Health Monitoring on Social Media over Time

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 monitoring people’s health over time. Recently, Temporal-LDA (TM–LDA) was proposed for efficiently modeling general-purpose topic transitions over time. In this paper, we propose Temporal Ailment Topic Aspect (TM–ATAM), a new latent model dedicated to capturing transitions that involve health-related topics. TM–ATAM learns topic transition parameters by minimizing the prediction error on topic distributions between consecutive posts at different time and geographic granularities. Our experiments on an 8-month corpus of tweets show that it largely outperforms its predecessors.

CSC Concepts

• Information systems → Social networks; Traffic analysis; Document topic models; Information systems applications; • General and reference → Measurement;

1. INTRODUCTION

Social media has become a major source of information for analyzing many aspects of daily life. In particular, public health monitoring can be conducted on Twitter to measure the well-being of different geographic populations. The ability to model transitions for ailments and detect statements such as “people talk about smoking and cigarettes before talking about respiratory problems”, or “people talk about headaches and stomach ache in any order”, has a range of applications in syndromic surveillance such as measuring behavioral risk factors and triggering public health campaigns.

Popular probabilistic topic modeling methods such as Latent Dirichlet Allocation [2] and pLSA [4] have a long history of successful application to news articles and academic abstracts. However, the small size of social media content poses serious challenges to the efficacy of such methods [9]. Dedicated methods, such as the Ailment Topic Aspect Model (ATAM), have thus been proposed to discover ailments from tweets [5].

While the primary goal of probabilistic topic modeling is to learn topic models, an equally interesting objective is to examine topic transitions. A temporal extension to LDA (TM–LDA) was hence developed for discovering the evolution of general-purpose topics in tweets [8]. In this paper, we examine the feasibility of measuring and predicting ailment transitions in Twitter, by combining ATAM and TM–LDA into a new model, coined TM–ATAM. Our model is different from dynamic topic models such as [1,7], as it is 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 by minimizing the prediction error on ailment distributions of consecutive periods at different temporal and geographic granularities.

The effectiveness of TM–ATAM requires to 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 cover a user population that is exposed to different weather conditions and miss meaningful transitions. Our experiments on a corpus of more than 500K health-related and geo-localized tweets collected over a period of 8 months, show that TM–ATAM outperforms ATAM, TM–LDA and LDA in estimating temporal health-related topic transitions of different geographic populations. The health-related topic transitions we unveiled can be broadly classified in 2 kinds: 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 Arizona revealed many self-transitions such as headaches and body pain. On the other hand, tweets about smoking, drugs and cigarettes in California, are followed by respiratory ailments.

2. MODEL, PROBLEM AND APPROACH

Table 1 summarizes the terminology we use throughout this paper. By using suitable geographic granularity g (country, state, county) and temporal granularity t (week, bi-week and months), we build our document sets $D_g^t$. While LDA is successful at uncovering generic topics, its limitations at discovering infrequent and specific topics such as health has already been shown [5]. The probabilistic Ail-
Table 1: Mapping tweets to documents

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P)</td>
<td>set of (tweet) posts</td>
</tr>
<tr>
<td>(G)</td>
<td>set of regions</td>
</tr>
<tr>
<td>(T)</td>
<td>set of time periods</td>
</tr>
<tr>
<td>(\mathcal{P}_g)</td>
<td>posts from region (g) during time (t)</td>
</tr>
<tr>
<td>(D_g)</td>
<td>document-set built by mapping the content of each post (p \in \mathcal{P}_g) to a document</td>
</tr>
<tr>
<td>(\Theta_g)</td>
<td>ailment distribution vector for document-set (D_g) of region (g) during time (t)</td>
</tr>
<tr>
<td>(m)</td>
<td>distance measure between distributions</td>
</tr>
</tbody>
</table>

“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:

<T> Sickness (0.45), Depression (0.27), Flu (0.09), Recovery (0.04), Sad (0.04), Headache (0.09)…<T>

Topic vector, \(\theta\), produced by LDA

<T> ThroatInfection (0.57), Generosity (0.14), Doctor (0.07), Weekend (0.07), Flu (0.07), Sick-Leave (0.07)…<T>

Topic vector, \(\theta\), produced by ATAM

Figure 1: LDA vs ATAM: Comparison of topic distributions for an example tweet.

**2.1 Ailment prediction problem**

In [8], TM–LDA was introduced to extend LDA with modeling topic evolution over time. However, while being quite elegant in modeling general-purpose topics TM–LDA is not specialized to capture health transitions over time.

Let \(\Theta_g^t\) be an ailment distribution vector where the weight of each ailment is representative of the discourse density of ailment in the tweets originating from region \(g\) during period \(t\). For a region \(g\), the interval of time spanning a set of consecutive time periods \(\{t_i, t_{i+1}, \ldots\}\) during which discovered ailment distributions \(\{\Theta_g^{t_i}, \Theta_g^{t_{i+1}}, \ldots\}\) do not change appreciably forms a homogenous time period w.r.t. ailments. By definition, a homogenous time period is (nearly) homogeneous in terms of ailments. In other words, the ailments evolve in a smooth fashion within a homogenous time period and change abruptly across homogeneous time periods. We posit that such homogenous time periods exist after which they encounter change-points in ailment topic discussions. 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. As an example, in Figure 2, 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). The sharp peaks obtained validate the existence of time intervals that are homogeneous w.r.t. ailments.

**Our problem:** Given a set of documents \(D_g^{t_i-1}\) formed by tweets originating from a region \(g \in G\) during time period \(t_i-1\), predict \(\Theta_g^t\), the ailment distribution of documents in \(D_g^t\), corresponding to posts from \(g\) in period \(t_i\) from \(\Theta_g^{t_i-1}\), the ailment distribution of document \(D_g^{t_i-1}\) corresponding to posts from \(g\) during period \(t_i-1\).

![Algorithm 1: TM-ATAM: change-point Detection and Training Ailment Distribution Predictor](image)

![Figure 2: Topic transitions over time.](image)
2.2 Modeling Health Topics over Time with TM–ATAM

To solve our problem, we propose TM–ATAM that builds on top of ATAM and TM–LDA. We first convert inferences of ATAM over a single document to associate with a given set of documents $D_g$, an ailment distribution, $\Theta_g^t$. We then go on to find **homogenous time periods**. We model ailment transitions within each homogenous time period and when a **change-point** is encountered we update these transitions. This is a fresh departure from existing solutions that operate in a **homogenous time period-agnostic** fashion [8]. TM–ATAM, at its heart, solves the following equation.

$$A_g^t \approx A_g^{t-1}M^*$$

where

$$A_g^{t-1} = \left( \begin{array}{c} \Theta_g^t \\ \vdots \\ \Theta_g^{t+1} \end{array} \right), A_g^t = \left( \begin{array}{c} \Theta_g^t \\ \vdots \\ \Theta_g^{t+1} \end{array} \right)$$

$M^*$ is an unknown transition matrix which is obtained by solving the following least squares problem.

$$M^* = \arg\min_M \|A_g^r - A_g^{r-1}M\|_F$$

Algorithm 1 contains the steps of our solution. It has two parts: **change-point detection** and **ailment prediction**.

**Change Point Detection.**

We use $\mathcal{Z}$ to refer to the set of all health-related and non-health related topics. 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 a topic distribution $\Theta_g^t$ (Lines 4 – 12). We then examine the **Bhattacharya Distance** between consecutive distributions $\Theta_g^{t-1}$ and $\Theta_g^t$ of the region $g$ to identify the most significant **change-point**, $t_c$, for region $g$ (Line 14). The time periods preceding and succeeding **change-point** are termed as **homogenous time periods**.

**Ailment Prediction.**

In the second module of TM–ATAM algorithm, we predict distribution of ailments in twitter discourse ahead of time for each **homogenous time period**. Lines 17–20 of Algorithm 1 outline the steps undertaken to identify the detection of ailments for intra-homogeneous periods.

### 3. EXPERIMENTS

We conducted experiments to evaluate the performance of TM–ATAM and to compare with the state of the art.

#### 3.1 Experimental setup

We employ Twitter’s Streaming API to collect tweets between 2014-Oct-8 and 2015-May-31. Collected tweets were subjected to two pre-processing steps as follows.

**Identifying health-related tweets:** We filter the tweets returned by the Decahose Stream to obtain health-related tweets. We say that a tweet is health-related if it contains a health keyword and passes our classification criteria. The process is automated with the help of an SVM classifier [3]. To this end, 5128 tweets were annotated through crowdsourcing efforts. The precision and recall of the classifier are 0.85 and 0.44. Table 2 shows that out of the 1.36B tweets we collected, 698K were health-related.

**Identifying geolocalized tweets:** The ability to operate seamlessly at varying geographic resolutions mandates that the exact location of each tweet be known to TM–ATAM. Twitter affords its users the option to share their geolocation. In our case, over half a million tweets are retained (569K as indicated in Table 2).

We examine various choices for the geographic granularities, temporal granularities and distance measures. TM–ATAM performs better on smaller regions and finer temporal granularity. We attribute this result to the fact that tweets from smaller regions and finer temporal granularity show less diversity in topics. We also used 2 distance measures to measure distribution difference namely, cosine similarity and bhattacharya distance. We observed that number of tweets at a given time period $t$ may affect the performance of Cosine Similarity. Finally, we chose to work with geographic granularity of **states**, temporal granularity of **months** and distance measure of **bhattacharya**.

**Test-bench and measures:** We run our experiments on a 32 core Intel Xeon @ 2.6Ghz CPU (with 20MB cache per core) system with 128 Gigs RAM running Debian GNU/Linux 7.9 (wheezy) operating system. All subsequently discussed components were implemented in Java 1.8.0_60. We used **perplexity** to compare between models [6].

#### 3.2 Experimental Results

Recall that the terms **change-point** and **homogenous-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. We divide each **homogenous time period** into training and test sets. ATAM is then **re-run** over the training set of each **homogenous time period**. We then model a **transition matrix** $M_{\text{atam}}$ on the training set of each **homogenous time period** as described in Section 2.2. We compute the probability of "health topic" $z$ for each tweet $p$ of the first month ($|T| - 1$) in the test set using the following formulas:

$$P(z|t_{|T|-1}) = \frac{\sum_{p \in t_{|T|-1}} P(z|p \text{ for } t_{|T|-1})}{\#\text{tweets for } t_{|T|-1}}$$

$$P(z|p) = \sum_{w} P(z|w)P(w|p) = \sum_{w} \frac{n(z,w)}{n(w)} P(w|p)$$

Here, values for $n(z,w), n(w)$ are taken from ATAM run on the training months. $P(w|p)$ is simply the number of times word $w$ occurs in the tweet $p$ divided by the total number of words in $p$. We then predict the future probability of each topic in the second month of the test data using the corresponding **transition matrix** $M_{\text{atam}}$. Probability of word $p(w_i)$ for any document set is calculated as follows:

$$p_n(w_i) = \sum_{z} P(w|z)P(z) = \sum_{z} \frac{n(z,w_i)}{n(z)} P(z)$$

<table>
<thead>
<tr>
<th>Table 2: Dataset Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>collection period (days)</td>
</tr>
<tr>
<td>#tweets</td>
</tr>
<tr>
<td>#tweets (health-related)</td>
</tr>
<tr>
<td>#tweets (health-related+geolocated)</td>
</tr>
</tbody>
</table>
Empirical comparison of ATAM, TM-LDA and LDA with TM-ATAM

![Graph showing comparison of ATAM, TM-LDA, and LDA with TM-ATAM]

Figure 3: Comparison for top 10 active regions. Histograms denote ratio of perplexities. TM–ATAM is always at 1.0.

Having computed $P(w)$, we can compute perplexity against the words of the tweets of second test month.

### 3.2.1 TM–ATAM vs ATAM, TM–LDA vs LDA

Figure 3 shows the perplexity ratio of TM–ATAM with state-of-the-art models. If ratio computes to be less than ‘1’ for competitor topic model, TM–ATAM is performing worse. If ratio is more than ‘1’ for competitor topic model, TM–ATAM is performing better. In order to assert the fact that health topics transit from one to another, we compute the perplexity of ATAM on words of the first month of the test set and not predicting any topic distribution using any transition matrix. Hence, this denotes the case where health topics stay static. As shown in Figure 3, TM–ATAM beats ATAM in all social media active regions (an active region is a region where the proportion of tweets if high enough). For training TM–LDA, we merge the training data (same as TM–ATAM) of each region and train a transition matrix of TM–LDA by solving least squares problem. For each tweet $p$ of the first month of the test set, we compute the probability of topics using LDA trained on merged training data (Formula 3). We then predict the future probability of each topic in the following month using $M'_{tmlda}$. We can then compute the perplexity of TM–LDA against words of actual tweets in the test set. Figure 3 shows that TM–ATAM consistently beats TM–LDA and LDA in predicting future health topics on the test month. Perplexity is indeed lower consistently beats TM–LDA and LDA in predicting future tweets in the test set. Figure 3 shows that TM–ATAM computes the perplexity of TM–LDA against words of actual tweets in the test set.

### 3.2.2 Homogenous Time Periods

In Figure 4 we show the top-2 sharpest change-points for the top regions. Those points can be explained with weather changes in those regions. Texas can be explained with a drop in temperature while Jervis Bay can be explained by an increase in rainfall. Dublin sees its lowest temperature in November. Singapore and Manila have very similar weather conditions and exhibit the same change point.

### 3.2.3 Topic Transitions

Entry $m_{ij}$ in the transition matrix $M$ produced by TM–ATAM, shows the degree that topic $z_i$ will contribute to topic $z_j$ in the subsequent homogenous time period. We adapt the threshold used in [8] to our settings: Threshold $= \mu + 2 \times \sigma_{non-diagonal}$. We identify two kinds of interesting transitions based on the above threshold: self-transitions: popular topics and one way transitions: $j^{th}$ topic is discussed before $j^{th}$ topic. As an example, one-way-transitions of California are summarized in Table 3.

### 4. CONCLUSION

We studied how to uncover ailment distributions over time in social media. We proposed a granularity-based model to conduct region-specific analysis that leads to the identification of time intervals characterizing homogeneous ailment discourse, per region. We modeled disease evolution within each homogeneous region and attempted to predict ailments. The fine-grained nature of our model results in significant improvements over state of the art methods.

### 5. REFERENCES