Performance of Public Health Monitoring on Social Media

[1]H.R.Vyshnavi, [2] A.Ravindra Kumar
[1]PG Scholar, Department of CSE, Kuppam Engineering College, Kuppam, Chittoor Dt. A.P.
[2] Associate Professor, Department of CSE, Kuppam Engineering College, Kuppam, Chittoor Dt. A.P.

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.**

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. 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 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

This paper is organized in five sections. After this introduction, in Section II, literature survey discussed of the paper, section III about the System Analysis, Section IV about System Design, as well as the novel feature of the proposed method. Finally, Sections V and VI provide the simulation results and the conclusions, respectively.

II. LITERATURE SURVEY

A. *Automated hate speech detection and the problem of offensive language, T. Davidson, D. Warmsley, M. W. Macy, and I. Weber*

A key challenge for automatic hate-speech detection on social media is the separation of hate speech from other instances of offensive language. Lexical detection methods tend to have low precision because they classify all messages containing particular terms as hate speech and previous work using supervised learning has failed to distinguish between the two categories. We used a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. We use crowd-sourcing to label a sample of these tweets into three categories: those containing hate speech, only offensive language, and those with neither. We train a multi-class classifier to distinguish between these different categories. Close analysis of the predictions and the errors shows when we can reliably separate hate speech from other offensive language and when this differentiation is more difficult. We find that racist and homophobic tweets are more likely to be classified as hate speech but that sexist tweets are generally classified as offensive. Tweets without explicit hate keywords are also more difficult to classify.

B. *TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media,Y. Wang, E. Agichtein, and M. Benzi*

Latent topic analysis has emerged as one of the most effective methods for classifying, clustering and retrieving textual data. However, existing models such as Latent Dirichlet Allocation (LDA) were developed for static corpora of relatively large documents. In contrast, much of the textual content on the web, and especially social media, is temporally sequenced, and comes in short fragments, including microblog posts on sites such as Twitter and Weibo, status updates on social networking sites such as Facebook and LinkedIn, or comments on content sharing sites such as YouTube. In this paper we propose a novel topic model, Temporal-LDA or TM-LDA, for efficiently mining text streams such as a sequence of posts from the same author, by modeling the topic transitions that naturally arise in these data. TM-LDA learns the transition parameters among topics by minimizing the prediction error on topic distribution in subsequent postings. After training, TM-LDA is thus able to accurately predict the expected topic distribution in future posts. To make these predictions more efficient for a realistic online setting, we develop an efficient updating algorithm to adjust the topic transition parameters, as new documents stream in. Our empirical results, over a corpus of over 30 million microblog posts, show that TM-LDA significantly outperforms state-of-the-art static LDA models for estimating the topic distribution of new documents over time. We also demonstrate that TM-LDA is able to highlight interesting variations of common topic transitions, such as the differencesin the work-life rhythm of cities, and factors associated with area-specific problems and complaints.

C. *Health monitoring on social media over time, S. Sidana, S. Mishra, S. Amer-Yahia, M. Clausel, and M. Amini*

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 overtime. 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.

D.The Joint Inference of Topic Diffusion and Evolution in Social Communities C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky

The prevalence of Web 2.0 techniques has led to the boom of various online communities, where topics spread ubiquitously among user-generated documents. Working together with this diffusion process is the evolution of topic content, where novel contents are introduced by documents which adopt the topic. Unlike explicit user behavior (e.g., buying a DVD), both the diffusion paths and the evolutionary process of a topic are implicit, making their discovery challenging. In this paper, we track the evolution of an arbitrary topic and reveal the latent diffusion paths of that topic in a social community. A novel and principled probabilistic model is proposed which casts our task as an joint inference problem, which considers textual documents, social influences, and topic evolution in a unified way. Specifically, a mixture model is introduced to model the generation of text according to the diffusion and the evolution of the topic, while the whole diffusion process is regularized with user-level social influences through a Gaussian Markov Random Field. Experiments on both synthetic data and real world data show that the discovery of topic diffusion and evolution benefits from this joint inference, and the probabilistic model we propose performs significantly better than existing methods

E. Anorexia on Tumblr: A Characterization Study, M. De Choudhury

Eating disorders, such as anorexia nervosa are a major health concern affecting many young individuals. Given the exten-sive adoption of social media technologies in the anorexia affected demographic, we study behavioral characteristics of this population focusing on the social media Tumblr. Aligned with observations in prior literature, we find the presence of two prominent anorexia related communities on Tumblr-pro-anorexia and pro-recovery. Empirical analy-ses on several thousand Tumblr posts show use of the site as a media-rich platform replete with triggering content for enacting anorexia as a lifestyle choice. Through use of com-mon pro-anorexia tags, the prorecovery community however attempts to \permeate" into the pro-anorexia community to educate them of the health risks of anorexia. Further, the communities exhibit distinctive affective, social, cogni-tive, and linguistic style markers. Compared with recover-ing anorexics, pro-anorexics express greater negative affect, higher cognitive impairment, and greater feelings of social isolation and selfharm. We also observe that these character-istics may be used in a predictive setting to detect anorexia content with ~80% accuracy. Based on our findings, clinical implications of detecting anorexia related content on social media are discussed.

III. SYSTEM ANALYSIS

A. Existing System

Existing, ATAM is effective at modelling health-related topics; it is not designed to model topic transitions over time. Also it does not formalize temporal and geographic granularity in the model. While pLSI and LDA have been shown to perform well on static documents, they cannot intrinsically capture topic evolution over time. LDA and ATAM formalize our health transition detection and prediction problems.

Dis-advantages:

- In the existing, the health transition detection is difficult because ATAM models don’t have temporal transitions of health-related topics. LDA represents each document as a probability distribution over topics.
- Ailment Topic Aspect Model (ATAM) was designed specifically to uncover latent health-related topics in a collection of tweets

B. Proposed System

We Propose 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. 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. 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.

.Advantages

- 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 modelling and predicting transitions of health-related tweets.

IV. SYSTEM DESIGN

A. System Architecture

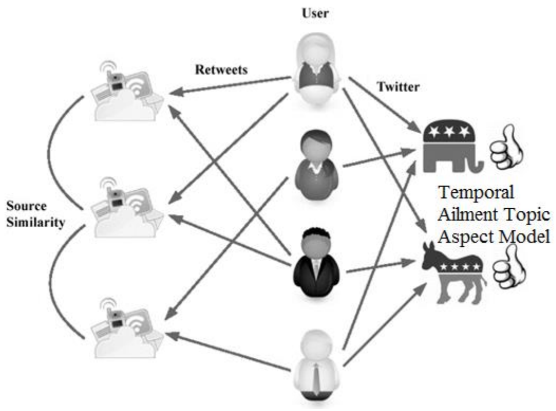


Fig. 1: Overview of the proposed approach,

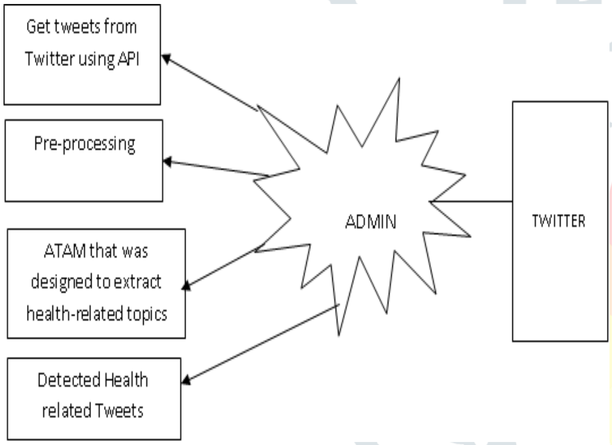

Get tweets from Twitter using API
Pre-processing
ATAM that was designed to extract health-related topics
Detected Health related Tweets
ADMIN
TWITTER

Fig .2: Block diagram

A. System Construction

In the first module we develop the System Construction module, to Health Monitoring on Social Media over Time. For this purpose we develop User entities. In User entity, a user can search information about Health related tweets in Twitter OSN. A user can able to search other user tweets about health related. A user can mine the information in big data like Twitter OSN about the health tweets. It minimizes the time and increase accuracy of results to know about the health tweets. A user can also view the other user id who tweet about the health. A user can able to search tweets tweeted by users about.

B. Pre-processing Tweets

In this module we implement about Preprocessing Tweets. Our technical contribution is to frame health leaning inference as a convex optimization problem that jointly maximizes tweet-retweet agreement with an error term, and user similarity agreement with a regularization term which is constructed to also account for heterogeneity in data. Our technique requires only a steady stream of tweets but not the Twitter social network, and the computed scores have a simple interpretation of "averaging," i.e., a score is the average number of positive/negative tweets expressed when re-tweeting the target user.

.

C. Detection with TM-ATAMs

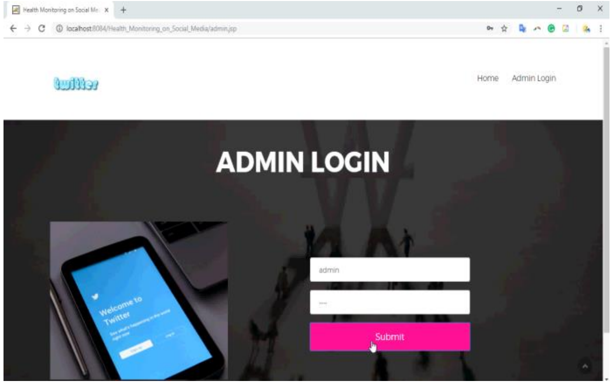
In this module, our objective is to model ailment transitions, that is potential change in time of the health topical content of our tweets. We do so by introducing a new model, TM–ATAM that we do in this module. This model is derived from TM–LDA. TM–ATAM, we need to do post processing in order to come up with homogeneous time periods, with respect to health-topics discussed in tweets. 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 are no substantial health topic transitions.

D. Quantitative Study

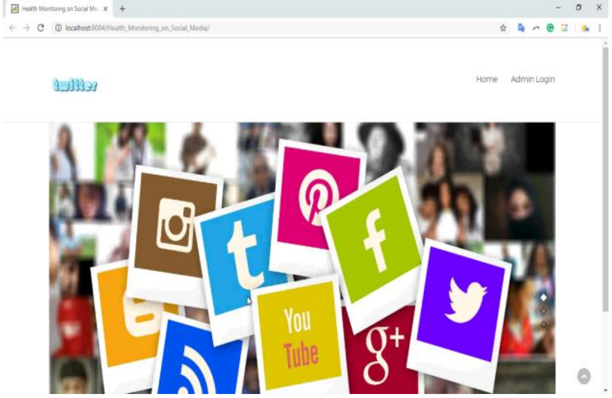
In this module we study the properties of the health leaning most popular tweets sources. The score histogram on the full set has a bimodal distribution. As an event is happening, the influx of Twitter users participating in the discussion makes the active population more liberal and less polarized.We employ Twitter's Streaming API to collect tweets. Since our interest lies in public health discourse on social media, we only keep tweets containing health-related keywords obtained from tweets. We focused on high precision as high quality health tweets is a pre-requisite for both TM–ATAM and T–ATAM to function efficiently.

V SIMULATION RESULTS

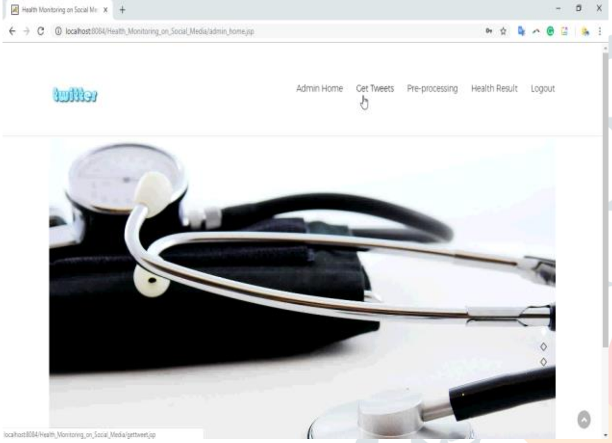
- ***Admin Login***

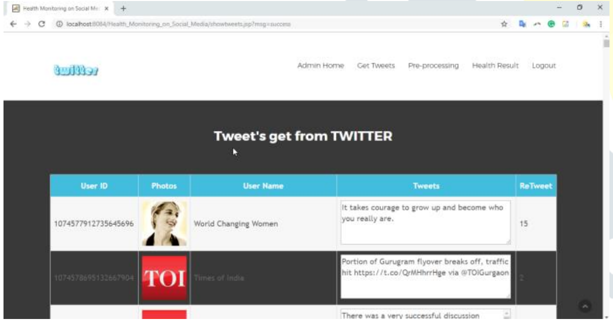


admin


Health Monitoring on Social Me
localhost:8084/Health_Monitoring_on_Social_Media/
twitter
Home
Admin Login
You Tube

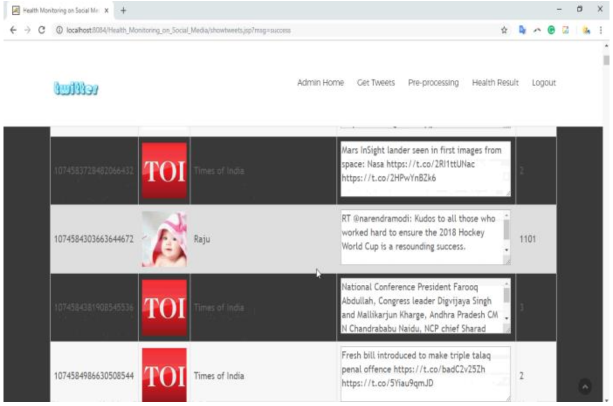
- ***Gets Tweets***


Health Monitoring on Social Me
localhost:8084/Health_Monitoring_on_Social_Media/admin_home.jsp
twitter
Admin Home
Get Tweets
Pre-processing
Health Result
Logout
localhost:8084/Health_Monitoring_on_Social_Media/gettweet.jsp

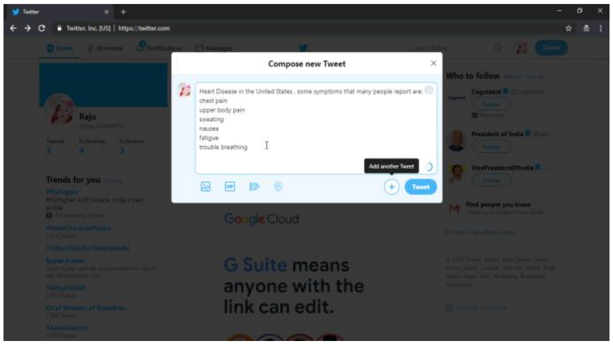

Health Monitoring on Social Me
localhost:8084/Health_Monitoring_on_Social_Media/showtweets.jsp?msg=success
twitter
Admin Home
Get Tweets
Pre-processing
Health Result
Logout

Tweet's get from TWITTER

User ID	Photos	User Name	Tweets	ReTweet
1074577912735645696		World Changing Women	It takes courage to grow up and become who you really are.	15
1074578695132667904	TOI	Times of India	Portion of Gurugram flyover breaks off, traffic hit https://t.co/QrMHhrrHge via @TOIGurgaon	2
			There was a very successful discussion	

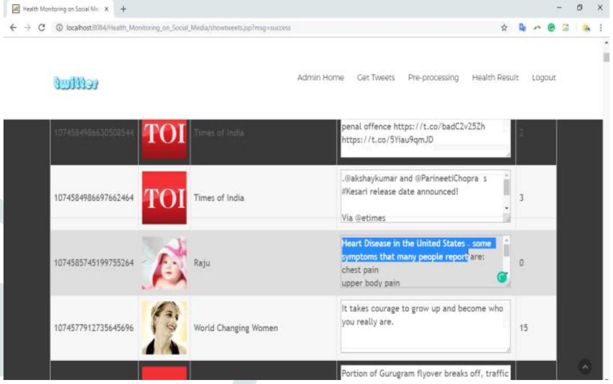

Health Monitoring on Social Me
localhost:8084/Health_Monitoring_on_Social_Media/showtweets.jsp?msg=success
twitter
Admin Home
Get Tweets
Pre-processing
Health Result
Logout

1074583728482066432	TOI	Times of India	Mars InSight lander seen in first images from space: Nasa https://t.co/2RI1ttUNac https://t.co/2HPwYnBZk6	2
1074584303663644672		Raju	RT @narendramodi: Kudos to all those who worked hard to ensure the 2018 Hockey World Cup is a resounding success.	1101
1074584381908545536	TOI	Times of India	National Conference President Farooq Abdullah, Congress leader Digvijaya Singh and Mallikarjun Kharge, Andhra Pradesh CM N Chandrababu Naidu, NCP chief Sharad	3
1074584986630508544	TOI	Times of India	Fresh bill introduced to make triple talaq penal offence https://t.co/badC2v25Zh https://t.co/5Yiau9qmJD	2

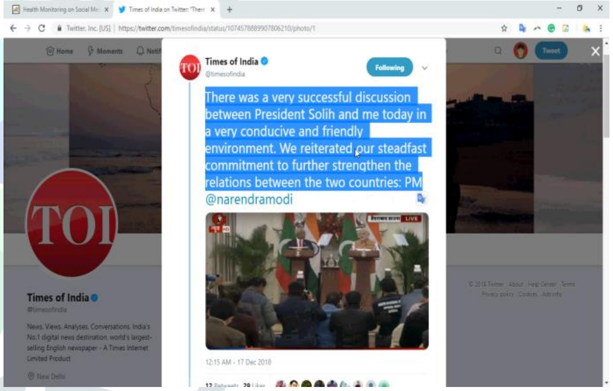


Compose new Tweet

Heart Disease in the United States . some symptoms that many people report are:
chest pain
upper body pain
sweating
nausea
fatigue
trouble breathing


Health Monitoring on Social Me
localhost:8084/Health_Monitoring_on_Social_Media/showtweets.jsp?msg=success
twitter
Admin Home
Get Tweets
Pre-processing
Health Result
Logout

1074584986630508544	TOI	Times of India	penal offence https://t.co/badC2v25Zh https://t.co/5Yiau9qmJD	2
1074584986697662464	TOI	Times of India	.@akshaykumar and @ParineetiChopra s #Kesari release date announced! Via @etimes	3
1074585745199755264		Raju	Heart Disease in the United States . some symptoms that many people report are: chest pain upper body pain	0
1074577912735645696		World Changing Women	It takes courage to grow up and become who you really are.	15
			Portion of Gurugram flyover breaks off, traffic	

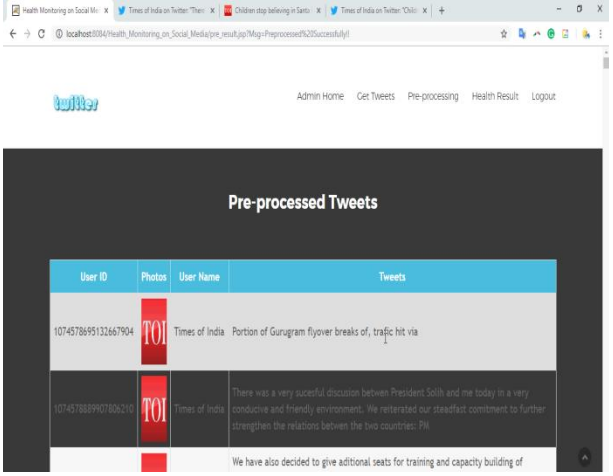

Health Monitoring on Social Me
Times of India on Twitter: "There
Twitter, Inc. [US] | https://twitter.com/timesofindia/status/1074578889907806210/photo/1
Home
Moments
Following
TOI
LIVE

Times of India
@timesofindia

There was a very successful discussion between President Solih and me today in a very conducive and friendly environment. We reiterated our steadfast commitment to further strengthen the relations between the two countries: PM @narendramodi

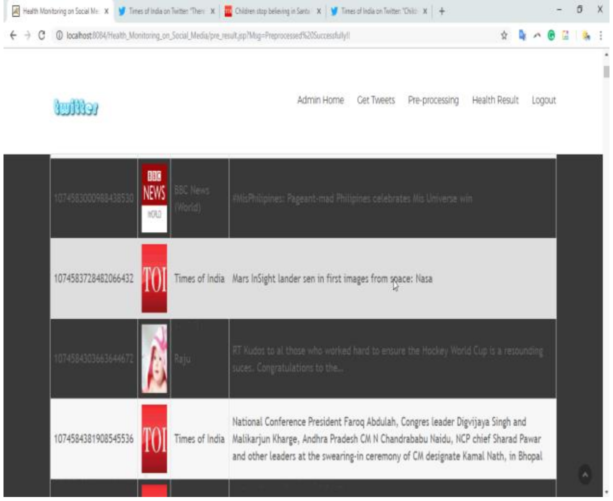
12:15 AM - 17 Dec 2018

- ***Pre-Processing***


Health Monitoring on Social Me
Times of India on Twitter: "There
Children stop believing in Santa
Times of India on Twitter: "Child
localhost:8084/Health_Monitoring_on_Social_Media/pre_result.jsp?Msg=Preprocessed%20Successfully!!
twitter
Admin Home
Get Tweets
Pre-processing
Health Result
Logout

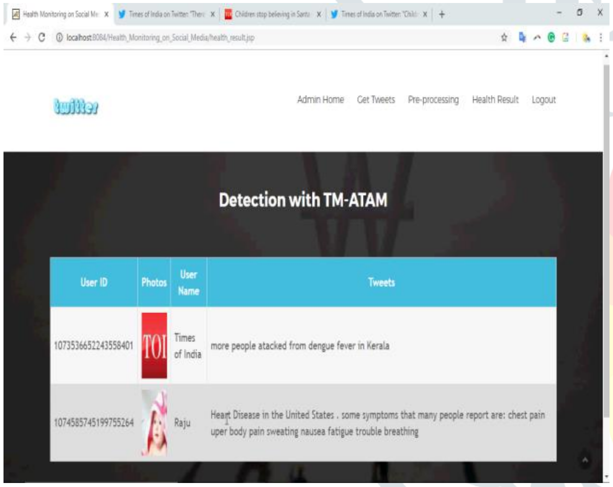
Pre-processed Tweets

User ID	Photos	User Name	Tweets
1074578695132667904	TOI	Times of India	Portion of Gurugram flyover breaks of, trafic hit via
1074578889907806210	TOI	Times of India	There was a very sucesful discusion betwen President Solih and me today in a very conducive and friendly environment. We reiterated our steadfast comitment to further strengthen the relations betwen the two countries: PM
			We have also decided to give aditional seats for training and capacity building of


Health Monitoring on Social Me
Times of India on Twitter: "Ther
Children stop believing in Santa
Times of India on Twitter: "Chil
localhost:8084/Health_Monitoring_on_Social_Media/pre_result.jsp?Msg=Preprocessed%20Successfully!!
twitter
Admin Home
Get Tweets
Pre-processing
Health Result
Logout

1074583000988438530	BBC NEWS WORLD	BBC News (World)	#MisPhilipines: Pageant-mad Philipines celebrates Mis Universe win
1074583728482066432	TOI	Times of India	Mars InSight lander sen in first images from space: Nasa
1074584303663644672		Raju	RT Kudos to al those who worked hard to ensure the Hockey World Cup is a resounding suces. Congratulations to the...
1074584381908545536	TOI	Times of India	National Conference President Faroq Abdulah, Congres leader Digvijaya Singh and Malikarjun Kharge, Andhra Pradesh CM N Chandrababu Naidu, NCP chief Sharad Pawar and other leaders at the swearing-in ceremony of CM designate Kamal Nath, in Bhopal

- ***Result Generation***


Health Monitoring on Social Me
Times of India on Twitter: "Ther
Children stop believing in Santa
Times of India on Twitter: "Chil
localhost:8084/Health_Monitoring_on_Social_Media/health_result.jsp
twitter
Admin Home
Get Tweets
Pre-processing
Health Result
Logout

Detection with TM-ATAM

User ID	Photos	User Name	Tweets
1073536652243558401	TOI	Times of India	more people atacked from dengue fever in Kerala
1074585745199755264		Raju	Heart Disease in the United States . some symptoms that many people report are: chest pain uper body pain sweating nausea fatigue trouble breathing

VI 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 insignificant 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.

REFERENCESS

[1] L. Manikonda and M. D. Choudhury, "Modeling and understandingvisual attributes of mental health disclosures in social media,"in Proceedings of the 2017 CHI Conference on Human Factors inComputing Systems, Denver, CO, USA, May 06-11, 2017., 2017, pp.170–181.

[2] S. R. Chowdhury, M. Imran, M. R. Asghar, S. Amer-Yahia, andC. Castillo, "Tweet4act: Using incident-specific profiles for classifying crisis-related messages," in 10th Proceedings of the International 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," inProceedings 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 SIGIR'99, 1999, pp. 50–57.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation,"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, July17-21, 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 JointInference of Topic Diffusion and Evolution in Social Communities,"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.

[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. Springer, 2011, pp.18–25.

[13] C. Cortes and V. Vapnik, "Support-vector networks," MachineLearning, 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 DH'15, 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 CHI'14, 2014, pp. 3563–3572.