Why Is It Difficult to Detect Outbreaks in Twitter?

Avaré Stewart[1], Nattiya Kanhabua[1], Sara Romano[2],
Ernesto Diaz-Aviles[1], Wolf Siberski[1], and Wolfgang Nejdl[1]

[1]L3S Research Center / Leibniz Universität Hannover, Germany
{stewart, kanhabua, diaz, siberski, nejdl}@L3S.de

[2]Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione / University Federico II Naples, Italy
sara.romano@unina.it

ABSTRACT

In this paper, we present an event-based Epidemic Intelligence (EI) system framework leveraging social media data, e.g., Twitter messages (or tweets) for providing public health officials the necessary tools to survey and sift through relevant information, namely, disease outbreak events. There exists three main research challenges in gathering epidemic intelligence from social media streams: 1) *dynamic classification* to enable message filtering, 2) *signal generation* producing reliable warnings based on observed term frequency changes in the filtered messages, and 3) providing *search and recommendation* functionalities to domain experts, for better assessment of the potential outbreak threats associated with the generated signals. We outline possible approaches to solve these important challenges as well as discuss areas where further research is required. The aim of this paper is to provide guidance for similar endeavors, and to give prospective event-based Epidemic Intelligence system builders a more realistic view on the benefits and issues of social media stream analysis.

1. INTRODUCTION

Social media, e.g., Facebook or Twitter messages, are valuable sources for providing real-time information, such as, status updates, opinions or news. Numerous real-time Web applications increasingly use Twitter for tasks, such as, detecting natural disaster [19], political persuasion [16, 20], or present trends [15]. In the medical domain, it has been shown that Twitter is capable of transmitting information faster than traditional media channels [8, 14], thus giving human experts a head start in dealing with health-related information. To exploit this timeliness potential, we present an event-based Epidemic Intelligence (EI) system, which has emerged as a type of intelligence gathering aimed to detect events of interest to the public health from unstructured text on the Web. In our proposed EI system, we detect public health events by mining and analyzing tweets; as well as provide support for public health officials to retrieve and explore the *signals* of infectious disease outbreaks. Signals represent a very dynamic type of information object, which are generated for each temporal anomaly found in time series data that occur when an infectious disease or its impact is above an expected level, for a particular time and place. Signals are monitored by public health authorities and help them assess the need for action, in response to potential threat. Note that, there are existing EI systems, such as, the BioCaster Global Health Monitor[1] or HealthMap [2]. However, they differ from our proposed system in the level of analysis, information sources, language coverage and visualization.

Although numerous approaches successfully detect relevant seasonal influenza outbreak events from Twitter [1, 5, 15], it seems that the challenges in building an EI system are easily underestimated, especially when it comes to detecting *emerging* (*unseen* or *non-seasonal*) health events from social media streams. Inspired by the outcome of collaborations conducted as part of the European research project Medical Ecosystem: Personalized Event-based Surveillance [3], with medical domain experts and epidemiologists, the aim of this paper is to describe an EI system that is better targeted towards the needs of real-world users to access public health information. This includes describing three main challenges associated with social media stream analysis systems, outlining possible approaches to handling such challenges and pointing out issues where more research is needed in order to achieve high-quality results coupled with wide-spread acceptance within the public health domain. Specifically, we have identified the following core challenges in event detection on Twitter data streams:

- **Adaptive Message Filtering.** Although the detection for well-known, recurring events (e.g., influenza) is mature, the detection of *novel* and *aperiodic* public health events requires *adaptive approaches* which take into account feature change over time, i.e., to enable the identification of new relevant terms.
- **Signal Generation from Noisy Data.** Time series data created from Twitter is usually noisy, incomplete and sparse. Given the imperfect data, it is important to consider measures for *assessing the reliability of signals*, i.e., the extent to which we can actually trust signals that have been generated for early warning.
- **Threat Assessment Support.** End users need assistance to cope with the cognitive challenges of *search and exploration* of outbreak signals. The effectiveness of straight-forward approaches to retrieval and collaborative filtering can be unsatisfied, given the dynamics of streaming data and the limited context of detected signals as well as their corresponding tweets.

[1]http://biocaster.nii.ac.jp/
[2]http://www.healthmap.org/en/
[3]http://www.meco-project.eu/

HSD 2013, August 1, 2013, Dublin, Ireland.

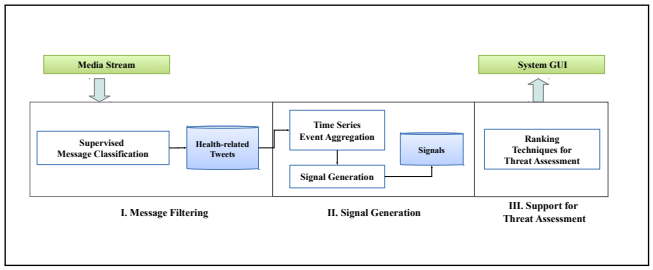

Figure 1: Event-based Epidemic Intelligence system.

2. SYSTEM OVERVIEW

The overview of our EI system is illustrated in Figure 1. We gather tweets relevant for outbreak surveillance using of multi-lingual list of terms consisting of not only infectious disease names and pathogens, but also, their synonyms and symptoms (provided by the experts).

The `Supervised Message Classification` module is responsible for filtering tweets irrelevant to the medical domain. Although existing works [4, 15, 18] have already addressed this problem in static settings. In this work, we propose an adaptive feature change detection method as we will discuss in more detail in Section 3.

In the `Signal Generation` module, signals (i.e., outbreak warnings) are generated using time series data, as was done in previous work [13], consisting of aggregated counts over the common entity tuples of the relevant messages. In general, input data used for anomaly detection is noisy (contain spurious events), incomplete (under-reporting of an event or using of acronyms and abbreviates not recognized) and sparse (low aggregation counts). In Section 4 we consider the impact these aspects for signal generation.

Even though successive stages of filtering are carried out within each module, domain experts may still be faced with potentially many signals and their associated tweets. Therefore, these signals are presented to the user via the `Support for Threat Assessment` module. In particular, we seek to employ time-aware ranking and recommendation techniques to tackle the problem of information overload (cf. Section 5).

3. CHALLENGE I: MESSAGE FILTERING

The online message classification continues to be a complex and challenging task for long term surveillance and intelligence gathering, in general. One reason for this is that given the evolution of real-world events, the variable to observe cannot always be known a priori. In EI, one such example is in the detection of food-borne illness, in which the contaminated food item is not known in advance.

Feature change detection. We need a way to: 1) dynamically detect when *new* and *relevant* terms in a stream appear; and then 2) subsequently incorporate the tweets containing these terms into the classification model.

Dynamic labeling. As new terminology evolves, the criteria for defining relevant tweets is also changing. However, expert labeling of classifier training instances is expensive and in practice difficult to obtain, especially for the rate and volume needed to build and maintain a good classifier.

Ambiguous and noisy data. When a tweet is matched with the defined keywords, the tweet itself may not refer to a public health event due to polysemy. For example, the term "fever" being used to express excitement, e.g., *Justin Bieber Fever* or *Royal Wedding Fever*. In addition, noise can be caused by spurious events in which an entity is correctly detected, but its role is not, namely: 1) *"A two hour train journey, Love In the Time of Cholera."* or 2) *"I liked a @YouTube video http://youtu.be/... a Metallica, Megadeth, & Anthrax - Helpless"*. Both mention infectious diseases Cholera and Anthrax, but their context is literature and music, respectively.

3.1 Proposed Approach

In order to capture aperiodic events with a high impact on accuracy over time, we propose the use of a method that takes into account that the natural language of the tweets in the stream changes constantly in response to the *temporal dynamics of real-world events*. Our dynamic classification consists of two main steps: 1) incorporating the use of an orthogonal vector, which is learned by a Support Vector Machine (SVM), as a description of the feature change; and 2) computing a novelty score that lets the system identify those tweets that contribute to the feature change, so that their true labels can be obtained. In this paper, we only focus on *text-based analysis* of Twitter messages. We plan to investigate the use of Web resources shared in social media, e.g., posted images and videos, as future work.

3.2 Implications

Identifying non-relevant tweets is difficult for multiple reasons. For example, automatically filtering a sarcastic and metaphoric tweets correctly is hard for limited context and remains to be tackled with this domain. Feature change detection should be able to identify new (entity and non-entity) keywords in the Twitter stream that are related to outbreaks of infectious diseases. These keywords could be used for automatically updating the list of search term that is used to collect the tweets. Consideration will also be given to when these newly discovered terms should be subject to decay. To this end, we plan to increase the scale of the analysis to include: the classification of more symptoms as well as polysemy terms for diseases. To get a better understanding of the impact of detected feature change on the classification accuracy, a larger set of expert labeled tweets for experimentation would be useful to further improve the significance of the results. Nonetheless, doing so, would still not address the need to experts to re-label each time feature change was detected and in practice, the overhead of such a task is too expensive and not timely enough.

4. CHALLENGE II: SIGNAL GENERATION

Health-related tweets obtained from the previous stage will be leveraged in order to generate an early warning signal (so-called *signal generation*). Signals represent each temporal anomaly found in time series data occurring when the impact of an infectious disease is above an expected level, and it is a difficult task because of the following challenges:

Incompleteness and sparsity of data. This implies that instances of an event are missing or under-reported. This may occur due to: 1) the presence of processing errors - an acronyms or abbreviations not recognized as medical conditions; 2) the fact that people who are actually suffering do not tweet; 3) the tweets which contain these mentions have not been collected by the system, i.e., based on the imbalance between the type of tweets collected (e.g., personal versus news tweets); and 4) the minimum required entity types are not present. Sparse time series data refers specifically to low aggregation counts, which impact the anomaly detection algorithm.

Temporal and spatial dynamics of diseases. The characteristics of infectious diseases are highly dynamic in time and space, and their behavior varies greatly among different regions and the time periods of the year. Some infectious diseases can be rare or aperiodic, while others occur more periodically. In addition, various diseases have different transmission rates and levels of prevalence within a region. For example, cholera infections vary greatly in frequency, severity, and duration. On the one hand, in some regions historically, only sporadic outbreaks occur in areas, such as, parts of South America and Africa. On the other hand, even in areas where cholera infections are endemic (the South Asian countries of Bangladesh and India) the epidemic levels change dramatically from one year to the next [9].

Given imperfect time series, we need to know the extent to which we can actually trust signals that have been generated for early warning. To this end, we aim at answering the question: *Are there ideal algorithms and/or parameter settings for signal generation using Twitter?*

4.1 Proposed Approach

Studying the usefulness of Twitter data in the medical domain requires real-world outbreak statistics. We previously built outbreak ground truths (historical baselines) by relying upon ProMED-mail[4], a global reporting system providing information about outbreaks of infectious diseases. We collected 3,056 ProMED-mail reports and identified 14 different outbreaks occurring during year 2011 as ground truths [12]. An important aspect of our work is that we consider the duration of each outbreak by analyzing temporal expressions mentioned in a ProMED-mail document, unlike aforementioned work [3] that assumes the publication date of a document as the estimated relevant time of an outbreak. The reason is that the events in ProMED-mail undergo moderation, so there is often a delay between the time of the actual outbreak and the publication date of the related report.

A basic approach to detect anomaly in health-related time series data is to exploit different state-of-the-art biosurveillance algorithms [2, 10]. These algorithms are already widely used in the existing Biosurveillance systems, so they can be used for assessing the reliability of signals from the perspective of the domain experts. The metrics used to assess generated signals are sensitivity, predictive positive value and F-measure. Sensitivity refers to the proportion of true signals correctly detected by a surveillance algorithm.

In our recent work [11], we sought a new feature by moving beyond using only keywords/medical conditions. We proposed to analyze the diversity metrics of tweets over time, so-called *temporal diversity*. The diversity statistics can capture a broad spectrum of topics, communities and knowledge that are evolving over time. In particular, analyzing temporal diversity can shed light on two aspects. First, an increase of content diversity over time indicates that a community is broadening its area of interest. Second, negative peaks in diversity can additionally reveal a temporary focus on specific events. To address an efficiency issue, we employed an algorithm based on sampling [6]. We performed a correlation analysis of the temporal diversity of 14 real-world events with their estimated event magnitudes during the known outbreak periods. Our analysis showed that correlation results are varied greatly among outbreaks reflecting the characteristics (severity and duration) of outbreaks.

[4]`http://www.promedmail.org`

4.2 Implications

As we aimed at detecting outbreak events for general diseases that are not only seasonal, but also sporadic diseases that occur in low tweet-density regions, some difficulties in constructing the outbreak ground truth still remain, which resulted in a dataset that was limited in terms of the number of outbreaks and their diversity. Particularly, the smaller the number of outbreaks we analyzed, the harder it was to generalize our solution. The process of creating the ground truth for disease outbreaks requires information extraction techniques, namely, different NLP tools for extracting relevant information. Unfortunately, the accuracy of such tools are not nearly 100%, which has a severe impact to the coverage (number) and quality of outbreak ground truth found. For example, place names are ambiguous and can be wrongly determined as the country of an outbreak as illustrated in this sentence *The **Uganda** Virus Research confirms Ebola virus Sudan species*. In addition, the accuracy of information extraction techniques as well as the noisiness of ProMED-mail data have also limited the coverage and quality of ground truth. For instance, there are many near-duplicate reports of outbreaks and many of irrelevant reports related to disease vaccines instead of outbreaks. Moreover, a report on historical statistics of a disease outbreak is irrelevant information, which should be carefully excluded from the ground truth. Similar to information about updates on an outbreak situation that must be avoided.

5. CHALLENGE III: SUPPORTING THREAT ASSESSMENT

For detected events, public health experts participating in its investigation face the overwhelming task of analyzing the large number of tweets associated to the corresponding signals [21]. The real-time nature of Twitter, on the one hand makes it attractive for public health surveillance; yet, on the other, the volume of tweets also makes it harder to: 1) capture the information transmitted, 2) compute sophisticated models on large pieces of the input, and 3) store the input data, which can be significantly larger than the algorithm's available memory [17].

5.1 Proposed Approach

To reduce this information overload and support the task of threat assessment, we explored to what extent recommender systems techniques can help to filter information items according to the experts' context and preferences. Our previous work [7] has shown the effectiveness of *Personalized Tweet Ranking for Epidemic Intelligence* for a case study of the 2011 EHEC outbreak in Germany. In particular, we focused on a personalized learning to rank approach that ultimately offers the user the most relevant and attractive tweets to support the task within her/his context. Our approach extended a learning to rank framework by considering a personalized setting that exploits a user's individual *context*. We considered such context as implicit criteria for selecting tweets of potential relevance, and guiding the recommendation process. We used the terms in the expanded context that correspond to medical conditions, locations or complementary context (that corresponds to the set of nouns, which are neither locations nor medical conditions) in order to to build a set of tweets by querying our collection. This step helped us to filter irrelevant tweets. Next, we elicited judgments from experts on a subset of the tweets retrieved in order to build a ranking function model. We then obtained

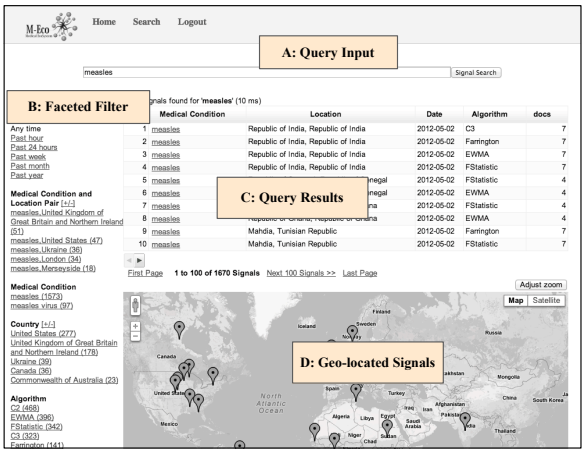

Figure 2: User Interface showing *A. Query Input*: for entering a search term, e.g., "measles", *B. Faceted Filter*: options for filtering signal search results by metadata, *C. Query results*: result set of signals, and *D. Geo-located Signals*: a map for visualizing signals' geo-location.

for each labeled tweet, a feature vector that help us to train our personalized ranking model. Finally, we applied a learning to rank algorithm to obtain the ranking function for the given user context.

In addition to the personalized learning to rank approach, we present the first prototype for support search and retrieval of signals. We envision the functionality of *signal-based* retrieval, that is, returning signals as results of a given query instead of only documents. Once the desired signals are obtained, the user is able to access the original tweets associated to each of them. Having signals as basic unit of information allows us to perform a focused indexing of only the tweets relevant to a particular signal. Figure 2 shows the user interface along with a brief description of its main panels. A possible solution is to implement a ranking model that: (1) extends a learning to rank framework by considering a personalized setting that exploits a user's individual *context*; (2) answers user's query by providing a list of relevant tweets ordered from newest to oldest, starting from the time the query was issued. When selecting tweets to include in the list, systems should favor both the *relevance* and *recency* of tweets.

5.2 Implications

However, the current load of experts in assessing these signals can be reduced significantly by employing personalized ranking techniques. Given that experts' interactions and explicit feedback are scarce in EI systems, the application of standard recommender system algorithms is not straightforward making it harder to build effective models for ranking or recommendation. By exploiting complementary context information, extracted from the social hash-tagging, and the latent topics discovered within the tweets, an effective ranking mechanism for messages associated with signals can be achieved. As a plan for future work, supporting temporal analytics for public health events will bring EI systems a big step forward, and also can provide useful guidance for other systems based on using social media data.

Acknowledgments This work was partially funded by the European Commission Seventh Framework Program (FP7 / 2007-2013) under grant agreement No.247829 for the Medical Ecosystem Project (M-Eco).

6. REFERENCES

[1] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, 2011.

[2] M. Basseville and I. Nikiforov. *Detection of abrupt changes: theory and application*. Prentice-Hall information and system sciences series. Prentice Hall.

[3] N. Collier. What's Unusual in Online Disease Outbreak News? *Journal of Biomedical Semantics*, 1(1):2, 2010.

[4] N. Collier and S. Doan. Syndromic classification of twitter messages. In *eHealth*, pages 186–195, 2011.

[5] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, 2010.

[6] F. Deng, S. Siersdorfer, and S. Zerr. Efficient jaccard-based diversity analysis of large document collections. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012.

[7] E. Diaz-Aviles, A. Stewart, E. Velasco, K. Denecke, and W. Nejdl. Epidemic intelligence for the crowd, by the crowd. In *International AAAI Conference on Weblogs and Social Media*, 2012.

[8] M. Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84, 2012.

[9] M. Emch, C. Feldacker, M. S. Islam, and M. Ali. Seasonality of cholera from 1974 to 2005: a review of global patterns. *International Journal of Health Geographics*, 7(1), 2008.

[10] C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society*, 159(3):pp. 547–563, 1996.

[11] N. Kanhabua and W. Nejdl. Understanding the diversity of tweets in the time of outbreaks. In *Proceedings of the 22nd international conference on World Wide Web companion*, 2013.

[12] N. Kanhabua, S. Romano, and A. Stewart. Identifying relevant temporal expressions for real-world events. In *SIGIR Workshop on Time-aware Information Access*, 2012.

[13] N. Kanhabua, S. Romano, A. Stewart, and W. Nejdl. Supporting temporal analytics for health-related events in microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012.

[14] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, 2010.

[15] V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *ACM Trans. Intell. Syst. Technol.*, 3(4):72:1–72:22, September 2012.

[16] C. Lumezanu, N. Feamster, and H. Klein. #bias: Measuring the tweeting behavior of propagandists. In *Proceedings of the Sixth International Conference on Weblogs and Social Media*, 2012.

[17] S. Muthukrishnan. *Data streams: algorithms and applications*. Now Publishers, 2005.

[18] M. Paul and M. Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 2011.

[19] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[20] E. T. K. Sang and J. Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, 2012.

[21] G. Shmueli and H. Burkom. Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. *Technometrics*, 52(1):39–51, February 2010.