Spatio-temporal Public Health Analysis and its Ethical Concerns

Debjyoti Paul
University of Utah
deb@cs.utah.edu

INTRODUCTION

Many research has revealed that analyzing tweets in volume can measure different population characteristics including public health measures [2, 8, 13, 17, 21, 22]. Research analysis like correlating influenza rates w.r.t geography (spatial) and time [23], state level food and health behavior analysis [18], predicting heart disease rate mortality rate based on twitter information [8]; are motivating examples to carry out such analysis for improving and building a good public health environment. All these above adhoc analysis inspire us to build a general system for comprehensive analysis. In this work, I will present an overview of the architechture and the desired features to build such system or tools.

1 PART 1A) SYSTEM ARCHITECTURE SPATIO-TEMPORAL ANALYSIS FOR HEALTH ANALYSIS:

A comprehensive system for spatio-temporal analysis requires the following components which can be braodly categorized based on their operations:

- **Data Ingestion**
  - *Data Collection Module*
- **Data Enrichment**
  - *Data Cleaning Module*
  - *Location Extraction Module*
- **AI/ML Models**
  - *Tweet/Document Classification Module*
  - *Sentiment Analysis Module*
  - *Image Classification Module (optional)*
- **Data Storage**
- **Data Processing Pipeline**
- **Analytics Processing Engine**
  - *Realtime Data Aggregation Support*
  - *Spatio-temporal Query Support*
- **Visualization**
  - *Interactive Dashboard*

In the follwong part I will throw some light on each component and discuss about challenges that it might have.

1.1 Data Ingestion

Data Collection Module: Twitter is the biggest social media data source for researchers. Twitter's 1% sample data stream API is the most common approach for data collection. Twitter statistics reveals that only 0.85% of tweets in the stream is geotagged [24] which is significantly lower. Utmost effort and care should be taken to collect more geotagged data. Twitter's location based API should be used for such purpose.

Challenges: Collecting data from location based API or any other keyword based search API are restrictive in nature with request limit per hour. Evading this problem might be challenging with limited resources. Multiple number of data collection servers collecting mutually exclusive geographical region can help to collect more geotagged data. For some social media sites, it is almost necessary to use proxy network to avoid IP block.

1.2 Data Enrichment:

Data Cleaning Module: Data collected from social media often needs to be cleaned (e.g. tokenize, language filter etc.) for processing. The common scenarios for cleaning operations are *(i) filtering english tweets, (ii) removing emoticons (iii) keywords extraction etc.*

Challenges: There are many good tools for data cleaning. The main concerns are *(i) which library tools to use for desired result. (ii) the library should have high processing throughput*

Location Extraction Module As mentioned earlier that the percentage of geotagged tweets is not high. However, a lot of attempt has been made to predict the location of the tweet based on user activity and history. Geotagging users is now a well studied problem and it has a median error of 6.38 km which might not be very significant for our analysis[5].

Challenges: Increasing need to collect more data about users. If we are interested in home location of users then the above mentioned [5] technique is satisfactory. However, if we want the dynamic location as the users move or travel then it becomes a challenging problem.

1.3 AI/ML Models:

Tweets/Document Classification Model: In order to distinguish between relevant (e.g. health, food, disease etc.) and non relevant tweets/documents we need a tweet classification component. Unsupervised methods like topic modeling with LDA [3], pLSA [11] and phrase LDA [9] and modified versions of them can help in classification problem. However, microblogs classification for targeted topic needs further attention. In our work [20] for *Spatio-temporal Sentiment Analysis for US Election*, we used political and non-poilitical tweet classification in a semi-supervised approach. The semi-supervised approach starts by creating training data for classification. Topic modeling act as a bootstrap method for creating training data that helps in learning tweet classification through context. This semi-supervised approach proved to be more robust [20].

Challenges: The semi-supervised approach used in [20] have not been used yet for classification in health realted topics. Previous works like *topic model for ailment* [22] (e.g. examples of word and topic realtion is shown in Figure 1) are extension of topic modeling with LDA and specially descigned for ailment tweet discovery. Remodeling semi-supervised classification for disease and health

University of Utah, Salt Lake City, USA.

Blood Pressure	Heart Attack	Diabetes Type II	Yoga	Alzheimer	Obesity	Diet and Exercise	Obesity
risk	heart	change	diabetes	medicine	diabetes	helps	health
blood	diabetes	diabetes	#yogafightsdiabetes	diseases	surgery	diabetes	diet
high	cardiovascular	#lifestyle	yoga	common	treatment	children	obesity
diabetes	attack	type	control	drugs	obesity	exercise	immune
pressure	stroke	ii	life	Alzheimer	cure	diet	syndrome
Vegetarian	**Pregnancy Diet**	**Celebrities Diet**	**Weight Loss Diet**	**Weight Loss Medicine**	**Religious Diet**	**Mental Health**	**Exercise& Diabetes**
diet	pregnancy	diet	weightlose	diet	burning	health	helps
eat	motherhood	beyonce	effective	#weightloss	#weightloss	nutrition	diabetes
fruits	diet	tips	morning	slimming	fasting	benefits	children
vegetables	baby	fatloss	dieting	pills	Ramadan	healing	exercise
fresh	motherhood	#angelinajolie	banana	#fatburners	diets	#mentalhealth	diet
Diet	**Daily Plan**	**Computer Games**	**Brain**	**Fitness**	**Diet& Diabetes**	**Obesity**	**Exercise**
diet	food	exercise	exercise	fitness	helps	workout	bellyfat
exercise	exercise	finding	brain	#gymlife	diabetes	burning	losing
protein	calorie	pokemon	improve	bodybuilding	children	exercise	exercise
beauty	goal	#pokemongo	memory	gym	exercise	fatburn	ways
muscle	completed	hour	performance	workout	diet	obesity	effective
Diet	**Alzheimer**	**Cancer**	**Children**	**Diabetes**			
health	study	cancer	obesity	diabetes			
diet	link	breast	kids	surgery			
obesity	Alzheimer	study	childhood	treatment			
immune	obesity	risk	rates	obesity			
syndrome	research	obesity	problem	cure			

Figure 1: An example for topic model on ailment[22].

are yet to experiment and might face challenges. For example tweets like "I feel like I'm going to die of Bieber Fever, No Joke!" and *"Web design class gives me a huge headache everytime"* both tweets does not talk about health condition. Hence learning context of words is desirable approach.

Sentiment Analysis Model: From past decade opinion mining on text data has been a popular research topic. Pang et. al. [19] gives a comprehensive survey on incipient opinion mining research. Twitter sentiment analysis with machine learning approaches like SVM [12], lexicon based [25], LDA [7, 14] and neural network [6, 26] etc. Our work on *"Spatio-temporal sentiment analysis on US Election"* used LSTM-RNN and FastText for achieving state-of-the-art result.

Challenges: To improve upon the existing state-of-the-art methods, we have to keep adopting and experiment new AI methods. Exploration will achieve more fruitful result if proper training/ground truth datasets on sentiment are available in future. Also sentiment anaysis with AI is limited to only a few language. To widespread the technology to different language we need resources (datasets), effort and experiments to achieve rewarding results.

Image Classification Model: *"A picture is worth thousand words"* is an English language-idiom that rightly characterize the scenario for tweets [1]. Twitter statistics reveals that 42% of tweets attach images [27]. Integrating image analysis module will reveal more information that is worth looking into. To the best of my knowledge we haven't yet researched a lot on public health by integrating twitter images.

Challenges: Recent works on object detection from images will be our starting point [10, 16]. We need to list the items we look out in images and provide enough example of them in training sample while training our model.

1.4 Data Storage:

Collected data enriched with information from ML/AI models are stored in databases. Spatio-temporal properties of data demands more attention with indexing. NoSQL key-value based databases can store all the information while the spatial indexes store object ids with location for accelerated access [4].

1.5 Data Processing Pipeline:

Creating the pipeline starting from data collection to data sink with so many components needs expertise in data processing. In our recent work *AI Pro: Data Processing Pipeline for AI models*[1] takes care of setting up the pipeline for end users just by configuration (which is popularly known as *code as configuration*). Researchers opting for customized processing pipeline will be able to create it with little or no effort.

Challenges: AI Pro processing pipeline needs adapt new technologies and support them, this requires community collaboration.

1.6 Analytics Processing Engine:

Realtime Data Aggregation Support: Enriched information stored in databases is ready to be analyzed by data scientist. Data scientists finds statistical significant information by aggregating on different attributes which is termed as slicing and dicing in data analytics world. Realtime operation of slicing and dicing with spatio-temporal operation is hard to achieve. A work from our lab by Li et. al (XDB) addresses this problem with online aggregation [15]. This work is vital to achieve realtime analytics support for our system.

Challenges: Scaling up the processing power in realtime environment is always a challenging task. The works mentioned above tries to solve the problem amicably. Sampling strategies on aggregate data guides data scientist to quickly evaluate the information/statistics vital for the analysis. It is essential that sampling strategies are good enough for intended data to fetch the truthful results. That poses a challenge for the system.

Spatio-temporal Query Support: As mentioned earlier in data storage section that geotagged data needs special attention, it is also true for query support. Spatial operations like distance range, nearest neighbour requires special support and they are costly toward resources. Another work from our lab Xie et al. integrated *"Spatial In-Memory Big data Analytics* [28] has extended Apache Spark SQL to support spatial queries by intoducing native indexing support over RDDs [**?**].

1.7 Visualization:

interactive Dashboard: Interactive dashboard for online analysis integrated with realtime processing APIs is a feature data scientist will love to see. Filtering and selection from visualization will make the system amiable and encouraging for user. Realtime dashboard to find the rate of mentions of different topics and other time series analysis might also be important for data scientist. STORM a data anlysis project from InitialDLab supports similar interface []. Interactive visualization and cohesive analysis might be able find correlation among health, food and physical activity.

Challenges: Interactive dashboard needs to be integrated with Analytics Processing API serve to fetch data with query. The tools like d3.js can help in designing the visualization but it requires a good skill and understanding of the subject.

Conclusion: All the above mentioned components are useful and necessary for a successful analytical system on spatio-temporal system. Our system is generic in nature but can be extended with custom topic related analysis module for special purposes. For example, spatio-temporal analysis from social media might be able to throw light on public health and ailment for collective benefit of the society.

[1]www.cs.utah.edu/deb/aipro/

2 PART B. ETHICAL CONCERNS:

Artificial intelligence is a boon only if we only use it for the benefit of mankind, same is true with all the technologies we use in our life. Spatio-temporal health analysis as a cumulative measure for public health is a boon for society where government can know the state of public health and act accordingly for the benefit of the state or country and its people. However if the same data is used to target individual person and try to know the concerns related to his/her health that will not be welcome by sane people.

Take another example if the browser history is used to know the health concern of an individual then it is highly unethical and a dangerous proposition. Take a hypothetical situation, say Bluecross BlueShield a health insurance provider for public wants to know the health concerns by analyzing their tweets/ browser history (data collected from 3rd party source). So that they can offer differ pricing on same health insurance plan based on the analysis they have obtained from personal information for monetary gain. The data source used here for such analysis irrespective of its truthfulness makes a judgement on an individual. This information may harm his/her mental, physical, financial health and it is highly unethical with detrimental consequence. As an ethical researcher I have issues in working under such circumstances and in such scenarios.

Taking another hypothtical situation where a government federal agency say *Center for Medicare and Medicaid Services* is trying to gauge public health.

If the cumulative and aggregate metrics on health and ailment is not targeting any individual but a collective society it will be welcome by the reearchers and by me. An example of such scenario; we know social media is very good at first hand reporting of public concerns. If there is an outbreak of a disease in a region then government can respond to the situation by sending doctors and health services team to learn its extent, gather facts and create awareness. Here spatio-temporal analysis just helped to identify concerns of public health; but facts and figures from on ground health workers are mandatory to take any actions like travel advisory and isolation procedures. Spatio-temporal snalysis act as a boon here.

However if the government agency also wants to use personal information to gauge health and habits of an individual then it is unethical and I will keep myself away from it.

REFERENCES

[1] S. G. S. Advice. Syracuse post standard. *Retrieved from Syracuse Post Standard Newspaper from March*, 28, 1911.
[2] J. M. Barros, J. Duggan, and D. Rebholz-Schuhmann. Disease mentions in airport and hospital geolocations expose dominance of news events for disease concerns. *Journal of biomedical semantics*, 9(1):18, 2018.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
[4] R. Christensen, L. Wang, F. Li, K. Yi, J. Tang, and N. Villa. Storm: Spatio-temporal online reasoning and management of large spatio-temporal data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1111–1116. ACM, 2015.
[5] R. Compton, D. Jurgens, and D. Allen. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE, 2014.
[6] C. dos Santos and M. Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, 2014.
[7] A. Duric and F. Song. Feature selection for sentiment analysis based on content and syntax models. *Decision support systems*, 53(4):704–711, 2012.
[8] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap, et al. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169, 2015.
[9] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3):305–316, 2014.
[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
[11] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.
[12] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.
[13] A. Karami, A. A. Dahl, G. Turner-McGrievy, H. Kharrazi, and G. Shaw. Characterizing diabetes, diet, exercise, and obesity comments on twitter. *International Journal of Information Management*, 38(1):1–6, 2018.
[14] E. Kouloumpis, T. Wilson, and J. D. Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164, 2011.
[15] F. Li, B. Wu, K. Yi, and Z. Zhao. Wander join and xdb: Online aggregation via random walks. *ACM SIGMOD Record*, 46(1):33–40, 2017.
[16] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.
[17] M. Mueller and M. Salathé. Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *arXiv preprint arXiv:1805.05491*, 2018.
[18] Q. Nguyen, H. Meng, D. Li, S. Kath, M. McCullough, D. Paul, P. Kanokvimankul, T. Nguyen, and F. Li. Social media indicators of the food environment and state health outcomes. *Public health*, 148:120–128, 2017.
[19] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.
[20] D. Paul, F. Li, M. K. Teja, X. Yu, and R. Frost. Compass: Spatio temporal sentiment analysis of us election what twitter says! In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1585–1594. ACM, 2017.
[21] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. *Icwsm*, 20:265–272, 2011.
[22] M. J. Paul and M. Dredze. A model for mining public health topics from twitter. *Health*, 11:16–6, 2012.
[23] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.
[24] L. Sloan, J. Morgan, W. Housley, M. Williams, A. Edwards, P. Burnap, and O. Rana. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological research online*, 18(3):1–11, 2013.
[25] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.
[26] D. Tang, B. Qin, and T. Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1422–1432, 2015.
[27] Thenextweb. Next web, 2018. https://thenextweb.com/socialmedia/2015/11/03/what-analyzing-1-million-tweets-taught-us/,[Online; accessed 27-Nov-2018].
[28] D. Xie, F. Li, B. Yao, G. Li, Z. Chen, L. Zhou, and M. Guo. Simba: Spatial in-memory big data analysis. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 86. ACM, 2016.