International Journal of Research
Available at https://pen2print.org/index.php/ijr
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 01
January 2019

=

Monitoring human health condition on OSN Overtime

ABBURI RAJANI & P.SAILAJA RANI
1PG Scholar, Dept of CSE, Malineni Lakshmaiah Engineering College, Singarayakonda, Prakasam (Dt), AP, India.
2Assistant Professor, Dept of CSE, Malineni Lakshmaiah Engineering College, Singarayakonda, Prakasam (Dt), AP, India.

Abstract_

Social media has become a major source for analyzing all as-pects 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 effi-ciently 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 predic-tion error on topic distributions between consecutive posts at different time and geographic granularities. Our experi-ments on an 8-month corpus of tweets show that it largely outperforms its predecessors.

Index Terms—Public health, Ailments, Social media, Topic models.

I.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.

2. MODEL, PROBLEM AND APPROACH

Table 1 summarizes the terminology we use throughout this paper. By using suitable

=

geographic granularity *g* (coun-try, 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 limita-tions at discovering infrequent and specific topics such as health has already been shown [5]. The probabilistic *Ail-*

Table 1: Mapping tweets to documents

Term	Description
P	set of (tweet) posts
G	set of regions
T_t	set of time periods
^{P}t	
g	posts from region *g* during time *t*
^{D}g	document-set built by mapping the content
t	of each post $p \in P_g^t$ to a document
Θ_g	ailment distribution vector for document-set D_g^t of region *g* during time *t*
m	distance measure between distributions

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

ment Topic Aspect Model (ATAM) was designed specifically to uncover latent health-related topics present in a collection of tweets [5]. 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 1. 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 ef-fective at modeling health-related topics, it is not designed to model topic transitions over time.

2 DATA MODEL, TOPIC MODELS AND THE TRANSI-TION DETECTION PROBLEM

We present our data and define a model that maps tweet posts to documents of different time and geographic gran-ularities. 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 P = fp1; p2:::png. 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, < id; coord; tstamp; content >.

Let G = fg1; g2; :::g represent a set of geographic regions around the world. We use P_g to refer to the set of posts in P that originate from a region g 2 G. The choice of a geographic granularity (country, state, county) is required to instantiate G.

In a similar fashion, with a suitable choice of tempo-ral granularity, we could divide up the entire time range spanned by posts in P into disjoint and consecutive periods,

T = ft1; t2:::g. Possible choices for instantiation of T are week, bi-week, month, etc. We use P_g^t to refer to the set of posts in P that originated from a region g during period t.

=

We consider D_g^t the document formed by the concatena-tion of the content of all posts belonging to the set P_g^t.

We use D_g = fD_g^t1 ; D_g^t2 ; : : :g to denote the set of all documents corresponding to the aggregation of tweets from region g for different time periods in 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) un-

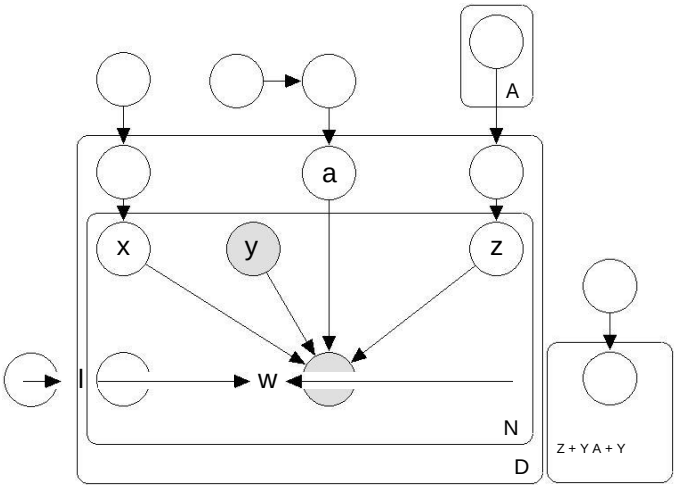

Fig. 3: Ailment Topic Aspect Model.

supervised generative models that describe the content of a document in a large collection D. In our case, 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 $_z$ over a set of words. LDA assumes that the topic distribution $_d$ of a document d and the vocabulary distribution $_z$ of a topic z are generated according to a Dirichlet distribution. Vectorial parameters and of these Dirichlet distributions are assumed to be common to the whole corpus.

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 col-lection of tweets [4]. The proposed method achieves re-markable improvements over LDA. Its novelty is that it distinguishes background words such as "home" and "watch-ing TV" from health-related words such as "hurts" and "al-lergy". For each document, these health-related words are considered to correspond to a unique ailment such as "obe-sity","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).

Figure 3 summarizes the process of ATAM. When generating a document (tweet), one first associates to it an ailment

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 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-point detection 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 time periods as well as health topical transitions.

3.3.1 Change-Point Detection with TM-ATAM

For each region g 2 G (Line 1), we first run ATAM over the full time period D_g (Line 2). Next for each period t 2 T (Line 3), we use the output of ATAM over D_g to generate ^t_g and deduce the ailment distribution $_g^t$ since we shall focus only on health-transitions (Lines 4– 12). Next, we examine the distance between consecutive distributions $_g^{t\ 1}$ and $_g^t$ 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, $[t_1\ t_{jT\ j}]$, 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 time-periods w.r.t ailments being discussed in Twitter. Quali-tatively, a homogeneous time period 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 dis-ease 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

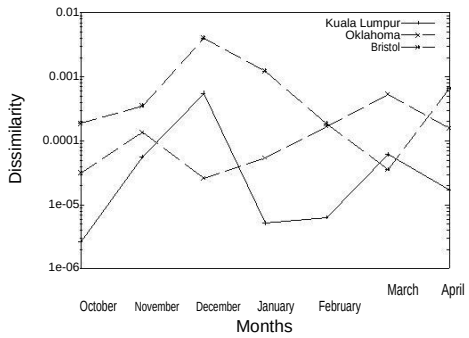

Fig. 5: Topic transitions over time.

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 Figure 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 Figure 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.

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 time period. This is a fresh de-parture from existing solutions that operate in a homogeneous time period-agnostic fashion. By defini-tion, a homogeneous time period is (nearly) homoge-neous in terms of ailments. In other words, the ailments evolve in a smooth fashion within a homogeneous time period 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.

We outline in Lines 17–21 of Algorithm 1 the steps un-dertaken. We use 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 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 .

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 $_g^t$ and $_g^{t+1}$ with the k times corresponding to the k-th top distances between consecutive vectors $_g^t$, $_g^{t+1}$ 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 $_g^t$ and $_g^{t+1}$ is above the threshold.

4.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.

4.1 Setup

4.1.1 Data

We employ Twitter's Streaming API to collect tweets be-tween 2014-Oct-8 and 2015-May-31. We use the Decahose

7

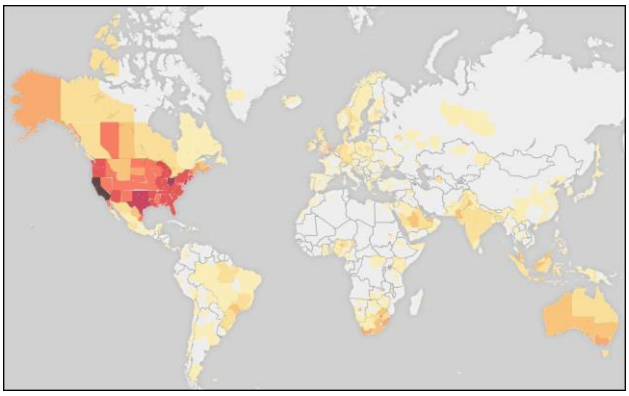

Fig. 7: Heatmap over collected health tweets. A major fraction of the tweets originate from various states in the US.

TABLE 2: Dataset Statistics

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

Stream[3] which gives a 10% random sample of the total tweets generated each day. The collected tweets were sub-jected 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 mes-sages about a user's health.) Since our interest lies in public health discourse on social media, we only keep tweets con-taining 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 1.36B tweets we collected, 698K were health-related.

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 (569K as indicated in Table 2). In Figure 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.

5.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 identifica-tion 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.

References

[1] L. Manikonda and M. D. Choudhury, "Modeling and understand-ing visual attributes of mental health disclosures in social media," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017., 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 clas-sifying crisis-related messages," in 10th Proceedings of the Inter-national Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, May 12-15, 2013., 2013.

[3] T. Davidson, D. Warmsley, M. W. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017., 2017, pp. 512–515.

[4] M. J. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health," in ICWSM'11, 2011.

[5] T. Hofmann, "Probabilistic Latent Semantic Indexing," in SI-GIR'99, 1999, pp. 50–57.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Alloca-tion," Journal of Machine Learning, 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 KDD'12, 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 Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016, 2016, pp. 849–852.

[9] D. M. Blei and J. D. Lafferty, "Dynamic Topic Models," in ICML'06, 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 Communi-ties," in ICDM'11, 2011, pp. 378–387.

[11] X. Wang and A. McCallum, "Topics Over Time: A Non-Markov Continuous-time Model

=

of Topical Trends," in KDD'06, 2006, pp. 424–433.

Malineni Lakshmaiah Engineering College, Singarayakonda, Prakasam(Dt), AP, India.

Author's Profile

A.Rajani Persuing Studying M.Tech in Malineni Lakshmaiah Engineering College, Singarayakonda, Prakasam(Dt), AP,India.

P.Sailaja Rani
Currently working as Assistant professor in Malineni Lakshmaiah Engineering College, Singarayakonda, Prakasam(Dt), AP, India.

She is Highly Passionate and Enthusiastic about Her Teaching and Believes that Inspiring Students to Give of Her Best in Order to Discover What He Already Knows is Better Than Simply Teaching..She is having 10+ experience in teaching field currently working as assistant professor in