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 [1, 3–5, 7, 8]. Research analysis like correlating influenza rates w.r.t geography (spatial) and time [9], state level food and health behavior analysis [6], predicting heart disease rate mortality rate based on twitter information [3]; are motivating examples to carry out such analysis for improving and create for public health. All these above adhoc analysis inspire us to build a general system for comprehensive analysis. In this work, I will present overview of architechture and desired features to build such system or tools.

1 PART A: 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**
- **Analytics Processing Engine:**
  - *Realtime Data Aggregation Support*
  - *Spatio-temporal Query Support*
- **Visualization**
  - *Realtime Dashboard*
  - *Dynamic Data Visualization Module*

In the follwong part I will throw some light on each component and discuss about challenges if they have any.

1.1 Data Ingestion

Data Collection Module: Twitter is 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 [10] 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.

University of Utah, Salt Lake City, USA.

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 limit 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 percentage of geotagged tweet is not high. However, a lot of attempts have 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[2].

Challenges: Need to collecting more data about users. If we are interested in home location of users then the above mentioned [2] technique is fine. However if we want the dynamic location as the users move or travel then it is 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 a document classification component. Unsupervised methods like topic modeling with LDA [], pLSA [] and phrase LDA [] and modified versions of them can help in classification problem. However, microblogs classification for targeted topic needs further attention. In our work [**?**] 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[**?**].

Challenges: The semi-supervised approach used in [**?**] have not been used yet for classification in health realted topics. Previous works like *topic model for ailment* [] are extension of topic modeling with LDA and specially descigned for ailment tweet discovery. Remodeling semi-supervised classification for disease and health 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:

Challenges:

2 PART B. PUBLIC HEALTH ANALYSIS AND ETHICAL CONCERNS:

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] M. Mueller and M. Salathé. Crowdbreaks: Tracking health trends using public social media data and crowdsourcing. *arXiv preprint arXiv:1805.05491*, 2018.

[6] 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.

[7] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. *Icwsm*, 20:265–272, 2011.

[8] M. J. Paul and M. Dredze. A model for mining public health topics from twitter. *Health*, 11:16–6, 2012.

[9] 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.

[10] 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.