Health Monitoring on Social Media over Time

Sumit Sidana, Sihem Amer-Yahia, Massih-Reza Amini,
Marianne Clausel, Shashwat Mishra
Univ. Grenoble Alps/CNRS
Grenoble, France
firstname.lastname@imag.fr

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.

CCS Concepts

•Information systems → Social networks; Traffic analysis; *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. In this paper, we propose TM–ATAM, a method that discovers temporal transitions of health-related topics in social media. TM–ATAM combines ATAM, a latent health-related topic model [16], with TM–LDA, that models general-purpose topic transitions [23].

SIGIR '16 July 17–21, 2016, Pisa, Italy

DOI:

Popular probabilistic topic modeling methods such as Latent Dirichlet Allocation [3] and pLSA [12] 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 [24]. Dedicated methods, such as the Ailment Topic Aspect Model (ATAM), have thus been proposed to discover ailments from tweets [16].

While the primary goal of probabilistic topic modeling is to learn sound topic models, an equally interesting objective is to examine *topic transitions.* A temporal extention to LDA (TM–LDA) was hence developed for discovering the evolution of general-purpose topics in tweets [23] for general topics. 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 [2,15,22], 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. For example, it has been shown that discussions on allergies break at different periods in different states in the USA [16]. 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 topic transitions of different geographic populations. The topics 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. Figure 1 shows example transitions we extracted for different states and cities in the world. Such

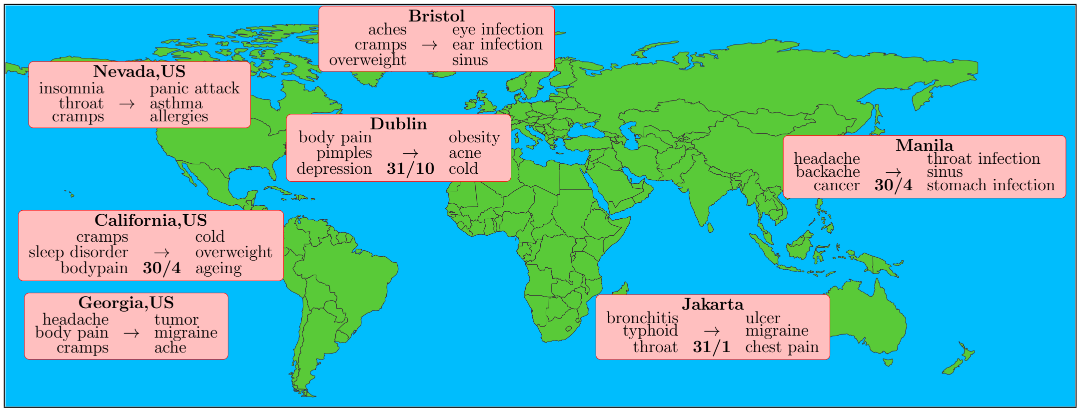

Figure 1: One-way-transitions obtained by TM–ATAM.

transitions are often due to external factors such as climate, health campaigns, nutrition and lifestyle of different world populations.

2. MODEL, PROBLEM AND APPROACH

2.1 Modeling topics with LDA and ATAM

Table 1 presents a summarized version of our terminology. 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 [16, 18]. The probabilistic *Ailment Topic Aspect Model* (ATAM) was designed specifically to uncover latent health-related topics present in a collection of tweets [16]. ATAM achieves remarkable improvement over LDA in discovering topics that correspond to ailments (in addition to discovering general topics). The topic distribution vector generated by ATAM for a sample tweet is shown in Figure 2. Note the stronger relevance to health–related matters in this vector than in the topic distribution vector generated by LDA for the same tweet. While ATAM is effective at modeling health-related topics, it is not designed to model topic transitions over time.

2.2 Ailment prediction problem

In [23], 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 Θ_g^t be a 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 con-

Table 1: Mapping tweets to documents

Term	Description
$\mathcal{P}$	set of (tweet) posts
$\mathcal{G}$	set of regions
$\mathcal{T}$	set of 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
Θ_g^t	ailment distribution vector for document-set D_g^t of region g during time t
m	distance measure between distributions - bhattacharya or cosine

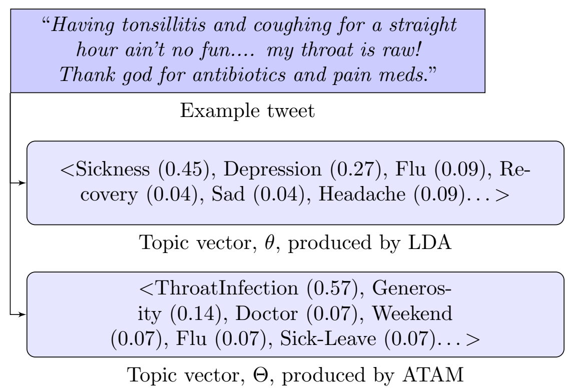

Figure 2: LDA vs ATAM: Comparison of topic distributions for an example tweet.

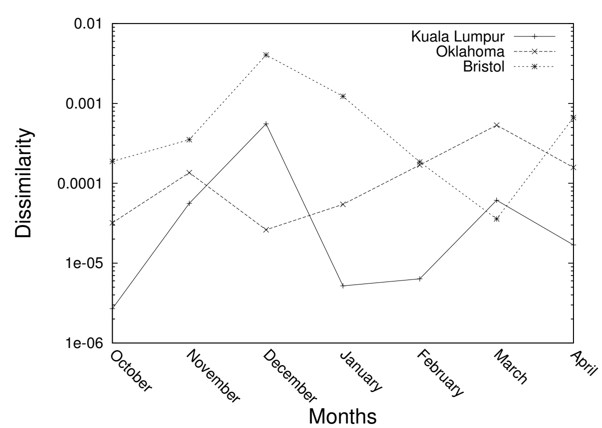

Figure 3: Topic transitions over time.

secutive 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 `homogenous time-period` boundary. 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 3, 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 $\Theta_g^{t_i}$, the ailment distribution of documents in $D_g^{t_i}$, 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}.

2.3 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^t, an ailment distribution, Θ_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 [23]. TM–ATAM, at its heart, solves the following equation.

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

where

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

M is an unknown transition matrix which has to be learned. To obtain the transition matrix M, we solve the following least squares problem.

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

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 Θ_g^t (Lines 4–12). We then examine the *Bhattacharya Distance* between consecutive distributions Θ_g^{t-1} and Θ_g^t of the region g to

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: $t_c = \underset{t}{\operatorname{argmax}}\ m(\Theta_g^{t-1},\ \Theta_g^t)$
15: $pre = [t_1\ ,\ t_{c-1}]$
16: $post = [t_c\ ,\ t_{|\mathcal{T}|}]$
17: **for all** $s \in \{pre,\ post\}$ **do**:
18: $A_g^t \approx A_g^{t-1}.M$
19: $M = (A_g^{t-1\intercal} A_g^{t-1})^{-1} A_g^{t-1\intercal} A_g^t$
20: **end for**
21: **end for**

identify the most significant change point, t_c (Line 14). The time period t_c is termed as the `change-point` for region g. The time periods preceding and succeding `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 on tweets and to compare with state-of-the-art topic models.

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 [7] with linear kernel and uni-gram, bi-gram and tri-gram word features. 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, 689K 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. It has been shown that a very small number of Twitter users choose to share their location. In our case, over half a

Table 2: Dataset Statistics

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

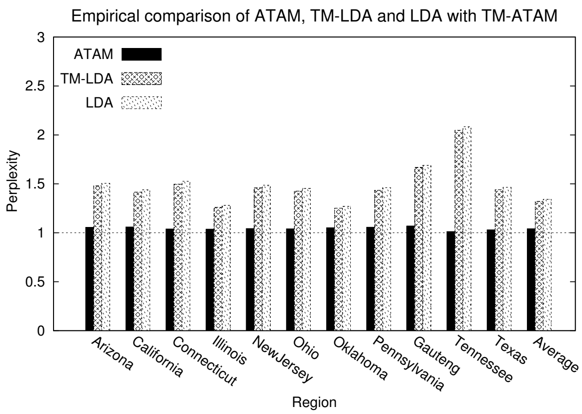

Figure 4: Empirical comparison of ATAM,TM–LDA and LDA with TM–ATAM for social media active regions. Histograms denote ratio of perplexities of `competitor topic model` and TM–ATAM. TM–ATAM is always at 1.0. If ratio is less than "1" for `competitor topic model`, TM–ATAM is performing worse. If ratio is more than "1" for `competitor topic model`, TM–ATAM is performing better.

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 tried with 2 distance measures namely, cosine similarity and bhattacharya distance. We observed that number of tweets at a given time granularity 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 Gig 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 [21].

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_{tmatam} on the training set of each *homogenous time period* as described in Section 2.3. We compute the probability of "health topic" z for each tweet p of the first month ($|\mathcal{T}|-1$) in the test set using the following formulas:

$$P(z|t_{|\mathcal{T}|-1}) = \frac{\sum_{p\ \in t_{|\mathcal{T}|-1}} P(z|p\ for\ t_{|\mathcal{T}|-1})}{\#tweets\ for\ t_{|\mathcal{T}|-1}} \qquad (3)$$

$$P(z|p) = \sum_w P(z|w)P(w|p) = \sum_w \frac{n(z,w)}{n(w)} P(w|p) \qquad (4)$$

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_{tmatam}. Probability of word $p_l(w_i)$ for any document set is calculated as follows:

$$p_l(w_i) = \sum_z P(w|z)P(z) = \sum_z \frac{n(z,w)}{n(z)} P(z) \qquad (5)$$

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

3.2.1 *Comparing TM–ATAM with ATAM, TM–LDA and LDA*

Histograms in Figure 4 show the perplexity ratio of TM–ATAM with state-of-the art models:

- ATAM: 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 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 4, TM–ATAM beats ATAM in all US active regions (an active region is a region where the proportion of tweets if high enough). In Non-US active regions, the performance of TM–ATAM is affected negatively due to no substantial change in health topics with time.

- TM–LDA: Each region can be viewed as a *virtual user*. We merge the training data (same as TM–ATAM) of each *homogenous time period* in 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 4 shows that TM–ATAM consistently beats TM–LDA in predicting future health topics on the test month. Perplexity is indeed lower for all words of the test month in all active states.

- LDA: This shows the perplexity of LDA with the assumption that topics do not transition. LDA (as shown by [23]) is outperformed by TM–LDA which in turn is beaten by TM–ATAM.

3.2.2 *Homogenous Time Periods*

In Figure 5 we show the top-2 sharpest *change-points* for the top regions. Those points can be explained with weather changes in those regions. For example, 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 [23] to our settings:

$$Threshold = \mu + 2 \times \sigma_{non-diagonal} \quad (6)$$

We identify two kinds of interesting transitions based on the above threshold:

- stable-topics: popular topics which are identified by diagonal entries being above threshold,
- one-way-transitions: i^{th} topic is discussed before j^{th} topic, identified by ij^{th} above threshold.

As an example, stable topics of Arizona are summarized in Table 4 and one-way-transitions of California are summarized in Table 3.

4. RELATED WORK

Social media has been used for a wide array of tasks including mental health assessment [8,9,17], inferring political affiliation [1, 4, 5, 11], and brand perception [13, 14]. Previous studies on syndromic surveillance have attempted to uncover ailment topics in online discourse [6, 16] or model the evolution of general topics [23]. In this paper, we combine the best of both worlds which leads to the discovery of ***ailment-homogenous time periods*** for social-media active regions.

Our approach, TM-ATAM, builds on TM-LDA for modeling general topic evolution over time [23]. Just like TM-LDA, TM-ATAM learns topic transitions over time and not topic trends. Other complementary approaches that learn the dynamicity of word distributions or topic trends have been proposed such as [2, 15, 22].

In [10], the authors model topics and their sentiments over time. Topic, sentiment and time are considered random variables. However, since different timestamps may be generated for every word in a single document, a multi-nomial distribution for modeling time is adopted. This choice may not be the best to model time. In [20], documents are modeled as recurrent sequential patterns called motifs using a generative process. The contribution of the model is 3-fold: capture temporal order of words, detect recurrent and concurrent temporal activities and use a sparsity constraint. However, no prior distribution is used which gives LDA advantage over PLSA. In [22], the authors propose a method that learns changing word distributions of topics over time and in [15], the authors leverage the structure of a social network to learn how topics temporally evolve in a community. Finally, in [19], non-negative factorization is used for learning topic trends.

Exploring the applicability of those complimentary approaches to the evolution of health topics in tweets, is a promising research direction.

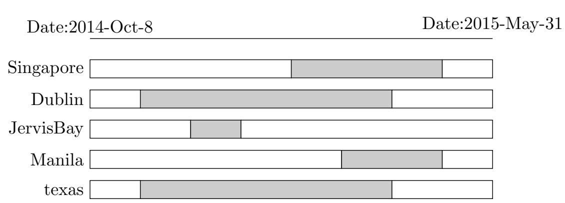

Figure 5: Top-2 Monthly *homogenous time period* for top active regions.

Table 3: One-Way Transitions for California (threshold: 0.815**)**

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

Table 4: stable-topics in Arizona (threshold: 0.035**)**

From Topic	To Topic	Weight
Stomach Infection	Stomach Infection	0.064
Headache	Headache	0.09
body pain	body pain	0.228

5. CONCLUSION AND FUTURE WORK

We study methods to uncover ailment distributions over time in social media. We propose a granularity-based model to conduct region-specific analysis that leads to the identification of time intervals characterizing homogeneous ailment discourse, per region. We model the temporal evolution of diseases within each homogeneous region and attempts to predict ailments. The fine-grained nature of our model results in significant improvements over state of the art methods. We envisage a use case where our ailment predictions can replace existing solutions for syndromic surveillance that are both costly and time consuming.

6. REFERENCES

[1] P. Barberá. Birds of The Same Feather Tweet Together: Bayesian Ideal Point Estimation using Twitter Data. *Political Analysis*, 23(1):76–91, 2015.

[2] D. M. Blei and J. D. Lafferty. Dynamic Topic Models. In *ICML'06*, pages 113–120, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning*, 3:993–1022, 2003.

[4] 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 *PLEAD'12*, pages 23–30. ACM, 2012.

[5] A. Ceron, L. Curini, and S. M. Iacus. Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters-Evidence from the United States and Italy. *Soc. Sci. Comput. Rev.*, 33(1):3–20, 2015.

[6] C. Chemudugunta, P. Smyth, and M. Steyvers. Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In *NIPS'06*.

[7] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[8] M. De Choudhury. Anorexia on Tumblr: A Characterization Study. In *DH'15*, pages 43–50, 2015.

[9] M. De Choudhury, A. Monroy-Hernández, and G. Mark. "narco" Emotions: Affect and Desensitization in Social Media During the Mexican Drug War. In *CHI'14*, pages 3563–3572, 2014.

[10] M. Dermouche, J. Velcin, L. Khouas, and S. Loudcher. A joint model for topic-sentiment evolution over time. In R. Kumar, H. Toivonen, J. Pei, J. Z. Huang, and X. Wu, editors, *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, pages 773–778. IEEE, 2014.

[11] L. Hemphill and A. J. Roback. Tweet Acts: How Constituents Lobby Congress via Twitter. In *CSCW'14*.

[12] T. Hofmann. Probabilistic Latent Semantic Indexing. In *SIGIR'99*, pages 50–57, 1999.

[13] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter Power: Tweets As Electronic Word of Mouth. *J. Am. Soc. Inf. Sci. Technol.*, 60(11):2169–2188, 2009.

[14] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent Twitter Sentiment Classification. In *HLT'11*, pages 151–160, 2011.

[15] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky. The Joint Inference of Topic Diffusion and Evolution in Social Communities. In *ICDM'11*, pages 378–387, 2011.

[16] M. J. Paul and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. In *ICWSM'11*, 2011.

[17] U. Pavalanathan and M. De Choudhury. Identity Management and Mental Health Discourse in Social Media. In *WWW'15*, pages 315–321, 2015.

[18] 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*, pages 18–25. Springer, 2011.

[19] A. Saha and V. Sindhwani. Learning Evolving and Emerging Topics in Social Media: A Dynamic nmf Approach with Temporal Regularization. In *WSDM'12*, pages 693–702, 2012.

[20] J. Varadarajan, R. Emonet, and J. Odobez. A sequential topic model for mining recurrent activities from long term video logs. *International Journal of Computer Vision*, 103(1):100–126, 2013.

[21] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, 2009.

[22] X. Wang and A. McCallum. Topics Over Time: A Non-Markov Continuous-time Model of Topical Trends. In *KDD'06*.

[23] Y. Wang, E. Agichtein, and M. Benzi. TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media. In *KDD'12*, pages 123–131, 2012.

[24] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In P. D. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Murdock, editors, *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, volume 6611 of *Lecture Notes in Computer Science*, pages 338–349. Springer, 2011.