A Survey on Health Monitoring on Social Media Over Time

Jacob Thaliyan[1], Ancy Mariya Jacob[2], Anju Shaju[3], Rinto Antony[4], Sandra Biju[5]
[1, 2, 3, 4, 5] Dept of Computer Science
[1, 2, 3, 4, 5]De Paul Institute of Science and Technology, Angamaly

Abstract- *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 , disaster management and politics .We are applying the Temporal Ailment Topic Aspect Model (TM–ATAM), a new latent model dedicated to solving the health-related topics. TM–ATAM is a non-obvious extension to ATAM that was designed to extract health-related topics.We conduct experiments to evaluate the performance of TM–ATAM and T–ATAM on real world data. The experimental setup including the datasets and test-bench.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 and a qualitative analysis of TM–ATAM's results.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 for the influenza rates in US. Finally, we highlight the key insights drawn from our experiments.*

Keywords- Data, Filtering health related tweets, Geolocation, Test bench

I. INTRODUCTION

Thanks to dedicated latent topic analysis methods such as the Ailment Topic Aspect Model, public health can now be observed on Twitter.We have systematically detected two problems: health transition detection and health transition prediction.

For solving the first problem we put forward a new latent model that captures transitions involving health-related topics called, TM-ATAM.TM-ATAM(Temporal Ailment Topic Aspect Model) is a non-obvious extension to ATAM(Ailment Topic Aspect Model) that was designed to extract health-related topics.It gains an understanding of health-related topic transitions by reducing the prediction error on topic distributions between consecutive posts at different time and geographic locations.

To solve the second problem, we develop T-ATAM (Temporal Ailment Topic Aspect Model) where time is considered as a random variable natively inside ATAM. Our observations on an 8-month data set of tweets show that TM-ATAM overtakes TM-LDA in calculating health-related transitions from tweets for different geographic populations.

We observe the effect of climate conditions in different geographic regions on the ability of TM-ATAM to detect transitions. 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 particular, we find that prediction accuracy for health topics is higher when operating TM-ATAM on finer spatial granularity and shorter time periods. Further, we go on to discover interesting region-specific intra and inter-homogeneous time period health related transitions.

While studying these transitions, we find that homogeneous time periods are continuous time periods for which people in the same region tweet about similar health issues. These results show that it is more logical to predict future ailments concerning people within the same homogeneous time period of a region than on any random health tweets.

Since in T-ATAM, time is considered a random variable following multinomial distribution, we expect it to outperform other models, TM-LDA and TM-ATAM in predicting health topics using perplexity measure.

According to our expectations, in most social-media active regions, in both US active regions and non-US active regions, T-ATAM outperforms TM-ATAM and ATAM. After analyzing T-ATAM's performance by changing various spatio-temporal parameters, we find that the prediction accuracy for health topics is higher when operating T-ATAM on finer spatial granularity and shorter time periods.

II. LITERATURE REVIEW

In this paper [1], We propose the recently introduced Ailment Topic Aspect Model to over one and a half million health related tweets and uncovered abundance of ailments, including allergies, obesity and insomnia.

This paper enquires: what public health information can be learned from Twitter? While previous studies are cocentrating on specific questions with specific models, we ask an open question with a general model: the newly introduced Ailment Topic Aspect Model.

From our perpective, this is the first implementations to use social media as a sources for public health informatics on a range of ailments, rather than a narrow set of applications on one more ailments. Public Health Informatics and the Web Syndromic surveillance, the observation on clinical malady that have notable influence on public health, impacts medical resource allocation, health policy and education.

While ATAM uncovers meaningful ailments, there are other public health resources that could be used to improve model output.To evaluate the output inferability, we compare the ailment distributions with distributions estimated from WebMD articles.

What Can be Learned from Twitter? Given our examinations suggesting that Twitter includes valuable knowledge, we can't help but wonder wheather Twitter can help us gain public health informations. With ATAM , we are observing several facets of public health, including temporal and geographic impacts on medical well being and more.

To differetiate, we measured the correlation by normalizing using just the health related tweets, which considers the percentage of health related messages that discuss the flu. For example: are users in some geographic areas more likely to exercise than others? We formulated a list of questions based on the behavioral risk factor surveillance system, run by the National Center for Chronic Disease Prevention and Health Promotion at the CDC. For each data set that could potentially be measured with one or more ATAM ailments, we measured the Pearson correlation coefficient between the ailments discovered in each US state with the state's risk factor rate.

Geographic Syndromic Surveillance So far we have demonstrated the ability to mine public health information both over time and by geographic region We now seek to combine these to track an ailment over time and geography.

Our research study on using Twitter to extract public health information focused on producing data that correlates with public health criterions and knowledge.

Here[2], We discuss the relationship of our findings to literature in visual sociology, in mental health self-disclosure, and implications for the design of health interventions. Social media have emerged to be instrumental means of social exchange and support seeking around stigmatized concerns like mental health. Specifically, we address the following two research questions: What visual features characterize images of mental health disclosures shared on social media? How do visual themes of mental health images complement and contrast with themes manifested in the language of these social media posts?

To address these questions, we provide a large dataset of over millions of public posts associated with ten mental health challenges shared on Instagram.

We additionally consulted the Diagnostic and Statistical Manual of Mental Health Disorders, that indicates these disorders to be prominent mental health challenges in populations.

We corroborated these observations with a licensed psychiatrist, and concluded that the users in our dataset are engaging in genuine mental health disclosures, tend to demonstrate disinhibition towards sharing their mental health experiences, and are appropriating the platform specifically for this purpose via the chosen account.

The researchers adopted a semi-open coding approach, borrowing from the literature on mental health self-disclosure and recent work in characterizing mental health images shared on different social media platforms. The annotators first independently coded all of the 20 clusters.

We believe our approach and findings can influence the design of new health interventions that leverage the rich information embedded in visual imagery of mental health disclosures.

In this paper[3], We report results in several kind of formats :document modeling, text classification, and collaborative filtering, comparing to a mixture of unigrams model and the probabilistic LSI model. To uphold the assertions regarding LSI, and to observe its relative strengths as well as weaknesses, it is useful to develop a generative probabilistic model of text corpora and to study the ability of LSI to rec given gover aspects of the generative model from data.

Stated a innovative model of text it is not clear why one should acquire the latent semmantic indexig approach one can attempt to proceed more directly, fitting the model to data using maximum likelihood .

Hoffmann who presented the probabilistic LSI model, also known as the aspect model, as an alternative to LSI presented was a significant step forward in this regard. The LSI approach, models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of contents.

Subsequently, if we want to consider covertible representations, we need to consider mixture models that records the exchangeability of both words and documents .

In contrast, the Latent Dirichlet allocation model allows documents to exhibit multiple topics to different degrees. Latent Dirichlet allocation is a well-defined generative model and generalizes easily to new documents. One of the advantages of Latent Dirichlet allocation over related latent variable models is that it provides well-defined inference procedures for previously unseen documents.

We trained a number of latent variable models, including Latent Dirichlet allocation, on two text corpora to compare the generalization performance of these models.

We could also consider partially exchangeable models in which we condition on exogenous variables; thus, for example, the topic distribution could be conditioned on features such as paragraph and sentence that provides a more powerful text model that makes use of information obtained from a parser.

In this paper[4],Latent Dirichlet allocation(LDA),is a technique used for the problem of modelling text corpora and other collections of discrete data.it is a generative probabilistic model.The objective of the model is to find short descriptions of the members in a data set that enable efficient processing of large collections. Also for preserving the important statistical relationships that are used for basic tasks such as classification, novelty detection,summarization,and similarity and relevance judgements.LDA is a three-level hierarchical Bayesian model,in which each of the element in a data set is formed as finite mixture over underlying set of topics.In text modeling the topic probabilities provide an explicit representation of a document.

The assumption of exchangeability is not equivalent to an assumption that the random variables are independent and identically distributed. Instead compatibility can substantially converted as meaning conditionally independent and identically distributed where the conditioning is with respect to an underlying latent parameters of a probability distribution. Conditionally the joint distributed random variables are the simple and factored. The joint distribution can be quite complex.Thus ,while the assumption of exchangeability is clearly a major simplifying assumptions do not necessarily lead to methods that are restricted to simple frequency counts or linear operations.

Here[5] ,The problem needing to be addressed can be formally named as Target-dependent Sentiment Classification of Tweets; namely, given a query, classifying the sentiments of the tweets as positive, negative or neutral consistent with whether or not they contain positive, negative or neutral sentiments that query.

Because people may mention multiple targets in one tweet or discuss a target during a tweet while saying many other unrelated things within an equivalent tweet, target-independent approaches are likely to yield unsatisfactory results: We can easily find many related tweets of a given tweet, like the tweets published by an equivalent person, the tweets replying to or replied by the given tweet, and retweets of the given tweet.

These related tweets provide rich information about what the given tweet expresses and can definitely be taken into consideration for classifying the sentiment of the given tweet.In the experiments, we consider the positive and negative tweets annotated by humans as subjective tweets, which amount to 727 tweets.Following, we balance the evaluation data set by randomly selecting 727 tweets from all neutral tweets annotated by humans and consider them as objective tweets.

Different from previous work using only information on the present tweet for sentiment classification, we propose to require the related tweets of the present tweet into consideration by utilizing graph-based optimization.

In this paper[6], A set of probabilistic time series models is developed to analyze the time evolution of topics in large document collections.

To state space models on the natural parameters of the multinomial distributions that represent the topics this approach can be used.Variational approximations supported Kalman filters and nonparametric wavelet regression are developed to hold out approximate posterior inference over the latent topics.

Dynamic topic models provide a qualitative window into the contents of an outsized document collection as an addition to giving quantitative, predictive models of a sequential corpus.

The models are signified by examine the Optical Character Reader archives of the journal Science from 1880 through 2000. These models are called "Topic models" because the discovered patterns often reflect the underlying topics which combined to make the documents.The mixing proportions are randomly drawn for each document; the mixture components, or topics, are shared by all documents.

These models are a strong method of dimensionality reduction for giant collections of unstructured documents.For many collections of interest the implicit assumption of exchangeable documents is inappropriate.In this paper, we develop a dynamic topic model which captures the evolution of topics during a sequentially organized corpus of documents.

Under this model, articles are grouped by year, and each year's articles arise from a set of topics that have evolved from the last year's topics. Then we develop systematic approximate posterior inference techniques for discovering the evolving topics from a sequential collection of documents. As a conclusion, we present qualitative results that demonstrate how dynamic topic models allow the exploration of an outsized document collection in new ways, and quantitative results that demonstrate more accuracy in predictions when compared with static topic models.

Here[7], Since introduction in 2006 ,the Twitter website has become popular hat it is currently ranked as the 10th most visited site over the world.Tweets exchanged over the internet are an important source of information even if their characteristics make them difficult to analyse.In this paper discuss the the problem of extracting relevant topics through tweets coming from different communities.And address the problem to select specific keywords for different communities over Tweet social media.The best measures to evaluate the most relevant terms for a specific community is TF-IDF,Okapi-BM25 which are the statistics have been proposed by the Information Retrieval or The Text Mining fields to extract the most representative words in documents.

People participating in on-line forums, micro blogging or discussing on social networks leave behind them digital traces of their opinion on a variety of topics. If we knew how to aggregate and cumulatively interpret this data.

An additional benefit of the applications is that they deliver the pulse of the community not only to decision makers,but to the community members themselves and will likely become one of the tools of e-democracy.

Mainly paper focus on the statistical measure TF-IDF weight.We can integrate this type of weight to enhance matrix representation of tweet data . After applying this kind of weight for tweets features,a process based on Latent Semantic Analysis can be performed.The result will be a compressed version of the original matrix of textual corpus.The weight proposed is useful to predict the community of the tweet.

Here[8], Content shared on social media platforms has been identified to be valuable in gaining insights into people's mental health experiences. Although there has been widespread adoption of photo-sharing platforms such as Instagram in recent years, the role of visual imagery as a mechanism of self-disclosure is less understood.We study the nature of visual attributes manifested in images relating to mental health disclosures on Instagram.

Employing computer vision techniques on a corpus of thousands of posts, we extract and examine three visual attributes: visual features, themes, and emotions in images.

Our findings indicate the use of imagery for unique self-disclosure needs, quantitatively and qualitatively distinct from those shared via the textual modality: expressions of emotional distress, calls for help, and explicit display of vulnerability. We discuss the relationship of our findings to literature in visual sociology, in mental health self-disclosure, and implications for the design of health interventions.The rich literature in visual sociology situates imagery to be a strong means of enabling emotional expression associated with mental illnesses, especially those feelings and experiences that individuals may struggle to express verbally or through written communication.

Given these considerations, sharing and reflecting on visual narratives are a known psychiatric approach to tackle mental illness.Specifically, we address the following three research questions: What visual features characterize images of mental health disclosures shared on social media? What are the sorts of visual themes manifested in these images, and what's the character of emotional expression related to these visual themes? How do visual themes of mental health images complement and contrast with themes manifested in the language of these social media posts? To address these research questions, we leverage an outsized dataset of over two million public posts related to ten psychological state challenges shared on Instagram.

Our findings indicate the prominence of a visual channel supporting candid and disinhibited social exchange

around mental health.Specifically, we find that the visual and emotional markers of mental health images capture unique characteristics of self-disclosure, beyond those expressed via the sharing of linguistic content.We situate our findings in literature on visual sociology and the role of visual narratives in mental well-being.We discuss how our work can inspire further research on visual cues of psychological state disclosures on social media.

III. 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 unravel them. These transition detections are corrected with TM–ATAM, a granularity-based model for conducting region-specific analysis that results in the identification of some 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 leads to significant 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.

REFERENCES

[1] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model," in NIPS'06, 2006, pp. 241–248.
[2] 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. ACM, 2012, pp. 23–30.
[3] D. M. Blei and J. D. Lafferty, "Dynamic Topic Models," in ICML'06, 2006, pp. 113–120.
[4] T. Hofmann, "Probabilistic Latent Semantic Indexing," in SIGIR'99, 1999, pp. 50–57.
[5] D. M. Blei and J. D. Lafferty, "Dynamic Topic Models," in ICML'06, 2006, pp. 113–120.
[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] L. Manikonda and M. D. Choudhury, "Modeling and understanding 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.
[8] M. J. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health," in ICWSM'11, 2011.