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 [1–4, 6, 7]. Research analysis like correlating influenza rates w.r.t geography (spatial) and time [8], state level food and health behavior analysis [5], predicting heart disease rate mortality rate based on twitter information [2]; 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 [9] 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 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.

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] 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.

[3] 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.

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

[5] 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.

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

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

[8] 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.

[9] 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.