**INTRODUCTION**

Social media has become a major source of information for analyzing all aspects of daily life. In particular, Twitter is used for public health monitoring to extract early indicators of the well-being of populations in different geographic regions. Twitter has become a major source of data for early monitoring and prediction in areas such as health [1], disaster management [2] and politics [3]. In the health domain, the ability to model transitions for ailments and detect statements like “people talk about smoking and cigarettes before talking about respiratory problems”, or “people talk about headaches and stomach ache in any order”, benefits syndromic surveillance and helps measure behavioral riskfactors and trigger public health campaigns. In this paper,we formulate two problems: the health transition detection problem and the health transition prediction problem. To address the detection problem, we develop TM–ATAM that models temporal transitions of health-related topics. To address the prediction problem, we propose T–ATAM, a novel method which uncovers latent ailment inside tweets by treating time as a random variable natively inside ATAM[4]. Treating time as a random variable is key to predicting the subtle change in health-related discourse on Twitter.

Common ailments are traditionally monitored by collecting data from health-care facilities, a process known as sentinel surveillance. Such resources limit surveillance, most especially for real-time feedback. For this reason, the Web has become a source of syndromic surveillance, operating

on a wider scale, near real time and at virtually no cost. Our challenges are: (i) identify health-related tweets, (ii)determine when health-related discussions on Twitter transitions from one topic to another, (iii) capture different such transitions for different geographic regions. Indeed, in addition to evolving over time, ailment distributions also evolve in space.

Therefore, to attain effectiveness, we must carefully model two key granularities, temporal and geographic. A temporalgranularity that is too-fine may result in sparse and spurioustransitions whereas a too-coarse one could miss valuable ailmenttransitions. Similarly, a too-fine geographic granularity may produce false positives and a too-coarse one may missmeaningful transitions, e.g., when it concerns users livingin different climates. For example, discussions on allergy break at different periods in different states in the USA [4].Therefore, processing all tweets originating from the USA together will miss climate variations that affect people’shealth. We argue for the need to consider different timegranularities for different regions and we wish to identify and model the evolution of ailment distributions between different temporal granularities.

While several latent topic modeling methods such as Probabilistic Latent Semantic Indexing (pLSI) [5] and LatentDirichlet Allocation (LDA) [6], have been proposed to effectively cluster and classify general-purpose text, ithas been shown that dedicated methods such as the AilmentTopic Aspect Model (ATAM) are better suited for capturing ailments in Twitter [4]. ATAM extends LDA to model how users express ailments in tweets. It assumes thateach health-related tweet reflects a latent ailment such as flu and allergies. Similar to a topic, an ailment indexes a word distribution. ATAM also maintains a distribution over symptoms and treatments. This level of detail provides a more accurate model for latent ailments.

On the other hand, while pLSI and LDA have been shown to perform well on static documents, they cannot intrinsically capture topic evolution over time. Temporal-LDA (TM–LDA) was proposed as an extension to LDA formining topics from tweets over time [7]. To address thehealth transition detection problem, we propose TM–ATAMthat combines ATAM and TM–LDA. A preliminary version of TM–ATAM was described in a short paper [8]. We show here that it is able to capture transitions of health-related discussions in different regions (see Figure 1). As a result, the early detection of a change in discourse in Nevada, USA into allergies can trigger appropriate campaigns.

In each geographic region, TM–ATAM learns transition parameters that dictate the evolution of health-related topics by minimizing the prediction error on ailment distributions of consecutive pre-specified periods of time. Our second problem, the health transition prediction problem, is to automatically determine those periods. We hence propose T–ATAM, a different and new model that treats time as a random variable in the generative model. T–ATAM discovers latent ailments in health tweets by treating time as a variable whose values are drawn from a corpus-specific multinomial distribution. Just like TM–LDA, TM–ATAM and T–ATAMare different from dynamic topic models [9], [10], [11],as they are designed to learn topic transition patterns from temporally-ordered posts, while dynamic topic models focus on changing word distributions of topics over time.

Our experiments on a corpus of more than 500K health related tweets collected over an 8-month period, show that TM–ATAM outperforms TM–LDA in estimating temporal topic transitions of different geographic populations. Our results can be classified in two kinds of transitions. Stable topics are those where a health-related topic is mentioned continuously. One-Way transitions cover the case where some topics are discussed after others. For example, our study of tweets from California revealed many stable topics such as headaches and migraines. On the other hand, tweeting about smoking, drugs and cigarettes is followed by tweeting about respiratory ailments. Figure 1 shows example one way transitions we extracted for different states and cities in the world. Such transitions are often due to external factors such as climate, health campaigns, nutrition and lifestyle of different world populations.