Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Ontology-based automatic identification of public health-related Turkish tweets

Emine Ela Küçük[a], Kürşad Yapar[b], Dilek Küçük[c,*], Doğan Küçük[d]

[a] Department of Public Health, Faculty of Health Sciences, Giresun University, Giresun, Turkey
[b] Department of Medical Pharmacology, Faculty of Medicine, Giresun University, Giresun, Turkey
[c] Electrical Power Technologies Group, TÜBİTAK Energy Institute, Ankara, Turkey
[d] Department of Computer Engineering, Gazi University, Ankara , Turkey

ARTICLE INFO

Keywords:
Public health
Twitter
Social media analysis
Health informatics
Automatic text processing

ABSTRACT

Social media analysis, such as the analysis of tweets, is a promising research topic for tracking public health concerns including epidemics. In this paper, we present an ontology-based approach to automatically identify public health-related Turkish tweets. The system is based on a public health ontology that we have constructed through a semi-automated procedure. The ontology concepts are expanded through a linguistically motivated relaxation scheme as the last stage of ontology development, before being integrated into our system to increase its coverage. The ultimate lexical resource which includes the terms corresponding to the ontology concepts is used to filter the Twitter stream so that a plausible tweet subset, including mostly public-health related tweets, can be obtained. Experiments are carried out on two million genuine tweets and promising precision rates are obtained. Also implemented within the course of the current study is a Web-based interface, to track the results of this identification system, to be used by the related public health staff. Hence, the current social media analysis study has both technical and practical contributions to the significant domain of public health.

1. Introduction

With the advent of social media tools like Twitter, Facebook, and Instagram; and their widespread use every day, social media analysis research is correspondingly boosted for different purposes. These purposes include using social media for trend analysis, opinion mining (sentiment analysis) for brands/products, and for tracking epidemics/diseases, to name a few. Similar to the last item, in this paper, we focus on the use of social media for automatic tracking of public health-related tweets in order to help the public health experts in determining the current public health concerns in a timely manner.

In a related review paper [1], it is emphasized that Twitter offers quite significant opportunities for monitoring public health compared to the traditional mechanisms employed in the discipline. For instance, instead of manual, slow, and time-consuming data collection methods that the conventional systems employ; real-time monitoring and statistics on public health can be achieved with reduced costs through the use of natural language processing and machine learning techniques on Twitter and other social media [1]. Furthermore, the locations of these public health events such as disease outbreaks, epidemics, and health threats can be localized through the related Twitter facilities [1].

Accordingly, related work on tweet analysis for disease or epidemic/pandemic surveillance includes [2] where a number of regression models are applied in order to identify influenza-related tweets and several conclusions are drawn after the comparison of these models. In [3], an SVM classifier is employed to detect health-related tweets where the training data of the classifier includes 5128 labeled tweets. After the application of this classifier to a set of about 11.7 million tweets, the authors produced a set of 1.63 million tweets related to public health, with high precision [3]. In [4], tweets are examined to track disease activity and public sentiment to influenza during an Influenza A H1N1 pandemic in US dated 2009. As a result of their analysis, the authors point out that Twitter traffic can be used to estimate disease activity in real-time and tweets can be used to measure disease-related public concern [4]. In another similar work published in the same year, tweets are analyzed to discover mentions of ailments, track illnesses over time, localize illnesses, analyze medication use and symptoms, among others [5]. They conclude that Twitter is a very promising application platform for public health research [5]. In [6], the authors analyze tweets for the purposes of Dengue (an infectious disease) surveillance. It is shown that spatial and temporal prediction of Dengue epidemics can be performed on Twitter [6]. Related tweets are also analyzed in a study to

[*] Corresponding author.

E-mail addresses: emine.kucuk@giresun.edu.tr (E.E. Küçük), kursad.yapar@giresun.edu.tr (K. Yapar), dilek.kucuk@tubitak.gov.tr (D. Küçük), dogan.kucuk@gazi.edu.tr (D. Küçük).

http://dx.doi.org/10.1016/j.compbiomed.2017.02.001
Received 1 July 2016; Received in revised form 1 February 2017; Accepted 3 February 2017
0010-4825/

reveal statistics related to dental pain and actions taken against the pain, hence it is also argued that Twitter is a fruitful platform to be utilized by dental professionals [7]. In another study, tweets are analyzed to track drug and alcohol use as a showcase to determine public health-related topics on Twitter by means of topic models [8]. Another research on influenza surveillance in social media attempts to distinguish between real influenza inflection mentions and Twitter chatter not reporting actual infections [9]. It is concluded that the employed approach leads to promising results revealed with an accuracy of about 85% for the detection of weekly change in the direction (increase or decrease) of influenza prevalence in the real influenza epidemic in 2012–2013 [9]. In [10], the authors point out that tweets on influenza may be reporting actual infections or may be stating awareness/fear. Hence, they propose a learning algorithm to classify tweets into these two classes and conclude that deep content analysis like theirs is necessary for influenza surveillance as well as other similar tasks on Twitter [10]. In [11], a minimally-supervised learning algorithm is developed to determine the everyday jargon used to express influenza-like illnesses and later used these to form Twitter queries targeting at tweets reporting these illnesses. High correlation rates are reported between their tweet trends and similar trends obtained by the traditional surveillance services [11]. An extension of this latter system to be applicable to other syndromes is presented in [12] and similar evaluation results are reported. In [13], a Naive Bayes classifier is used to train a classifier for influenza-like illness detection in Portuguese tweets. They compare their results with the corresponding data from Influenzanet [14] and report high correlation rates [13]. More recently, in [15], the authors propose a machine learning approach which combines the data from social media, search, and conventional data sources to nowcast and forecast influenza activity. And in [16], an online tool called FluOutlook is presented which performs influenza forecasting by combining current and historical influenza data, social media, and different forecast models.

In this paper, we propose an ontology-based system for automatic tracking of public health-related tweets in Turkish[1]. The main contributions of the proposed system are listed below:

- The system is based on a public health ontology covering significant concepts including disease names, related symptoms, and medication categories. This ontology is created semi-automatically within the course of the current study and during the construction procedure the ontology concepts are expanded through linguistically-motivated expansion schemes so that a large list of terms related to public health is obtained. This term list distilled from the ontology is made publicly available for research purposes.
- The evaluations are performed on two randomly compiled tweets sets (of one million tweets each, collected during two distinct consecutive 20-day periods) without providing any search criterion except the language. We discuss the evaluation results and provide error analyses which can be utilized as a guide during further studies.
- To the best of our knowledge, this is the first study to analyze Turkish tweets for public health surveillance purposes. Most of the previous work is carried out on English tweets, only one of them is on Portuguese tweets [13], but it is a significant research issue to carry out analysis experiments on social media content in other languages, or on multilingual content.
- As a proof-of-concept, a Web-based interface is implemented so that the related public health experts can use the interface for public health surveillance through Twitter.

The rest of the paper is organized as follows: In Section 2, the details of building the public health ontology from scratch through a semi-automated procedure is described. The ultimate system based on this ontology for public health surveillance is presented in Section 3 together with its overall evaluation results on the data sets and a brief description of the Web interface of the system. Section 4 includes more detailed evaluation results regarding the system performance. Discussions on these evaluation results and also on the results of an SVM-based experiment are provided in Section 5. Finally, Section 6 concludes the paper with a summary of main points and future research directions.

2. Semi-automatic construction of the Turkish public health ontology

Domain ontologies are important semantic information sources covering the concepts, relations, and rules within the domain under consideration [18]. Various significant domain ontologies have been proposed so far in the literature, such as the gene ontology [19], the bioinformatics ontology [20], and the protein ontology [21].

In order to be used in our ultimate Twitter tracking system, we have constructed a public health ontology in Turkish comprising significant public health concepts. Our approach is a semi-automated one including manual, automatic, and semi-automatic stages performed in a pipelined manner, as depicted in Fig. 1. The manual stages of the ontology development process are depicted with white boxes while those fully-automated stages are shown with green boxes, to differentiate the two. The only semi-automatic stage, Stage 6, is depicted with a light-green box.

Similar to the study described in [22], we have used Wikipedia as a semantic resource together with other Web resources to facilitate our ontology development process. The main motivation for building this ontology is to compile a wide-coverage list of public health terms to be used by our ultimate system.

The details of the development stages of our ontology, given in Fig. 1, are provided below. The number of ontology terms present at the end of each stage is shown within boxes connected to the stages with dashed arrows.

- *Stage 1*: The authors of the study have determined the main concepts of our public health ontology as *GeneralPublicHealth*, *Disease*, *Symptom*, and *Medication*, where some of these authors are domain experts.[2] The subconcepts of some of these main concepts are manually determined, such as *Epidemic*, *Hospital*, *Emergency*, and *Vaccination* as the subconcepts of *GeneralPublicHealth*. Similarly, under the concept of *Medication*, the generic medication types such as *Antibiotic* and *Antihistamine*.
- *Stage 2*: In this automatic stage, Wikipedia articles and other Web resources are automatically processed to determine the subconcepts of *Disease* and *Symptom*. Especially those pages including lists of related concepts such as ⟨https://tr.wikipedia.org/wiki/Hastal%C4%B1k_isimleri_listesi⟩ are considered where this Wikipedia page provides a list of disease names in Turkish. At this stage, we also include English resources as diseases may sometimes be expressed with their English names within tweets.
- *Stage 3*: The automatic extraction procedure might have extracted terms with writing errors mainly due to character encoding problems, particularly during the extraction of terms having one of the six Turkish characters with diacritics (*ç, ğ, ı, ö, ş,* and *ü*). Another source of errors propagated from the previous stage is that the community-created data sources like Wikipedia might include erroneous information which needs manual correction. These errors are corrected during this manual stage.

[1] A preliminary version of this paper is presented in [17].

[2] To improve the readability and comprehensibility of our paper, the ontology concepts are given in English. However, when there is a need for actual examples, actual concepts in Turkish will be utilized.