Published online: 2019-08-16

Appendix: Content Summaries of Best Papers for the Public Health and Epidemiology Informatics Section of the 2019 IMIA Yearbook

Arsevska E, Valentin S, Rabatel J, de Goër de Hervé J, Falala S, Lancelot R, Roche M

Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System

PLoS One 2018 Aug 3;13(8):e0199960

Animal infection outbreaks are a major public health threat. In this paper, the researchers developed a platform to detect automatically animal infection outbreaks from online news sources, the Platform for Automated extraction of Disease Information from the web (PADI-web). Information is retrieved from Google News in the English language and a free news aggregator of about 4,500 news sites. Five types of information are extracted in each news article with natural language processing techniques and machine learning: the number of cases, the location, the date, the disease name, and the affected host (cattle, pig...). An automatic rule discovery module based on the frequent item discovery algorithm, a data mining technique, discovers rules that are later used as features for a machine learning classifier. As an example of rule automatically discovered, when the word « killed » is found in one of three words preceding a number then this number is likely to be the number of cases. A support vector machine classifier combines different rules to predict the class (number of cases, date…) of a term in the news article. The authors achieved good performances in these information extraction tasks ranging from 80 % to 95 % as measured by the F-score, the harmonic mean of precision and recall.

Although many limitations hamper the feasibility of an exhaustive surveillance worldwide, this platform detected in 2016 several animal infectious outbreaks days before it was notified by the World Organization for Animal Health. Once again, free online content on the Internet contains key information for public health surveillance that can be extracted with artificial intelligence methods. This paper is a very good example of a full implementation and evaluation of such a web-based system.

Effland T, Lawson A, Balter S, Devinney K, Reddy V, Waechter H, Gravano L, Hsu D

Discovering foodborne illness in online restaurant reviews

J Am Med Inform Assoc 2018 Dec 1;25(12):1586-92

A foodborne illness is an infectious intestinal disease caused by a pathogen (bacteria, virus, or parasite) that enters the body through food or drink consumption. Common symptoms include abdominal cramps, nausea, and vomiting. In this paper, the authors used data from consumer reviews obtained from the Yelp website, a social networking site that records user's rate and review on restaurants. In collaboration with the New York City (NYC) Department of Health and Mental Hygiene (DOHMH), the authors developed a system to identify restaurant reviews on Yelp indicating foodborne illness. The classification algorithm uses a bag-of-words approach with a logistic regression classifier. The system is parameterized to favor recall over precision to reduce the risk of missing true positives. Yelp reviews of NYC restaurants are pulled every day and DOHMH epidemiologists validate each signal in a user interface. If a signal is validated, a Yelp message is sent to the author of the review to gather more information for further investigation. The system identified 10 outbreaks and 8,523 reports of foodborne illness associated with NYC restaurants since July 2012. Interestingly, only 3% of the illness incidents had been reported to DOHMH, which highlights the value of social media as an important source for foodborne illness surveillance. A very interesting aspect of the work presented is the use of biased adjusted data for improving the performance of the classifier in a continuous approach where validated data coming from previous alerts are used to further improve the algorithm. Obviously, the true performance of the system is depending on the quality of the information available on the social media and any manipulation of these data by the provider (in regards of the allegations reported in the news about Yelp) could have an influence on the results but this is not discussed in the paper.

Wakamiya S, Kawai Y, Aramaki E

Twitter-Based Influenza Detection After Flu Peak via Tweets With Indirect Information: Text Mining Study

JMIR Public Health Surveill 2018 Sep 25;4(3):e65

Social media data such as Twitter can be used to detect Influenza outbreak. It requires natural language processing (NLP) techniques and a classifier to perform tweet classification. However, differentiating direct and indirect information is crucial for Public Health surveillance that needs a good approximation of the number of cases. For example, "I got the flu today" is a direct information (D) tweet whereas "there is a major outbreak in Okinawa" provides indirect information (I). The authors developed a specific module to handle this task by applying a binary classifier based on support vector machine under the bag-of-words representation of tweets. A location module tries to infer the user location using the user profile, the GPS coordinates when available, and the content of tweets. Furthermore, the authors introduced the concept of "trapped sensors" to better predict the number of cases per area. The idea is that after the onset of an epidemic, people become deactivated by the event and they do not report having flu. Such deactivated people are called "trapped sensors". Statistical models that take into account these "trapped sensors" to predict the number of cases showed better correlation with the reference (information from Health Authorities) than over models. This paper explains the many pitfalls that exist while using Twitter data for flu monitoring and proposes a new approach to make good approximation of Influenza cases.