Epidemic Intelligence for the Crowd, by the Crowd*

Ernesto Diaz-Aviles[1], **Avaré Stewart**[1], **Edward Velasco**[2],
Kerstin Denecke[1], and **Wolfgang Nejdl**[1]
[1]L3S Research Center / University of Hannover. Hannover, Germany
{*diaz, stewart, denecke, nejdl*}*@L3S.de*
[2]Robert Koch Institute. Berlin, Germany
VelascoE@rki.de

Abstract

Tracking Twitter for public health has shown great potential. However, most recent work has been focused on correlating Twitter messages to influenza rates, a disease that exhibits a marked seasonal pattern. In the presence of sudden outbreaks, how can social media streams be used to strengthen surveillance capacity? In May 2011, Germany reported an outbreak of *Enterohemorrhagic Escherichia coli* (EHEC). It was one of the largest described outbreaks of EHEC worldwide and the largest in Germany. In this work, we study the *crowd*'s behavior in Twitter during the outbreak. In particular, we report how tracking Twitter helped to detect key user messages that triggered signal detection alarms before MedISys and other well established early warning systems. We also introduce a personalized learning to rank approach that exploits the relationships discovered by: (i) latent semantic topics computed using Latent Dirichlet Allocation (LDA), and (ii) observing the social tagging behavior in Twitter, to rank tweets for epidemic intelligence. Our results provide the grounds for new public health research based on social media.

1 Epidemic Intelligence Based on Twitter

Public health officials are faced with new challenges for outbreak alert and response. This is due to the continuous emergence of infectious diseases and their contributing factors such as demographic change, or globalization. Early reaction is necessary, but often communication and information flow through traditional channels is slow. *Can additional sources of information, such as social media streams, provide complements to the traditional epidemic intelligence mechanisms?*

Epidemic Intelligence (EI) encompasses activities related to early warning functions, signal assessments and outbreak investigation. Only the early detection of disease activity, followed by a rapid response, can reduce the impact of epidemics. Recent works have shown the potential of using Twitter for public health, and its real-time nature makes it even more attractive for public health surveillance. These works have either focused on: the text classification and filtering of tweets, e.g., (Sofean et al. 2012); or finding predictors for diseases that exhibit a seasonal pattern (i.e., influenza-like illnesses) by correlating selected keywords with official influenza statistics and rates, e.g., (Lampos and Cristianini 2011; Signorini, Segre, and Polgreen 2011). Still others have focused on mining Twitter content for topic and aspect modeling (Paul and Dredze 2011). Furthermore, these existing approaches have all focused on countries where the tweet density is known to be high (e.g., the UK, or U.S.). In contrast these studies, ours focuses on a sudden outbreak of a disease that does not involve any seasonal pattern. Moreover, our work shows the potential of Twitter in countries where the tweet density is significantly lower, such as Germany (Semiocast 2012).

In this paper, we seek to address the issues that can help deliver a public health surveillance system based on Twitter, by taking into account two important stages in epidemic intelligence: *Early Outbreak Detection* and *Outbreak Analysis and Control*, and take up the following questions:

1. *Early Outbreak Detection*: Is it possible, by only using Twitter, to find early cases of an outbreak, before well established systems?
2. *Outbreak Analysis and Control*: Is it possible to use Twitter to understand the potential causes of contamination and spread? and How can we provide support for public health official to analyze and assess the risk based on the available social media information?

The contributions of this paper are summarized as follows:

- We provide an example of the application of standard surveillance algorithms on Twitter data collected in real-time during a major outbreak of EHEC in Germany, and provide insights showing the potential of Twitter for early warning.
- This paper presents an innovative personalized ranking approach that offers decision makers the most relevant and attractive tweets for risk assessment, by exploiting latent topics and social hash-tagging behavior in Twitter.

2 Twitter for Early Warning

The continuous emergence of infectious diseases and their contributing factors impose new challenges to public health officials. Early reaction is necessary, but often communication and information flow through traditional channels is slow. Additional sources of information, such as social media streams, provide complements to the traditional reporting mechanisms.

*The full version of this work is available at http://arxiv.org /abs/1203.1378

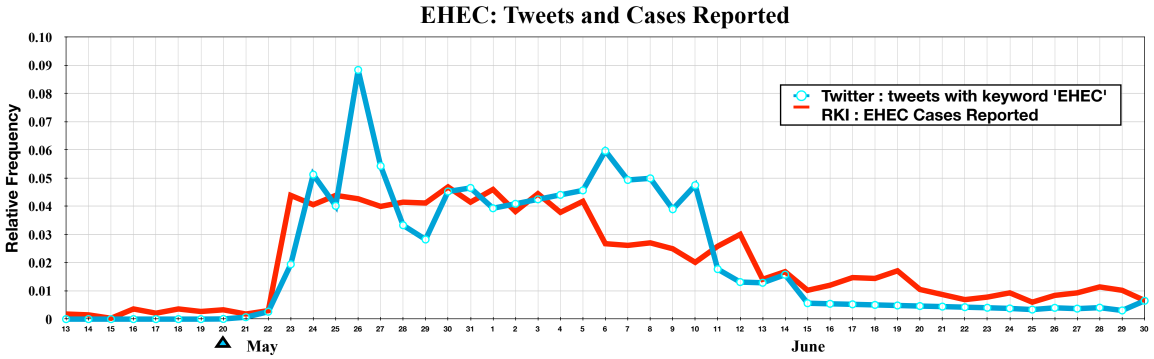

Figure 1: **Relative frequency of cases reported to RKI and the number of tweets mentioning the name of the disease: *EHEC*. Monitoring Twitter allowed us to generate the first signal on Friday, May 20th, 2011, using standard biosurveillance methods, before well established early warning systems.**

For example, if we observe Figure 1, we can see two plots, one of them corresponds to the relative frequency of EHEC cases as reported by RKI (Robert Koch Institute (RKI) 2011), and the other to the relative frequency of mentions of the keyword "EHEC" in the tweets collected during the months of May and June 2011. We can appreciate the high correlation of the curves, which corresponds to a Pearson correlation coefficient of 0.864. We can also observe the *inertia* of the crowd that continued tweeting about the outbreak, even though the number of cases were already declining (e.g., June 5 to 11).

Twitter has shown potential as a source of information for public health event monitoring, but could it be possible to generate an early warning signal before well established systems by only tracking Twitter? In this section, we have a closer look to the time period of the EHEC outbreak in Germany, and address this question.

Data Collection As part of the European project *M-Eco* (meco-project.eu), we currently monitor over 500 diseases and symptoms, which include "EHEC". During May and June, 2011, we incrementally collected 7,710,231 tweets related to medical conditions, 456,226 of them were related to the EHEC outbreak in Germany, and were produced by 54,381 distinct users.

Detection Methods The surveillance algorithms we used are well documented in the disease aberration literature e.g. (Khan 2007). The objective of these algorithms is to detect aberration patterns in time series data when the volume of an observation variable exceeds an expected threshold value. In our case, for example, the observation variable corresponds to mentions of medical condition "EHEC" withing the tweets.

The five biosurveillance algorithms we used for early detection are: the Early Aberration Reporting System (EARS) (1) C1, (2) C2, and (3) C3 algorithms, (4) F-statistic, and (5) Exponential Weighted Moving Average (EWMA). Please refer to (Khan 2007) for a detailed introduction. We signal an alarm if the test statistic reported by the detection methods exceeds a threshold value, which is determined experimentally. Detection methods' parametrization follows the guidelines of N. Collier (Collier 2010) and is detailed in (Diaz-Aviles et al. 2012).

Using any of the detection methods, a daily count less than five tweets was enough to signal an alert on May 20th, 2011. The Early Warning and Response System (EWRS) of the European Union received a first communication by the German authorities on Sunday May 22. MedISys detected the first media report in the German newspaper *Die Welt* on Saturday May 21 (Linge et al. 2011) and ProMED-mail and all other major early alerting systems (e.g., ARGUS, Biocaster, GPHIN, HealthMap, PULS) covered the event on Monday May 23.

Why was this early detection possible with respect to well established early warning systems? We tracked only Twitter as source of information, in contrast to MedISys for example, that tracks hundreds of news sources on the Internet. We consider Twitter's *diversity* was the key element that helped in the earlier detection of the event.

Twitter is a diverse stream of *multiple sources*. In Twitter converges the *contribution from the crowd* - millions of individual users obscure and renown; big and small media outlets; global and local newspapers, etc. Our work and that of MedISys focus on an analysis at a national level, but there are cases where support for the *local perspective* is important, for example local and smaller news papers reaching a broader audience through Twitter.

3 Twitter for Outbreak Analysis and Control

For public health officials, who are participating in the investigation of an outbreak, the millions of documents produced over social media streams represent an overwhelming amount of information for risk assessment.

To reduce this overload we explore to what extent recommender systems techniques can help to filter information items according to the public health users' context and preferences (e.g., disease, symptoms, location). In particular, we focus on a personalized learning to rank approach that ultimately offers the user the most relevant and attractive tweets for risk assessment. In this section, we introduce our approach and report an experimental evaluation on the EHEC dataset collected from Twitter.

Background: Learning to Rank for IR

In learning to rank for information retrieval (Liu 2009), given a query q, we rank a document d using a ranking model obtained by training over a set of labeled query-document pairs using a learning algorithm. An unseen query-document pair (q, d) will be ranked according to a weighted sum of feature scores: $f(q, d) = \vec{w} \cdot \phi(q, d)$, where $\phi(q, d)$ are the

Algorithm 1 Personalized Tweet Ranking algorithm for Epidemic Intelligence (PTR4EI)

```
Input: User Context $C_u = (t, MC_u, L_u)$,
    Inverted index $\mathcal{T}$ of tweets collected for epidemic intelligence before time $t$
Output: Ranking Function $f_{C_u}$ for User Context $C_u$
Compute LDA topics ($topicsLDA$) on $\mathcal{T}$
Consider each $mc \in MC_u$ as a hash-tag, and extract from $\mathcal{T}$ all co-occurring hash-tags: $coHashTags$
Classify the terms in $topicsLDA$ and the hash-tags in $coHashTags$ as Medical Condition $MC_x$, Location $L_x$ or Complementary Context[1] $CC_x$
Build a set of queries as follows:
    $Q = \{q \mid q \in MC_u \times \mathcal{P}(\{L_u \cup MC_x \cup L_x \cup CC_x\})\}$
For each query $q_i \in Q$ obtain tweets $D$ from the collection $\mathcal{T}$
Elicit relevance judgments $Y$ on a subset $D_y \subset D$
For each tweet $d_j \in D$, obtain the feature vector $\phi(q_i, d_j)$ w.r.t. $(q_i, d_j) \in Q \times D$
Apply learning to rank to obtain a ranking function for the user context $C_u$: $f_{C_u}(q, d) = \vec{w} \cdot \phi(q, d)$
return $f_{C_u}(q, d)$
```

different features extracted from the (q, d) pair. The goal of the algorithm is to learn the weight vector $\vec{w}$ using a training set of queries and documents, in order to minimize a given loss function (e.g., error rate, degree of agreement between the two rankings, classification accuracy or mean average precision).

Ranking Tweets for Epidemic Intelligence

We propose to use the user *context* as implicit criteria to select tweets of potential relevance, that is, we will rank and derive a short list of tweets based on the user context. The user context C_u is defined as a triple

$$C_u = (t, MC_u, L_u) \ , \tag{1}$$

where t is a discrete time interval, MC_u the set of medical conditions, and L_u the set of locations of user interest.

Our learning approach, PTR4EI, is shown in Algorithm 1. We build upon a learning to rank framework by considering a personalized setting that exploits user's individual context.

More precisely, we consider the context of the user, C_u, and prepare a set of queries, Q, for a target event (e.g., a disease outbreak). We first compute LDA (Blei, Ng, and Jordan 2003) on an indexed collection $\mathcal{T}$ of tweets for epidemic intelligence, where not all tweets are necessarily interesting for the target event.

We also extract the hash-tags that co-occur with the user context by considering the medical conditions and locations in C_u as hash-tags themselves, and find which other hash-tags co-occur with them within a tweet, and how often they co-occur, which will help us to select the most representative hash-tags for the target event.

The set Q is constructed by expanding the original terms in C_u with the ones in the LDA topics and co-occurring hash-tags, which are previously classified as medical condition, location or complementary context. This phase of the algorithm can be considered a particular case of the *query expansion* task in information retrieval, where search terms are named entities (i.e., medical conditions, locations, name of people, organizations, etc.), whose implicit correlations are discovered in the reduced dimensional space induced by the latent topics and top co-occurring hash-tags.

We build the set D of tweets by querying index $\mathcal{T}$ using $q \in Q$ as query terms. Next, we elicit judgments from experts on a subset of the tweets retrieved, in order to construct $D_y \subset D$, where $|D_y| \ll |D|$. For each tweet $d_j \in D$, we obtain the features vector $\phi(q_i, d_j)$, with respect to the pair $(q_i, d_j) \in Q \times D$.

Finally, with these elements, we apply a learning to rank algorithm to obtain the ranking function for the given user context. The ranking function is applied to rank existing and new incoming tweets.

In the rest of the section, we evaluate our approach considering as event of interest the EHEC outbreak in Germany, 2011.

Experiments and Evaluation To support users in the assessment and analysis during the EHEC outbreak, we set the user context as C_u = ([{2011-05-23; 2011-06-19}], {“EHEC”}, {“Lower Saxony”}), in this way, we are taking into account the main period of the outbreak, the disease of interest, and the German state having the most reported cases.

Following Algorithm 1, we computed LDA and extracted the co-occurring hash-tags using the indexed collection $\mathcal{T}$ described in Section 2. Table 1a shows, as example, four LDA topics for week 22 of the time period of interest, and Table 1b presents the hash-tags co-occurring with *#EHEC*.

We asked three experts: one from the Robert Koch Institute and the other two from the Lower Saxony State Health Department to provide their individual judgment on a subset D_y of 240 tweets, evaluating for each tweet, if it was relevant or not to support their analysis of the outbreak. Any disagreement in the assigned relevance scores were resolved by majority voting.

For each tweet, we prepared five binary features: F_{MC}, F_L, $F_{\text{#-tag}}$, F_{CC}, and F_{URL}. We set the corresponding feature value equal *true* if a medical condition, location, hash-tag, complementary context term, or URL were present in the tweet, and*false* otherwise. For learning the ranking function, we used Stochastic Pairwise Descent algorithm (Sculley 2009).

We compared our approach, that expand the user context with latent topics and social generated hash-tags, against two ranking methods:

1. **RankMC**: It learns a ranking function using only medical conditions as feature, i.e., F_{MC}. Please note, that this baseline also considers related medical conditions to the ones in MC_u, which makes it stronger than non-learning approaches, such as BM25 or TF-IDF scores, that use only the MC_u elements as query terms.
2. **RankMCL**: It is similar to RankMC, but besides the medical conditions, it uses a local context to perform the ranking (i.e., features: F_{MC} and F_L).

We randomly split the dataset into 80% training tweets, which will be used to compute the ranking function, and 20% testing tweets. To reduce variability, we performed the

[1] **Complementary Context** CC is defined as the set of nouns, which are neither Locations nor Medical Conditions, e.g., names of persons, organizations or affected organisms. $CC \cap (L \cup MC) = \emptyset$

Topic 1	Topic 2	Topic 3	Topic 4
EHEC *(MC)*	EHEC *(MC)*	EHEC *(MC)*	EHEC *(MC)*
dead *(MC)*	intestinal germ *(MC)*	cucumbers *(CC)*	cucumbers*(CC)*
Germany *(L)*	source *(-)*	pathogen *(MC)*	salad *(CC)*
people *(-)*	search *(-)*	Spain *(L)*	pain *(MC)*
live *(-)*	Hamburg *(L)*	farmers *(CC)*	women *(CC)*

(a) LDA topics

Medical Condition	Location	Complementary Context	
bacteria	berlin	cucumbers	bild
diarrhea	germany	obst	fdp
ehec_pathogen	hamburg	salad	n24
hus	lübeck	terror	rki
intestinal_infection	spain	tomatoes	rtl

(b) Hash-tags co-occurring with *#EHEC*

Table 1: **Example of (a) four LDA topics (columns) and (b) hash-tags co-occurring with *#EHEC* computed between May 30 and June 5, 2011 (week 22). We classify terms within each topic as *Medical Condition (MC)*, *Location (L)*, or *Complementary Context (CC)* entities. Terms outside these categories are ignored.**

Method	P@1	P@3	P@5	P@10
RankMC (baseline)	90 %	73.34 %	64 %	69 %
RankMCL (baseline)	90 %	83.33 %	88 %	85 %
PTR4EI	**100 %**	**90 %**	**94 %**	**96 %**

Table 2: **Ranking Performance in terms of P@$\{1, 3, 5, 10\}$**

experiment using ten different 80/20 partitions. The test set is used to evaluate the ranking methods. The reported performance is the average over the ten rounds.

Evaluation Measures For evaluation, we used three evaluation measures widely used in information retrieval, namely precision at position n ($P@n$), mean average precision (MAP), and normalized discount cumulative gain (NDCG) (Baeza-Yates and Ribeiro-Neto 2011).

Results[2] The ranking performance in terms of precision is presented in Table 2, MAP and NDCG results are shown in Figure 2. As we can appreciate PTR4EI outperforms both baselines. Local information helps RankMCL to beat RankMC, for example MAP improves from 71.96% (RankMC) up to 81.82% (RankMCL). PTR4EI, besides local features, exploits complementary context information and particular Twitter features, such as the presence of hashtags or URLs in the tweets, this information allows it to improve its ranking performance even further, reaching a MAP of 91.80%. A similar behavior is observed for precision and NDCG, where PTR4EI is statistically significantly better than RankMC and RankMCL.

4 Conclusion and Future Directions

We have shown the potential of Twitter to trigger early warnings in the case of sudden outbreaks and how personalized ranking for epidemic intelligence can be achieved. We believe our work can serve as a building block for an open early warning system based on Twitter, and hope that this paper provides some insights into the future of epidemic intelligence based on social media streams. Currently we are working closely with German and international public health institutions to help them integrate monitored social media into their existing surveillance systems.

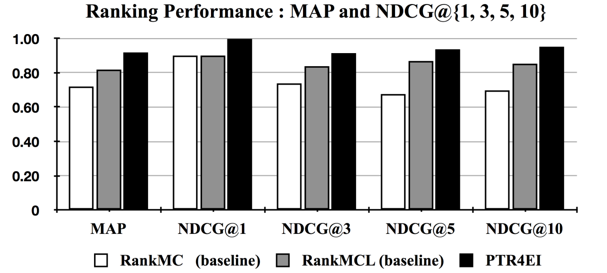

Figure 2: **MAP and NDCG Results.**

As future work, we plan to scale up our experiments, and to apply techniques of online ranking in order to update the model more efficiently as the outbreak develops.

We hope that this paper provides some insights into the future of epidemic intelligence based on social media streams.

Acknowledgments This work was funded, in part, by the European Commission FP7/2007-2013 under grant agreement No.247829 for the M-Eco Medical Ecosystem Project.

References

Baeza-Yates, R., and Ribeiro-Neto, B. 2011. *Modern Information Retrieval.* Addison Wesley, 2nd edition.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3:993–1022.

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

Diaz-Aviles, E.; Stewart, A.; Velasco, E.; Denecke, K.; and Nejdl, W. 2012. Epidemic Intelligence for the Crowd, by the Crowd (full version). http://arxiv.org/abs/1203.1378.

Khan, S. A. 2007. Handbook of Biosurveillance. *Journal of Biomedical Informatics.*

Lampos, V., and Cristianini, N. 2011. Nowcasting Events from the Social Web with Statistical Learning. *ACM Transactions on Intelligent Systems and Technology (TIST).*

Linge, J.; Mantero, J.; Fuart, F.; Belyaeva, J.; Atkinson, M.; and Van Der Goot, E. 2011. Tracking Media Reports on the Shiga Toxin-Producing Escherichia coli O104:H4 Outbreak in Germany. In *ICST Conference on eHealth, 2011.*

Liu, T.-Y. 2009. Learning to Rank for Information Retrieval. *Found. Trends Inf. Retr.* 3:225–331.

Paul, M., and Dredze, M. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM).*

Robert Koch Institute (RKI). 2011. Final Presentation and Evaluation of Epidemiological Findings in the EHEC O104:H4 Outbreak, Germany 2011. Technical report. http://goo.gl/9tciB.

Sculley, D. 2009. Large Scale Learning to Rank. In *NIPS 2009 Workshop on Advances in Ranking.*

Semiocast. 2012. Countries on Twitter. http://goo.gl/RfxZw.

Signorini, A.; Segre, A. M.; and Polgreen, P. M. 2011. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE.*

Sofean, M.; Stewart, A.; Denecke, K.; and Smith, M. 2012. Medical case-driven classification of microblogs: Characteristics and annotation. In *ACM IHI 2012.*

[2] The improvements between PTR4EI and the baselines for all measures reported are statistically significant based on pairwise t-tests (p-value < 0.02).