Health Monitoring on Social Media over Time

Joint work with S. Sidana, S. Mishra, S. Amer–Yahia, M.R. Amini
1. The CrowdHealth project

2. Topic modeling and tweet analysis
   - What is topic modeling?
   - Time evolving topic models
   - The tweet case

3. TM-ATAM: approach
   - ATAM
   - Taking time into account
   - Challenges
   - Distribution vectors
   - Ailment prediction

4. Experiments
   - Data
   - Summary of experiments

5. Results
   - Contributions
The CrowdHealth project

Topic modeling and tweet analysis
- What is topic modeling?
- Time evolving topic models
- The tweet case

TM-ATAM: approach
- ATAM
- Taking time into account
- Challenges
- Distribution vectors
- Ailment prediction

Experiments
- Data
- Summary of experiments

Results
- Contributions
The CrowdHealth project

Objectives

- Our datas: geolocalized tweets collected from Oct 2014 to May 2015:
- Aims: extract health-related tweets from this data base and verify correlations between demographics, nutrition and health
The CrowdHealth project
Several steps

- Preprocessing of the data to identify geotagged and health–related tweets ($\approx 500000$)
- Study of the evolution with time and the influence of localization on the health–related content of these tweets using topic modeling.
The CrowdHealth project
More about our datas

An example of tweet

“Having tonsillitis and coughing for a straight hour ain’t no fun... my throat is raw! Thanks god for antibiotic and pain meds”

Several possible informations may be in the tweets

- time, localization
- content: symptoms (tonsillitis, coughing, throat infection), treatment (antibiotic), contextual information...
The CrowdHealth project
What can be done?

- Health topics evolve over region and time
- If India is about “tuberculosis”, then U.S. is about “obesity”
- If Summer is about “sun-burns”, then Winter is about “allergies”
- Can we predict health topics in first place?
  - Prediction of chronic obesity in New York can trigger health campaigns
  - Prediction of depression in Missouri can lead to measure behavioral risk factors
- Need of **time-aware and geo-aware topic model** to capture, model and predict health-topic transitions
1. The CrowdHealth project

2. Topic modeling and tweet analysis
   - What is topic modeling?
   - Time evolving topic models
   - The tweet case

3. TM-ATAM: approach
   - ATAM
   - Taking time into account
   - Challenges
   - Distribution vectors
   - Ailment prediction

4. Experiments
   - Data
   - Summary of experiments

5. Results
   - Contributions
What is topic modeling?

Objectives

- Problem: modeling text corpora.
- Aim: find short description of each document.
- Approach: use of generative probabilistic models involving some latent variables (not observed) related to the underlying topics of the corpus.
What is topic modeling?
The model LDA [Blei 2003]
What is topic modeling?

The model LDA [Blei 2003]

Generative process:
For each topic $k \in \{1, \cdots, K\}$, generate $\phi_k \sim Dir(\beta)$ a distribution probability of the words on each topic

For each document $d$ in a corpus $D$:

1. Choose $N \sim Poisson(\xi)$
2. Choose $\theta_d \sim Dir(\alpha)$
3. For each of the $N$ words $w_n$
   1. Choose a topic $z_{d,n} \sim Multinomial(\theta_d)$
   2. Choose a word $w_{d,n}$ from $p(w_{d,n}|z_{d,n}, \phi_{z_{d,n},w_{d,n}})$, a multinomial probability conditioned on the topic $z_n$
What is topic modeling?
The LDA model [Blei 2003]
Many possibilities to define dynamic topic models!

- Dynamic Topic Models [Blei et al. 2006]
- Topic Over Time [Wang et al. 2006]
- TM LDA [Wang et al. 2012]
- Topic Sentiment Model [Dermouche et al. 2014]
Time evolving topic models
TM LDA [Wang et al. 2012]

- Models evolution of latent topics with time

\[ \theta_i \approx \frac{\theta_{i-1} \cdot M}{\|\theta_{i-1} \cdot M\|_{\ell_1}} \]

- LDA is not good for summarizing health topics [K.W. Prier et al. 2011]
And what about health–related tweets?

- Health–related topic cannot be separated from the other ones. Confusion with health–related words and others

> “damn flu, home with a fever watching TV”

Our approach use a well adapted topic model for our health–related tweets: ATAM [Drezde et al. 2011] and combine it with TM–LDA.
The CrowdHealth project

Topic modeling and tweet analysis
- What is topic modeling?
- Time evolving topic models
- The tweet case

TM-ATAM: approach
- ATAM
- Taking time into account
- Challenges
- Distribution vectors
- Ailment prediction

Experiments
- Data
- Summary of experiments

Results
- Contributions
Topic modeling for health-tweets
The model ATAM [Drezde et al. 2011]

\[ \eta : \text{ailment distribution}, \ \theta : \text{LDA-topic}, \ x : \text{background/health-related word}, \ y : \text{aspect (general/symptom/treatment)} \]
Topic modeling for health–tweets
The model ATAM [Drezde et al. 2011]

“Neck pain and lower back pain for a pelvis and knee injury? Word ?!”

Example tweet

<Body Pain (0.88), Comparison (0.11)…>

Topic vector, $\Theta$, produced by ATAM

<BackPain (0.30), Allergies (0.06), Anxiety (0.01)…>

Ailment vector, $\eta$, produced by ATAM

$\Theta = [\theta, \eta]$
The CrowdHealth project
Topic modeling and tweet analysis
ATAM and TM–ATAM
Experiments
Results

Topic modeling for health–tweets
The model ATAM [Drezde et al. 2011]

“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, $\theta$, produced by LDA

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

Topic vector, $\Theta$, produced by ATAM

$\Theta = [\theta, \eta]$
Taking time into account

- Pick up a *region* of interest, say, California
- Run ATAM (health-aware topic model) over health tweets of California
- Aggregate inference over a *time* period, say, a month
- Given the health-topic distribution of current month, can we predict the health-topic distribution of next month?
- TM-LDA way: $A \times T = B$
- $T$: Predictor, $A$: current distribution, $B$: Future Distribution
Challenges

- Various Challenges in above approach:
  - Region Instantiation?
  - Time Instantiation?
  - Distribution Vectors?

- Geo:
  - Level-2 administrative divisions: Too microscopic to exhibit independent health patterns
  - Level-0 administrative divisions: Too coarse to start with
  - Level-1 administrative divisions: Natural choice

- Time:
  - Weeks: Too fine and sparse to have enough tweets
  - Years: Too big to start with
  - Months: Natural time granularity
The CrowdHealth project

Topic modeling and tweet analysis

ATAM and TM–ATAM

Experiments

Results

Sampled variables to distribution vectors

Algorithm 1 Ailment Distributions Generation

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: $\Theta^t_g[z] \leftarrow 0$
6: end for
7: for all $d \in D^t_g$ do :
8: for all $w \in d$ do :
9: $z \leftarrow \text{topic}(w)$
10: $\Theta^t_g[z] \leftarrow \Theta^t_g[z] + \frac{1}{|d| \times |D^t_g|}$
11: end for
12: end for
13: end for

Health Monitoring on Social Media over Time
Intra-Season Ailment Prediction

- TM-ATAM is a fresh departure from existing solutions which operate in a season-agnostic fashion
- Key-idea: Predict evolution of ailments within each season

**Algorithm 2 Intra-Season Ailment Prediction**

1. **for all** $g \in G$ **do**
2. $t_c = \arg \max_t m(\Theta_{g}^{t-1}, \Theta_{g}^{t})$
3. $pre = [t_1, t_c-1]$
4. $post = [t_c, t_{\mid T\mid}]$
5. **for all** $s \in \{pre, post\}$ **do**:
6. $A_g^t \approx A_g^{t-1} \cdot M$
7. $M = (A_g^{t-1\top} A_g^{t-1})^{-1} A_g^{t-1\top} A_g^t$
8. **end for**
9. **end for**
| 1 | The CrowdHealth project |
| 2 | Topic modeling and tweet analysis |
|   |   | What is topic modeling? |
|   |   | Time evolving topic models |
|   |   | The tweet case |
| 3 | TM-ATAM: approach |
|   |   | ATAM |
|   |   | Taking time into account |
|   |   | Challenges |
|   |   | Distribution vectors |
|   |   | Ailment prediction |
| 4 | Experiments |
|   |   | Data |
|   |   | Summary of experiments |
| 5 | Results |
|   |   | Contributions |
Dataset Statistics

**Table:** Dataset Statistics

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>collection period (days)</td>
<td>235</td>
</tr>
<tr>
<td>#tweets</td>
<td>1,360,705,803</td>
</tr>
<tr>
<td>#tweets (health-related)</td>
<td>698,212</td>
</tr>
<tr>
<td>#tweets (health-related + geolocated)</td>
<td>569,408</td>
</tr>
</tbody>
</table>
Summary of Experiments

- TM-ATAM outperforms TM-LDA in predicting perplexity of future tweets
- TM-ATAM has a higher prediction accuracy when operating on finer spatial granularity and shorter time periods
- TM-ATAM: A dedicated method to model *intra-season* and *full* health-related transitions
Performance of TM-ATAM vs. TM-LDA in U.S. states

**Figure:** TM-ATAM and TM-LDA prediction accuracies for top-10 active U.S. states
Performance of TM-ATAM by changing spatial granularity

**Figure**: Variation in performance of TM-ATAM with geographic granularity over regions. "States" and "Counties" correspond to first and second level administrative divisions.
Performance of TM-ATAM by changing temporal granularity

**Figure:** Variation in perplexity achieved by TM-ATAM at different temporal granularities. Results for top-10 social media active regions.
Figure: Variation in perplexity achieved by TM-ATAM with different distance measures. Results computed over top-10 active U.S. regions.
Seasons in U.S. states

- Seasons exhibit homogenous ailment distribution and across seasons ailment distributions are expected to fluctuate drastically.

*Texas can be explained with drop in temperature.*

**Figure:** Monthly season boundaries for top-10 active U.S. regions.
### Seasons in Non U.S. states

<table>
<thead>
<tr>
<th>Date:2014-Oct-8</th>
<th>Date:2015-May-31</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singapore</td>
<td></td>
</tr>
<tr>
<td>Dublin</td>
<td></td>
</tr>
<tr>
<td>Gauteng</td>
<td></td>
</tr>
<tr>
<td>JervisBay</td>
<td></td>
</tr>
<tr>
<td>Manila</td>
<td></td>
</tr>
</tbody>
</table>

**Figure:** Monthly season boundaries for top-10 active non-U.S. regions.

- Jervis Bay can be explained with increase in rainfall
- Dublin sees its lowest temperature in November
- Singapore and Manila have very similar weather conditions
## Transition Matrices

**Table:** $M_{\text{full}}$ : Transition Stats for California

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>#tweets</td>
<td>55475</td>
</tr>
<tr>
<td>$\mu_{\text{full}}$</td>
<td>0.015</td>
</tr>
<tr>
<td>$\sigma_{\text{non-diagonal}}$</td>
<td>0.4</td>
</tr>
<tr>
<td>$\mu_{\text{diagonal}}$</td>
<td>-0.002</td>
</tr>
<tr>
<td>Threshold</td>
<td>$0.015 + 2 \times 0.4 = 0.815$</td>
</tr>
</tbody>
</table>

**Table:** Transitions Stats for Kuala Lumpur

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{\text{diagonal}}$</td>
<td>0.0025</td>
</tr>
<tr>
<td>$\mu_{\text{diagonal}}$</td>
<td>0.01</td>
</tr>
<tr>
<td>$\mu_{\text{diagonal}}$</td>
<td>0.024</td>
</tr>
<tr>
<td>$\sigma_{\text{non-diagonal}}$</td>
<td>0.09</td>
</tr>
<tr>
<td>$\sigma_{\text{non-diagonal}}$</td>
<td>0.068</td>
</tr>
<tr>
<td>$\sigma_{\text{non-diagonal}}$</td>
<td>0.018</td>
</tr>
</tbody>
</table>

Health Monitoring on Social Media over Time
Interesting Transitions

- Entry \( m_{ij} \) in the transition parameter matrix \( M \) produced by TM-ATAM, shows the degree that topic \( z_i \) will contribute to topic \( z_j \).
- We analyze 3 kinds of transition matrices: two intra-season transition matrices and one full transition matrix.
## Interesting Transitions

**Table:** $M_{\text{full}}$ (Threshold : 0.45) Transitions for Kuala Lumpur

<table>
<thead>
<tr>
<th>Type</th>
<th>From Topic</th>
<th>To Topic</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Way Transition</td>
<td>Heart Disease/Blood Pressure</td>
<td>Coughing/runny nose/watery eyes</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Brain Disorder</td>
<td>Body pains/Weight Loss</td>
<td>0.599</td>
</tr>
<tr>
<td></td>
<td>Urinary Infection/Intestine/Tract</td>
<td>Stomach Pain/Blood Pressure</td>
<td>0.51</td>
</tr>
</tbody>
</table>

**Table:** $M_{\text{pre}}$ (Threshold : 0.039) and $M_{\text{post}}$ (Threshold : 0.13) Transitions for Kuala Lumpur

<table>
<thead>
<tr>
<th>Type</th>
<th>From Topic</th>
<th>To Topic</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self Transition $M_{\text{pre}}$</td>
<td>headache</td>
<td>headache</td>
<td>0.19</td>
</tr>
<tr>
<td>Self Transition $M_{\text{post}}$</td>
<td>body pain</td>
<td>body pain</td>
<td>0.228</td>
</tr>
</tbody>
</table>
Interesting Transitions

**Table:** $M_{pre}$ transitions in Arizona (threshold : 0.035)

<table>
<thead>
<tr>
<th>Transition Type</th>
<th>From Topic</th>
<th>To Topic</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Transition</td>
<td>Stomach Infection</td>
<td>Stomach Infection</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>Headache</td>
<td>Headache</td>
<td>0.09</td>
</tr>
<tr>
<td>Symmetric-Transition</td>
<td>Stomach Infection</td>
<td>Headache</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Headache</td>
<td>Stomach Infection</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Stomach Infection</td>
<td>Pneumonia</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Pneumonia</td>
<td>Stomach Infection</td>
<td>0.04</td>
</tr>
</tbody>
</table>
The CrowdHealth project

Topic modeling and tweet analysis
- What is topic modeling?
- Time evolving topic models
- The tweet case

TM-ATAM: approach
- ATAM
- Taking time into account
- Challenges
- Distribution vectors
- Ailment prediction

Experiments
- Data
- Summary of experiments

Results
- Contributions

Health Monitoring on Social Media over Time
Contributions

- We come up with seasons, homogenous time-intervals, within which it makes sense to model health transitions.
- We design a dedicated geo-aware, time-aware topic model and prove its ability to model and update health-related transitions by beating previous state of the art.
- We test the effects of changing various parameters on the designed model.
Thank you for your attention!