VALIDATING SEARCH PROTOCOLS FOR MINING OF HEALTH AND DISEASE EVENTS ON TWITTER

Aditya Lia Ramadona[1,2], Lutfan Lazuardi[3], Sulistyawati[1,4], Anwar Dwi Cahyono[5], Åsa Holmner[6], Hari Kusnanto[3], Joacim Rocklöv[1]

[1] Department of Public Health and Clinical Medicine, Epidemiology and Global Health, Umeå University, Umeå, Sweden
[2] Center for Environmental Studies, Gadjah Mada University, Yogyakarta, Indonesia
[3] Department of Public Health, Faculty of Medicine, Gadjah Mada University, Yogyakarta, Indonesia
[4] Department of Public Health, Ahmad Dahlan University, Yogyakarta, Indonesia
[5] District Health Office, Yogyakarta, Indonesia
[6] Department of Radiation Sciences, Umeå University, Umeå, Sweden

ABSTRACT

BACKGROUND: Twitter is a free social networking and micro-blogging service that enables its users to read and share information with user and media communities in messages no longer than 140-character. In the year of 2016, there were more than 24 million Indonesian twitter users sharing news, events, as well as personal feelings and experiences on Twitter. This study seeks to validate a search protocol of health related terms using real-time Twitter data which can later be used to understand if, and how, twitter can reveal information on the current health situation in Indonesia. In this validation study of mining protocols, we: 1) extracted geo-located conversations related to health and disease postings on Twitter using a set of pre-defined keywords, 2) assessed the prevalence, frequency and timing of such content in these conversations, and 3) validated how this search protocol was able to detect relevant disease tweets.
SUBJECT AND METHODS: Groups of words and phrases relevant to disease symptoms and health outcomes were used in a protocol developed in the Indonesian language in order to extract relevant content from geo-tagged Twitter feeds. A supervised learning algorithm using Classification and Regression Tree´s (CART) was used to validate search protocols of disease and health hits comparing to those identified by a team of human experts. The experts categorized tweets as positive or negative in respect to health events. The model fit was evaluated based on prediction performance.
RESULTS: 390 tweets from historical Twitter feeds and 1,145,649 tweets from Twitter stream feeds during the period July 26th to August 1st, 2016. Only twitter hits with health related keywords in the Indonesian language were obtained. The accuracy of predictions of mined hits versus expert validated hits using the CART algorithm showed good validity with AUC beyond 0.8.

https://doi.org/10.26911/theicph.2016.074

CONCLUSION: Monitoring of public sentiment on Twitter, combined with contextual knowledge about the disease, can detect health and disease tweets and potentially be used as a valuable real-time proxy for health events over space and time.

Keywords: social networking, disease detection, disease early warning, digital epidemiology, big data analysis

https://doi.org/10.26911/theicph.2016.074