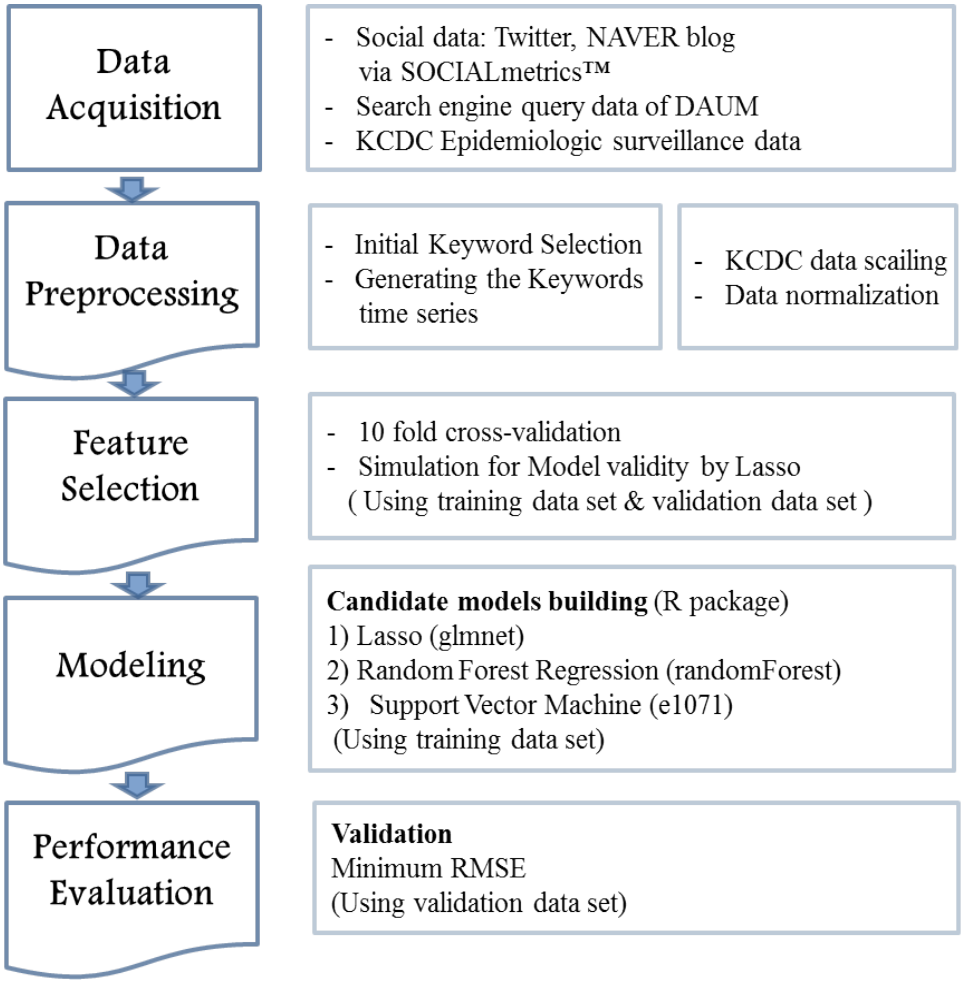


Figure 1.6 Main analysis methods in this study.

The remainder of this paper is organized as follows. Chapter 2 describes the methodology for identifying keywords as predictors for detecting influenza epidemics in Korea using social media data. Chapter 3 presents the methodology for constructing a forecast model for detecting influenza epidemics, using the search engine query and social media data. Chapter 4 presents a general discussion and addresses limitations, public health implications, and conclusions.

References

1. Winslow, C.E., *THE UNTILLED FIELDS OF PUBLIC HEALTH.* Science, 1920. **51**(1306): p. 23-33.
2. Scanfeld, D., V. Scanfeld, and E.L. Larson, *Dissemination of health information through social networks: twitter and antibiotics.* Am J Infect Control, 2010. **38**(3): p. 182-8.
3. Prieto, V.M., et al., *Twitter: a good place to detect health conditions.* PLoS One, 2014. **9**(1): p. e86191.
4. Kass-Hout, T.A. and H. Alhinnawi, *Social media in public health.* Br Med Bull, 2013. **108**: p. 5-24.
5. Milinovich, G.J., et al., *Internet-based surveillance systems for monitoring emerging infectious diseases.* Lancet Infect Dis, 2014. **14**(2): p. 160-8.
6. Yuan, Q., et al., *Monitoring influenza epidemics in china with search query from baidu.* PLoS One, 2013. **8**(5): p. e64323.
7. Park, S. and E. Cho, *National Infectious Diseases Surveillance data of South Korea.* Epidemiol Health, 2014. **36**: p. e2014030.
8. ITU, *The World in 2014: ICT Facts and Figures, 2014*, in *ICT Facts and Figures, 2014*. 2014, International Telecommunication Union: Geneva, Switwerland.
9. KISA, M.a., *Survey on the Internet Usage, Korea*, M.a. KISA, Editor. 2013. 12, MSIP and KISA: Seoul, Korea.
10. Rice, R.E., *Influences, usage, and outcomes of Internet health information searching: multivariate results from the Pew surveys.* Int J Med Inform, 2006. **75**(1): p. 8-28.
11. Keller, M., et al., *Use of unstructured event-based reports for global infectious disease surveillance.* Emerg Infect Dis, 2009. **15**(5): p. 689-95.

12. Brownstein, J.S., et al., *Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project.* PLoS Med, 2008. **5**(7): p. e151.
13. Freifeld, C.C., et al., *HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports.* J Am Med Inform Assoc, 2008. **15**(2): p. 150-7.
14. Lazer, D., et al., *Twitter: big data opportunities--response.* Science, 2014. **345**(6193): p. 148-9.
15. Pawelek, K.A., A. Oeldorf-Hirsch, and L. Rong, *Modeling the impact of twitter on influenza epidemics.* Math Biosci Eng, 2014. **11**(6): p. 1337-56.
16. Santos, J.C. and S. Matos, *Analysing Twitter and web queries for flu trend prediction.* Theoretical Biology and Medical Modelling, 2014. **11**(Suppl 1): p. S6.
17. Prieto, V.M., et al., *Twitter: a good place to detect health conditions.* PloS one, 2014. **9**(1): p. e86191.
18. Chew, C. and G. Eysenbach, *Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak.* PloS one, 2010. **5**(11): p. e14118.
19. Paul, M.J., M. Dredze, and D. Broniatowski, *Twitter improves influenza forecasting.* PLoS Curr, 2014. **6**.
20. Broniatowski, D.A., M.J. Paul, and M. Dredze, *National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic.* PloS one, 2013. **8**(12): p. e83672.
21. Althouse, B.M., Y.Y. Ng, and D.A. Cummings, *Prediction of dengue incidence using search query surveillance.* PLoS neglected tropical diseases, 2011. **5**(8): p. e1258.
22. Hulth, A. and G. Rydevik, *Web query-based surveillance in Sweden during the influenza A (H1N1) 2009 pandemic, April 2009 to February 2010.* Euro

Surveill, 2011. **16**(18): p. -.

23. Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data.* Nature, 2009. **457**(7232): p. 1012-4.
24. Milinovich, G.J., et al., *Using internet search queries for infectious disease surveillance: screening diseases for suitability.* BMC Infect Dis, 2014. **14**(1): p. 3840.
25. Heymann, D.L. and G.R. Rodier, *Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases.* Lancet Infect Dis, 2001. **1**(5): p. 345-53.
26. Madoff, L.C., *ProMED-mail: an early warning system for emerging diseases.* Clin Infect Dis, 2004. **39**(2): p. 227-32.
27. Mykhalovskiy, E. and L. Weir, *The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health.* Can J Public Health, 2006. **97**(1): p. 42-4.
28. Collier, N., et al., *BioCaster: detecting public health rumors with a Web-based text mining system.* Bioinformatics, 2008. **24**(24): p. 2940-1.
29. Herman Tolentino, M., et al., *Scanning the emerging infectious diseases horizon-visualizing ProMED emails using EpiSPIDER.* Advances in disease surveillance, 2007. **2**: p. 169.
30. Velasco, E., et al., *Social media and internet-based data in global systems for public health surveillance: a systematic review.* Milbank Q, 2014. **92**(1): p. 7-33.
31. Corley, C.D., et al., *Using Web and social media for influenza surveillance.* Adv Exp Med Biol, 2010. **680**: p. 559-64.
32. Gu, H., et al., *Importance of Internet surveillance in public health emergency control and prevention: evidence from a digital epidemiologic study during avian influenza A H7N9 outbreaks.* J Med Internet Res, 2014. **16**(1): p. e20.

33. Hulth, A. and G. Rydevik, *GET WELL: an automated surveillance system for gaining new epidemiological knowledge*. BMC Public Health, 2011. **11**: p. 252.
34. Polgreen, P.M., et al., *Using internet searches for influenza surveillance.* Clin Infect Dis, 2008. **47**(11): p. 1443-8.
35. Hulth, A., G. Rydevik, and A. Linde, *Web queries as a source for syndromic surveillance*. PloS one, 2009. **4**(2): p. e4378.
36. Cho, S., et al., *Correlation between national influenza surveillance data and google trends in South Korea.* PLoS One, 2013. **8**(12): p. e81422.
37. Seo, D.W., et al., *Cumulative query method for influenza surveillance using search engine data.* J Med Internet Res, 2014. **16**(12): p. e289.
38. Lazer, D., et al., *Big data. The parable of Google Flu: traps in big data analysis*. Science, 2014. **343**(6176): p. 1203-5.
39. Cook, S., et al., *Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic.* PloS one, 2011. **6**(8): p. e23610.

Chapter2. Identification of keywords from Twitter and web blog posts to predict influenza epidemics

2.1 Introduction

Seasonal influenza epidemics present a significant public health challenge, and early detection is crucial for disease control. In the last few years, the availability of big data from novel sources has contributed substantially to influenza surveillance. Previous studies have proposed the use of big data analysis of online news reports [1-3], search engine queries [4-6], and social media [7-14] to detect influenza epidemics. Although the predictive utility of big data have been debated in light of the recent errors in Google Flu Trends (GFT) [15, 16], the majority of researchers agree that this novel data paradigm offers significant possibilities for infectious disease surveillance, in accordance with the development of more sophisticated methodologies [17].

Here, I propose an advanced method for the rapid detection of influenza outbreaks by analyzing social media data. My approach to identifying influenza-related keywords from social media is distinctly from that used for search engine query data, which are submitted with the sole purpose of obtaining information. Social media plays host to the expression of a wide variety of personal experiences pertaining to health, including disease

symptoms, coping and recovery [7, 8]. Therefore, it represents a highly contextual data source, with greater breadth compared to search engine query data, and can also be used to disseminate information [17, 18].

Words that consistently appear with the word “influenza” in social media contexts can be used as keywords for the prediction of influenza outbreak. Previous studies have demonstrated that data from Twitter [7, 8, 11-14] and web blog [9, 10] accord with official surveillance data. In a recent study, use of data from Twitter was not associated with outbreak overestimation, unlike GFT. Moreover, the rate of specific tweets was strongly correlated with CDC influenza-like illness rates [7, 19]. However, the use of social media data for influenza surveillance is a relatively novel approach, and several recent studies have relied solely on the keyword “influenza”, despite the rich source of available information associated with the various phases and manifestations of influenza epidemics. Furthermore, in several countries, Twitter use is not widespread, such as Korea, although a number of people use social network services (SNS). In Korea, only 19.4% of SNS users use Twitter [20]. Thus, to represent a larger proportion of the Korean population, we focused on multiple data sources.

This study identified keywords predicting influenza epidemics, taken from both Twitter and blog data. My method included the following steps: initial keyword selection to identify keywords in posts discussing influenza that serve as latent predictors; generation of a keyword time series using a pre-processing approach based on extract conditions; and optimal feature selection using algorithms that improve prediction performance.

2.2 Methods

2.2.1 Data Sources

Social Media Data

The social media data were collected from daily NAVER blog posts and Twitter posts from September 1, 2010, to June 30, 2014, using the social Big Data mining system, SOCIALmetrics™ Academy (Daumsoft [http://www.daumsoft.com/eng/]). The SOCIALmetrics™ system contains social media data crawlers that collect posts from Twitter and NAVER blog and processes texts using state-of-the-art natural language-processing and text-mining technologies. The Twitter crawler utilizes streaming application programmer's interface (API) (http://dev.twitter.com/docs/streaming-apis) for data collection using so-called "track keywords". The system used a few thousands of empirically selected and tuned track keywords that can maximize the coverage of the crawler operating in near-real-time fashion. It was estimated that the daily coverage of the Twitter crawler is over 80%. The collected posts were fed into a spam-filtering module that checks for posts containing spam keywords and written by known spammers. The lists of spam keywords and spammers were semi-automatically monitored and managed. The NAVER blog is a weblog service offered by the biggest portal site in South Korea (http://www.section.blog.naver.com). The NAVER blog crawler resembles general-purpose web crawlers. The big difference is that the system maintains a list of active bloggers for post collection. The active blogger list is automatically expanded. The estimated coverage of the Naver

blog crawler is also over 80%. It was applied an extensive spam-filtering process similar to that of the Twitter crawler on the collected blog posts. I and data mining company conducted the search according to the Twitter and blog post website Terms and Conditions of use. All Twitter and NAVER blog posts were publicly available and the information collected did not reveal the identity of the social media users; thus, user confidentiality was preserved.

Epidemiological Surveillance Data

Official case count data were obtained from the Korea Centers for Disease Control and Prevention (KCDC), which routinely collects epidemiological data. The KCDC publishes statistical information from the national influenza surveillance system on a weekly basis, typically with a 1-week reporting lag. The clinical data were the rate of physician visits for ILI between September 1, 2010, and June 30, 2014. Data were obtained from the weekly reports on influenza surveillance from the KCDC infectious disease web statistics system (http://www.is.cdc.go.kr/nstat/index.jsp). The epidemiological surveillance data we obtained was publicly available and did not identify individual subjects.

2.2.2 Initial Keyword Selection

Keywords were obtained using the following steps:

Identification of specific keywords or terms associated with influenza

Keywords were words associated with the presence of influenza that appeared frequently with the word “influenza” in social media posts. My database comprised words most likely to be associated with *dokgam* (the Korean word for influenza) and *inpeulruenja* (the Korean pronunciation of the English word “influenza”) in the accumulated Twitter and NAVER blog posts over a 43-month tracking period (September 1, 2010, to March 31, 2014). Because the data were based on keyword priority according to the number of accumulated posts for each week, I combined the keyword data and converted it into time series variables between September 1, 2010, and March 31, 2014. In this way, I identified 2,065 associated keywords.

Keyword filtering

Of the 2,065 keywords associated with influenza, several were not related to influenza seasons; thus, it was necessary to filter the keywords. First, I excluded keywords that occurred infrequently during the influenza season and those that showed non-sequential patterns in the time series throughout the tracking period. Then, I selected keywords whose correlation with the epidemiology surveillance data was at least 0.4. After filtering, 32 of the 2,065 keywords associated with the word “influenza” remained.

Classifying complex and simple keywords

Keywords that remained after filtering were classified as “complex keywords”, which represent a combination of the core keyword (CKW) and

32 associated keywords (AKW). The CKW was the synonym for influenza commonly used in Korea to describe influenza (*dokgam, inpeulruenja*, *peulru* [Korean pronunciation of words for flu], influenza [in English], and flu [in English]). Additionally, of the 32 AKW, keywords related to flu symptoms and the five synonyms of influenza were classified as “simple keywords”. Thus, overall, I obtained 32 complex and 17 simple keywords associated with influenza epidemics.

2.2.3 Generating the keyword time series

The keyword data were presented as a time series. Based on the 49 seed keywords, I generated a keyword time series, i.e., the weekly volume of tweets and blog posts mentioning the keywords, which were obtained from Twitter and blog posts independently. Because the social media keyword data were available on a daily basis, whereas the KCDC official case count data were reported weekly, I converted the keyword data to weekly counts for the analysis. Several extraction conditions were applied to generate the keyword time series according to keyword type.

Extraction condition for complex keywords (CKW + AKW)

First, *dokgam*, *inpeulruenja* (*influenza*), and the English words "influenza" and "flu" were treated as synonyms. A post with at least one of these synonyms was considered to contain the CKW. The volume of complex keywords was derived from the number of posts in which CKW and AKW occurred concurrently on a daily basis. Second, I excluded posts with words related to influenza vaccination to maximize the possibility of detecting an influenza epidemic. Since the standard inoculation period for influenza is October to December, words related to influenza vaccination frequently appeared during this time period regardless of the influenza epidemic season. Common Korean synonyms for "influenza vaccine" are *dokgam jeopjong*, *dokgam jusa*, *yebang jusa*, *yebang jeopjong*, and *baeksin* (Korean pronunciation of terms for "vaccine"). Posts with CKW and AKW that contained any of these synonyms were excluded. The Korean film "The Flu,

2013” (Korean title, *Gamgi*) was released on August 14, 2013, during tracking period. To control for the influence of discourse on this film, I excluded posts containing the word *younghwa* (the Korean word for film) after July 1, 2013.

Extract condition for simple keywords

For simple keywords, the time series volume was derived from the number of posts containing AKW regardless of whether CKW appeared. The conditions of exclusion were consistent with those used for complex keywords.

2.2.4 Feature Selection and Model Building

I divided the data into training and validation sets. Data from January 8, 2011, to August 31, 2013 were used as the training set for modeling, and data from September 1, 2013 to June 30, 2014 were used as the validation set to test the model. Because my objective was to develop a method for identifying predictors of an influenza epidemic using social media data, I used a prediction model that reflected the incidence of ILI with a 1-week lag.

To identify the best predictor feature subset, I first used the least absolute shrinkage and selection operator (lasso) algorithm to select the best dataset and features following data normalization. The primary objectives of

feature selection are to avoid overfitting caused by irrelevant features, improve prediction performance of the predictors, and identify faster and more cost-effective predictors [21, 22]. Lasso is useful for efficient and simple feature selection because it tends to assign zero weights to most irrelevant or redundant features [23]. Feature selection processing was performed on the training set using 10-fold cross validation.

I used several machine learning techniques to construct models that predicted influenza epidemics using the best features. I built the candidate models using lasso, support vector machine (SVM), and random forest regression (RFR) using the training set based on the features we selected. To find the model having the best performance, I evaluated the root mean square error (RMSE) of the predicted values and ILI incidence using the validation set. All statistical tests were conducted using R version 3.0.3 (R Development Core Team).

2.3 Results

I identified 2,065 keywords that frequently appeared with the word “influenza” in Twitter and blog posts accumulated during the 43-month tracking period. From those, I created 32 complex and 17 simple keywords associated with influenza epidemics in South Korea [Table 2-1]. The keyword time series were more consistent with the ILI trends following the application of several extraction conditions [Figure 2-1]. Blog_A and Twitter_A (blue) indicate trends based on keyword frequencies per 100,000 daily Twitter and web blog posts mentioning influenza (including all synonyms). Blog_B and Twitter_B (red) indicate trends after the elimination of noise created by irrelevant information from Blog_A and Twitter_A. The use of several extract conditions to remove irrelevant information smoothed the keyword time series and increased the correlation with ILI data.

Table 2.1 Keywords associated with an influenza epidemic

Complex keywords			Simple keywords		
CKW	**AKW**	**English**	**CKW**	**AKW**	**English**
FLU*	심하다	severe	none	독감유행	influenza epidemic
FLU*	아프다	be sick	none	독감증상	influenza symptom
FLU*	유행	epidemic	none	독감환자	influenza patient
FLU*	기침	cough	none	감기	cold
FLU*	감기	cold	none	기침	cough
FLU*	병원	hospital	none	몸살	body aches
FLU*	조심하다	be careful	none	고열	high fever
FLU*	좋은	good	none	콧물	runny nose
FLU*	정상	symptom	none	근육통	muscular pain
FLU*	몸살	body aches	none	신종플루	new influenza
FLU*	환자	patient	none	바이러스	virus
FLU*	검사	check	none	독감	influenza
FLU*	고열	high fever	none	인플루엔자	influenza
FLU*	몸	body	none	플루	flu
FLU*	상태	condition	none	influenza	
FLU*	의사	doctor	none	flu	
FLU*	입원	hospitalization	none	FLU*	
FLU*	신종플루	new influenza			
FLU*	건강	health			
FLU*	면역력	immunity			
FLU*	목	throat			
FLU*	콧물	runny nose			
FLU*	입	mouth			
FLU*	근육통	muscular pain			
FLU*	날씨	weather			
FLU*	엄마	mother			
FLU*	바이러스	virus			
FLU*	머리	head			
FLU*	아이	child			
FLU*	진료	medical treatment			
FLU*	뉴스	news			
FLU*	약	medicine			

CKW, core keyword; AKW, associated keyword
FLU*, synonym of the word “influenza” commonly used by Koreans.

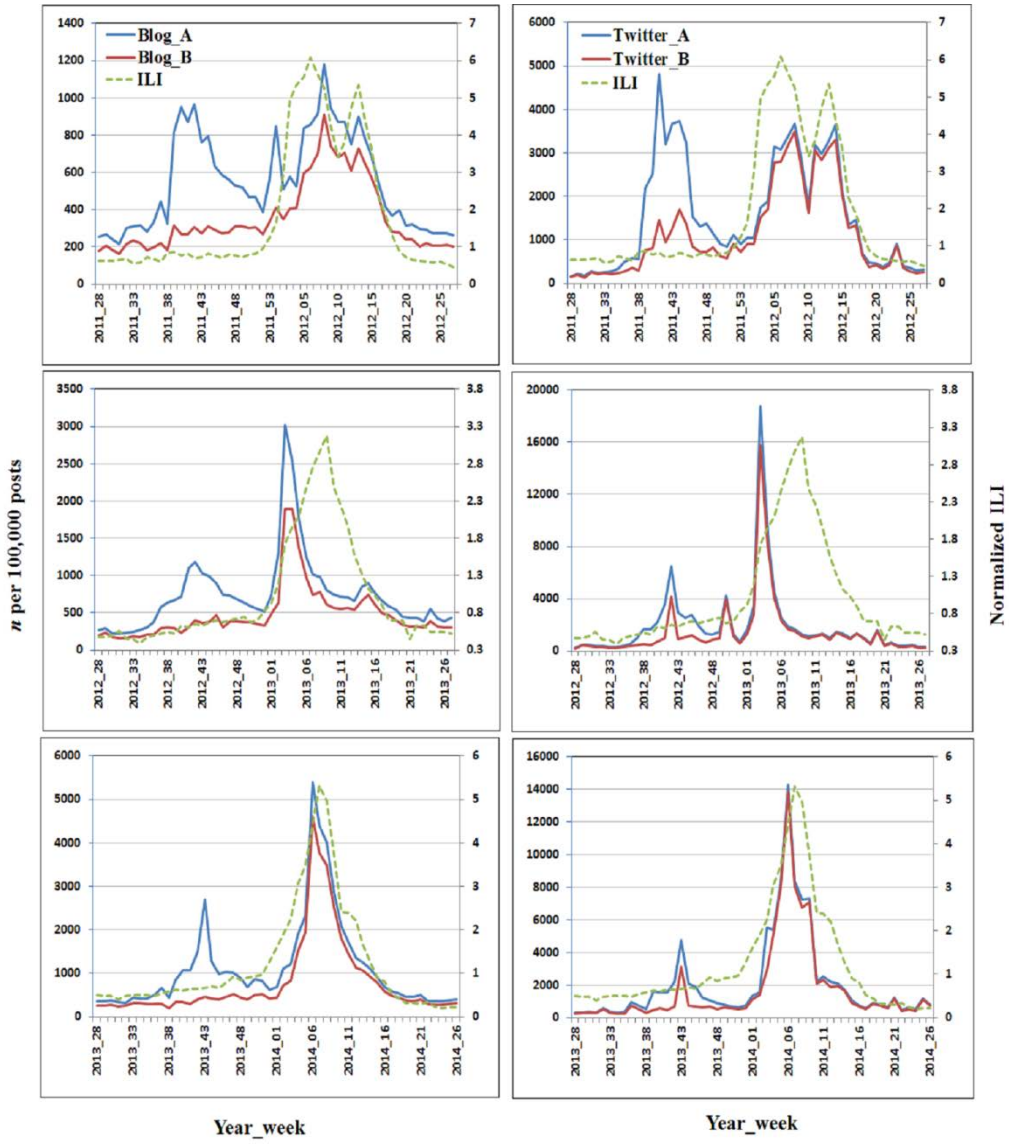


Figure 2.1 Comparison of trends in social media and influenza-like illness data before and after preprocessing to eliminate irrelevant information.

ILI, influenza-like illness.

The best dataset was chosen by repeating random sampling 10 times using the lasso method [Figure S2-1], and 147 variables were ultimately generated for analysis. The results of feature selection processing are shown in Figure 2 and Table 2. Of the 147 variables, 15 principle features had the minimum lambda value in the lasso algorithm [Figure 2-2]. The best predictive features for early detection of influenza epidemics were derived from five variables from the Twitter or blog data sources and from a combination of both [Table 2-2].

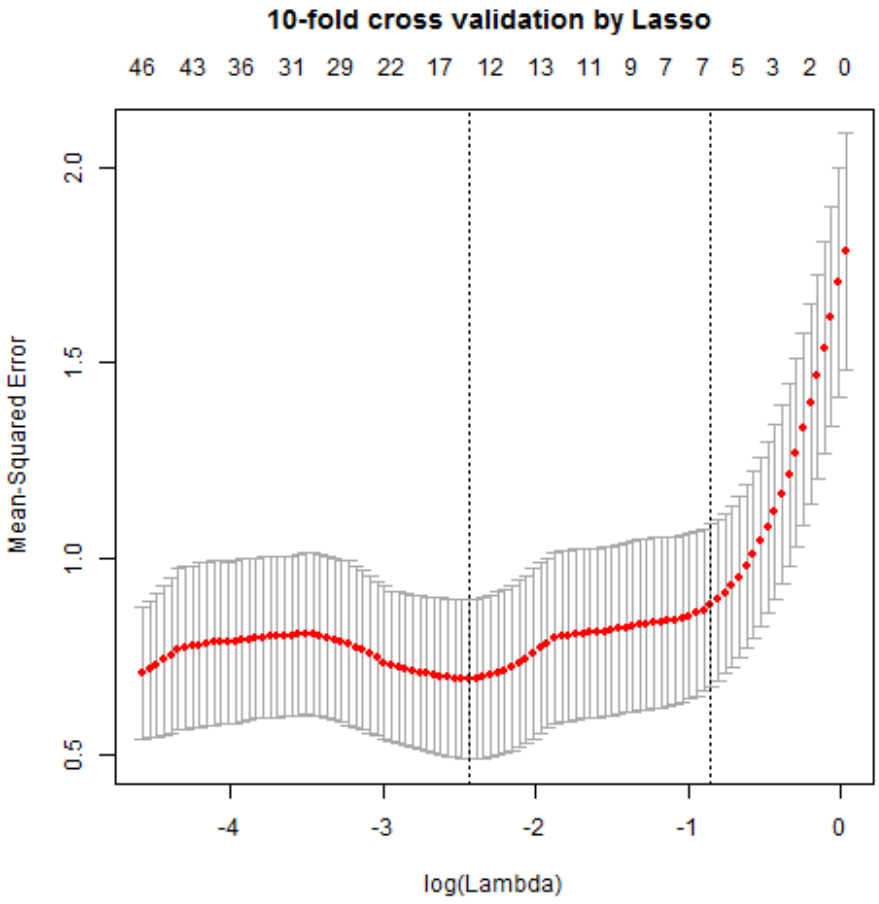


Figure 2.2 Optimal feature selection. This figure shows the results of feature selection processing. Of the 147 variables identified, 15 best features were chosen by repeating random sampling 10 times using the lasso method.

Table 2.2 Best predictive features for model building.

Feature			English	10cv.lasso_coef
(Intercept)				-0.107354349
Combination set	FLU*	입원	hospitalization	3.540934576
	FLU*	아이	child	2.411909888
	none	독감증상	influenza symptom	2.663244975
	none	고열	high fever	0.019603955
	none	Flu(English)		-0.116806973
Blog set	FLU*	아프다	be sick	1.816563906
	FLU*	환자	patient	-0.212374067
	FLU*	고열	high fever	1.893500106
	FLU*	콧물	runny nose	0.090843549
	FLU*	아이	child	1.159064819
Twitter set	FLU*	좋은	good	-0.082907135
	FLU*	검사	check	1.013703003
	FLU*	바이러스	virus	-1.956746459
	FLU*	진료	medical treatment	1.211738323
	none	몸살	body aches	2.466706959

10 CV, 10-fold cross validation; coef, coefficient.
FLU*, synonyms of the word “influenza” commonly used by Koreans.

I compared the performance of candidate models using the validation set of KCDC surveillance data to predict the next observation. As a result, the SVM model (cost=2.7; gamma=0.0002) performed well, having a minimum RMSE and a correlation between predictions and observed cases of 0.92 [see Figure 2-2]. The RFR and GBR models did not perform better than the SVM model [Figures S2-2 to S2-3].

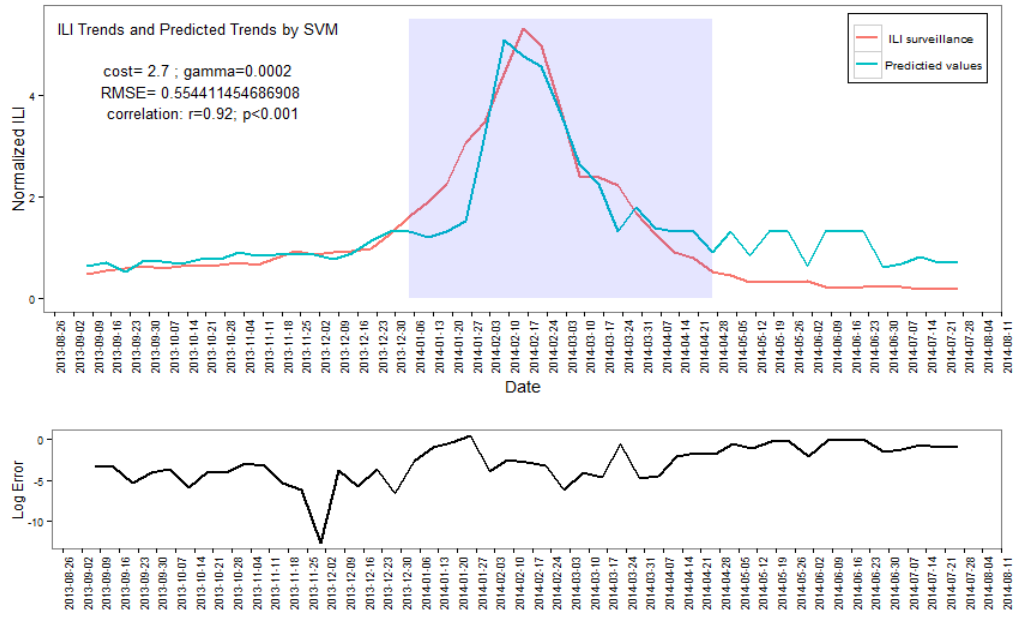


Figure 2.3 Prediction and error based on the SVR model. This figure shows the performance of the SVR model using the validation set of KCDC surveillance data to predict the next observation. The SVM model (cost=2.7; gamma=0.0002) performed well, the predictions were highly correlated with the observed ILI (0.92, $p < 0.001$).

ILI, Influenza-like illness; SVR, Support vector machine for regression

Note: Log Error = $\log((\text{Obs-Exp})^2/\text{abs}(\text{Exp}))$

2.4 Discussion

I have developed an advanced method to enable early detection of influenza epidemics through pre-processing of social media data. The best predictor features for detecting influenza epidemics in Korea included 15 keywords derived from Twitter and web blog posts. The prediction based on best features was highly correlated with the incidence of influenza.

The findings contribute to growing evidence suggesting that social media data are useful for detection of influenza outbreaks. Furthermore, my approach has several significant implications in terms of methodology for influenza surveillance using social media data. First, to my knowledge, this study is the first to use data from both Twitter and web blogs for influenza surveillance. Most investigations of social media intending to monitor influenza epidemics have examined either tweets [7, 8, 11-14] or blog posts [9, 10]. However, I found that the best predictive features were evenly distributed between Twitter and blog posts. It may be that consideration of pertinent social media data will be required for influenza surveillance studies in the future. Second, my findings suggest that the rate of Twitter use, amount of traffic, and number of Twitter users in the population would not matter in developing the influenza surveillance model. Twitter is not widely used by Koreans; yet, a considerable body for our model was derived from Twitter or a combination of Twitter. This study examined social media text collected from SNS in contrast to traditional surveillance studies based on individuals. The text content included information about family, friends, and the communities of the social media users as well as about the users

themselves. Therefore, the aggregated texts including health-related conversations on the internet may be representative of a large segment of the population. Thus, these findings provide a counterargument to the contention that internet-driven data, such as social media and search queries, cannot be used to represent an entire population.

Third, although the weight of various keywords is likely to deviate from one influenza season to another [4, 15], prediction model based on a combination of best features from Twitter and blog posts performed well for the recent influenza season. Whether predictors will be consistent in the future remains to be seen. However, I am reasonably confident that our approach to keyword selection through the pre-processing of social media data (i.e., analyzing associations, generating keyword time series using several extraction conditions, and selecting optimal features from a supervised learning framework) can be used widely. Furthermore, my approach offers clues for understanding such predictors and their weight, which may be changeable over time. Online surveillance detects disease by tracking the behavior of individuals through massive amounts of aggregated data on the internet; however, individual behavior is constantly changing, and thus keywords change. My approach can help reflect time-varying keywords as predictors. Fourth, although model was created to predict the incidence of influenza throughout the year including high- and low-incidence seasons, I found a strong correlation between the influenza surveillance data reported by the KCDC and our predictions using a model that included 15 keywords ($r = 0.92$; $p < 0.001$). My predictions were more highly correlated with KCDC data than were those reported in previous studies. In the

majority of previous studies, the correlations between prediction and observation were not as high as those reported for approaches based on search engine queries [17]. Although a recent investigation of a filtering algorithm developed to estimate an influenza epidemic using Twitter revealed a strong correlation between the Twitter and United States CDC surveillance data ($r = 0.93$), the authors studied only the specific influenza season as defined by the United States CDC [7]. The GFT was not able to provide an accurate prediction of non-seasonal influenza outbreaks during the 2009 influenza virus A (H1N1) pandemic. The original GFT model was not correlated with the United States Outpatient Influenza-like Illness Surveillance Network data ($r = 0.290$) [24], suggesting that the ability of a model to predict non-seasonal influenza outbreaks is necessary. Finally, best predictive features included six symptom-related and four healthcare-related keywords, two nomenclature terms, and three other keywords. Prediction was derived from contextual keywords such as “good,” “child,” “be sick,” “hospitalization,” and “patient” that are not commonly used in search engine queries. This finding suggests that social media incorporate a wide range of contextual health-related information not included in search engine queries that are submitted to gather information.

Most agree that social media constitute a highly informative data source for real-time monitoring of emerging epidemics. However, the conversion of these informative data into a novel and reliable body of information requires sophisticated methods. Noise from irrelevant information, uncertainty about the representativeness of the social media users, changing rates of internet use over time are several examples of the limitations of social media as a

surveillance tool. This study has several of these limitations; nevertheless, the proposed method is useful for the preliminary detection of early signs of an influenza outbreak. Furthermore, the basic principles of my approach can be applied to other countries, languages, infectious diseases, and types of social media.

References

1. Collier, N., et al., *BioCaster: detecting public health rumors with a Web-based text mining system.* Bioinformatics, 2008. **24**(24): p. 2940-1.
2. Freifeld, C.C., et al., *HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports.* J Am Med Inform Assoc, 2008. **15**(2): p. 150-7.
3. Herman Tolentino, M., et al., *Scanning the emerging infectious diseases horizon-visualizing ProMED emails using EpiSPIDER.* Advances in disease surveillance, 2007. **2**: p. 169.
4. Yuan, Q., et al., *Monitoring influenza epidemics in china with search query from baidu.* PloS one, 2013. **8**(5): p. e64323.
5. Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data.* Nature, 2009. **457**(7232): p. 1012-4.
6. Hulth, A. and G. Rydevik, *Web query-based surveillance in Sweden during the influenza A (H1N1) 2009 pandemic, April 2009 to February 2010.* Euro Surveill, 2011. **16**(18): p. -.
7. Broniatowski, D.A., M.J. Paul, and M. Dredze, *National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic.* PloS one, 2013. **8**(12): p. e83672.
8. Santos, J.C. and S. Matos, *Analysing Twitter and web queries for flu trend prediction.* Theoretical Biology and Medical Modelling, 2014. **11**(Suppl 1): p. S6.
9. Corley, C.D., et al., *Using Web and social media for influenza surveillance.* Adv Exp Med Biol, 2010. **680**: p. 559-64.
10. Gu, H., et al., *Importance of Internet surveillance in public health emergency control and prevention: evidence from a digital epidemiologic study during avian influenza A H7N9 outbreaks.* J Med Internet Res, 2014. **16**(1): p. e20.
11. Paul, M.J., M. Dredze, and D. Broniatowski, *Twitter improves influenza forecasting.* PLoS Curr, 2014. **6**.

12. Chew, C. and G. Eysenbach, *Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak.* PloS one, 2010. **5**(11): p. e14118.
13. Prieto, V.M., et al., *Twitter: a good place to detect health conditions.* PloS one, 2014. **9**(1): p. e86191.
14. Pawelek, K.A., A. Oeldorf-Hirsch, and L. Rong, *Modeling the impact of twitter on influenza epidemics.* Math Biosci Eng, 2014. **11**(6): p. 1337-56.
15. Lazer, D., et al., *Big data. The parable of Google Flu: traps in big data analysis.* Science, 2014. **343**(6176): p. 1203-5.
16. Lazer, D., et al., *Twitter: big data opportunities--response.* Science, 2014. **345**(6193): p. 148-9.
17. Milinovich, G.J., et al., *Internet-based surveillance systems for monitoring emerging infectious diseases.* Lancet Infect Dis, 2014. **14**(2): p. 160-8.
18. Kass-Hout, T.A. and H. Alhinnawi, *Social media in public health.* Br Med Bull, 2013. **108**: p. 5-24.
19. Broniatowski, D.A., M.J. Paul, and M. Dredze, *Twitter: big data opportunities.* Science, 2014. **345**(6193): p. 148.
20. KISDI, *KISDI STAT Report (13-04): Current use of SNS* 2013: Seoul, Korea.
21. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* The Journal of Machine Learning Research, 2003. **3**: p. 1157-1182.
22. Saeys, Y., I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics.* bioinformatics, 2007. **23**(19): p. 2507-2517.
23. Li, F., Y. Yang, and E.P. Xing. *From lasso regression to feature vector machine*. in *Advances in Neural Information Processing Systems*. 2005.
24. Cook, S., et al., *Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic.* PloS one, 2011. **6**(8): p. e23610.

Supporting Information of the Chapter 2

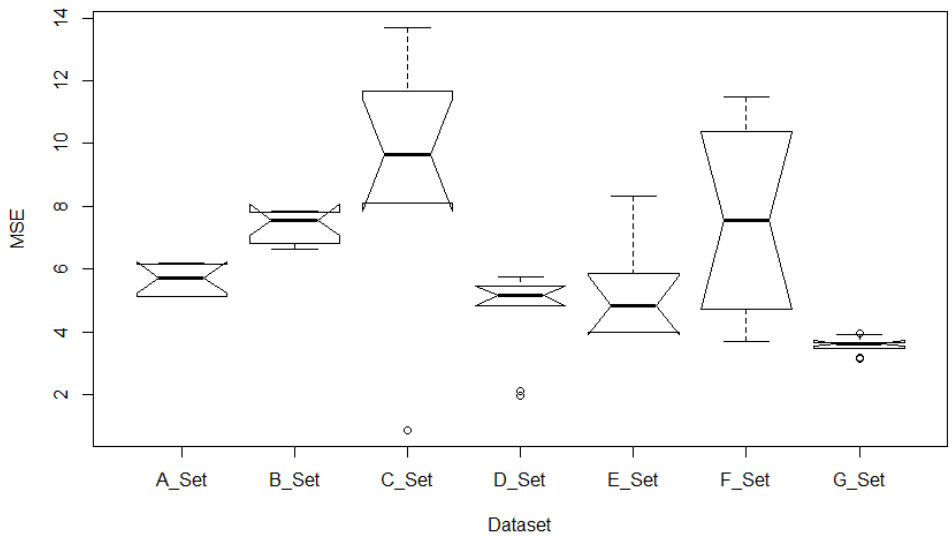


Figure S2.1 Identification of the optimal dataset by 10 times experiments of Lasso.

MSE, mean square error

A Set : (Blog + Twitter)/2

B Set : (Blog)

C Set : (Twitter)

D Set : A Set & B Set

E Set : A Set & C Set

F Set : B Set & C Set

G Set: A Set & B Set & C Set

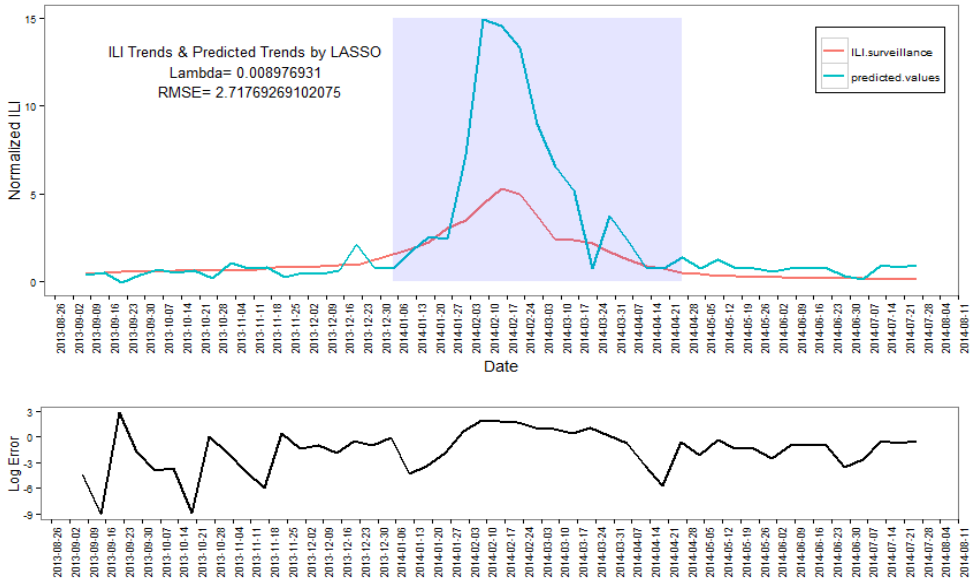


Figure S2.2 Prediction and error based on the lasso model.

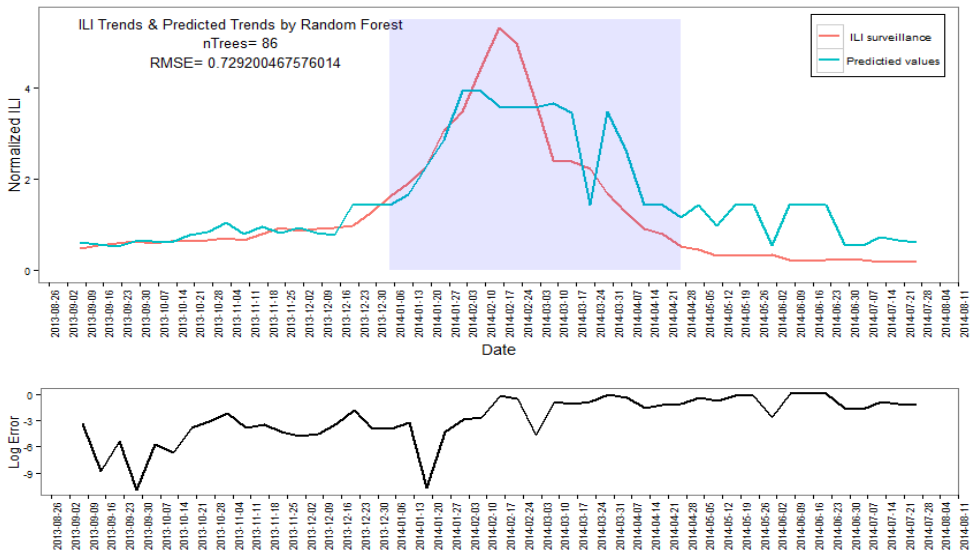


Figure S2.3 Prediction and error based on the random forest model.

Chapter3. Estimating influenza outbreaks using both search engine query data and social media data in South Korea

3.1 Introduction

An initial and well-known example of utilizing internet data for a health-related application came from the estimation of influenza incidence using anonymous logs of web search engine queries. Numerous studies have provided evidence of a correlation between search query data from Google [1-3], Yahoo! [4], Baidu [5], or other medical web sites [6] and traditional data used for influenza surveillance, such as influenza-like illness (ILI) and/or laboratory-confirmed data. These studies indicate that individuals faced with disease or ill health will search information on the Internet regarding their state of health and possible countermeasures to illness; logs of queries submitted to search engines by individuals seeking this information are potential sources of information for detecting emerging epidemics, as it is possible to track changes in the volumes of specific search queries. However, the recent errors arising from Google Flu Trends (GFT), which has been predominantly used in previous studies, serves as a reminder to investigators that this novel data paradigm calls for critical assessment and the development of more empirical methodologies to explore the predictive utility of big data [7, 8]. It is clear that upcoming studies need to focus on methods for more precisely identifying the particular phases associated with

influenza epidemics based on data from these highly informative sources.

Selecting the queries that are most likely to be associated with influenza epidemics poses a particular challenge for the generation of improved predictions. In previous studies, researchers have utilized queries selected by various methods, such as specific keyword tools offered by particular web sites [5], surveys of patients who visited the emergency room [1, 9], or common knowledge about influenza including the definition of ILI [9, 10] as well as fully automated methods for identifying queries related to influenza from search logs [3, 4, 6]. Since researchers do not have full access to search logs, an approach using social media data may also be helpful for obtaining information for query selection. Recently, social media data have been highlighted as an additional potential data source for disease surveillance because they contain a greater variety of contextual health information with diverse descriptions of health states. Thus, it could be a useful reference point for researchers who wish to select initial target queries in query-based prediction.

In South Korea, there is no currently reliable forecasting system for infectious disease based on search query data [1, 9], despite the high availability and use of the Internet in Korea [11]. Moreover, few studies have thus far evaluated whether such data could be of value in national influenza forecasting [1, 9], and a recent study has suggested that Google Trends in the Korean language is insufficient for use as a model for influenza prediction in South Korea [1]. I need to proactively determine whether queries of search engines that are more widely used by Koreans have the capacity to enhance traditional influenza surveillance systems in South Korea. I consider the use

of social media data to select queries that are most likely to be associated with influenza epidemics in a situation involving limited access to search logs.

The purpose of this study was to further explore two concerns: (a) First, I describe a methodological extension for detecting influenza outbreaks using search query data, providing a new approach for query selection through the exploration of contextual information obtained from social media data. (b) Second, I evaluate whether it is possible to use these queries for monitoring and predicting influenza epidemics in South Korea.

3.2 Methods

3.2.1 Data Sources

- **Epidemiological Surveillance Data**

National influenza surveillance data were obtained from the Korea Center for Disease Control and Prevention (KCDC), which routinely collects epidemiological data and national statistics pertaining to influenza incidence, typically with a 1-week reporting lag (http://www.is.cdc.go.kr/nstat/index.jsp). I used clinical data and virological data between April 3, 2011, listed as week 32, and April 5, 2014, listed as week 14. As clinical data, we used the rates of physician visits for ILI and, as virological data, the rates for positive results for the influenza virus in

laboratory tests. The data obtained were anonymous and publicly available.

- **Social Media Data**

In developing an approach for query selection, I drew on social media data. Social media data were collected from the daily Naver blog and Twitter posts between September 1, 2010, and August 31, 2013 (3 years), using the social "big data" mining system, SOCIALmetrics™ Academy (Daumsoft; http://www.daumsoft.com/eng/). The SOCIALmetrics™ system contains social media data crawlers that collect posts from Twitter and the NAVER blog. The system also processes text using state-of-the-art natural language processing and text mining technologies. The Twitter crawler utilizes a streaming application programmer's interface (API) (http://dev.twitter.com/docs/streaming-apis) for data collection using the so-called "track keywords" function. It tracked several thousand keywords that were empirically selected and tuned to maximize the coverage of the crawler operating in near real-time fashion. The system was estimated that the daily coverage of the Twitter crawler was over 80%. The collected posts were fed into a spam-filtering module that checks for posts containing spam keywords written by known spammers. The lists of spam keywords and spammers were semi-automatically monitored and managed. The NAVER blog is a weblog service offered by the biggest portal site in South Korea (http://www.section.blog.naver.com). The NAVER blog crawler resembles general-purpose web crawlers, the main difference being that a list of active bloggers for post collection is maintained and automatically expanded. The

estimated coverage of the Naver blog crawler was also over 80%. The system was applied an extensive spam-filtering process similar to that of the Twitter crawler on the collected blog posts.

I and data mining company conducted the search according to the Twitter and blogging web site Terms and Conditions of use. All Twitter and NAVER blog posts were publicly available and the information collected did not reveal the identity of the social media users; thus, user confidentiality was preserved.

- **Search Engine Query Data**

The query data originate from the search engine on the Korean web site Daum (http://www.daum.net/). Google is the most used search engine in the world, but it is not dominant in South Korea. Local search engines based on the Korean language such as Daum are more widely used than Google. Daum is the second largest search engine in the portal sites market of South Korea [12]. Since the query data of Korean web sites are not publicly available, I sent the list of target queries to Daum and received scaled volume data pertaining to the queries listed. Weekly relative volumes of queries submitted to the search engine between April 3, 2011, and April 5, 2014, were used for analysis. The relative volumes were calculated by dividing the number of each query by the total number of search queries in any given week. The web site Daum is written in Korean, thus the submitted queries are primarily in Korean. No information was available that could have potentially revealed the identity of a web site visitor, such that complete

confidentiality was maintained.

3.2.2 Query Selection

To obtain queries related to influenza that were submitted to the Daum search engine by the Korean population at large, several approaches were applied. Search queries were obtained using the following methods:

- **Seed keyword for exploring the queries**

Although *influenza* is the official term used by the KCDC, *dokgam*, *inpeulruenja, peulru* and *sinjongpeulru* are the words typically used in Korea to describe influenza. Since the 2009 pandemic of influenza virus A (H1N1), the term *sinjongpeulru* to describe the new strain of flu has been more popular in Korea than the term *influenza A (H1N1)*. Thus, *dokgam, inpeulruenja, peulru, sinjongpeulru, influenza,* and *flu* were defined as seed keywords for exploring the queries. Since web search queries typically consist of word combinations of an average of two or three terms [13, 14], these seed keywords were also used as essential keywords in word combinations.

- **Exploring the influenza-related words through social media data**

To obtain search queries related to influenza, I considered the words that usually appear with the word *influenza* in the accumulated

posts submitted to *Twitter* and blogs. I first conducted synonym processing for the seed keywords of *dokgam*, *inpeulruenja*, *peulru*, *sinjongpeulru*, *influenza*, and *flu*, and named the resulting application FLU. Then, I investigated the words most likely to be associated with FLU using the accumulated posts during the critical 3-year period (between September 1, 2010 and August 31, 2013). Association analysis was performed to identify tuples of topic keyword and associated keywords. This analysis resulted in a total of 157 associated words.

Certain words associated with influenza were not related to influenza seasons or were not commonly entered into search engines. Therefore, I excluded words considered as inadequate candidates for search query following the keyword filtering; in our first phase, I generated 103 candidate queries of single words or word combinations consisting of seed keywords and/or words associated with influenza as determined using social media data.

- **Identifying the chief complaints related to influenza**

Some additional queries related to influenza were obtained through a review of influenza symptoms referring to patients' chief complaints. The influenza surveillance system of the KCDC defines ILI as the sudden onset of high fever (38℃ or greater) accompanied by a cough and/or sore throat. These symptoms, based on the definition of ILI, were included. Additionally, I included queries related to influenza symptoms suggested by the CDC [15] and a consultative committee of medical

doctors: this second phase generated 29 candidate queries of single words or word combinations consisting of seed keywords and associated words in reference to chief complaints relating to influenza.

- **Using web query recommendations**

  Internet search users often require multiple iterations of query refinement to find the desired results from a search engine [13]. Users of search engines can improve their web search through the help of query recommendations that suggest lists of related queries, allowing users to improve the usability of web search engines and to access queries that better represent their search intent [14]. I considered queries suggested by keyword recommendations from the Korean web site Daum and Naver. In this third phase, entering FLU into the search engines allowed us to identify 75 related queries in the form of single words or word combinations.

3.2.3 Feature Selection and Prediction Model

I divided the data into training and validation sets. Data from April 3, 2011, to June 29, 2013, were used as the training set for modeling, and data from June 30, 2013, to April 5, 2014, were used as the validation set for the

model test. Volumes of 6 seed queries and 146 related queries, obtained after duplicate queries were eliminated from the set of 216 candidate queries, were used for analysis. To identify optimal predictors, I applied a least absolute shrinkage and selection operator (Lasso) algorithm following data normalization. Feature selection can be used to avoid overfitting of irrelevant features and to improve predictive performance (i.e., resulting in more rapid and cost-effective predictions) [16, 17]. The lasso algorithm benefits from a tendency to assign zero weights to irrelevant or redundant features [18]. Because I aimed to identify predictors of influenza epidemics, feature selection processing was performed at three time points (defined as lag -2, -1, and 0) on the training set portion of the influenza surveillance data using 10-fold cross validation. I considered all optimal features selected in each lag for model building.

Support Vector Machine for regression (SVR) was conducted to construct a model predicting influenza epidemics with selected features. Grid search and 10-fold cross-validation were performed to select the optimal SVR parameter settings, including the penalty parameter C and the kernel function parameter such as the gamma for the radial basis function kernel. Ranges of values for grid search can be summarized as follows: penalty parameter C [0.01, 10, 0.01]; gamma [0.0001, 1, 0.0001], where elements in each list denote the beginning, end, and number of samples to generate, respectively. I assessed the root mean square error (RMSE), particular log errors, and the correlation between predicted values and influenza surveillance data using the validation set. All statistical analyses were

performed using the R software package (ver. 3.0.3; R Development Core Team, Auckland, New Zealand).

Ethics Statement

This study was exempted from ethical review by the Institutional Review Board of Seoul National University.

3.3 Results

In total, 146 queries related to influenza were generated through my initial queries selection approach (see Figure S3.1). Feature selection was performed based on 152 queries including 6 seed keywords, and optimal features for prediction of influenza incidence was chosen by 10 times experiment of Lasso. Table 3.1 presents the results of feature selection based on ILI surveillance data. Of the 152 queries, 15, 14, and 29 principle features (the total number of the features without duplication = 36) had the minimum lambda value in lag -2, -1, and 0, respectively. The optimal features for prediction of ILI incidence were derived from queries in reference to the social media data (a, 29/36), query recommendation (b, 24/36), chief complaint of influenza (c, 4/36) and seed keyword (s, 1/36) (Table 3.1).

Table 3.1 Optimal feature for ILI surveillance

Query	In English	Query Ref.*	10-fold CV Coefficient		
			lag -2	lag -1	lag 0
(Intercept)			0.3320	0.3208	0.4971
a 형 influenza	a type influenza	ab	0.7447	0	0.1086
a 형 독감	a type flu	ab	4.9278	20.1538	21.5026
a 형 인플루엔자	a type influenza	ab	0.0645	0.7606	1.1272
b 형 influenza	b type influenza	ab	0	0	0.3448
b 형 독감	b type flu	ab	0	0.0291	1.4471
influenza a		ab	2.3447	0.0857	0
influenza a 형	influenza a type	ab	1.8935	0.9271	0.0287
vaccine		a	0	0	-0.1151
건강	health	a	0.3933	0.3951	0.1089
독감 감염	flu infection	a	0.0518	0	0
독감 검사	flu check	ab	4.3029	8.8931	4.4023
독감 격리기간	flu isolation period	b	0	0	0.1774
독감 기침	flu cough	ac	0	0	1.1055
독감 바이러스	flu virus	ab	0	0	-0.2200
독감 열	flu fever	c	0.3906	0	0
독감 예방	flu prevention	ab	0	0	-0.1519
독감 예방접종	flu vaccination	ab	0	0	-0.1174
독감 입원	flu hospitalization	ab	0	0	1.4696
독감 전염	flu infection	ab	0	0	2.5687
독감 전파	flu dissemination	ab	0.5472	0.3224	0.0173
독감 폐렴	flu pneumonia	ac	0	0	0.0054
독감 학교	flu school	a	0	0.1223	0
독감 환자	flu patient	a	0.0661	0	0
소아 독감증상	child flu symptoms	b	0.8107	0.3225	0.1351
신종플루 증상	new flu symptoms	ab	55.9795	46.1557	58.4149
심한감기	severe cold	a	0	0	0.0307
어린이 독감유행	child flu epidemic	b	0	0	0.0019

온몸이 아픔	whole body pain	c	0	0.0384	0.0720
인플루엔자 검사	influenza check	ab	0	0.2329	0
인플루엔자 약	influenza medicine	ab	0	0	-0.0047
인플루엔자 유행	influenza epidemic	a	0	0	0.0032
인플루엔자 증상	influenza symptoms	ab	6.2541	0	0
인플루엔자 증세	influenza symptoms	ab	0	0	0.2090
중국독감	china influenza	b	0	0	-0.0556
타미플루	tamiflu	ab	0	0	0.5174
플루	flu	s	0.6209	0.5623	0.3390

* a, social media; b, query recommendation; c, chief complaint of influenza; s, seed keyword
CV, cross validation

I evaluated the performance of prediction model created in training set for ILI surveillance with the validation set. As a result, the SVR model (*C*=1.32; gamma= 0.0002) performed well, the prediction values were highly correlated with the recent observed ILI incidence rate (r=0.956; $p<0.001$) (Figure 3.1).

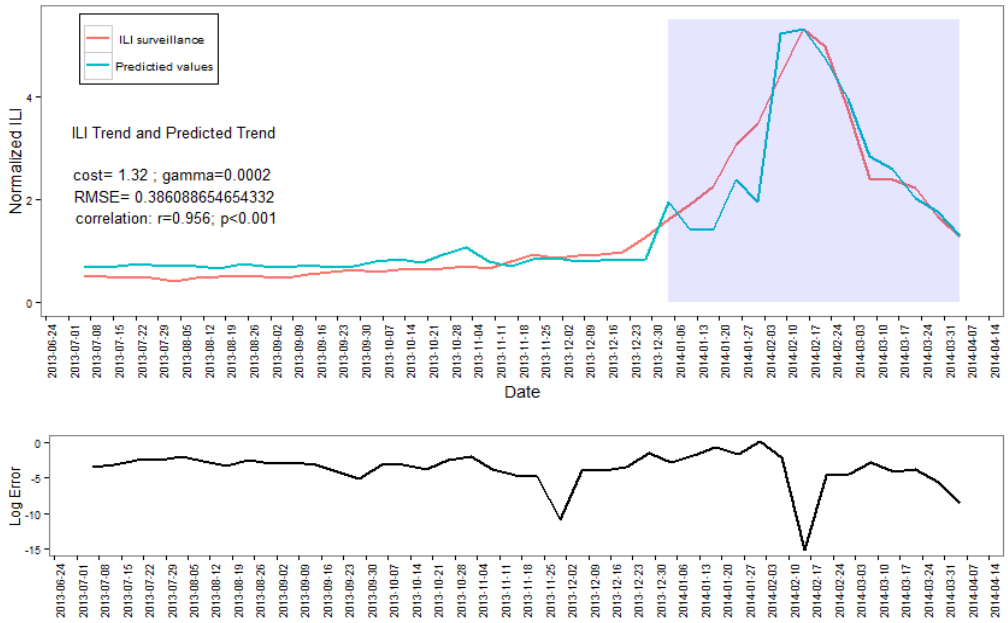


Figure 3.1 SVR Prediction and error for ILI surveillance in Korea. This figure shows the performance of the SVR model using the validation set of KCDC surveillance data to predict the next observation. The SVR model (C=1.32; gamma=0.0002) performed well, the prediction values were highly correlated with the observed ILI (r=0.956; p<0.001).

ILI, Influenza-like illness

Note: Log Error = $\log((\text{Obs}-\text{Exp})^2/\text{abs}(\text{Exp}))$

I adopted the same principle with regard to prediction of virological surveillance as with ILI. Table 3.2 presents the results of feature selection based on virological surveillance data. Of the 152 queries, 28, 26, and 45 principle features (the total number of the features without duplication = 53) had the minimum lambda value in lag -2, -1, and 0, respectively. The optimal features for prediction of virological incidence were also derived from queries in reference to the social media data (a, 42/53), query recommendation (b, 31/53), chief complaint of influenza (c, 7/53) and seed keyword (s, 1/53) (Table 3.2).

Table 3.2 Optimal feature for virological surveillance

Query	In English	Query Ref.*	10-fold CV coefficient		
			lag -2	lag -1	lag 0
(Intercept)			-1.4587	-3.1243	-2.1469
a 형 influenza	a type influenza	ab	26.4130	18.8985	22.5794
a 형 독감	a type flu	ab	0	0	379.0405
b 형 독감	b type flu	ab	6.0071	15.3239	24.0387
b 형 독감증상	b type flu symptoms	ab	0	0	0.2293
influenza a		ab	37.9534	25.0212	17.4485
influenza a 형	influenza a type	ab	24.1144	19.3422	11.4256
감기 바이러스	cold virus	a	0	0	4.8980
감기 빨리 낫는 법	how to cure a flu quickly	b	5.3654	4.2615	2.3426
감기 예방	cold prevention	ab	0	0	-0.4497
감기 예방법	how to prevent a cold	a	-0.1549	-2.7363	-4.1397
건강	health	a	4.0911	3.5621	3.3903
근육통	muscle pain	ac	0	0	-0.2651
날씨	weather	a	0	0	-0.1106
독감 a 형	flu a type	ab	0	0	22.7720
독감 감염	flu infection	a	12.2360	1.4485	0
독감 검사	flu check	ab	38.2543	31.8781	0
독감 격리기간	flu isolation period	b	0	0	12.1449
독감 고열	flu high fever	ac	0	0	1.7450
독감 기침	flu cough	ac	0	0	25.9113
독감 노인	flu in the elderly	a	0	0	-3.7387
독감 바이러스	flu virus	ab	0	0	-0.7774
독감 아이	flu child	a	0	0	2.6944
독감 어린이	flu child	a	0	0	-0.4767
독감 예방	flu prevention	ab	-2.4665	-9.7599	-12.1914
독감예방법	how to prevent a flu	b	0	0	-0.6382
독감 유행	flu epidemic	ab	0	0	-0.1094
독감 입원	flu hospitalization	ab	8.1559	0	13.7929
독감 전염	flu infection	ab	38.1841	81.8303	9.7623

독감 전파	flu dissemination	ab	2.5960	5.6131	3.9730
독감 주사	flu injection	ab	-3.9069	0	0
독감 주의보	flu watch	b	0.8833	0.3100	0
독감 학교	flu school	a	9.2678	0	0
독감 합병증	flu complication	a	0	0	3.5133
독감 환자	flu patient	a	7.0242	5.0266	3.2047
돼지 독감	swine flu	b	0.3584	0	0
마스크	mask	a	8.0527	0	0
몸살	body aches	ac	0	1.3872	3.9124
소아 독감증상	child flu symptoms	b	4.7366	8.0579	9.0412
아동 독감 유행	child flu epidemic	ab	0	0	-5.2728
어른 독감 증상	adult flu symptoms	b	5.1556	1.4849	0.6098
얼굴통증	face pain	c	-1.0571	0	0
온몸이 아픔	whole body pain	c	2.9618	3.7248	4.7913
의사	doctor	a	-3.1527	-0.4355	-0.7115
인플루엔자 a 형	influenza a type	ab	0	8.3492	5.8367
인플루엔자 사망자	influenza the death	ab	0	-0.3633	-5.1928
인플루엔자 약	influenza medicine	ab	0	0	-0.5600
인플루엔자 증상	influenza symptoms	ab	3.0387	2.0506	5.3029
입원	hospitalization	a	0	0	-0.2128
조류독감	avian flu	b	3.9715	4.2387	3.4924
타미플루	tamiflu	ab	0	65.6177	75.4623
폐렴	pneumonia	abc	0	0	-1.2882
플루	flu	s	15.9915	13.4057	5.9239
환자	patient	a	-4.5427	-3.1704	-2.9216

* a, social media; b, query recommendation; c, chief complaint of influenza; s, seed keyword
CV, cross validation

Figure 3.2 shows the result of the performance of prediction model for virological surveillance. The SVR model (*C*=2.14; gamma= 0.0006) performed well, the prediction values were highly correlated with the recent observed virological incidence rate (r=0.963; p<0.001) (Figure 3.2).

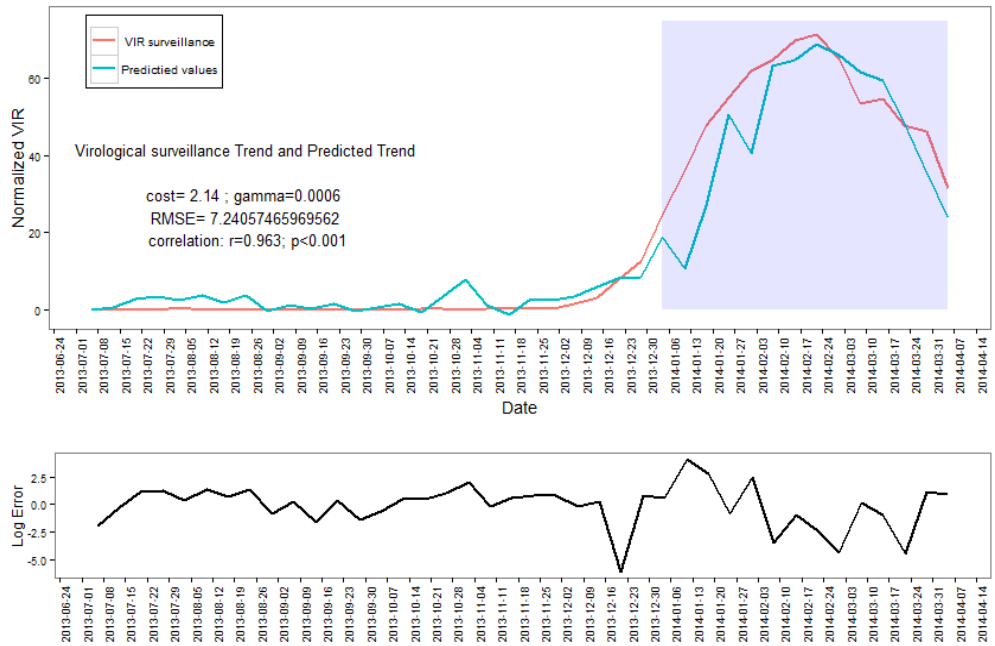


Figure 3.2 SVR Prediction and error for Virological surveillance in Korea. This figure shows the performance of the SVR model using the validation set of KCDC surveillance data to predict the next observation. The SVR model (C=2.14; gamma=0.0006) performed well, the prediction values were highly correlated with the observed ILI (r=0.963; p<0.001).

VIR, Virological positive rate

Note: Log Error = $\log((Obs-Exp)^2/abs(Exp))$

3.4 Discussion

The present study investigated whether search queries have the capacity to enhance the traditional influenza surveillance system in South Korea. To select queries most likely to be associated with influenza epidemics, I adopted an approach that explored contextual information available in social media data. A considerable proportion of optimal features for my final models were derived from queries with reference to the social media data. my best model for South Korean ILI data included 36 queries and was highly correlated with observed ILI incidence rates. My model for virological data, which included 53 queries generated through the same principles as the ILI model, performed equally well in terms of its correlation with observed virological incidence rates. Hence, my models for detecting national influenza incidence have the power to predict. These results demonstrate the feasibility of search queries in enhancing influenza surveillance in South Korea.

Although created to predict the incidence of influenza throughout the year, including during high- and low-incidence seasons, my model performed as well as previous models that benefitted from full access to search logs to predict influenza incidence using search queries [3, 4, 6]. Researchers who do not have full access to search logs need to choose the most pertinent queries, but these may be difficult to determine [1]. My current approach for query selection using social media data appears to be ideal for supporting influenza surveillance based on search query data. In generating a prediction model using search query data, it is important to note

that search queries change over time. An individual's search behavior changes constantly and keywords submitted by individuals may be influenced by numerous factors such as media-driven interest or various events [5, 19, 20]. These changes alter or degrade the performance of search query-based surveillance. The recent GFT overestimation can also be understood in the same context [7, 8]. Constructing a model that is flexible over time is probably the most difficult, but also the most important, task to complete in the future creation of robust surveillance systems. The systematic exploration of changing predictors in social media data may help to update models based on search queries within a statistical learning framework.

Internet usage is strongly associated with behaviors related to health information seeking and sharing. Some users write expositions about their health through various social media channels such as blogs and Twitter, while some users leave query logs of health-related questions on the Internet search engines of web sites. These types of activities may provide complementary information; it is likely that social media data contain diverse descriptions of personal experiences and information, whereas search engine query data specifically relate to queries, which are submitted for the sole purpose of obtaining information. Starting with studies that have exploited search trends [3, 4, 6], the notion of detecting influenza activity using internet-based data has been extended to experimentation with social media data [21]. Thus far, several studies have tried to separately evaluate the scientific potential of each type of novel data for detecting emerging influenza incidence. Although previous empirical studies have reported some

significant results, this domain of inquiry is still very much in its infancy [5, 19, 20] and several limitations pertaining to data sources can be identified [7, 8]. Beyond simply conducting experiments to replicate the findings of previous studies using each type of novel data, perhaps it is time to consider a new strategy, one that adopts mutually reinforcing measures of the valuable information contained in each type of data.

I have used query data obtained from Daum, a Korean local web site. The market share of Daum is only 17.4% despite being the second largest search engine in South Korea; nevertheless, my prediction exhibited strong congruence with national ILI incidence rates. Previous research using query data from Daum has found that some cumulative queries selected by means of survey were also strongly correlated with national influenza surveillance data in South Korea between September 6, 2009, and September 1, 2012 [9]. The findings jointly suggest the possibility of developing an influenza surveillance system using a non-dominant search engine.

However, changes in internet usage rates and health information seeking rates may constitute a somewhat central limitation on the use of search query data. Noise from irrelevant information and uncertainty regarding the representativeness of the sample of health information seekers are also significant limitations. These limitations exist in the data used in my study; thus, optimal features of my model may need to be updated over time.

3.5 Conclusion

Despite several limitations, the current study provides further evidence, based on a new approach, for linkages between the use of Internet-based data and the surveillance of emerging influenza incidence in South Korea. I found that internet-based influenza surveillance that combines search engine query data with social media data has the power to predict influenza outbreaks, exhibiting strong congruence with traditional surveillance data. Such an approach may provide valuable support in preparing for severe pandemics, like the 2009 influenza A (H1N1) pandemic, and in controlling seasonal influenza epidemics. Furthermore, in an attempt to exploit the complementary nature of two types of data sources, in this study I fused information drawn from social media with a methodology for query-based influenza surveillance. My results imply that these new data sources may be compatible and complementary in predicting influenza incidence.

References

1. Cho, S., et al., *Correlation between national influenza surveillance data and google trends in South Korea.* PLoS One, 2013. **8**(12): p. e81422.
2. Cook, S., et al., *Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic.* PloS one, 2011. **6**(8): p. e23610.
3. Ginsberg, J., et al., *Detecting influenza epidemics using search engine query data.* Nature, 2009. **457**(7232): p. 1012-4.
4. Polgreen, P.M., et al., *Using internet searches for influenza surveillance.* Clin Infect Dis, 2008. **47**(11): p. 1443-8.
5. Yuan, Q., et al., *Monitoring influenza epidemics in china with search query from baidu.* PloS one, 2013. **8**(5): p. e64323.
6. Hulth, A., G. Rydevik, and A. Linde, *Web queries as a source for syndromic surveillance.* PloS one, 2009. **4**(2): p. e4378.
7. Lazer, D., et al., *Big data. The parable of Google Flu: traps in big data analysis.* Science, 2014. **343**(6176): p. 1203-5.
8. Lazer, D., et al., *Twitter: big data opportunities--response.* Science, 2014. **345**(6193): p. 148-9.
9. Seo, D.W., et al., *Cumulative query method for influenza surveillance using search engine data.* J Med Internet Res, 2014. **16**(12): p. e289.
10. Kang, M., et al., *Using Google trends for influenza surveillance in South China.* PloS one, 2013. **8**(1): p. e55205.
11. ITU, *The World in 2014: ICT Facts and Figures, 2014*, in *ICT Facts and Figures, 2014*. 2014, International Telecommunication Union: Geneva, Switwerland.
12. WebCite. *Market share of search engine in South Korea*. . 2015 [cited 2015 April 05]; Available from: http://www.webcitation.org/6QrnPuRwJ.
13. He, Q., et al. *Web query recommendation via sequential query prediction*. in *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference*

on. 2009. IEEE.

14. Baeza-Yates, R., C. Hurtado, and M. Mendoza. *Query recommendation using query logs in search engines*. in *Current Trends in Database Technology-EDBT 2004 Workshops*. 2005. Springer.
15. CDC. *Seasnoal Influenza (Flu):Flu Basics-Influenza Symptoms*. 2015 March 1, 2015]; Available from: http://www.cdc.gov/flu/about/disease/symptoms.htm.
16. Guyon, I. and A. Elisseeff, *An introduction to variable and feature selection.* The Journal of Machine Learning Research, 2003. **3**: p. 1157-1182.
17. Saeys, Y., I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics.* bioinformatics, 2007. **23**(19): p. 2507-2517.
18. Li, F., Y. Yang, and E.P. Xing. *From lasso regression to feature vector machine*. in *Advances in Neural Information Processing Systems*. 2005.
19. Milinovich, G.J., et al., *Internet-based surveillance systems for monitoring emerging infectious diseases.* Lancet Infect Dis, 2014. **14**(2): p. 160-8.
20. Althouse, B.M., Y.Y. Ng, and D.A. Cummings, *Prediction of dengue incidence using search query surveillance.* PLoS neglected tropical diseases, 2011. **5**(8): p. e1258.
21. Bernardo, T.M., et al., *Scoping review on search queries and social media for disease surveillance: a chronology of innovation.* J Med Internet Res, 2013. **15**(7): p. e147.

Supporting Information of the Chapter 3

Table S3.1 Queries related to influenza generated by initial queries selection approach

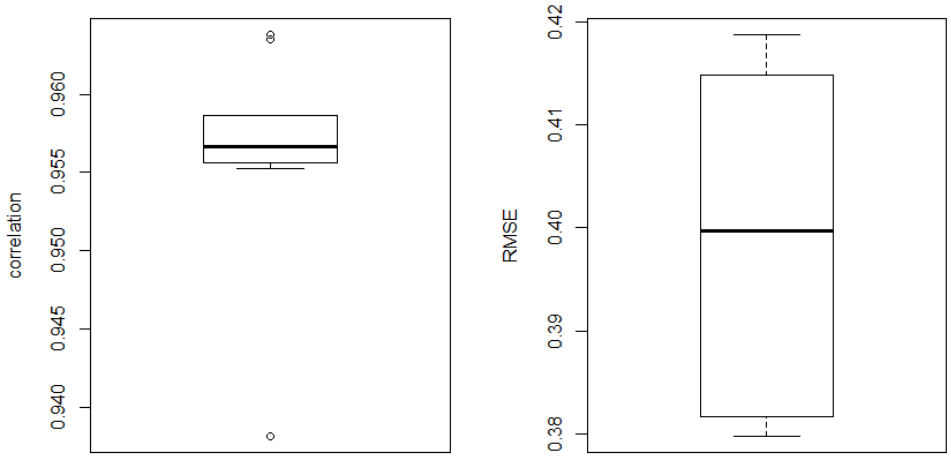
Ref. Chief complaint of influenza		Ref. Social media data					
Query	**In English**	**Query**	**In English**	**Query**	**In English**	**Query**	**In English**
dokgam yeol	flu fever	*goyeol*	high fever	*gamgi baireoseu*	cold virus	*dokgam josim*	flu be careful
guto	vomiting	*geunyuktong*	muscle pain	*gamgi yebangbeop*	how to prevent a cold	*dokgam joeun*	flu good
dokgam guto	flu vomiting	*gichim*	cough	*geongang*	health	*dokgam jinryo*	flu medical treatment
dokgam mokapeuma	flu throat pain	*dokgam goyeol*	flu high fever	*nalssi*	weather	*dokgam chiryobeop*	flu treatment
moksseurarim	sore throat	*dokgam geunyuktong*	flu muscle pain	*dokgam gamgi*	flu cold	*dokgam hakgyo*	flu school
mokibueum	swollen throat	*dokgam gichim*	flu cough	*dokgam gamyeom*	flu infection	*dokgam hapbyeongjeung*	flu complication
moktongjeunga	throat pain	*dokgam dutong*	flu headache	*dokgam geongang*	flu health	*dokgam hoheup*	flu breath
momsalgiun	body aches symptoms	*dokgam balyeol*	flu fever	*dokgam gyeoul*	flu winter	*dokgam hwanja*	flu patient
eolgultongjeung	face pain	*dokgam komul*	flu runny nose	*dokgam nalssi*	flu weather	*dokgamuisa*	flu doctor
yeol	fever	*dokgam pyeryeom*	flu pneumonia	*dokgam noin*	flu in the elderly	*maseukeu*	mask
onmomi apeum	whole body pain	*dutong*	head ache	*dokgam maseukeu*	flu mask	*myeonyeokryeok*	immunity
		momsal	body aches	*dokgam myeonyeokryeok*	flu immunity	*baireoseu*	virus
		balyeol	fever	*dokgam mok*	flu throat	*baeksin*	vaccine
		inhutonga	Sore throat	*dokgam mom*	flu body	*simhan gamgi*	severe cold
		komul	runny nose	*dokgam mom sangtae*	flu body state	vaccine	
		gigwanjiyeom	bronchitis	*dokgam byeongwon*	flu hospital	*inpeulruenja yuhaeng*	influenza epidemic
		*pyeryeom**	pneumonia	*dokgam samang*	flu death	*inpeulruenja samangja*	people who died of influenza
		dokgam momsal	flu body aches	*dokgam samangja*	people who died of flu	*imsanbu dokgam*	pregnant women flu
				dokgam sangtae	flu state	*ipwon*	hospitalization
				dokgam i	flu child	*uisa*	doctor
				dokgam yak	flu medicine	*hwanja*	patient
				dokgam eorini	flu child		
				dokgam imsanbu	flu pregnant women		

*** Seed keywords :** *dokgam*(flu), *inpeulruenja* (influenza), *peulru*(flu), *sinjongpeulru*(new flu), influenza, flu

Table S3.1 Queries related to influenza generated by initial queries selection approach (continued)

Ref. Social media data					
Ref. Query recommendation					
Query	**In English**	**Query**	**In English**	**Query**	**In English**
a *hyeong* influenza	a type influenza	*dokgam jeonyeom*	flu infection	*gamgi pparri natneunbeop*	how to cure a cold quickly
a *hyeong dokgam*	a type flu	*dokgam jeonpa*	flu dissemination	*dokgam gamgi chai*	differences flu cold
a *hyeong dokgam jeungsang*	a type flu symptoms	*dokgam jusa*	flu injection	*dokgam gyeokrigigan*	flu isolation period
a *hyeong inpeulruenja*	a type influenza	*dokgam jeungsang*	flu symptoms	*dokgam pparri natneunbeop*	how to cure a flu quickly
b *hyeong* influenza	b type influenza	*dokgam chiryo*	flu treatment	*dokgam yuhaengjuuibo*	flu watch
b *hyeong dokgam*	b type flu	*dokgam e joeun eumsik*	good food for flu	*dokgam yebanghaneun bangbeop*	how prevent a flu
b *hyeong dokgam jeungsang*	b type flu symptoms	*seongin dokgam jeungsang*	adult flu symptoms	*dokgam jambokgi*	flu incubation period
b *hyeong inpeulruenja*	b type influenza	*sinjong peulru*	new flu	*dokgam juuibo*	flu watch
influenza A	influenza A	*sinjongpeulru jeungsang*	new flu symptoms	*dokgam jeungse*	flu symptom
influenza a *hyeong*	influenza a type	*agi dokgam jeungsang*	baby flu symptoms	*dwaejidokgam*	swine flu
influenza B	influenza B	*adong dokgam jeungsang*	child flu symptoms	*soa dokgamjeungsang*	child flu symptoms
influenza b *hyeong*	influenza b type	*i dokgam jeungsang*	child flu symptoms	*sinjong inpeulruenja*	new influenza
tamiflu		*eorini dokgam jeungsang*	child flu symptoms	*sinjongdokgam*	new flu
gamgi	cold	*inpeulruenja* a *hyeong*	influenza a type	*sinjongpeulru samang*	new flu death
gamgiyebang	cold prevention	*inpeulruenja* b *hyeong*	influenza b type	*sinjongpeulru samangja*	new flu the death
gyeouldokgam	winter flu	*inpeulruenja geomsa*	influenza check	*sinpeul**	new flu
dokgam ahyeong	flu a type	*inpeulruenja samang*	influenza death	*eorini dokgamyuhaeng*	child flu epidemic
dokgam bhyeong	flu b type	*inpeulruenja yak*	influenza medicine	*eoreun dokgamjeungsang*	adult flu symptoms
dokgam geomsa	flu check	*inpeulruenja yebang*	influenza prevention	*yojeum dokgamjeungsang*	these days flu symptom
dokgam baireoseu	flu virus	*inpeulruenja jeungsang*	influenza symptoms	*yua dokgamjeungsang*	toddler flu symptoms
dokgam baeksin	flu vaccine	*inpeulruenja jeungse*	influenza symptoms	*inpeulruenja yuhaengjuuibo*	influenza watch
dokgam sinjongpeulru	flu new flu	*inpeulruenja chiryo*	influenza treatment	*joryudokgam*	avian influenza
dokgam yebang	flu prevention	*joryudokgam jeungsang*	avian flu symptoms	*junggukdokgam*	china influenza
dokgam yebangjeopjong	flu vaccination	*jilbyeonggwanribonbu*	KCDC		
dokgam yuhaeng	flu epidemic	*tamipeulru*	tamiflu		
dokgam ipwon	flu hospitalization	*pyeryeom**	pneumonia		

*** Seed keywords :** *dokgam*(flu), *inpeulruenja* (influenza), *peulru*(flu), *sinjongpeulru*(new flu), influenza, flu



	N	Mean	S.D.	Min	Max
Correlation	1000	0.9579031	0.010608	0.9212078	0.9840354
RMSE	1000	0.3778032	0.0455562	0.2588971	0.4244483

Figure S3.1 1,000 times validation of the final result.

RMSE, Root Mean Squared Error

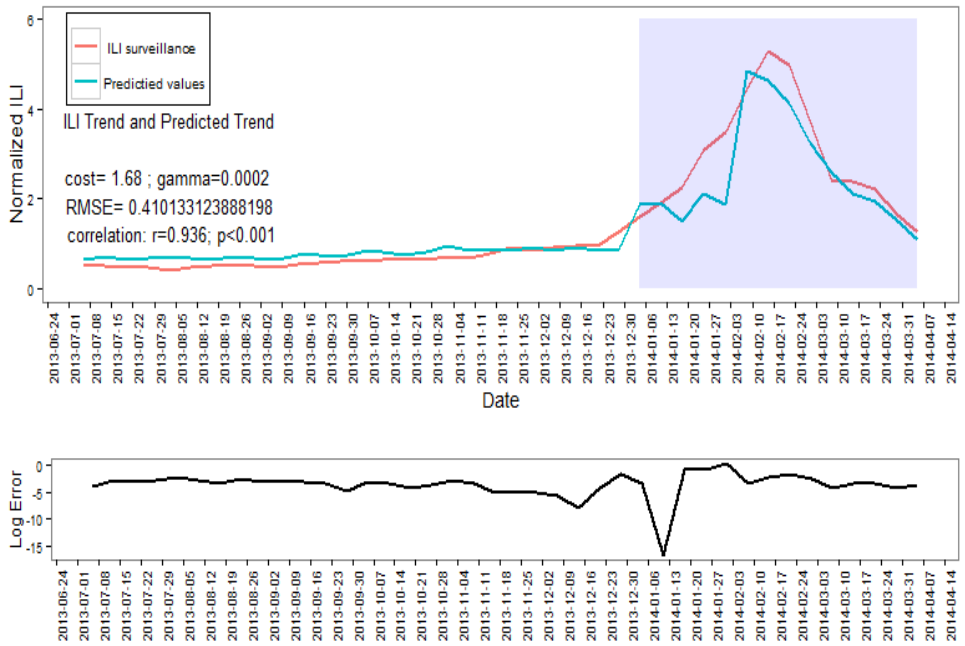


Figure S3.2 SVR Prediction without initial queries selection through social media data. This figure shows the performance of the SVM regression model using the validation set of KCDC surveillance data to predict the next observation.

ILI, Influenza-like illness

Note: Log Error = $\log((\text{Obs}-\text{Exp})^2/\text{abs}(\text{Exp}))$

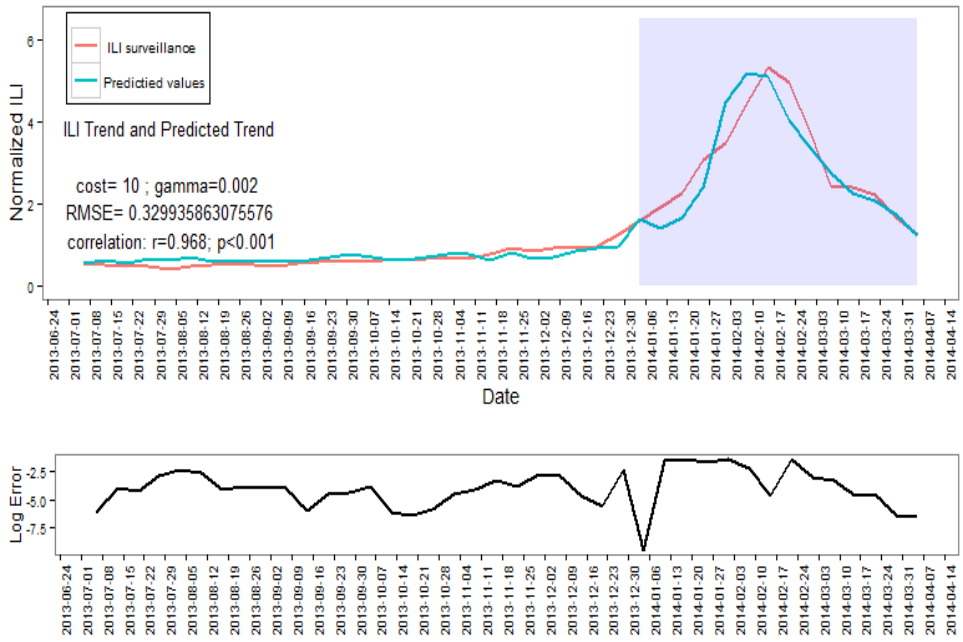


Figure S3.3 SVR Prediction without Feature selection by Lasso. This figure shows the performance of the SVM regression model using the validation set of KCDC surveillance data to predict the next observation.

ILI, Influenza-like illness

Note: Log Error = $\log((\text{Obs}-\text{Exp})^2/\text{abs}(\text{Exp}))$

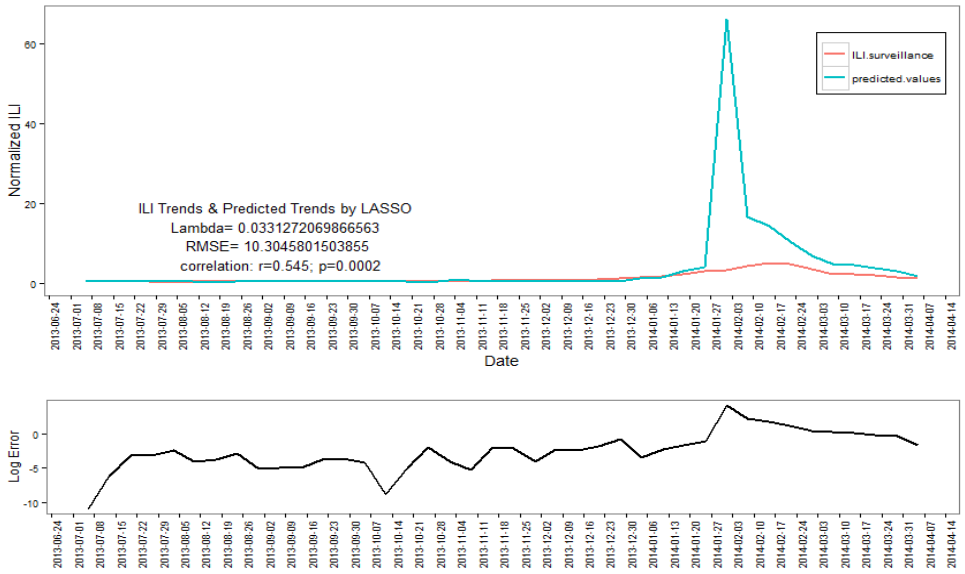


Figure S3.4 Prediction and error based on the lasso model.

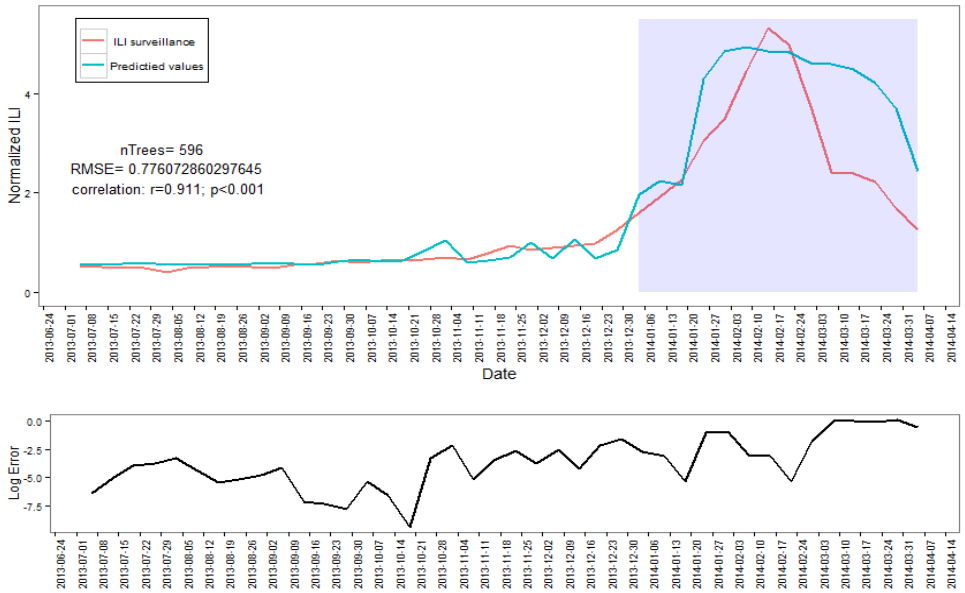


Figure S3.5 Prediction and error based on the random forest model.

Chapter 4 General Discussion

4.1 Discussion

This study investigated whether Internet-based online surveillance can be used to complement and augment the traditional surveillance system in South Korea using Internet-based big data, particularly data from social media and search engine queries. I first identified a set of keywords that served as predictors for detecting influenza epidemics using data from Twitter and blogs. A set of 15 keywords optimally predicted influenza epidemics, evenly distributed across Twitter and blog data sources. The predictions based on these keywords were highly correlated with the incidence of influenza, exhibiting a root-mean-square error (RMSE) of 0.55 and a correlation between predictions and recent influenza incidence data of 0.92 ($p < .001$) (see Figure 2.3 in the Chapter 2). In the chapter 3, I then investigated whether search queries have the capacity to augment the data obtained by the traditional influenza surveillance system in South Korea. To select queries most likely to be associated with influenza epidemics, I explored contextual information available in social media data. A considerable proportion of optimal features for final models were derived from queries generated in reference to the social media data. The best model for South Korean data, which included 36 queries, showed a strong correlation with observed ILI incidence rates ($r = 0.956$; $p < .001$) (see

Figure 3.1 in the Chapter 3). These results demonstrate that the models for detecting national influenza incidence have the power to accurately predict outbreaks, confirming the feasibility of using data from search queries to enhance influenza surveillance in South Korea.

Table 4.1 compares the performance of each model according to the data source used. In an attempt to serve as complementary one to the other, information of social media data has fused into a methodology for query-based influenza surveillance in Model 3. As a result, Model 3 was optimal, exhibiting a minimal RMSE and a strong correlation between predictions and observed cases. Model 1, which was based solely on social media data and Model 2, which was based solely on internet search queries, did not perform as well as Model 3 (see Table 4.1). These results indicate that an approach for query selection using social media data may be ideal for supporting influenza surveillance based on search query data.

To identify a subset of the best predictor features, I used the Lasso algorithm to select the best dataset and features, subsequent to data normalization. The primary objectives of feature selection are to avoid over-fitting that may be caused by irrelevant features, to improve the performance of the predictors, and to identify faster and more cost-effective predictors. LASSO is useful for efficient and simple feature selection because it tends to assign zero weights to most irrelevant or redundant features. In addition, I compared the performance of Model 3 to that of Model 4, which did not involve initial feature selection by LASSO. The results are presented in Table 4.2, and indicate that Model 4 performed slightly better than Model 3. Constructing a model that is flexible over time is probably the most difficult

but important task to solve in the design of a robust surveillance system for the future. The central assumption when using a feature-selection technique is that the data contain many redundant or irrelevant features. Redundant features are those that provide no more information than the currently selected features, and irrelevant features provide no useful information at all. Thus, Model 3, which applied initial feature-selection techniques, is likely to provide significant benefits over Model 4 in the construction of a robust predictive model (Table 4.2).

I used several machine learning techniques to construct models that predicted influenza epidemics using the best available features. I built the candidate models using LASSO, Support Vector Machine (SVM), and Random Forest Regression (RFR), using the training set based on the selected features. I compared the performance of candidate models using the validation set of KCDC surveillance data to predict the next observation of an influenza epidemic. Table 4.3 presents the results. The SVM model (cost = 1.32; gamma = 0.0002) performed well, exhibiting a minimal RMSE and a correlation between predictions and observed cases of 0.956. The RFR and LASSO models did not perform better than the SVM with respect to the regression model (Table 4.3).

I conducted similar analyses with respect to predictions based on virological surveillance as those I performed with ILI. Table 4 presents the results of feature selection based on virological surveillance data. The SVR model (C = 2.14; gamma = 0.0006) performed well; the prediction values were strongly correlated with recently observed virological incidence rates (r = 0.963; $p < 0.001$) (Table 4.4). I conclude that the basic principles

underpinning the approach used in this research may be applied to other diseases or data sources.

Table 4.1 Final SVR model comparison according to data source

	Model 1	**Model 2**	**Model 3***
Main Data Source	Social Media Data (only)	Internet Search Queries (only)	Social Media data & Internet Search Queries
Data Source for Initial Keyword Selection	1. Exploring the influenza-related words through social media data	1. Identifying chief complaint of influenza 2. Using web query recommendation	1. Exploring the influenza-related words through social media data 2. Identifying chief complaint of influenza 3. Using Web query recommendation
Data source and feature selection for prediction model	Twitter/ Naver blog and a combination dataset of both	Queries originate from the search engine of Daum	Queries originate from the search engine of Daum
Model Performance (prediction and error)	1. Correlation between predictions and recent influenza incidence data of 0.92 (p<.001) 2. RMSE = 0.554	1. Correlation between predictions and recent influenza incidence data of 0.936 (p<.001) 2. RMSE = 0.410	1. Correlation between predictions and recent influenza incidence data of 0.956 (p<.001) 2. RMSE = 0.386
Reference	Figure 2. 3 (Chapter2)	Figure S3.2 (Chapter3; supporting information)	Figure 3.1 (Chapter3)

* Final Optimal Model

Table 4.2 SVR model comparison according to feature selection method

	Model 3*	**Model 4**
Model Building Process	Initial query selection → feature selection by LASSO → SVR model building based on best feature index	Initial query selection → SVR model building based on initial query index (without feature selection by LASSO)
Input Features for Model building	Total number of features without duplication = 36 (of the 152 queries, 15, 14, and 29 principle features had the minimum lambda value in lags -2, -1, and 0, respectively)	Total number of features = 155 (full index of initial query)
Model Performance (prediction and error)	Correlation between predictions and recent influenza incidence data of 0.956 (p<0.001) RMSE = 0.386	Correlation between predictions and recent influenza incidence data of 0.968 (p<0.001) RMSE = 0.329
CPU cycle	594.3765/(2.93*10^9) Hz	1773.5184/(2.93*10^9) Hz
Reference	Figure 3.1 (Chapter 3)	Figure S3.3 (Chapter3; supporting information)

* Final Optimal Model

Table 4.3 Model comparison according to machine learning techniques

	Model 3*	**Model 5**	**Model 6**
Machine learning algorithm	Support Vector Machine for regression (SVR)	Least Absolute Shrinkage and Selection Operator (LASSO)	Random Forest Regression (RFR)
Model building process	Feature selection by LASSO → SVR model building based on best feature index	Feature selection by LASSO → LASSO model building based on best feature index	Feature selection by LASSO → RFR model building based on best feature index
Parameter settings #	Penalty parameter C [0.01, 10, 0.01]; Gamma [0.0001, 1, 0.0001], where elements in list denote the beginning, end, and number of samples to generate, respectively (BEST C=1.32, BEST G=0.0002)	Type.measure= 'mse' BESTLAMDA= lambda.min (BESTLAMDA=0.033127)	MaxTree= 1000 BESTNTREE=which.min(RMSE) ntree = BESTNTREE (ntree=596)
Model Performance (prediction and error)	Correlation between predictions and recent influenza incidence data of 0.956 (p<0.001); RMSE = 0.386	Correlation between predictions and recent influenza incidence data of 0.545 (p=0.0002); RMSE = 10.304	Correlation between predictions and recent influenza incidence data of 0.911 (p<0.001); RMSE = 0.776
Reference	Figure 3.1 (Chapter 3)	Figure S3.4 (Chapter3; supporting information)	Figure S3.5 (Chapter3; supporting information)

* Final optimal model; # To select the optimal SVR parameter settings, 10-fold cross-validation was performed.

Table 4.4 SVR Model Comparison according to National Influenza surveillance data

	ILI surveillance model (Model 3*)	**Virological surveillance model**
Main data sources	Clinical data: rates of physician visits for ILI	Virological data: positive rates for the influenza virus through laboratory tests
Input features for model building	Total number of features without duplication = 36 (of the 152 queries, 15, 14, and 29 principle features had the minimum lambda value in lags -2, -1, and 0, respectively)	Total number of features without duplication = 53 (of the 152 queries, 28, 26, and 45 principle features had the minimum lambda value in lags -2, -1, and 0, respectively)
Model performance (prediction and error)	Correlation between predictions and recent influenza incidence data of 0.956 (p<0.001) RMSE = 0.386	Correlation between predictions and recent influenza incidence data of 0.963 (p<0.001) RMSE = 7.24
Reference	Figure 3.1 (Chapter 3)	Figure 3.2 (Chapter3)

* Final Optimal Model

4.2 Limitations

Online surveillance detects disease by tracking the behavior of individuals through massive amounts of aggregated data on the Internet; however, novel and developing approaches using these data have led to calls for sophisticated methodologies that result in reliable surveillance. Several limitations arise due to the nature of the data, including noise from irrelevant information, uncertainty regarding the representativeness of social media users, and changes in the quantity and quality of Internet use over time. In particular, changes in Internet usage rates and health information-seeking rates are likely to represent fairly central limitations regarding the use of search query data. Because the search behavior of individuals is constantly changing, keywords submitted by individuals may be influenced by various factors, including media-driven interest or other events. These changes may alter or degrade the performance of search query-based surveillance. The recent GFT overestimation can be understood in the same context. These limitations exist in the data used in my study; thus optimal features of the model may need to be updated over time.

4.3 Significance for public health

Despite the aforementioned limitations, the results of this study are significant from a public health standpoint. I found that Internet-based influenza surveillance using both search engine query data and social media data has the power to predict influenza outbreaks, exhibiting strong congruence with traditional surveillance data. Thus, my results demonstrate the feasibility of data from search queries and social media for enhancing influenza surveillance in South Korea. In addition, this study proposed several approaches for achieving the near real-time surveillance of emerging influenza outbreaks. It is important to prepare for the next severe pandemic, such as the 2009 H1N1 pandemic, and to properly control seasonal influenza epidemics. Approaches adopted in this study for influenza forecasting can support and extend the capacity of traditional surveillance systems for detecting emerging influenza outbreaks. Epidemiologists and policy makers may find it useful to apply my approach for further influenza surveillance. In addition, the basic principles underpinning the approach to this research may be applied to other countries, languages, diseases, and data sources.

My results also point to the desirability of future studies. I suggest several strategies for the expansion of online infectious disease surveillance methods based on my approach. The relevant concerns and specific strategies for future studies may be summarized as follows:

1) The first strategy is to develop a national disease information system based on Internet data for the early detection of the incidence and spread of infectious diseases. This should be done by:

- developing a predictive model for the incidence and spread of other infectious diseases through the use of unstructured event-based news reports, search engine queries, social media data, and other necessary tools;
    - considering additional data sources to increase the validity and reliability of prediction, such as those related to weather, air pollution, and environmental measurements;
- designing a real-time information provider system for the public about the incidence and spread of infectious diseases based on the predictive model and implementing such a system on the Web and in the mobile environment;
- systematizing a platform-based analysis process for the provision of a real-time information service pertaining to the incidence and spread of infectious diseases;
- offering strategies and methods for the practical use of an advanced supplementary backup system for the surveillance and monitoring of diseases at the national level.

2) The second strategy is to prepare and make provisions for the advanced and efficient management of an online surveillance system for infectious diseases. This should be achieved by:

- identifying and theorizing about the socio-demographic risk factors pertaining to the incidence and spread of infectious diseases through explanatory variables of a standardized analysis model used for the surveillance of infectious diseases;

- suggesting efficient measures for infectious disease management by analyzing individual health-seeking behaviors related to infectious diseases.

4.4 Conclusion

This study provides further evidence based on a new approach linking the use of Internet-based data and surveillance of emerging influenza outbreaks in South Korea. I found that Internet-based influenza surveillance using data from both search engine queries and social media has the power to predict influenza outbreaks, exhibiting strong congruence with traditional surveillance data. Hence, such an approach may be valuable for supporting preparation for the next severe pandemic, similar to the 2009 H1N1 pandemic, and for the proper control of seasonal influenza epidemics. Furthermore, I established that information gleaned from social media data can be productively fused with a methodology for query-based influenza surveillance. My results suggest that these new data sources can be quite compatible with and complementary to each other with respect to predicting influenza incidence.

It is evident that approaches using these novel data sources for influenza surveillance are more immediate and efficient in terms of time and cost than traditional surveillance approaches, although it is difficult to conclude that they can or should replace traditional surveillance systems. Looking ahead to the future, there is no doubt that Internet-based big data can play a

foundational role with respect to information sources regarding public health. Internet-based online influenza surveillance could potentially complement and augment traditional surveillance systems. In particular, fusing Internet search query data and SNS data based on big data analysis can result in the rapid and efficient prediction of the occurrence of diseases and their proliferation, thereby allowing for better recognition of diseases and initiation of preventive measures. I suggest that these data sources offer opportunities for monitoring public health in general, beyond their specific application here to outbreaks of influenza.

Abstract in Korean

국문초록

인터넷검색쿼리와 소셜미디어 데이터를 활용한 사회인구학적 독감 감시모형개발

서울대학교 대학원

보건학과

우혜경

연구목적: 감염성 질환의 발생을 조기에 확인하는 것은 공공보건 차원에서 정부 및 개인이 시기 적절하게 중재 및 대처할 수 있게 함으로서 질병발생과 확산으로 인한 혼란과 피해를 최소화 할 수 있는 가장 효과적인 방법이다. 최근 국내·외적으로 웹 및 모바일 기능과 정보를 질병감시에 응용하고자 하는 학문적, 정책적 논의들이 제기되고 있다. 의학 및 보건학 분야에서 정보기술의 응용은 질병발생과 확산에 대한 빠른 정보획득을 가능하게 하고, 실시간 정보제공이나 대응을 위한 기반을 제공해 줄 수 있기 때문이다. 본 연구는 전 세계에 걸쳐 계절마다 수천에서 수 만 명의 사망자를 발생시키는 대표적 감염성 질환인 독감을 중심으로 인터넷 기반의 데이터를 활용한 감염병 감시모형을 개발하고자 하였다.

연구방법 및 결과: 독감 감시모형의 개발을 위해서 1) 독감발생 및 확산을 예측하는 키워드 검색 연구와 2) 독감 예측모형 개발 및 평가연구를 수행하였다.

(연구 1) 독감발생 및 확산을 예측하는 키워드 검색 연구: 트위터 및 네이버 블

로그의 소셜미디어 데이터와 질병관리본부의 인플루엔자 표본감시 자료인 인플루엔자 의사환자 발생건수 데이터를 활용하여 독감발생 및 확산을 예측하는 키워드를 확인하고자 하였다. 초기 키워드 선정을 위해 ① 지난 43개월 동안 트위터 및 블로그 포스트에서 독감과 함께 자주 등장하는 연관어를 찾았고 (총 2,065 개), ② 실제 질병관리본부의 인플루엔자 의사환자 발생건수와 상관관계가 있는 키워드들로 필터링 하여 ③ 총 49개로 구성된 단일 키워드 및 조합된 키워드 리스트를 생성하였다. 키워드의 속성에 따라 몇 가지 제외조건을 적용하여 선별된 키워드를 포함하는 포스트들의 시계열 볼륨의 데이터 셋을 구축하였고, 모델링과 평가를 위한 데이터 셋을 구분하였다. 데이터 정규화 이후 Least absolute shrinkage and selection operator (Lasso) algorithm을 사용하여 10-fold cross validation 방법으로 최적 데이터 서브 셋과 모델링을 위한 피쳐들을 선택하였다. 마지막으로 최종 선택된 키워드들이 실제 독감의 발생과 확산을 예측하는지 평가해 보기 위해서 Lasso, Support Vector Machine for Regression (SVR) 및 Random Forest Regression (RFR) 등 머신 러닝(machine learning) 방법을 활용하여 모델링 및 평가를 수행하였다. 분석결과, 총 15개의 키워드가 독감의 발생 및 확산을 예측하는 키워드로 선택되었고, 모형평가 결과에서 최종 예측모형은 질병관리본부의 최근 독감발생률과 매우 높은 상관관계를 나타냈다(SVR model correlation: r=0.92, p<.001; RMSE=0.55).

연구2) 독감 예측모형 개발 및 평가연구: 트위터 및 네이버 블로그의 소셜미디어 데이터, 포털 사이트 Daum의 검색엔진 쿼리데이터, 그리고 질병관리본부의 인플루엔자 표본감시 자료인 인플루엔자 의사환자 발생건수 및 실험실 검사 양성건수 데이터를 활용하여 독감 예측모형을 개발하고자 하였다. 모형을 위한 쿼리선택 방법으로써 ① 트위터 및 블로그 데이터를 이용한 독감 연관어 키워드 탐색(총 103개의 단일키워드 또는 조합키워드), ② 인플루엔자 유사질환(ILI)의 공식정의,

미국 질병관리본부(CDC)에서 제공하는 독감증상자료 및 전문가 자문을 통한 환자들의 독감관련 주 호소(chief complaints) 정보 취합(총 29개의 키워드), ③ 우리나라 대표 인터넷 포털사이트인 Naver와 Daum의 웹 쿼리 추천시스템을 통한 추천 쿼리 리스트 탐색(총 75개 키워드) 등을 모두 고려하였다. 선택된 총 216개의 후보쿼리 중에서 중복된 쿼리를 제외하고, 총 6개의 씨앗쿼리와 총 146개의 독감 연관쿼리 리스트를 생성하였다. 선별된 152개의 쿼리들의 시계열 볼륨을 daum으로부터 제공받아 데이터 셋을 구축하였고, 모델링과 평가를 위한 데이터 셋을 구분하였다. 데이터 정규화 이후 Lasso algorithm을 사용하여 10-fold cross validation 방법으로 모델링을 위한 피쳐들을 선택하였다. 마지막으로 독감 발생 및 확산을 예측하는 모형을 구축하고 평가하기 위해서 Lasso, SVR 및 RFR 등 머신 러닝 방법을 채택하여 분석하였다. 분석결과, 총 36개의 검색쿼리가 질병관리본부의 인플루엔자 의사환자건수에 기반한 독감발생률을 잘 예측하는 쿼리로 선택되었고, 모형평가 결과에서 최종 예측모형은 질병관리본부의 최근 독감발생률과 매우 높은 상관관계를 나타냈다(SVR model의 correlation: r= 0.956, p<.001; RMSE=0.39). 같은 연구 절차로 인플루엔자 실험실 검사 양성건수에 기반한 독감발생률 예측모형을 구축하여 실험해 본 결과, 총 53개의 검색쿼리가 예측모형에 적합한 것으로 분석되었고, 모형평가 결과에서 모형의 예측력과 성능이 높은 것으로 나타났다(SVR model의 correlation: r= 0.963, p<.001; RMSE=7.24).

결론: 인터넷 검색엔진쿼리와 소셜 미디어 데이터를 활용하여 개발한 독감 감시 예측모형은 질병관리본부에서 공표하는 독감 발생률과 높은 상관관계를 보였다. 본 연구의 결과에 따라 검색쿼리와 소셜미디어 데이터는 독감 감시를 위한 자료원으로써 충분히 타당성이 있다는 사실을 확인하였다. 아울러, 본 연구는 최적의

예측모델을 개발하기 위해서 성격이 다른 두 데이터를 융합한 새로운 방법론을 시도하였다. 인터넷 검색엔진쿼리와 소셜 미디어 데이터 정보의 융합은 상호 보완적으로 독감 발생 및 확산을 예측하는 모형의 성능을 향상시켰다. 본 연구에서 사용된 방법론은 독감뿐만 아니라 다른 감염성 질환의 예측모형 개발에도 유연하게 적용될 수 있을 것이며, 향후 감염병 국가감시체계의 기능을 보완 및 강화할 수 있는 보완시스템 개발에 유용하게 활용될 수 있을 것이다.

핵심어: 독감, 감시, 인구 감시, 정보역학, 정보감시, 인터넷 검색, 쿼리, 소셜 미디어, 빅 데이터, 예측, 역학, 조기대응