Health Monitoring on Social Media over Time

Sumit Sidana, Sihem Amer-Yahia, Marianne Clausel,
Majdeddine Rebai, Son T. Mai, and Massih-Réza Amini

Abstract—Social media has become a major source for analyzing all aspects of daily life. Thanks to dedicated latent topic analysis methods such as the Ailment Topic Aspect Model (ATAM), public health can now be observed on Twitter. In this work, we are interested in using social media to monitor people's health over time. The use of tweets has several benefits including instantaneous data availability at virtually no cost. Early monitoring of health data is complementary to post-factum studies and enables a range of applications such as measuring behavioral risk factors and triggering health campaigns. We formulate two problems: *health transition detection* and *health transition prediction.* We first propose the Temporal Ailment Topic Aspect Model (TM–ATAM), a new latent model dedicated to solving the first problem by capturing transitions that involve health-related topics. TM–ATAM is a non-obvious extension to ATAM that was designed to extract health-related topics. It learns health-related topic transitions by *minimizing the prediction error on topic distributions between consecutive posts at different time and geographic granularities.* To solve the second problem, we develop T–ATAM, a Temporal Ailment Topic Aspect Model where time is treated as a random variable *natively* inside ATAM. Our experiments on an 8-month corpus of tweets show that TM–ATAM outperforms TM–LDA in estimating health-related transitions from tweets for different geographic populations. We examine the ability of TM–ATAM to detect transitions due to climate conditions in different geographic regions. We then show how T–ATAM can be used to predict the most important transition and additionally compare T–ATAM with CDC (Center for Disease Control) data and Google Flu Trends.

Index Terms—Public health, ailments, social media, topic models

1 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 risk factors 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 temporal granularity that is too-fine may result in sparse and spurious transitions whereas a too-coarse one could miss valuable ailment transitions. Similarly, a too-fine geographic granularity may produce false positives and a too-coarse one may miss meaningful transitions, e.g., when it concerns users living in 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's health. We argue for the need to consider different time granularities 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 Latent Dirichlet Allocation (LDA) [6], have been proposed to effectively cluster and classify general-purpose text, it has been shown that dedicated methods such as the Ailment Topic 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 that each health-related tweet reflects a latent ailment such as flu and

- *S. Sidana, S. Amer-Yahia, M. Clausel, S. T. Mai, and M.-R. Amini are with the University Grenoble Alpes/CNRS, Grenoble 38400, France. E-mail: {Sumit.Sidana, Sihem.Amer-Yahia, Marianne.Clausel, mtson, Massih-Réza.Amini}@univ-grenoble-alpes.fr.*
- *M.-R. Rebai is with the University Paris-Saclay, Paris 91190, France. E-mail: majdeddine.rebai@ensta-Paristech.fr.*

Manuscript received 17 Aug. 2017; revised 9 Jan. 2018; accepted 12 Jan. 2018. Date of publication 18 Jan. 2018; date of current version 5 July 2018.
(Corresponding author: Son T. Mai.)
Recommended for acceptance by Y. Zhang.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2018.2795606

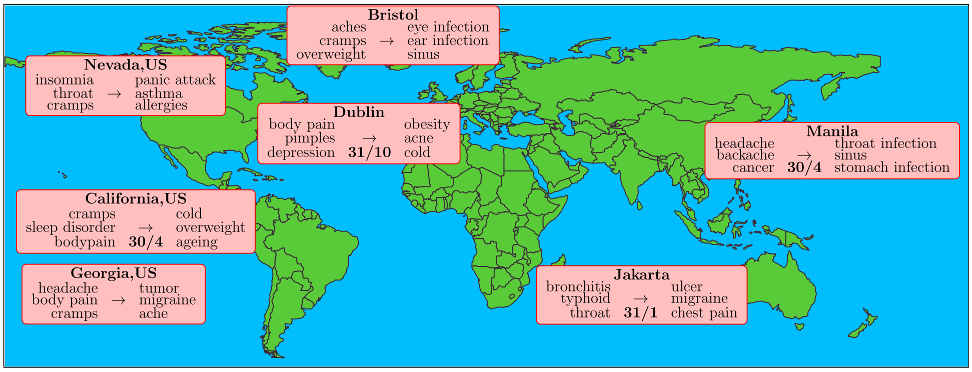

Bristol
aches cramps overweight → eye infection ear infection sinus
Nevada,US
insomnia throat cramps → panic attack asthma allergies
Dublin
body pain pimples depression 31/10 → obesity acne cold
Manila
headache backache cancer 30/4 → throat infection sinus stomach infection
California,US
cramps sleep disorder bodypain 30/4 → cold overweight ageing
Georgia,US
headache body pain cramps → tumor migraine ache
Jakarta
bronchitis typhoid throat 31/1 → ulcer migraine chest pain

Fig. 1. One-Way ailment transitions obtained by TM–ATAM for various regions. For each location, the time period is divided into two parts, preceding and following the most significant change-point discovered for that location. We show the most popular ailments on either side of this boundary.

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 for mining topics from tweets over time [7]. To address the health transition detection problem, we propose TM–ATAM that 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 Fig. 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–ATAM are 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 500 K 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. Fig. 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.

Our empirical evaluation relies on two approaches: *perplexity* as a measure to predict future ailments, and a comparison against a ground truth. Using perplexity, we show that by modeling transitions in the same `homogeneous time period`, TM–ATAM consistently outperforms TM–LDA in predicting health topics in all social-media active regions. By outperforming TM–LDA in predicting future health topics, we effectively show that it is essential to use a dedicated method that separates health-related topics from other topics. We also find that prediction accuracy for health topics is higher when operating TM–ATAM on finer spatial granularity and shorter time periods. That could be explained with more focused discourse, and hence less noise, in finer spatio-temporal granularities. T–ATAM is the big winner as it largely outperforms the other models, TM–LDA and TM–ATAM, in predicting health topics in both US and non-US regions. Finally, by a comparison with CDC "flu" data (the rates of the positive tests of influenza measured by the Center of Disease Control and Prevention in the US) and Google Flu Trends data, T–ATAM shows very good correlations.

We summarize our contributions as follows:

(1) TM–ATAM, a model able to *detect* health-related tweets and their evolution over time and space. TM–ATAM learns, for a given region, transition parameters by minimizing the prediction error on ailment distributions of pre-determined time periods.
(2) T–ATAM, a new model able to *predict* health-related tweets by treating time as a variable whose values are drawn from a corpus-specific multinomial distribution.
(3) Extensive experiments that show the superiority of T–ATAM for predicting health transitions, when compared against TM–LDA and TM–ATAM, and its effectiveness against a ground truth.

To the best of our knowledge, this is the first paper that effectively enables the early detection of evolving health-related topics in tweets. Section 2 defines our data model, and the two existing topic models for tweets LDA and ATAM, and formalizes our health transition detection and prediction problems. In Sections 3 and 4, we describe the construction of our two models TM–ATAM and T–ATAM. Section 5 contains experiments. Related work is provided in Section 6, and conclusion in Section 7.

TABLE 1
Mapping Tweets to Documents

Term	Description
$\mathcal{P}$	posts
$\mathcal{G}$	regions
$\mathcal{T}$	time periods
$\mathcal{P}_g^t$	posts from region g during time t
D_g^t	document-set built by mapping the content of each post $p \in \mathcal{P}_g^t$ to a document

2 Data Model, Topic Models and the Transition Detection Problem

We present our data and define a model that maps tweet posts to documents of different time and geographic granularities. We follow that with a background section that describes LDA and ATAM. Then we introduce the problems we are addressing in this work.

2.1 Mapping Tweets to Documents

We consider a set of posts $\mathcal{P} = \{p_1, p_2 ... p_n\}$. A *post* is the smallest unit of user-activity on a social media platform, such as a tweet, a tumblr post, or a facebook status update. In addition to a unique identifier and content, we assume the existence of two attributes, geographic coordinates and timestamp, for each post, $\langle id,\ coord,\ tstamp,\ content\rangle$.

Let $\mathcal{G} = \{g_1, g_2, \ldots\}$ represent a set of geographic regions around the world. We use $\mathcal{P}_g$ to refer to the set of posts in $\mathcal{P}$ that originate from a region $g \in \mathcal{G}$. The choice of a geographic granularity (country, state, county) is required to instantiate $\mathcal{G}$.

In a similar fashion, with a suitable choice of temporal granularity, we could divide up the entire time range spanned by posts in $\mathcal{P}$ into *disjoint and consecutive* periods, $\mathcal{T} = \{t_1, t_2 ...\}$. Possible choices for instantiation of $\mathcal{T}$ are week, bi-week, month, etc. We use $\mathcal{P}_g^t$ to refer to the set of posts in $\mathcal{P}$ that originated from a region g during period t.

We consider D_g^t the document formed by the concatenation of the content of all posts belonging to the set $\mathcal{P}_g^t$.

We use $\mathcal{D}_g = \{D_g^{t_1}, D_g^{t_2}, \ldots\}$ to denote the set of all documents corresponding to the aggregation of tweets from region g for different time periods in $\mathcal{T}$. Table 1 contains our terminology.

2.2 Background: Uncovering Latent Topics in Tweets

We review the principles general-purpose as well as health-related topic modeling. Existing models are (generally) unsupervised generative models that describe the content of a document in a large collection $\mathcal{D}$. In our case, $\mathcal{D}$ shall correspond to the set of documents built from tweets originating from one given region during a fixed time period.

2.2.1 Uncovering Latent Topics with LDA

Latent Dirichlet Allocation (LDA) represents each document as a probability distribution over k topics [6]. Each topic z in turn is represented as a probability distribution ϕ_z over a set of words. LDA assumes that the topic distribution θ_d of a document d and the vocabulary distribution ϕ_z of a topic z are generated according to a Dirichlet distribution. Vectorial parameters α and β of these Dirichlet distributions are assumed to be common to the whole corpus.

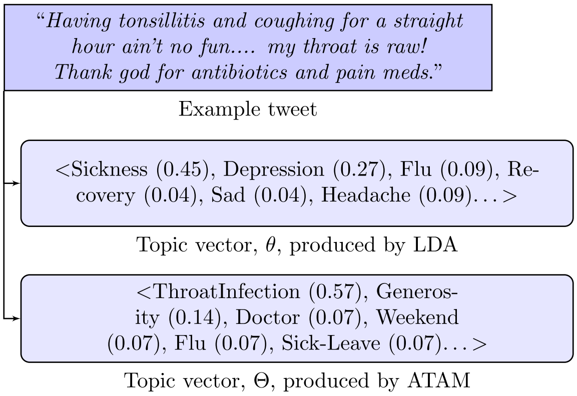

"Having tonsillitis and coughing for a straight hour ain't no fun.... my throat is raw! Thank god for antibiotics and pain meds."
Example tweet
<Sickness (0.45), Depression (0.27), Flu (0.09), Recovery (0.04), Sad (0.04), Headache (0.09)...>
Topic vector, θ, produced by LDA
<ThroatInfection (0.57), Generosity (0.14), Doctor (0.07), Weekend (0.07), Flu (0.07), Sick-Leave (0.07)...>
Topic vector, Θ, produced by ATAM

Fig. 2. LDA versus ATAM: Topic distributions for an example tweet.

While LDA is successful at uncovering generic topics, such as "healthcare", "obesity", "substance abuse", infrequent topics that may be related to specific subjects, such as "tobacco use", pose a challenge to LDA. Furthermore, for an excessively frequent topic, such as "weight loss", LDA adds noise, in the form of words such as "gardening", "oils", "anti-ageing", "muscle gain", that are not related to the topic [4], [12]. LDA is therefore not a good choice for modeling latent topics in health-related data.

2.2.2 Uncovering Health Topics with ATAM

The probabilistic *Ailment Topic Aspect Model* was designed specifically to uncover latent health-related topics in a collection of tweets [4]. The proposed method achieves remarkable improvements over LDA. Its novelty is that it distinguishes *background words* such as "home" and "watching TV" from *health-related words* such as "hurts" and "allergy". For each document, these health-related words are considered to correspond to a unique ailment such as "obesity","insomnia" or "injuries". The word could be associated to the ailment as its symptom (e.g., the word "weight" is clearly a symptom related to the ailment "obesity"), a treatment (the word "diet" is clearly a symptom related to the ailment "obesity") or a general word (the word "dentist" is not a background word and belongs to the vocabulary of the ailment "dental" but is neither a symptom nor a treatment).

Fig. 3 summarizes the process of ATAM. When generating a document (tweet), one first associates to it an ailment such as "allergy", "insomnia" or "injury". Thereafter, the document is generated word by word. Using two auxiliary random variables ℓ and x, one chooses if the word is a background word or a general-purpose word ($\ell = 0$ or $\ell = 1, x = 0$). The word can then be drawn from a vocabulary distribution common to the whole corpus (case $\ell = 0$) or generated from an underlying Dirichlet distribution topic z (case $\ell = 1, x = 0$). When the word is related to health (case $\ell = 1, x = 1$), another random variable y enables to choose if this word is a aspect-neutral (case $y = 0$), a symptom (case $y = 1$) or a treatment (case $y = 2$). *Words are hence drawn depending on the ailment a which has been associated to the document.* Fig. 2 shows the topic distribution vectors for a sample tweet. Note the stronger relevance to health-related matters in the ATAM vector compared to its LDA counterpart. Note that each tweet is associated with one single ailment and one topic distribution whereas at the corpus level, we have one ailment distribution η. A topic distribution can also be associated to a corpus by concatenating all tweets in that corpus and considering them as a single document.

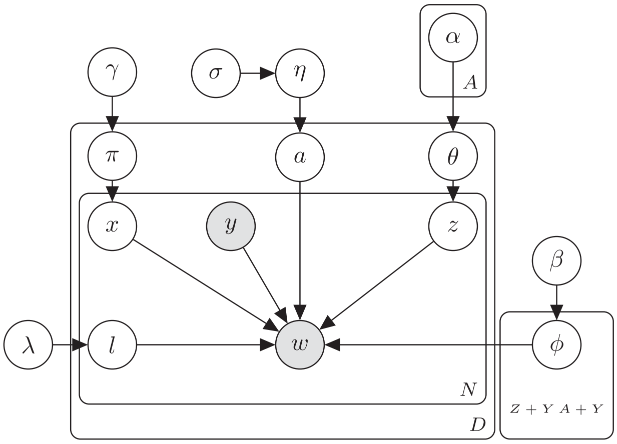


Fig. 3. Ailment topic aspect model.

2.3 Transition Prediction and Detection Problems

Using ATAM over a region and a period, we associate with a given aggregated document D_g^t, an aggregated topic distribution, Θ_g^t which is a mix of general-purpose and health-related topics. More precisely, this topic distribution has the following components :

- η_g^t : Distribution over ailments of the corpus of tweets used to build the aggregated document D_g^t
- θ_g^t : Distribution over general topics in the document D_g^t, considered as a single document

We then define $\Theta_g^t = \big((1-\pi)\theta_g^t \quad \pi\eta_g^t\big)$ where π is the proportion of non-health words related to some aspect of an ailment. We show Θ for an example tweet in Fig. 4.

Transition Prediction Problem. Given our documents $D_g^{t_{i-1}}$ formed by tweets originating from a region $g \in G$ during time period t_{i-1}, predict the ailment distribution η_g^t of documents in $D_g^{t_i}$, corresponding to posts from g in period t_i from the topic distribution $\Theta_g^{t_{i-1}}$ of document $D_g^{t_{i-1}}$ corresponding to posts from g during period t_{i-1}. To solve this problem, we develop a model TM–ATAM in Section 3.

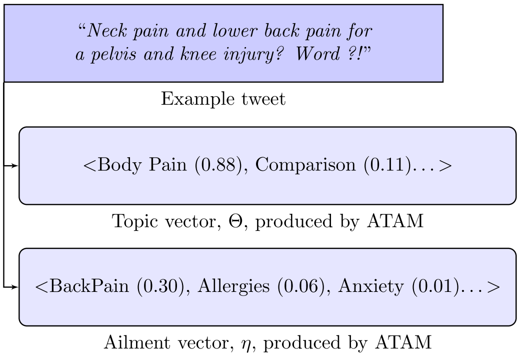
Transition Detection Problem. Our second problem aims to detect health-topic transitions, that is `change-points`in the ailment distribution vector η_g^t. More precisely, considering the evolution of the content of our tweets on the whole period, that is the successive documents $D_g^{t_1}, \ldots, D_g^{t_N}$, we want to detect when the ailment distribution η_g^t is changing. A general formulation of this problem would return the k `change-points`with the highest distance between two successive ailment distributions. To solve this problem, we develop a model T–ATAM in Section 4.

3 A First Model for Ailment Transitions: TM–ATAM

Our first objective is to model ailment transitions, that is potential change in time of the health topical content of our tweets. We do so by introducing a new model, TM–ATAM that we define in this section. This model is derived from TM–LDA that we describe first.

3.1 General-Purpose Topic Modeling over Time with TM–LDA

In order to take into account the evolution of the underlying topics of a dynamic collection of documents with time (e.g., a microblog or a facebook page), Wang et al. (2012) introduced a modified version of the LDA model, TM–LDA [7]. In [7], TM–LDA was introduced to extend LDA with modeling topic evolution of dynamic collection of documents over time. Topic distribution of the ith document, θ_i is assumed to depend linearly on the topic distribution of the previous document, θ_{i-1}. At the heart of the algorithm lies the following equation.



"Neck pain and lower back pain for a pelvis and knee injury? Word ?!"
Example tweet
<Body Pain (0.88), Comparison (0.11)... >
Topic vector, Θ, produced by ATAM
<BackPain (0.30), Allergies (0.06), Anxiety (0.01)... >
Ailment vector, η, produced by ATAM

Fig. 4. θ is predicted by ATAM on an example tweet, η predicted by ATAM on all posts containing the example tweet.

$$\theta_i \approx \frac{\theta_{i-1}.M}{\|\theta_{i-1}.M\|_{\ell_1}}, \tag{1}$$

where M is a $k \times k$ matrix, called the transition matrix, and k is the number of topics. To obtain the transition matrix, the authors propose to solve the following least squares problem ($\|\cdot\|_F$ denotes the Frobenius norm and X denotes the search space)

$$M = \operatorname*{argmin}_X \|A.X - B\|_F, \tag{2}$$

where A and B are as specified below.

$$A = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{i-1} \end{pmatrix},\ B = \begin{pmatrix} \theta_2 \\ \vdots \\ \theta_i \end{pmatrix}, \tag{3}$$

However, while being quite elegant in modeling general purpose topics TM–LDA is not specialized to capture `health` transitions over time.

3.2 TM–ATAM: Modeling Health Topics Transition over Time

While ATAM is effective at modeling health-related topics, it is not designed to model topic transitions over time. We hence propose TM–ATAM that builds on top of ATAM and TM–LDA. TM–ATAM computes the aggregate topic distribution, Θ_g^t, of a set of documents D_g^t and learns the evolution with time of the vector Θ_g^t.

TM–ATAM, at its heart, solves following equation.

$$A_g^t \approx A_g^{t-1}.M, \tag{4}$$

where

$$A_g^{t-1} = \begin{pmatrix} \Theta_g^1 \\ \vdots \\ \Theta_g^{t-1} \end{pmatrix},\ A_g^t = \begin{pmatrix} \Theta_g^2 \\ \vdots \\ \Theta_g^t \end{pmatrix}. \tag{5}$$

M is an unknown transition matrix which is obtained by solving the following least square problem

$$\min_M \| A_g^t - A_g^{t-1}.M \|_F,$$

TM–ATAM thus learns a transition matrix which is used to model health topics. It will be our main tool in our transition learning task.

3.3 Learning Transitions with TM-ATAM

We now focus on the transition learning problem and explain how we solve it using TM-ATAM. Algorithm 1 contains the steps of our solution. It has two main parts: `change-pointdetection` and `transition learning`. We first describe how `change-points` are detected and then go on to show how this last step will be used to predict the evolution of ailment-topic distribution over time within `homogeneous timeperiods`as well as health topical transitions.

3.3.1 Change-Point Detection with TM-ATAM

For each region $g \in \mathcal{G}$ (Line 1), we first run ATAM over the full time period D_g (Line 2). Next for each period $t \in \mathcal{T}$ (Line 3), we use the output of ATAM over D_g to generate Θ_g^t and deduce the ailment distribution η_g^t since we shall focus only on `health-transitions` (Lines 4–12). Next, we examine the distance between consecutive distributions η_g^{t-1} and η_g^t of the region g to identify the most significant health-related `change-point`, t_c (Line 14). We treat the choice of distance measure m as black box, which could be *Bhattacharya Distance*[1] or *Cosine Similarity*.[2] The time period t_c is termed as the `change-point` for region g. The entire span of time, $[t_1 \; t_{|\mathcal{T}|}]$, is divided into two intervals, *pre*, consisting of all time periods prior to the `change-point` (Line 15), and *post*, consisting of all time periods after the `change-point` (Line 16).

We term these intervals as `homogeneous timeperiods` w.r.t ailments being discussed in Twitter. Qualitatively, a `homogeneous timeperiod` is a time interval (collection of consecutive time periods) during which the tweets originating from the region are homogeneous in terms of ailment topics. The `change-point`characterizes a significant change point in the evolution of ailments. We posit that such change points exist. These change points in ailment topic discussions may be caused by onset of the disease or some other external factors. Nevertheless, they are the interesting points for analyzing purposes. Such analysis may lead to various insights into onset of diseases. Onset of disease is usually affected by several factors, such as weather, which may cause a sudden onslaught of ailments different from the ones that were in circulation previously. The pervasive nature of communicable diseases is also a contributing factor. Note that the results in Fig. 5 support our assumption, where we show the difference between ailment distributions of consecutive months for 3 different regions Kuala Lumpur (a city in Indonesia), Oklahoma (a state in the USA), and Bristol (a city in the UK). In Fig. 5, dissimilarity on Y-axis denotes the Bhattacharya distance between ailments distributions (inferred by TM–ATAM) of consecutive months for the 3 regions. The sharp peaks obtained validate the existence of time intervals that are homogeneous w.r.t. ailments.

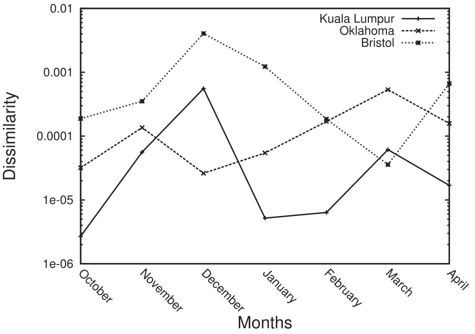


Fig. 5. Topic transitions over time.

3.3.2 Ailment Prediction and Transition Learning

The key idea in TM–ATAM is after these `change-points` detection, is to predict evolution of health topics within each `homogeneous timeperiod`. This is a fresh departure from existing solutions that operate in a `homogeneous time period`-agnostic fashion. By definition, a `homogeneous timeperiod` is (nearly) homogeneous in terms of ailments. In other words, the ailments evolve in a smooth fashion within a `homogeneous timeperiod` and change abruptly across `homogeneous time period` boundaries. In this study, we set k to 1 and find a single `change-point` for each region g. While this may not be true for all regions, we obtain significant improvement in terms of prediction accuracy over the state-of-the-art with just a single boundary.

Algorithm 1. TM-ATAM: `change-point` Detection and Training Ailment Distribution Predictor

```
1: for all $g \in G$ do
2:   Run ATAM on $D_g$
3:   for all $t \in \mathcal{T}$ do:
4:     for all $z \in \mathcal{Z}$ do:
5:       $\Theta_g^t[z] \leftarrow 0$
6:     end for
7:     for all $d \in D_g^t$ do:
8:       for all $w \in d$ do:
9:         $z \leftarrow topic(w)$
10:        $\Theta_g^t[z] \leftarrow \Theta_g^t[z] + \frac{1}{|d| \times |D_g^t|}$
11:      end for
12:    end for
13:  end for
14:  $t_c = \text{argmax} \; m(\eta_g^{t-1}, \; \eta_g^t)$
15:  $pre = [t_1^t \; , \; t_{c-1}]$
16:  $post = [t_c \; , \; t_{|\mathcal{T}|}]$
17:  for all $s \in \{pre, \; post\}$ do:
18:    Run ATAM on the period $s$ and infer for each time-period of the homogeneous time period $s$, the vectors $\Theta_g^t$ which includes the ailment vector $\eta_g^t$ for each period of the season and then form its aggregation: $A_g^{t_s}$
19:    $A_g^{t_s} \approx A_g^{t_s-1}.M_s$
20:    Estimate the matrix transition related to season $s$, $M_s = (A_g^{t_s-1} A_g^{t_s-1})^{-1} A_g^{t_s-1} A_g^{t_s}$
21:  end for
22: end for
```

We outline in Lines 17–21 of Algorithm 1 the steps undertaken. We use $\mathcal{Z}$ to refer to the set of all health and non-health topics. The key step is the estimation of the unknown transition matrix for each season s (the *pre*-`change-point` season and the *post*-`change-point` one), that can be used to

[1] https://en.wikipedia.org/wiki/Bhattacharyya_distance
[2] https://en.wikipedia.org/wiki/Cosine_similarity

predict the content of our set of tweets. The *pre-*`change-point` season and the *post-*`change-point` are the time intervals on which we run our tests. We also use it further to learn transitions as explained in Section 5.3.

To make easier comparison between regions we focus on the case where we estimate only one change point. We emphasize that one can easily modify the algorithm to estimate several change points. One has only to replace the estimation of the time t corresponding to the maximal distance between two consecutive vectors η_g^t and η_g^{t+1} with the k times corresponding to the kth top distances between consecutive vectors η_g^t, η_g^{t+1} if we want to estimate k change points. Another possible alternative is to set a threshold common to all regions and to keep times t such that the distance between η_g^t and η_g^{t+1} is above the threshold.

4 An Alternative Model : Time-Aware Ailment Topic Aspect Model (T–ATAM)

TM–ATAM assumes that there is a common linear relation between all the aggregate topic distributions at a given period t and the one at the period just before. TM–ATAM fails to perform optimally when operated in regions where there are no substantial transitions in health topics, as also shown in Section 5. In particular, TM–ATAM does not take into account the potential seasonality effect, which maybe very different according to the disease of interest. Also, in TM–ATAM, we need to do post processing in order to come up with `homogeneous time periods`, with respect to health-topics discussed in tweets.

We now introduce a second *time-aware* model, coined the term, T–ATAM, where the timestamp t of each tweet is considered as a random variable, depending on the ailment associated to the post. Note that since time is now a random variable, we shall now aggregate our tweets only by region and run our new model on the different sets of posts $\mathcal{P}_g$ to have a deep understanding on the time evolution of the health-related content of our set of tweets. It is highly expected there is a strong dependence of the content of our posts with respect to time but also to the ailment of interest. For example, tweets associated to flu are probably mainly concentrated in winter and that those associated to sunburns mainly posted in summer whereas some ailments maybe non seasonal ones. T–ATAM learns `homogeneous time periods` by itself and no post-processing is needed in order to come up with `change-point` in ailments being discussed in tweets. This is because, in generative process, time-stamp is generated conditioned on the ailment assigned to the tweet. Therefore, ailments learned are already time (season)-aware after the model has run its course. Fig. 6 shows the graphical representation of T–ATAM. This model adds three extra random variables to the graphical model of ATAM(Fig. 3): t, ψ and μ.

Let us now describe the generative process of T–ATAM. Basically, the generative process of each document is exactly the same as that of ATAM, that we have already described in Section 2.2, except that now a time stamp is generated for each document depending on the ailment associated to the considered tweet (steps (3) and (5-III) of the generative process described below). Time stamp is an observed random variable. To generate the time stamp associated to a given tweet, we first generate one time-distribution per ailment $\{\psi_a, a \in A\}$ (step (3)). Thereafter, depending on the ailment a associated to the concerned document, we generate its time stamp according to a multinomial distribution with parameter ψ_a (step (5-III)). Additionally, ψ_a is drawn from a Dirichlet distribution parameterized by a vector μ_a, specific to each ailment. This is intuitive because different ailments have their own specific chance of breaking out at different time periods. We summarize the generative process of our new model T–ATAM just below :

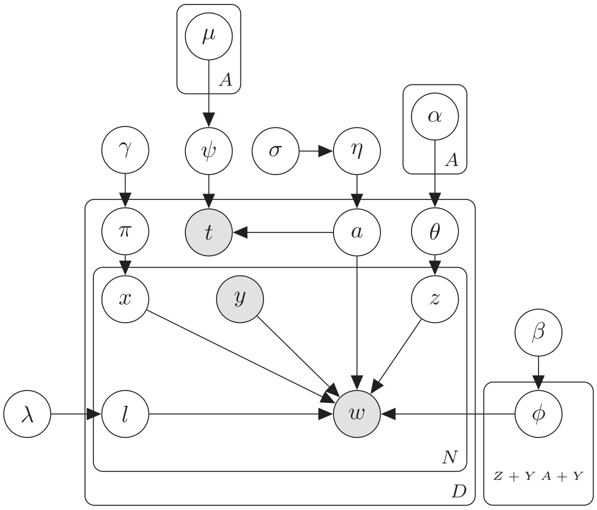


Fig. 6. Time-aware ailment topic aspect model.

Generative Process.

```
(1)  Set the background switching binomial $\lambda$
(2)  Draw an ailment distribution $\eta \sim Dir(\sigma)$
(3)  Draw A multinomials $\psi_A \sim Dir(\mu)$
(4)  Draw word multinomials $\phi \sim Dir(\beta)$ for the topic,
     ailment, and background distributions
(5)  For each message $1 \ \leq m \leq D$
      (I)    Draw a switching distribution $\pi \sim Beta(\gamma_0, \gamma_1)$
      (II)   Draw an ailment $a \sim Mult(\eta)$
      (III)  Draw a time stamp $t \sim Mult(\psi_a)$
      (IV)   Draw a topic distribution $\theta \sim Dir(\alpha_a)$
      (V)    For each word $w_i \in N_m$
              (A)  Draw aspect $y_i \in \{0, 1, 2\}$(observed)
              (B)  Draw background switcher $l \in \{0, 1\} \sim Bi(\lambda)$
              (C)  if l == 0:
                     (i)  Draw $w_i \sim Mult(\phi_{B,y})$(a background)
              (D)  Else:
                     (i)   Draw $x_i \in \{0, 1\} \sim Bi(\pi)$
                     (ii)  If $x_i == 0$ :(Draw word from topic z)
                           (a) Draw topic $z_i \sim Mult(\theta)$
                           (b) Draw $w_i \sim Mult(\phi_z)$
                     (iii) Else:(draw word from ailment a
                           aspect y)
                           (a) Draw $w_i \sim Mult(\phi_{a,y})$
```

It should be noted that token level sampling for $\mathbf{y}$, $\mathbf{x}$ and l for T–ATAM stays the same as ATAM. Document-level sampling for ailment a for T–ATAM changes and is given by following equation:

$$
\begin{aligned}
&P(a_m|\boldsymbol{a_{-m}}, \mathbf{w}, \mathbf{t}, \mathbf{y}, \mathbf{x}, l) \\
&\qquad \propto P(a_m|\boldsymbol{a_{-m}}) P(t_m|\boldsymbol{t_{-m}}, \mathbf{a}, \mu) \\
&\qquad \prod_{n}^{N_m} p(w_{m,n}|\mathbf{a}, \boldsymbol{w_{-(m,n)}}, \mathbf{y}, \mathbf{x}, l).
\end{aligned} \tag{6}
$$

TABLE 2
Dataset Statistics

collection period (days)	235
#tweets	1,360,705,803
#tweets (health-related)	698,212
#tweets (health-related+geolocated)	569,408

Factor which is to be multiplied with existing factors at document level sampling of ATAM for posterior distribution of ailment a: $P(t_m|\mathbf{t_{-m}}, \mathbf{a}, \mu)$

$$P(t_m|\mathbf{t_{-m}}, \mathbf{a}, \mu) = \frac{n^{i,t_m}_{-m} + \mu}{n^{i}_{-m} + T\mu}, \tag{7}$$

Superscript i indexes over ailments. In particular, n^i denotes number of times an ailment occurs in the corpus and n^{i,t_m} denotes number of times an ailment occurs with a time stamp t_m.

As proved in the experimental results, this new model is much more accurate than the previous one both in terms of perplexity measure and in agreement with ground truth. This model also beats ATAM in many of the regions where there is no substantial health topic transitions. Note that in the case of T–ATAM, we can infer `change-point` and transitions as in the case of TM–ATAM.

5 Experimental Evaluation

We conduct experiments to evaluate the performance of TM–ATAM and T–ATAM on real world data. Section 5.1 describes the experimental setup including the datasets and test-bench. In Section 5.2, we compare TM–ATAM and T–ATAM against state-of-the-art approaches. That is followed by a detailed study of the behavior of TM–ATAM in Section 5.4.1 and a qualitative analysis of TM–ATAM's results in Section 5.3. Then in Section 5.4.2, the effect of changing parameters in T–ATAM is studied. Finally, we study the correlations between T–ATAM's results with CDC data and Google Flu Trends in Section 5.5 for the influenza rates in US. Finally, we highlight the key insights drawn from our experiments in Section 5.6.

5.1 Setup

5.1.1 Data

We employ Twitter's Streaming API to collect tweets between 2014-Oct-8 and 2015-May-31. We use the *Decahose Stream*[3] which gives a 10 percent random sample of the total tweets generated each day. The collected tweets were subjected to two pre-processing steps.

Filtering Health-Related Tweets. We removed retweets and tweets containing URLs; they were almost always false positives(e.g., news articles about the flu, rather than messages about a user's health.) Since our interest lies in public health discourse on social media, we only keep tweets containing one of 20,000 health-related keywords obtained from wrongdiagnosis.com. This website lists detailed information about ailments, symptoms and treatments. Resulting tweets were given to an SVM classifier [13] with linear kernel and uni-gram, bi-gram and tri-gram word features. To train the classifier, a modest-sized sample of the original corpus was annotated through crowdsourcing efforts where annotators were asked to label 5,128 tweets. The precision and recall of the employed classifier are 0.85 and 0.44. In our case, we focused on high precision as high quality health tweets is a pre-requisite for both TM–ATAM and T–ATAM to function efficiently. Table 2 shows that out of the 1.36 B tweets we collected, 698 K were health-related.

TABLE 3
Default Parameters

Term	Description	Value
$\mathcal{G}$	Geographic granularity	*states*
$\mathcal{T}$	Temporal granularity	*months*
m	Distance measure	*Bhattacharya*

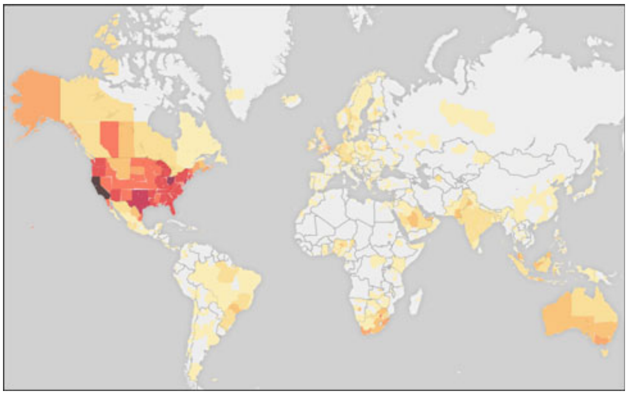


Fig. 7. Heatmap over collected health tweets. A major fraction of the tweets originate from various states in the US.

Geolocation. The ability to operate seamlessly at varying geographic resolutions mandates that the exact location of each tweet be known to TM–ATAM and T–ATAM. Twitter affords its users the option to share their geolocation. It has been shown that a very small number of Twitter users choose to share their location. While this artefact results in significant reduction in the number of tweets, in absolute terms, we retain more than half a million tweets (569 K as indicated in Table 2). In Fig. 7, we present a heatmap that shows the geographic spread of these tweets. The darker the color, the higher the number of tweets. The top-10 regions (at spatial granularity *state*) with the highest number of health tweets lie exclusively in the US. We choose to work with geographic granularity of `states`, temporal granularity of `months` and distance measure of `Bhattacharyya` as the default parameters in subsequent experiments as shown in Table 3.

5.1.2 Test-Bench

We run our experiments on a 32 core Intel Xeon @ 2.6 Ghz CPU (with 20 MB cache per core) system with 256 Gig RAM running Debian GNU/Linux 7.9 (wheezy) operating system. All subsequently discussed components were implemented in Java 1.8.0_60.

5.2 Comparison between Models

5.2.1 Perplexity Measure

We use *perplexity*, an empirical measure often used in NLP.[4] Perplexity of a language model measures how accurately the model can explain previously unseen data/documents.

[3] https://dev.Twitter.com/streaming/overview

[4] https://en.wikipedia.org/wiki/Perplexity

Given a language model l and a document d, perplexity is defined as below.

$$Perplexity(l) = 2^{-\sum_{w_i \in d} \log p_l(w_i)}, \tag{8}$$

This formula of perplexity for a document d can be converted to a formula of perplexity for a set of documents D_g^t as follows:

$$Perplexity\_D_g^t(l) = 2^{-\sum_{w_i \in d} \log \frac{\sum_{d \in D_g^t} p_l(w_i)}{|D_g^t|}}. \tag{9}$$

It denotes the perplexity of language model l on a document-set at geo-granularity g and temporal granularity t. Higher probability of words that occur in unseen documents results in lower perplexity and is hence better. Here, $p_l(w_i)$ is the probability of occurrence of word w_i as estimated by the language model l in the document set. Previously unseen words can result in infinite perplexity. We use add-one smoothing to overcome this fact.[5] $p_l(w_i)$, probability of word, for any document set is calculated using

$$p_l(w_i) = \sum_z P(w|z)P(z) = \sum_z \frac{n(z,w)}{n(z)} P(z). \tag{10}$$

Here $P(z)$ is the probability of topic. The key point in Equation (10), is that, first term $P(w|z)$ does not change and only the second term $P(z)$ changes with topic models. This is because we are in scope of those topic models where topic probabilities (and not the word probabilities themselves within each topic) change with time as ours is not the domain of dynamic topic models where word probabilities per topic change with time. Having computed $P(w)$, we can compute perplexity using the Formula (9). We compute perplexity of TM–ATAM and T–ATAM and then compare perplexity of TM–ATAM and T–ATAM with that of TM–LDA on the same test document set D_g^t as explained in the next section. All TM–ATAM, T–ATAM and TM–LDA are treated as language models as all give out probability of each word for any document-set D_g^t.

5.2.2 Comparing TM–ATAM and T–ATAM with TM–LDA and ATAM

We present results on the comparison of prediction accuracy of TM–ATAM and T–ATAM against ATAM and TM–LDA. Recall that the terms change-pointand homogeneous time period refer to the point in time at which discourse density of ailments changes substantially, and the time period before and after that point, respectively.

TM–ATAM. We first divide the postings of each region into two .homogeneous time periods as inferred by change-pointt_c. We then divide each homogeneous time periodinto train and test set as follows. *Pre-*homogeneous time periodis divided into train ($[t_1\ , t_{c-3}]$) and test ($[t_{c-2}, t_{c-1}]$) set. *Post-*homogeneous time periodis divided into train ($[t_{c+1}, t_{|\mathcal{T}|-2}]$) and a test ($[t_{|\mathcal{T}|-1}, t_{|\mathcal{T}|}]$) set. For example, if t_c for a region is between January to February, then train set and test set of *pre-*homogeneous time periodare tweet posts of the months in the set [October, December] and [January, February] respectively. Train and test set of *post-*homogeneous time periodare tweet posts of months in the set [March, April] and [May, June] respectively. We obtain 69 homogeneous time periods for 66 regions. ATAM is *re-run* over train set of each homogeneous time period. It should be noted that though computing change-pointt_c required access to full dataset, perplexity calculations are done within each homogeneous time periodand clear distinction is made between train and test set while computing it. We then model a transition matrix M_{tmatam} on the training data of each homogeneous time period as described in Section 3.3. For each tweet p of the first month in the test set (t_{c-2} for the *pre* homogeneous time periodand $t_{|\mathcal{T}|-1}$ for the *post* homogeneous time period), we compute the probability of "health topic" z using the Formulas:

$$P(z|t_{c-2}) = \frac{\sum_{p\ \in\ t_{c-2}} P(z|p\ for\ t_{c-2})}{\#p \in \quad t_{c-2}} \tag{11}$$

$$P(z|t_{|\mathcal{T}|-1}) = \frac{\sum_{p\ \in\ t_{|\mathcal{T}|-1}} P(z|p\ for\ t_{|\mathcal{T}|-1})}{\#p \in \ t_{|\mathcal{T}|-1}}. \tag{12}$$

Here $P(z|p)$ is computed simply by (w is the word of tweet p)

$$P(z|p) = \sum_w P(z|w)P(w|p) = \sum_w \frac{n(z,w)}{n(w)} P(w|p). \tag{13}$$

Here, values for $n(z,w), n(w)$ are taken from ATAM run on the training months. If we encounter an unseen word, we use add-one smoothing to avoid $P(z|w)$ to shoot to infinity and hence perplexity to shoot to infinity. $P(w|p)$ is simply the number of times word w occurs in the tweet p divided by the total number of words in the tweet p. We then predict the future probability of each topic in the second month of the test data ($P(z|t_{c-1})$ for *pre* homogeneous time period and $P(z|t_{|\mathcal{T}|})$ for the *post* homogeneous time period) using the corresponding transition matrix M_{tmatam}. The perplexity of TM–ATAM can now be computed against the words of the tweets of second test month (t_{c-1} and $t_{|\mathcal{T}|}$) using the Formula 9. This gives 69 values of perplexity, one for each homogeneous time period of each region. We compare our results with following competitors:

ATAM. Underlying assumption of atam is that topics stay static with respect to time. In order to assert the fact that health topics transit from one to another, we compare performance of TM–ATAM with ATAM by computing perplexity of ATAM on words of first month of test set and not predicting any topic distribution using transition matrix. For each tweet p of the first month in the test set (t_{c-2} for the *pre* change-point and $t_{|\mathcal{T}|-1}$ for the *post* change-point), we compute the probability of "health topic" z using the Formulas (11), (12), and (13). It should be noted that in this case *we do not model a transition matrix to predict probability of topics for second month of test set*. Hence, this denotes model where health topics stay *static*. We can then compute perplexity of ATAM against words of actual tweets of the second months of test month (t_{c-1} and $t_{|\mathcal{T}|}$). As shown in Fig. 8, TM–ATAM beats ATAM in all US active regions. In Non-US active regions, the performance of TM–ATAM gets affected due to no substantial change in health topics with time. That means there is no substantial change in health topics discussed in those tweets. This may mean limitation of Twitter and sparsity of tweets in these regions but not necessarily a limitation of our model. In fact, our model could be applied to other microblogs such as Reddit or Google search queries. This also means that these are the regions where many diseases are prevalent and discussed all over the year. As a result, there are no transitions of health topics

[5] https://en.wikipedia.org/wiki/Additive_smoothing

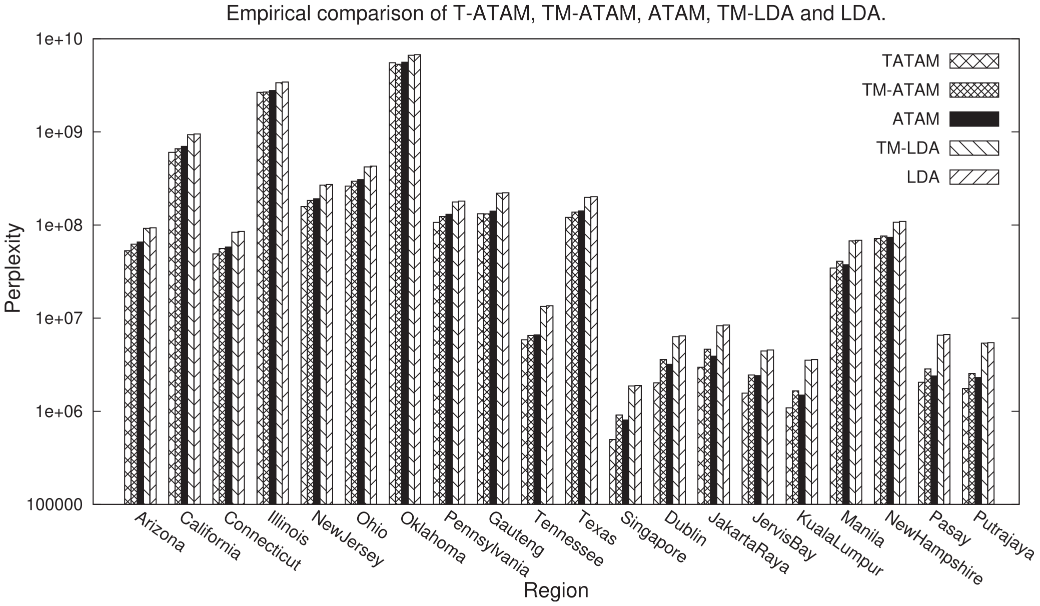


Fig. 8. Perplexity comparison of T–ATAM, TM–ATAM, TM–LDA, and ATAM for top 20 social media active regions.

and since TM-ATAM is meant to model transitions of health topics, its performance is affected negatively.

T–ATAM. In order to assert the fact that considering time as a random variable for T–ATAM is more efficient, we compare T–ATAM with ATAM and TM–ATAM by computing perplexity of T–ATAM on words of second month of test set. After getting the `homogeneous time periods` for each region and dividing each `homogeneous time period` into train and test set, T–ATAM is run over train set of each `homogeneous time period`. Then, for each tweet p of the second month of the test set (t_{c-1} and $t_{|\mathcal{T}|}$), we compute the probability $P(z|t_{c-1})$ and $P(z|t_{|\mathcal{T}|})$ using the Formulas (11), (12), and (13). In case of T–ATAM, we do not model any transition matrix and directly compute $P(z)$ on second month as model itself learned ailments using the knowledge of time in-built in the model. This tests T–ATAM's capability using time as a random variable model for coming up with ailment distributions which are actual representative of words tweeted about in the time of interest. It should also be noted that $n(z,w), n(w)$ for Formula (13) are calculated from T–ATAM run over train set. Now, perplexity can be calculated against the words of the tweets of second test month (t_{c-1} and $t_{|\mathcal{T}|}$). As shown in Fig. 8, T–ATAM beats both ATAM and TM–ATAM in all active regions. Especially, for Non-US active regions, while, TM–ATAM's performance gets affected, T–ATAM shows a good ability to predict future tweets based on its better capability to incorporate knowledge of time within the model itself. T–ATAM overcomes the shortcomings of no-substantial change in health topics as diseases inferred from health tweets are time-aware. As such, health topics inferred by T–ATAM are not limited to short-lived topics but also cover topics that are regularly discussed on Twitter. This could explain why T–ATAM performs better even in regions where health topics are stable over time.

Predicted TM–LDA. Each region can be viewed as a `virtual user` and the transition matrix M_{tmlda} of TM–LDA is trained by solving least squares problem in the following manner. We merge the training data of each `homogeneous time period` in each region and train a transition matrix of TM–LDA. So, training data for TM–LDA is the same as that of TM–ATAM and T–ATAM: ($[t_1\ , t_{c-3}]$) for the *pre* `homogeneous time period` and ($[t_{c+1}, t_{|\mathcal{T}|-2}]$) for the *post* `homogeneous time period`. For each tweet p of the first month of the test months (t_{c-2} and $t_{|\mathcal{T}|-1}$), we compute the probability $P(z|t_{c-2})$ and $P(z|t_{|\mathcal{T}|-1})$ using LDA trained on merged training data (Formulas (11), (12), and (13). We then predict the future probability of each topic in following month (t_{c-1} and $t_{|\mathcal{T}|}$) using corresponding M_{tmlda}. We can then compute the perplexity of TM–LDA against words of actual tweets of the test months (t_{c-1} and $t_{|\mathcal{T}|}$) using Formula (9).

We take average over both `homogeneous time periods` (*pre* and *post*) and get a perplexity value for each region. Fig. 8 shows that TM–ATAM and T–ATAM consistently beats TM–LDA and ATAM in predicting future health topics on the test month by computing lower perplexity on the words of the tweets of the test month in all social media active states.

5.3 Qualitative Analysis of TM–ATAM

5.3.1 Change Points

The central idea in TM–ATAM is to identify `homogeneous time periods`, i.e., time intervals that exhibit homogeneous ailment distributions, as well as transitions between them. A natural question that emerges is *how and why* ailments differ across `change-point` boundaries. In Fig. 9 we show the sharpest change point, representing the strongest transition, for the non-US regions respectively. Those points can be explained with weather changes in those regions. Jervis Bay can be explained by an increase in rainfall. Dublin sees its lowest temperature in the November

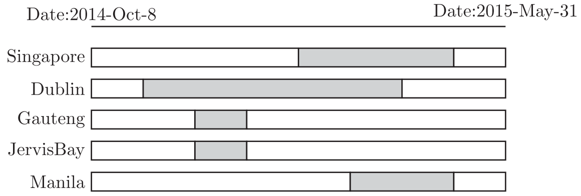


Fig. 9. Monthly `change-point` boundaries for top-10 active non-U.S. regions.

TABLE 4
M_{full} Transitions for California (Threshold: 0.815)

Transition Type	From Topic	To Topic	Weight
One-Way Transitions	smoking/junkies/drugs/cigarettes	respiratory diseases	2.70
	depression/complaining/cursing/slangs/self-pity	joint pains/body pains	3.25

period. Singapore and Manila have very similar weather conditions and exhibit the same change point. A deeper look at these transitions would provide more insights.

5.3.2 Topic Transitions

Entry m_{ij} in the transition parameter matrix M produced by TM–ATAM, shows the degree that health topic z_i will contribute to health topic z_j in the following predicted ailment distribution. We analyze 3 kinds of transition matrices corresponding to our setting: *intra*-`homogeneous time period`: M_{pre}, M_{post} and *inter*-`homogeneous time period`: M_{full}. Let mean be denoted by μ and standard deviation be denoted by σ for further discussion. We adapt the threshold used in [7] to our settings:

$$Threshold = \mu + 2 \times \sigma_{non-diagonal}, \tag{14}$$

Here μ is the mean of the corresponding transition matrix. $\sigma_{non-diagonal}$ is the standard deviation of non-diagonal entries. We choose this threshold because 95.45 percent of the values lie within two standard deviations of the mean.[6] We identify three kinds of interesting transitions based on the threshold defined in [7]:

- Self transitions: Diagonal entries above threshold
- Symmetric Transitions: Both m_{ij} and m_{ji} is higher than threshold
- One-Way Transitions: Only one of m_{ij} and m_{ji} is higher than threshold

Table 4 lists interesting one-way health topic transitions observed in California for the full time period. Self-Transitions are hard to find in *full* time periods as topics change a lot between *`homogeneous time periods`*. Mean of diagonal entries is the quantification of how stable the transitions are and standard deviation of non-diagonal entries is the quantification of how much the topics fluctuate in the given time granularity.

Further, we analyze M_{pre}, M_{post} and M_{full} of Kuala Lumpur. Various statistics are summarized in Table 5. $\mu_{diagonal}$ is higher for both M_{pre} and M_{post} than M_{full}. $\sigma_{non-diagonal}$ is higher for M_{full} than both *intra*-`homogeneous time period` transition matrices. These statistics go on to show that health topics do not drastically change and are coherent within the same *`homogeneous time period`* and transform into one another a lot across the *`homogeneous time periods`*. This further re-instates this fact that it is more sensible to model topic transition matrices within the same *`homogeneous time period`* and update them once the *`homogeneous time period`* has ended and *`change-point`*is encountered. Further, we analyze the interesting self transitions of *intra*-`homogeneous time period` (M_{pre} and M_{post}) and one-way transitions of M_{full}. Further, we found interesting self-transitions and one-way transitions in Kuala Lumpur and Arizona.

5.4 Effect of Parameters

5.4.1 TM–ATAM: Effect of Parameters

Geographic Granularity. We examine two different choices for the geographic granularity i.e., *states* and *counties* which correspond to first and second level administrative divisions[7]. While TM–ATAM can be instantiated at varying granularities of space, learning accurate ailment distributions requires a certain minimum number of tweets. Selecting larger than optimal sized regions would introduce errors into the prediction algorithm. Choice of geographic granularity is non-trivial. Predicted perplexity in counties is lower, hence better, than perplexity at the level of states as shown in Fig. 10. This is due to the fact that tweets from smaller regions show less diversity in topics.

Temporal Granularity. We examine two different temporal granularities, *months* and *weeks*. Analogous to geographic granularity, choice of temporal granularity should not be too fine or too coarse. We show performance of TM–ATAM on time granularities in Figs. 11 and 12. This is also attributed to the fact that prediction of health topics in smaller temporal granularity is more accurate as health topics do not transform by a substantial amount in shorter periods.

5.4.2 T–ATAM: Effect of Parameters

Geographic Granularity. We choose to compare T–ATAM's performance in two different cases: The first when it does not consider any geographic granularity (Global) and the second case when T–ATAM is instantiated at the first level administrative division which is *states*. While T–ATAM can be instantiated at varying space granularities, learning accurate ailment distributions requires legitimate initialization of geographic granularity and a certain minimum number of tweets.

Results in Fig. 13 show that operating in smaller geographic granularity yields smaller perplexity and hence, better prediction for health topics. We attribute this result to the fact that tweets from finer geographic granularity have less diversity in topics. Also, per-state breakdown for detecting health topics is better since people in the same region are exposed to the same weather conditions and are more likely to have similar eating habits and hence develop similar diseases.

Temporal Granularity. To study the effect of time granularity on T–ATAM's performance, we run it using different time granularities: *weeks* and *months*. The results are shown in Figs. 14 and 15 for US and Non-US active regions. Clearly, the smaller the time granularity, the lower perplexity we obtain. As in the case of TM–ATAM, this result can be attributed to the fact that when considering smaller time granularity, we have less noise in tweets and prediction of future words' probabilities is better.

[6] https://en.wikipedia.org/wiki/68-95-99.7_rule

[7] https://en.wikipedia.org/wiki/Table_of_administrative_divisions_by_country

TABLE 5
Transitions Stats for Kuala Lumpur

Statistic	Value
$\mu_{diagonal}\ M_{full}$	0.0025
$\mu_{diagonal}\ M_{post}$	0.01
$\mu_{diagonal}\ M_{pre}$	0.024
$\sigma_{non-diagonal}\ M_{full}$	0.09
$\sigma_{non-diagonal}\ M_{post}$	0.068
$\sigma_{non-diagonal}\ M_{pre}$	0.018

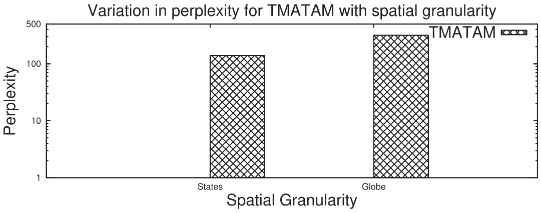


Fig. 10. Variation in performance of TM–ATAM with geographic granularity over regions. ”States” and ”Counties” correspond to first and second level administrative divisions.

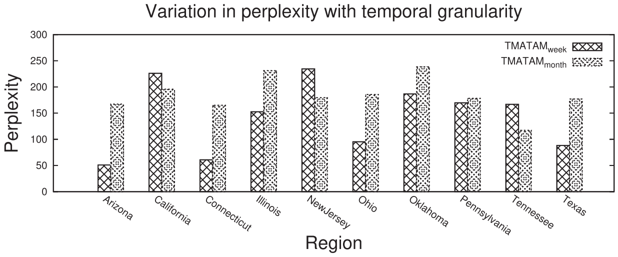


Fig. 11. Variation in perplexity for TM–ATAM at different temporal granularities. Results for top-10 active regions.

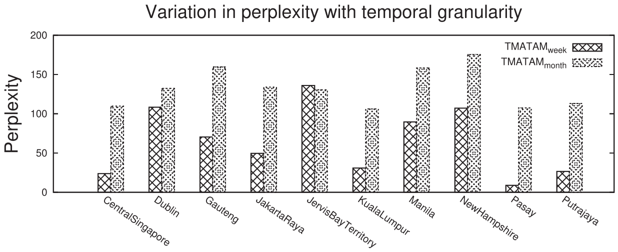


Fig. 12. Variation in perplexity for TM–ATAM at different temporal granularities. Results for top-10 non-US active regions.

5.4.3 ATAM: Effect of Parameters

Geographic Granularity and Temporal Granularity. Please note that ATAM does not formalize temporal and geographic granularity in the model. We run it just to see effect of varying various parameters on its performance. The performance of ATAM improves when run on each individual state separately as shown in Fig.16.

Temporal Granularity. As in the case of TM–ATAM, ATAM also gets better results when run on shorter time periods. Shorter periods capture subtle change points in ailment distributions that might be missed when ATAM is run on longer periods. But as time granularity gets smaller, data gets sparser and for some regions health topics inferred do not make sense. These observations on ATAM are evident in Figs. 17 and 18.

5.4.4 TM–LDA: Effect of Parameters

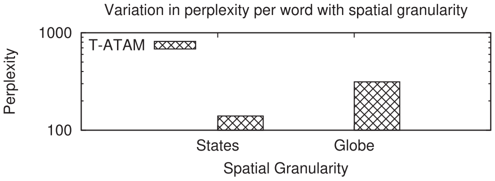


Fig. 13. Variation in performance of T–ATAM with geographic granularity over regions. ”States” correspond to first level administrative divisions.

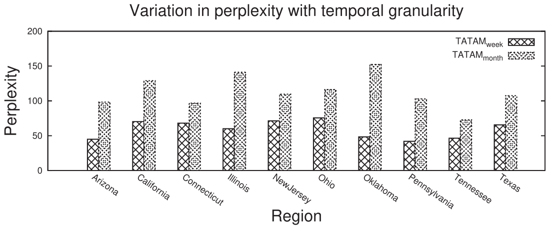


Fig. 14. Variation in perplexity for T–ATAM at different temporal granularities. Results for top-10 social media active regions.

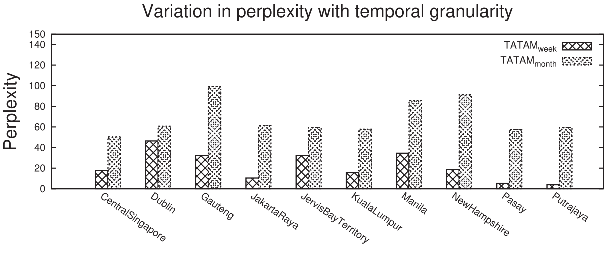


Fig. 15. Variation in perplexity for T–ATAM at different temporal granularities. Results for top-10 non-US active regions.

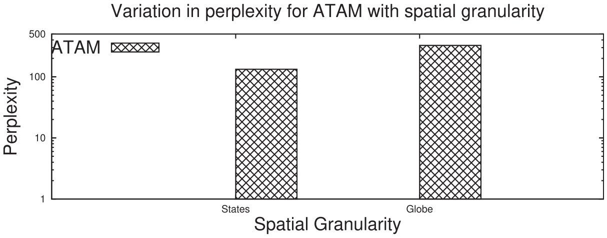


Fig. 16. Variation in perplexity achieved by ATAM at different spatial granularities.

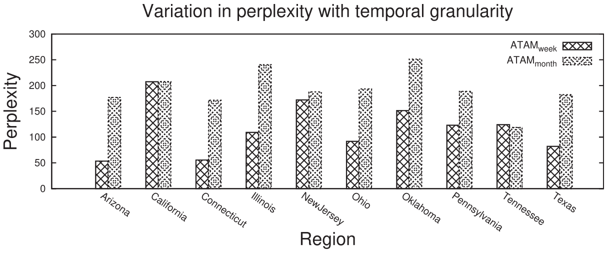


Fig. 17. Variation in perplexity for ATAM at different temporal granularities. Results for top-10 active regions.

Geographic Granularity. Recall that for TM–LDA and LDA each region is a `virtual user`, and analysis by varying geographic granularity is not possible - in the case of the whole globe, we are left with a single `virtual user`. In case of a single user, modeling the transition matrix of TM–LDA is not qualitative as TM–LDA at its heart relies on tweet content of many users to model its transition matrix. So, we confine to analysis of varying temporal granularity to TM–LDA and LDA.

Temporal Granularity. Results on varying temporal granularity are not stable in case of TM–LDA. In some regions we get better results for weeks and in others we get better

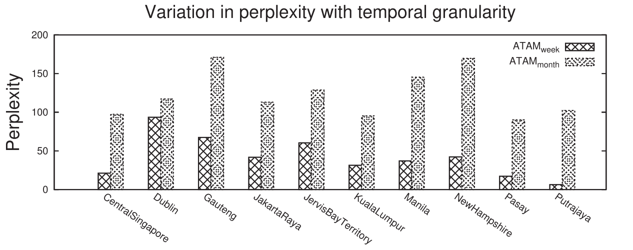


Fig. 18. Variation in perplexity for ATAM at different temporal granularities. Results for top-10 non-US active regions.

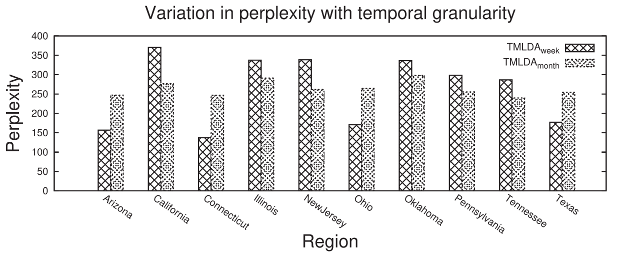


Fig. 19. Variation in perplexity for TM–LDA at different temporal granularities. Results for top-10 active regions.

results for months. Results of TM-LDA on varying temporal granularity for top-U.S and Non-U.S. regions are shown in Figs. 19 and 20.

5.4.5 LDA: Effect of Parameters

Temporal Granularity. We found in our experiments that Temporal analysis of LDA is same as that TM-LDA. Due to lack of space we do not show it.

5.5 T–ATAM versus CDC and Google Flu Trends

Qualitative analysis of TATAM was not done as in the case of TM-ATAM. Qualitative analysis of TM-ATAM in Section 5.3 was highly dependent on the transition matrix which is an inherent part of TM-ATAM. As there is no transition matrix for T-ATAM, we compare its correlation with CDC which is the ground truth.

To evaluate the output of T–ATAM and its ability to explore several aspects of public health, we focus on syndromic surveillance. Since T–ATAM discovers many ailments such as "flu", we use it to track influenza in the US. The CDC provides the rate of Influenza positive tests for the whole US and by region (10 Standard federal regions) reported by Public Health Laboratories.

We gather data from CDC site[8] between 8-Oct-2014 and 31-June-2015 and we measure the correlation between the probability of the flu ailment for each month (ailment distribution produced by T–ATAM) and the influenza rate in the United States measured by the CDC. We study T–ATAM's correlation for a range of topics (parameter Z) and ailments (parameter A). The best correlation is obtained with $Z = 10$ and $A = 25$ for both T–ATAM and ATAM. Results for the whole US (c.f. Fig. 21) with T–ATAM yield a correlation coefficient of 0.9465, while ATAM obtains a correlation of 0.829. T–ATAM is hence a very good candidate to track the flu rate in tweets.

We study also T–ATAM's performance for smaller geographic granularity and we calculate correlation in the 10 standard federal US regions.[9] T–ATAM yields good correlations for these regions. For instance, we obtain a correlation coefficient of 0.8700 in the Central South of the US and a correlation coefficient of 0.8592 in the Mid Atlantic.

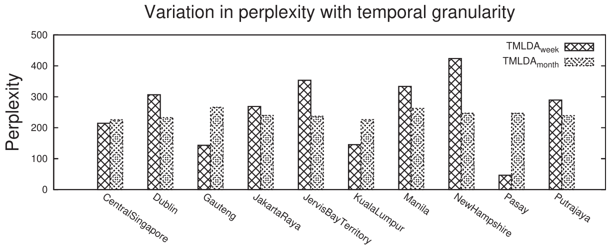


Fig. 20. Variation in perplexity achieved by TM–LDA at different temporal granularities. Results for 10 non-US active regions.

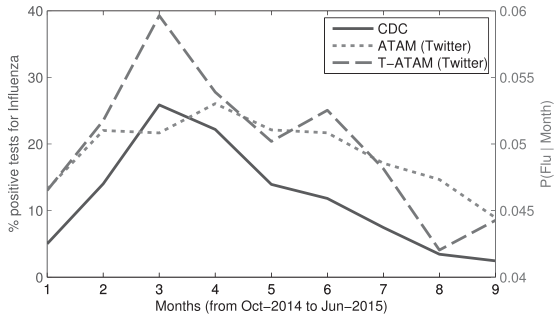


Fig. 21. T-ATAM's and ATAM's flu probability given the month and influenza rate calculated by CDC as a percentage of positive tests for flu in US.

The fact that T–ATAM's correlation for flu is better than ATAM's correlation shows the importance of modeling time natively in the model to capture seasonal diseases.

We make a second comparison against Google Flu Trends which provides estimates of influenza activity in many countries such us the US. It builds on aggregating Google search queries. We gather data from Google Flu Trends[10] for the same time period (from 8-Oct-2014 to 31-June-2015) and we aggregate it into months then we calculate the correlation between the Google Flu Trends rates and both T–ATAM and ATAM's flu probabilities. We obtain a correlation of 0.8906 for T–ATAM and 0.8391 for ATAM. This confirms again the superiority of T–ATAM over all methods.

5.6 Summary of Results

By modeling transitions in the same `homogeneous time period`, TM–ATAM consistently outperforms TM–LDA in predicting health topics in all social-media active regions.

We analyze the performance of TM–ATAM by changing spatio-temporal parameters. In particular, we find that prediction accuracy for health topics is higher when operating TM–ATAM on finer spatial granularity and shorter time periods.

Further, we go on to discover interesting region-specific intra and *inter-*`homogeneous time period` health-related transitions. While studying these transitions, we find that `homogeneous time periods` are continuous time periods for which people in the same region tweet about similar health issues. When those `homogeneous time periods` end, we found that ailments discussed in Twitter transition

[8] https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

[9] https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States

[10] https://www.google.org/flutrends/about/data/flu/us/data.txt

into other ailment topics. These results show that it is more logical to predict future ailments concerning people within the same `homogeneous time period` of a region than on any random health tweets.

By outperforming TM–LDA in predicting future health topics, we show that it is essential to use a dedicated method that separates health-related topics from other topics.

Since in T–ATAM, $time$ is considered a random variable following multinomial distribution, we expect it to outperform other models, TM–LDA and TM–ATAM in predicting health topics using perplexity measure. According to our expectations, in most social-media active regions, in both US active regions and non-US active regions, T–ATAM outperforms TM–ATAM and ATAM (c.f Fig. 8).

After analyzing T–ATAM's performance by changing various spatio-temporal parameters, we find that (as in the case of TM–ATAM) the prediction accuracy for health topics is higher when operating T–ATAM on finer spatial granularity and shorter time periods.

Finally, T–ATAM shows good correlations with the CDC's flu data (the rates of the positive tests of influenza measured by the Center of Disease Control and Prevention in the US) and Google Flu Trends data for a syndromic surveillance study.

6 Related Work

Proliferation of social media platforms such as *Twitter*, *pinterest*, *facebook*, *tumblr* has led to their application to a wide array of tasks including mental health assessment [14], [15], [16], inferring political affiliation [17], [18], [19], [20], brand perception [21], [22] etc.

Social media, especially Twitter, are good sources of personal health [23], [24], [25], [26]. Previous studies on public health surveillance have attempted to uncover ailment topics on online discourse [4], [27] or model the evolution of general topics [7]. In this paper, we combine the best of both worlds which leads to the discovery of *disease*-`change-points` for social-media active regions. We model the evolution of diseases within `change-points` and obtain significant improvement over the state-of-the-art for public health surveillance using social media.

Just like TM-LDA, TM-ATAM and T–ATAM learn topic transitions over time and not topic trends. Such transitions the purpose of answering questions such as people talk about fever before talking about stomach ache. Other complementary approaches that learn the dynamicity of word distributions or topic trends have been proposed. That is the case of [9] that models topic evolution over time as a discrete chain-style process where each piece is modeled using LDA. In [11], the authors propose a method that learns changing word distributions of topics over time and in [10], the authors leverage the structure of a social network to learn how topics temporally evolve in a community. TM–ATAM and T–ATAM are however different from dynamic topic models such as [9] and [10], and from the work of Wang et al. [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. TM–ATAM learns transition parameters that dictate the evolution of health-related topics by minimizing the prediction error on ailment distributions of consecutive periods at different temporal and geographic granularities. T–ATAM on the other hand discovers latent ailments in health tweets by treating time as a corpus-specific multinomial distribution. Classical approaches [28] have been applied to mining topics for inferring citations. Other discriminative approaches [29], [30] have been applied to do an empirical study on topic modeling and time-based topic modeling respectively. None of those are directly applicable to health data.

Finally, in [31], Non-negative Factorization is used for learning topic trends. Exploring the applicability of that complimentary approach to the evolution of health topics in tweets, is a promising research direction.

7 Conclusion

We develop methods to uncover ailments over time from social media. We formulated health transition detection and prediction problems and proposed two models to solve them. Detection is addressed with TM–ATAM, a granularity-based model to conduct region-specific analysis that leads to the identification of time periods and characterizing homogeneous disease discourse, per region. Prediction is addressed with T–ATAM, that treats time *natively* as a random variable whose values are drawn from a multinomial distribution. The fine-grained nature of T–ATAM results in significant improvements in modeling and predicting transitions of health-related tweets. We believe our approach is applicable to other domains with time-sensitive topics such as disaster management and national security matters.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments that helped us greatly improve our paper. The authors would also like to thank Shashwat Mishra for the valuable discussions they had with him during the course of this work. This work is supported by the CDP LIFE Project.

References

[1] L. Manikonda and M. D. Choudhury, "Modeling and understanding visual attributes of mental health disclosures in social media," in *Proc.CHI Conf. Human Factors Comput. Syst.*, 2017, pp. 170–181.

[2] S. R. Chowdhury, M. Imran, M. R. Asghar, S. Amer-Yahia, and C. Castillo, "Tweet4act: Using incident-specific profiles for classifying crisis-related messages," in *10th Proc. Int. Conf. Inform. Syst. Crisis Response Manag.*, 2013.

[3] T. Davidson, D. Warmsley, M. W. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. 11th Int. Conf. Web Social Media*, 2017, pp. 512–515.

[4] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health," in *Proc. Int. Conf. Weblogs Social Media*, 2011.

[5] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 50–57.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn.*, vol. 3, pp. 993–1022, 2003.

[7] Y. Wang, E. Agichtein, and M. Benzi, "TM-LDA: Efficient online modeling of latent topic transitions in social media," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 123–131.

[8] S. Sidana, S. Mishra, S. Amer-Yahia, M. Clausel, and M. Amini, "Health monitoring on social media over time," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2016, pp. 849–852.

[9] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 113–120.

[10] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky, "The joint inference of topic diffusion and evolution in social communities," in *Proc. 11th Int. Conf. Data Mining*, 2011, pp. 378–387.

[11] X. Wang and A. McCallum, "Topics over time: A non-markov continuous-time model of topical trends," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 424–433.

[12] K. W. Prier, M. S. Smith, C. Giraud-Carrier, and C. L. Hanson, "Identifying health-related topics on twitter," in *Social Computing, Behavioral-Cultural Modeling and Prediction*. Berlin, Germany: Springer, 2011, pp. 18–25.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995. [Online]. Available: http://dx.doi.org/10.1007/BF00994018

[14] M. De Choudhury, "Anorexia on tumblr: A characterization study," in *Proc. 5th Int. Conf. Digital Health*, 2015, pp. 43–50.

[15] M. De Choudhury, A. Monroy-Hernández, and G. Mark, ""narco" Emotions: Affect and desensitization in social media during the mexican drug war," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2014, pp. 3563–3572.

[16] U. Pavalanathan and M. De Choudhury, "Identity management and mental health discourse in social media," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 315–321.

[17] F. Bouillot, P. Poncelet, M. Roche, D. Ienco, E. Bigdeli, and S. Matwin, "French presidential elections: What are the most efficient measures for tweets?" in *Proc. 1st Edition Workshop Politics Elections Data*, 2012, pp. 23–30.

[18] L. Hemphill and A. J. Roback, "Tweet acts: How constituents lobby congress via twitter," in *Proc. ACM Conf. Comput. Supported Cooperative Work Social Comput.*, 2014, pp. 1200–1210.

[19] A. Ceron, L. Curini, and S. M. Iacus, "Using sentiment analysis to monitor electoral campaigns: Method matters-evidence from the united states and italy," *Social Sci. Comput. Rev.*, vol. 33, no. 1, pp. 3–20, 2015.

[20] P. Barberá, "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter Data," *Political Anal.*, vol. 23, no. 1, pp. 76–91, 2015.

[21] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol.*, 2011, pp. 151–160.

[22] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 11, pp. 2169–2188, 2009.

[23] S. Wang, M. J. Paul, and M. Dredze, "Exploring health topics in Chinese social media: An analysis of sina weibo," in *Proc. AAAI Workshop World Wide Web Public Health Intell.*, 2014.

[24] A. Culotta, "Estimating county health statistics with twitter," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2014, pp. 1335–1344.

[25] N. Kanhabua and W. Nejdl, "Understanding the diversity of tweets in the time of outbreaks," in *Proc. Int. Conf. World Wide Web*, 2013, pp. 1335–1342.

[26] O. J. Dyar, E. Castro-Sánchez, and A. H. Holmes, "What makes people talk about antibiotics on social media? a retrospective analysis of twitter use," *J. Antimicrobial Chemotherapy*, vol. 69, pp. 2568–2572, 2014.

[27] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2006, pp. 241–248.

[28] R. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 542–550.

[29] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proc. 3rd Workshop Social Netw. Mining Anal.*, 2010, pp. 80–88.

[30] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis, "A time-dependent topic model for multiple text streams," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 832–840.

[31] A. Saha and V. Sindhwani, "Learning evolving and emerging topics in social media: A dynamic NMF approach with temporal regularization," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, 2012, pp. 693–702.

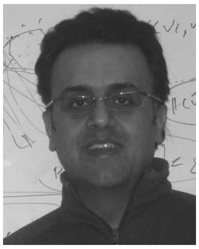
Sumit Sidana received the bachelor's degree in information technology from SVIET, India, and the master's degree in computer science from IIIT, Hyderabad. He is working toward the PhD degee at Grenoble Alpes University, Grenoble.

Sihem Amer-Yahia is research director in the CNRS, LIG, in Grenoble where she leads the SLIDE team. Before that, she was a principal scientist in the QCRI, and senior scientist in Yahoo! Research and AT&T Labs. She is PC chair for VLDB 2018.

Marianne Clausel received the PhD degree in applied mathematics from the University of Créteil, France, in 2008. She is a professor with the University of Lorraine. Her research is in statistics and probabilistic modeling, in particular topic modeling and data mining.

Majdeddine Rebai is an intern in the CNRS, LIG, in Grenoble. He is a third-year engineering student in ENSTA ParisTech.

Son T. Mai received the PhD degree in computer science from the Ludwig Maximilian University of Munich (LMU), Germany. He is a postdoctoral fellow at LIG, Grenoble. He was a postdoc with Aarhus University, Denmark.



Massih-Réza Amini is a professor with the University of Grenoble Alpes and head of the Machine Learning Group. His research is in statistical machine learning. He is interested in the learnability of statistical models where only a small portion of data is labeled to make prediction. He has also contributed in developing machine learning techniques for information retrieval and text mining.