Detecting and Tracking Disease Outbreaks by Mining Social Media Data

Yusheng Xie[1,2] **Zhengzhang Chen**[1,2] **Yu Cheng** **Kunpeng Zhang** **Kathy Lee** **Ankit Agrawal** **Wei-keng Liao** **Alok Choudhary**[2]

Northwestern University, Evanston, IL USA

[1]: authors contributed equally.

[2]: {yxi389,zzc472,choudhar}@eecs.northwestern.edu

Abstract

The emergence and ubiquity of online social networks have enriched web data with evolving interactions and communities both at mega-scale and in real-time. This data offers an unprecedented opportunity for studying the interaction between society and disease outbreaks. The challenge we describe in this data paper is how to extract and leverage epidemic outbreak insights from massive amounts of social media data and how this exercise can benefit medical professionals, patients, and policymakers alike. We attempt to prepare the research community for this challenge with four datasets. Publishing the four datasets will commoditize the data infrastructure to allow a higher and more efficient focal point for the research community.

1 Introduction

"Social Media" is producing massive amounts of data, so called "BIG DATA", with *volume*, *velocity*, *variety* and *veracity* (the four Vs in big data challenges) at an unprecedented scale. Of the four challenges associated with this data, *volume* refers to the size of the data, *velocity* refers to the speed at which the data is being generated, *variety* refers to the heterogeneity and complexity of the data (e.g., text in multiple languages, images, videos, voice, activities and demographics of the producers, context, characteristics of consumers etc.), and *veracity* of the data refers its resistance to the now pervasive spam users or other abuses. In addition, the connectivity of the networks through which the data (e.g., "retweets" in Twitter) propagate can drastically increase the data velocity and volume.

But one may ask the omnipresent question, "So what?" Having and producing big data by itself is of little value if it is dormant. The data mining and artificial intelligence research community must enable the next big step that "derives accurate, actionable, predictive and timely knowledge from heterogeneous, complex, online, high-dimensional, and massive data[1]."

Detecting and tracking epidemic outbreaks in a population is an important and challenging task in the course of preserving a sustainable world. The data we have curated over the past few years offers an unprecedented opportunity to study this challenge.

In this paper, we identify the real-time challenge in detecting and tracking epidemic outbreaks. By releasing relevant datasets from the social media, we aim to provide a platform for researchers in the domain to work on the real-time challenge. "Social Media" data here is characterized by publicly available data from multi-way communications, interactions, crowd sourcing (e.g., Facebook), discussion forums, expert blog sites (e.g., USA Today Health blog), general knowledge repositories (e.g., Wikipedia), etc.

2 Motivation and Challenge

Early detection and tracking mechanisms are critical in reducing the impact of epidemics and preventing the epidemics from becoming unmanageable by making a rapid response. For example, the cholera epidemic killed over 100,000 people worldwide, and sickened 35 million people during the year 2010 [Enserink, 2010].

Traditional epidemic surveillance systems are primarily built from virology and clinical data, which is manually collected. For example, in order to detect the outbreaks of influenza, the Japanese Infection Disease Surveillance Center collected influenza patient data from nearly 5,000 clinics [Aramaki *et al.*, 2012], while Germany used 420 public health departments to collect the infectious diseases data and process the reports in 2009. The U.S. Centers for Disease Control and Prevention (CDC) uses the same way to gather the data. Other traditional methods include conducting randomized telephone polls, or using sensor networks to detect pathogens in the atmosphere [Metcalf *et al.*, 1995]. However, these surveillance systems have various drawbacks, such as the delay due to slow reporting mechanisms, information aggregation and processing times. Typically, there is a 1-2 weeks reporting lag. Thus, the traditional surveillance systems often fail to report rapidly emerging diseases like the lung disease SARS in 2002 [Lew *et al.*, 2003].

Real-time feedback in situations like epidemic outbreaks is indispensable. But, due to human intervention and the lag it introduces, traditional surveillance systems simply cannot be consulted, even if such systems can eventually produce accurate tracking or detection. In other words, insights of the situation, no matter how precise, are rendered useless unless

[1]http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

they are derived in a timely fashion. Therefore, the major challenge in tracking is how to do it in real-time.

Indeed, in [Ginsberg *et al.*, 2009] the authors get accurate and responsive prediction and tracking results by building proprietary models based on proprietary data, which is Google's own search history. Although Google generously provide their flu trend for free via a website[2], it is difficult for the research community to contribute to this project efficiently. Publishing high quality and rich datasets will commoditize the data infrastructure to allow a higher level focal point for the research community.

3 Problem Statement

In contrast to tradition, social media, which has so far not been widely adopted in surveillance systems, has the potential to provide the necessaries for a real-time epidemic detection system because of its immediate responsiveness and crowdsourcing power. The problem of detecting and tracking epidemic outbreaks through social media can be defined as follows. Given a disease or epidemic of interest, a time window, and a stream of textual or multimedia data from social media, the task is how to extract relevant spatial and temporal knowledge about the epidemic in the real world.

For example, Twitter users may post about an illness, and the social links in the network, which may correspond to real-world relationships, can give us clues about who these ill people will most likely come in contact with. And there are more than a thousand tweets each day that include the keywords like "influenza" and "flu". Furthermore, for social services like Twitter and Instagram, user activities retrieved from public API often come with accurate GPS-based location tags, which can be accurately and instantaneously consumed by prediction and detection mechanisms built upon social media data. By using social media surveillance system policy makers can know the critical point for epidemics control, the effectiveness of the policy, whether the real status is different from the reported scenarios, etc; the medical professionals can work on the treatment method earlier; and public health officers could more effectively allocate healthcare workers, resources and drugs.

Despite recent advances in mining techniques, there is a gap to bridge between state-of-the-art and a desirable real-time surveillance system. Not only the massive size of the datasets challenge current data mining approaches, structurelessness, heterogeneity, and spam users also pose real challenges. Simply applying existing data mining algorithms to the data cannot bridge this gap. Designing a system that can collect massive unstructured web-scale big data in minimal delay and process it efficiently is at the core of this real-time challenge. Our contribution is to freely provide the appropriate datasets, on which prospective methods can be tested.

4 Datasets

With this work, four distinct social media datasets are made publicly available. Respectively, they are health-related blogs (**BL**), Facebook public walls of medical brands and organizations (**FBW**), Facebook public posts mentioning health-related terms (**FBP**), and media uploads from Instagram (**INST**). Table 1 lists important metrics of the four datasets. Our system will keep updating the four datasets daily.

Table 1: Overview of four datasets.

Source Metric	BL	FBW	FBP	INST
Entries	79K	29K	3.2M	154K
Dimensions	17	16	16∼70	26∼40
Start date	2008/7	2010/7	2012/10	2012/9
Sentiment	labeled	labeled	labeled	labeled
GPS tag	none	none	some	all
Format	CSV	CSV	JSON	JSON

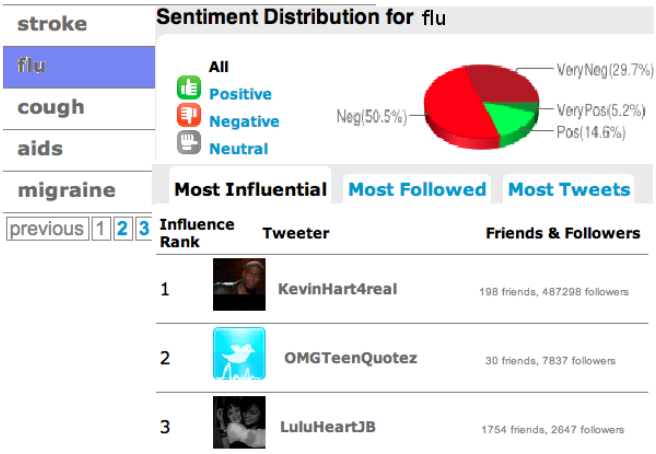

Figure 1: Visualization of the curating and preprocessing steps including sentiment classification and user statistics.

Blogs (BL)

The BL dataset includes blog articles and user comments from "Bites" (`http://bites.today.com`), "Life Inc." (`http://lifeinc.today.com`), "The Body Odd" (`http://bodyodd.msnbc.msn.com`), and "USA Today Health" (`http://yourlife.usatoday.com`). Each blog associates two files, the blog file and the comment file. The blog file contains articles published on that blog website, while the comment file contains all users' comments to all articles in the blog file.

Facebook Public Walls (FBW)

FBW dataset contains public data from 19 Facebook walls that are related to health issues such as "National Institutes of Health". Each Facebook wall is associated with three csv files: the post file that contains official posts made by the wall, the comment file that contains user comments to the posts, and the post_like file that contains user's Facebook "Post Likes"[3].

[2] http://www.google.org/flutrends/

[3] http://developers.facebook.com/docs/reference/api/post/

Facebook Public Posts (FBP)
In addition to public walls on Facebook, we have also retrieved public messages on Facebook directly. Using a list of common medical conditions and diseases from Wikipedia[4], we are able to retrieve any public Facebook messages/posts relevant to each item on that list.

Two aspects characterize the FBP dataset. First, unlike other sources, the records from FBP are strongly heterogenous. An average record would have 20 dimensions while some records may have over 70 dimensions, some of which are nested. The second aspect is how FBP is formed. Since FBP is collected by searching, instead of listening to particular Facebook walls or Twitter handles, it can eventually evolve into a real-time searchable knowledge base that comprises heterogenous data from heterogeneous sources.

Instagram Data (INST)
A main reason why INST is included as part of this package is the geo-spatial information it provides. Most social network services allow their users to enable GPS-based tag, and Instagram is not the only one. But much more documents and profiles opt for enabling GPS tag than any other services. Since epidemics spread mostly depending on spatial constrains such as distance, altitude, accessibly, etc., GPS coordinates for the happening event would be immensely valuable.

4.1 Data Curation

For the research community to better focus on building high-performance analytics and mining infrastructures for outbreak detection, a series of complicated yet efficient steps, as shown in Figure 1, is taken in creating our datasets. By publishing and curating our datasets, the authors aim to standardize the preprocessing steps in organizing and labeling the data.

Our datasets contain only the relevant information. Finding and selecting what is relevant and interesting from the inundation of web data is the first step in tackling big data. In preparing our datasets, the authors collect Facebook data only from health-/disease-related walls or posts, tweets only relevant to epidemics and symptoms, etc.

Web-scale social media data lose half its value if they stay static since publication. Maintaining a dataset over time entails much more than calling the downloading scripts every once in while. Live datasets requires not only dedicated, persistent, and durable storage infrastructure but also compliance with data provider's APIs and policies. Both requirements are exacting and expensive for individual researchers or small research groups. In our collection, Bl is collected through web crawling; FBW and FBP are collected through Facebook's Graph API[5]; INST is collected through the Instagram API[6].

The proposed datasets also contain sentiment labels produced by high-accuracy algorithms [Hu *et al.*, 2013]. High-level semantic meanings are the cornerstone in understanding social conversations [Zhang *et al.*, 2011]. Its automation is a necessary step towards real-time surveillance based on big data.

[4]http://simple.wikipedia.org/wiki/List_of_diseases

[5]http://developers.facebook.com/docs/reference/api/

[6]http://instagram.com/developer/

Our datasets resemble the network properties of social networks. To address the increasing importance of social graphs and to fully leverage the rich network information present in our social media datasets, all repeating entities (e.g. user IDs) are identified by globally consistent IDs. Every tweet or comment is an identifiable part of a web-scale social graph.

Our datasets contain purely public information and are distributed freely via GitHub[7].

5 Conclusion

Detection of epidemic outbreaks is a tremendous opportunity for our research community to create real world impacts. Early detection and tracking epidemics outbreaks is crucial to policy makers, medical professionals, patients and public health officials.

By publishing and curating large amount of social media data, the authors solicit researchers to build data analytics platform through scalable algorithms and architectures, which then can constantly provide collective intelligence and awareness of epidemic outbreaks in real time.

6 Acknowledgements

This work is supported in part by the following grants: NSF awards CCF-0833131, CNS-0830927, IIS-0905205, CCF-0938000, CCF-1029166, and OCI-1144061; DOE awards DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DESC0005340, and DESC0007456; AFOSR award FA9550-12-1-0458.

References

[Aramaki *et al.*, 2012] E Aramaki, S Maskawa, and M Morita. Influenza patients are invisible in the web: Traditional model still improves the state of the art web based influenza surveillance. 2012.

[Enserink, 2010] M Enserink. No vaccines in the time of cholera. *Science*, 329(5998):1462–1463, 2010.

[Ginsberg *et al.*, 2009] J Ginsberg, M Mohebbi, R Patel, L Brammer, M Smolinski, and L Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.

[Hu *et al.*, 2013] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013.

[Lew *et al.*, 2003] TK Lew, T Kwek, and D Tai. Acute respiratory distress syndrome in critically ill patients with severe acute respiratory syndrome. *JAMA*, 290(3):374–380, 2003.

[Metcalf *et al.*, 1995] T G Metcalf, J L Melnick, and M K Estes. Environmental virology: From detection of virus in sewage and water by isolation to identification by molecular biology-a trip of over 50 years. *Annual Review of Microbiology*, 49(1):461–487, 1995. PMID: 8561468.

[Zhang *et al.*, 2011] Kunpeng Zhang, Yu Cheng, Yusheng Xie, Daniel Honbo, Ankit Agrawal, Diana Palsetia, Kathy Lee, Weikeng Liao, and Alok Choudhary. Ses: Sentiment elicitation system for social media data. In *ICDMW '11*, pages 129–136, Washington, DC, USA, 2011. IEEE Computer Society.

[7]https://github.com/yvesx/IJCAI2013.git