# Tracking Twitter for Epidemic Intelligence

## Case Study: EHEC/HUS Outbreak in Germany, 2011

**Ernesto Diaz-Aviles**
L3S Research Center/University of Hannover
Hannover, Germany
diaz@L3S.de

**Avaré Stewart**
L3S Research Center/University of Hannover
Hannover, Germany
stewart@L3S.de

### ABSTRACT

In the presence of sudden outbreaks, how can social media streams be used to strengthen surveillance capacity? In May 2011, Germany reported one of the largest described outbreaks of *Enterohemorrhagic Escherichia coli* (EHEC). The Shiga toxin-producing strain O104:H4 infected several thousand people, frequently leading to haemolytic uremic syndrome (HUS) and gastroenteritis (GI). By the end of June, 47 persons had died. In this work, we study the *crowd*'s behavior in Twitter during the outbreak. In particular, we present how Twitter can be exploited to support Epidemic Intelligence (EI) in the tasks of early warning, signal assessment and outbreak investigation. A user study with experts from the Robert Koch Institute, Germany's national-level public health authority, and from Lower Saxony State Health Department (NLGA) provide important insights towards the realization of an open early warning system based on Twitter, helping to realize the vision of *Epidemic Intelligence for the Crowd, by the Crowd*.

**Author Keywords:** Epidemic Intelligence; Medicine 2.0; Twitter.

**ACM Classification Keywords:** H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*Information Filtering*; K.4 [**Computer and Society**]: General

**General Terms:** Algorithms, Performance, Experimentation

### 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. In this paper, we explore the behavior of twitter users during the EHEC/HUS outbreak in Germany. 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*.

In contrast to previous works, e.g., [7], ours focuses on a *sudden outbreak* of a disease that does not involve any seasonal pattern. Moreover, our study shows the potential of Twitter in countries where the tweet density is significantly lower, such as Germany [9]. The contributions of this paper are summarized as follows:

- We provide insights showing the potential of Twitter for early warning by tracking Twitter in real-time during a major outbreak of EHEC/HUS in Germany.
- We present the results of a user study with field experts, and show how the information extracted from Twitter can support the tasks of outbreak analysis and control, by exploiting latent topics and social hash-tagging.

### TWITTER FOR EARLY OUTBREAK DETECTION

In the onset of an outbreak, 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. 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/HUS outbreak in Germany, and address this question.

#### Data Collection

In the context of the European project M-Eco[1], we monitor over 500 diseases and symptoms on Twitter. 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.

#### Tracking the Outbreak on Twitter

Figure 1 shows the epidemiological curves (relative frequency) during the main period of the outbreak, i.e., the months of May and June 2011. The curves correspond to: (i) EHEC cases as reported by RKI [8], and (ii) tweets containing the keyword "EHEC". We can 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).

[1] **Medical Ecosystem Personalized Event-Based Surveillance – M-Eco**: `meco-project.eu`

*WebSci'12*, June 22–24, 2012, Evanston, Illinois, USA.
Copyright 2012 ACM 978-1-4503-1228-8. ...$10.00

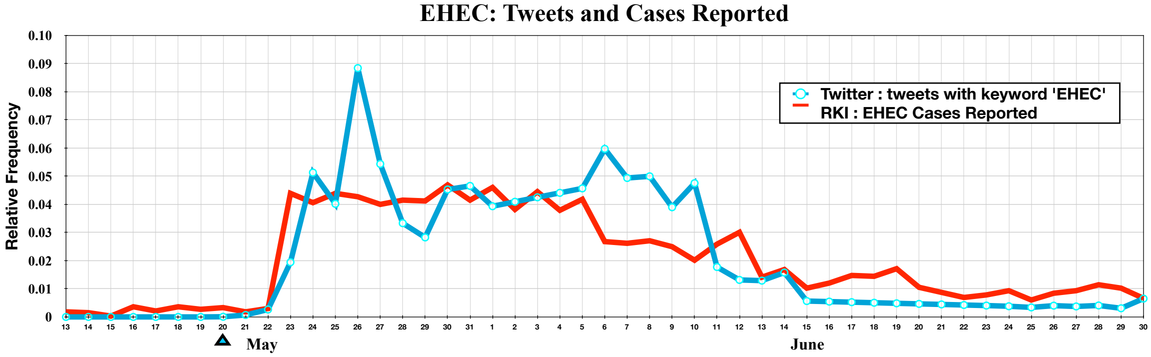

**Figure 1. Relative frequency of: (i) cases reported to RKI and (ii) tweets mentioning the name of the disease: *EHEC*. Tracking Twitter for epidemic intelligence allowed us to detect the first tweets of the sudden outbreak on Friday, May 20th, 2011 (triangle on the time axis).**

| No | Tweet | User | Location in Germany |
|---|---|---|---|
| 1 | Hannover: Gefahr durch EHEC-Erreger http://bit.ly/l2bJwN | BS_Zeitung | Braunschweig |
| 2 | Hannover: Gefahr durch EHEC-Erreger http://bit.ly/l2bJwN | WN_Wolfsburg | Wolfsburg |
| 3 | Hannover: Gefahr durch EHEC-Erreger http://bit.ly/l2bJwN | SZ_Zeitung | Salzgitter |
| 4 | 20. Mai 2011, 19:21 Uhr - Infizierte mit lebensgefhrlichem EHEC-Erregern http://on.bild.de/kDqgVr | BILD_Hannover | Hannover |
| 5 | 20. Mai 2011, 19:48 Uhr - Mehrere Hamburger mit EHEC-Erreger infiziert http://on.bild.de/kuMnCV | BILD_Hamburg | Hamburg |
| 6 | 20. Mai 2011, 19:48 Uhr - Mehrere Hamburger mit EHEC-Erreger infiziert: In Hamburg haben sich mehrere Menschen m... http://bit.ly/jEfoHw | Hamburg_ | Hamburg |
| 7 | Mehrere Hamburger mit EHEC-Erreger infiziert: Hamburg (dpa/lno) - In Hamburg haben sich mehrere Menschen mit dem... http://bit.ly/lRM5Kr | Lokales_Hamburg | Hamburg |
| 8 | Mehrere Hamburger mit EHEC-Erreger infiziert: Hamburg (dpa/lno) - In Hamburg haben sich mehrere Menschen mit dem... http://bit.ly/m7ZQWp | Insel Hiddensee | Hiddensee Island |
| 9 | Mehrere Hamburger mit EHEC-Erreger infiziert http://bit.ly/jt9uHA | schleswigbiz | Schleswig |

**Table 1. The early tweets gathered related to the outbreak on May 20, 2011. These few tweets were enough to trigger an alert using a *moving average* biosurveillance detection method. The Early Warning and Response System (EWRS) of the European Union received a first communication by the German authorities on Sunday May 22.**

If we fed the Twitter data to a *moving average* biosurveillance detection method [5][2], a daily count of nine tweets was enough to signal an alert on May 20th, 2011. The Early Warning and Response System (EWRS) [3] of the European Union received a first communication by the German authorities on Sunday May 22nd. MedISys [4] detected the first media report in the German newspaper *Die Welt* [5] on Saturday May 21st [6] and ProMED-mail [6] and all other major early alerting systems (e.g., ARGUS, Biocaster, GPHIN, HealthMap, PULS) covered the event on Monday May 23rd.

A closer look to the day of May 20th, reveals that the first alarm was triggered based on nine tweets, the actual messages are shown in Table 1, all of them generated from sources in northern Germany, not far from where the first cases of the outbreak were reported. Observe that the early tweets detected were generated by local newspapers, those Twitter users acted as *local sensors*, producing tweets that spread the news faster than the official channels and main stream media.

## TWITTER FOR OUTBREAK ANALYSIS AND CONTROL

For public health officials, who are participating in the investigation of an outbreak, the millions of tweets generated represent an overwhelming amount of information for risk assessment. The investigator must drill down into the tweets to learn more about the source of information, the content of the documents, the locations mentioned, etc., which contributed to the aberration. There may be potentially many tweets for an investigator to examine, so support mechanisms are required to help them during the document analysis stage.

In order to reduce this overload, we explore to what extent dimensionality reduction and ranking techniques can help to filter information items according to the public health users' context and preferences (e.g., disease, symptoms, location). Formally, we define the user context $C_u$ the triple:

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

[2] Moving average with parameters: training window of 4 days, a buffer of 1 day, and upper control limit equal to a mean plus three standard deviations

[3] **EWRS**: ewrs.ecdc.europa.eu

[4] **MedISys**: medusa.jrc.it/medisys

[5] **Die Welt**: welt.de

[6] **ProMED-mail**: promedmail.org

where $t$ is a discrete time interval, $MC_u$ the set of medical conditions, and $L_u$ the set of locations of user interest. In our case, the user context corresponds to: $C_u$ = ([{2011-05-23; 2011-06-19}], {"EHEC"}, {"Lower Saxony"}).

In particular, we focus on discovering cues derived from the social interactions in Twitter. To this end, we compute a low-dimensional representation of the data using:

1. **The topics result of applying Latent Dirichlet Allocation (LDA)** on our collection of tweets. LDA [2] is a generative probabilistic model for collections of discrete data such as text corpora. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over terms. Four LDA topics are shown in Table 2.

2. **Hash-tagging behavior in Twitter** [7]. We extract the hash-tags that co-occur with the medical conditions and locations under investigation, and how often they co-occur, which will help us to select the most representative hash-tags for the target event. Hash-tags co-occuring with *#EHEC* are shown in Table 3.

We classify the terms in the resulting low-dimensional space as: *Medical Condition*, *Location*, or *Complementary Context*[8].

| | | | |
|---|---|---|---|
| **Week 21** | | | |
| EHEC *(MC)* | fever *(MC)* | EHEC *(MC)* | EHEC *(MC)* |
| cucumbers *(CC)* | pain *(MC)* | casualty *(-)* | pathogen *(MC)* |
| Spain *(L)* | headache *(MC)* | women *(CC)* | Northern Germany *(L)* |
| tomatoes *(CC)* | sniff *(MC)* | intestinal germ *(MC)* | diarrhea *(MC)* |
| salad *(CC)* | pain *(MC)* | panic *(MC)* | dead *(MC)* |
| **Week 22** | | | |
| 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)* |
| **Week 23** | | | |
| EHEC *(MC)* | headache *(MC)* | EHEC *(MC)* | EHEC *(MC)* |
| cucumber *(CC)* | pain *(MC)* | cucumbers *(CC)* | sprout *(CC)* |
| eu *(CC)* | fever *(MC)* | sprout *(CC)* | source *(-)* |
| crisis management *(-)* | people *(-)* | pathogen *(MC)* | suspicion *(-)* |
| farmers *(CC)* | cough *(MC)* | salad *(CC)* | hus *(MC)* |
| **Week 24** | | | |
| EHEC *(MC)* | headache *(MC)* | stomach ache *(MC)* | pain *(MC)* |
| germ *(MC)* | fever *(MC)* | sniff *(MC)* | bellyache *(MC)* |
| sprout *(CC)* | slept *(-)* | pain *(MC)* | cough *(MC)* |
| health *(MC)* | sniff *(MC)* | regions *(-)* | throat *(CC)* |
| all-clear *(CC)* | head *(CC)* | examined *(-)* | sniff *(MC)* |

**Table 2. Four LDA topics (columns) computed weekly during the main period of the outbreak: from May 23 to June 19, 2011. We classify terms within each topic as *Medical Condition (MC)*, *Location (L)*, or *Complementary Context (CC)*. Terms outside these categories are ignored.**

We found that both, LDA and co-occurring hash-tags, reflect the key aspects of the outbreak. For example, the main locations affected in northern Germany, the reporting on alleged E. coli contaminations in Spanish cucumber, tomatoes and salad in early stages of the investigation (Weeks 21 and 22), and the announcements by German authorities that bean sprouts were the source of infection (Week 23).

| Medical Condition | Location | Complementary Context | |
|---|---|---|---|
| **Week 21** | | | |
| bacteria | bremen | cucumber_salad | cdu |
| diarrhea | cuxhaven | cucumbers | edeka |
| ehec_victim | hamburg | ehec_vegetable | fdp |
| hus | münster | tomatoes | merkel |
| intestinal_infection | northern_germany | vegetables | rki |
| **Week 22** | | | |
| bacteria | berlin | cucumbers | bild |
| diarrhea | germany | obst | fdp |
| ehec_pathogen | hamburg | salad | n24 |
| hus | lübeck | terror | rki |
| intestinal_infection | spain | tomatoes | rtl |
| **Week 23** | | | |
| bacteria | bavaria | cucumbers | ehec_freei |
| diarrhea | berlin | salad | fdp |
| ehec_pathogen | germany | sojasprout | merkel |
| hus | hamburg | sprout | n24 |
| intestinal_infection | lower_saxony | | rki |
| **Week 24** | | | |
| bacteria | lower_saxony | donate_blood | |
| died | | ehec_free | |
| health | | sojasprout | |
| hus | | | |

**Table 3. Hash-tags co-occurring with *#EHEC* during May 23 and June 19, 2011, the main period of the outbreak. The hash-tags are classified as entities of type *Medical Condition*, *Location*, or *Complementary Context*, hash-tags out of these categories are discarded.**

How can this extracted information help in the investigation of the outbreak? In order to address this question, we conducted a user study with experts from the Robert Koch Institute and from the Lower Saxony State Health Department (NLGA). The details of the study are discussed in the next section.

## User Study

The user study asked the following question: *What is the best approach to rank small elements of textual content, such as Twitter messages, to support the investigation of an outbreak?*. The study considered a time sensitive split of the data, where no future evidence is used. The training data corresponds to week 22, and the test was conducted in week 23. This corresponds to a more realistic approach, where models are computed online with data that arrives on a stream.

Given the user context specified in Eq. 1, we evaluated two ranking strategies:

1. **Strategy 1: Re-rank tweets based on LDA topics** . We use the medical conditions, locations and complementary context identified in the LDA topic terms, and boost tweets containing this entities.

2. **Strategy 2: Re-rank tweets based on hash-tags** . We use hash-tags related to the user context to re-rank tweets.

As a baseline, we used TF-IDF scores to rank tweets (documents), for the disjunctive query built based on the terms of the user context $C_u$, i.e., "EHEC *OR* Lower Saxony".

We asked three experts: one from the Robert Koch Institute and the other two from the Lower Saxony State Health Department (NLGA)[9] to provide their individual judgment on the relevancy of tweets in a ranked list. We presented to each assessor a ranked list of 30 tweets, per strategy and entity, and 60 for the baseline. In total, each assessor evaluated 8 lists of 30 tweets each. Each assessor provided a judgment

[7] Hashtags are words or phrases prefixed with the symbol #, e.g., *#WebSci2012*, a form of metadata tag used to mark keywords or topics in a tweet. Hashtags were created by Twitter users as a way to categorize messages, the practice is now a Twitter standard.

[8] **Complementary Context** is defined as the set of nouns, which are neither Locations nor Medical Conditions. Complementary Context may include named entities such as names of persons, organizations, affected organisms, expressions of time, quantities, etc. We denote the set of named entities that represents the complementary context as $CC$, where $CC \cap (L \cup MC) = \emptyset$

[9] **NLGA**: nlga.niedersachsen.de

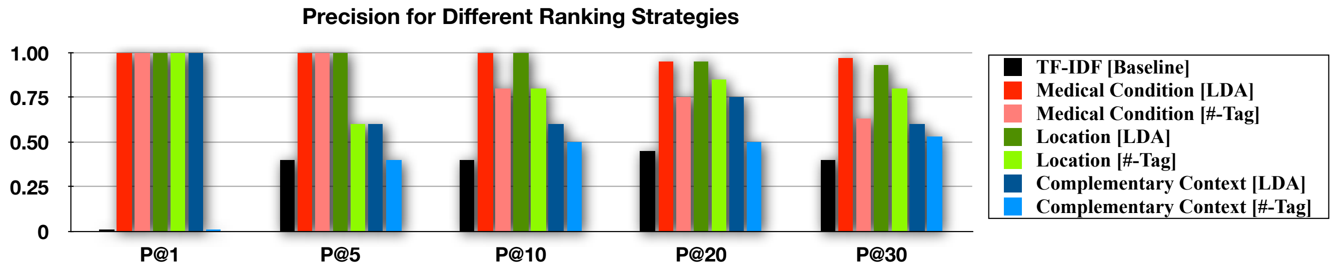

**Figure 2. Precision@N computed at different cut-off values of the list of tweets.**

of value of 1 if the tweet was relevant to support the investigation of the outbreak, 0 otherwise. Any disagreement in the assigned relevance scores were resolved by majority voting. All tweets where selected from the indexed collection discussed earlier in this paper.

The ranking metric used was *precision* at different cuts of the list or $P@N$ [1], where $N \in \{1, 5, 10, 20, 30\}$, which is defined as follows:

$$P@N = \frac{\text{\# of relevant tweets top-N results}}{N} \tag{2}$$

### Results of the Evaluation

Figure 2 shows the results of the user study. We can observe that both strategies largely outperform the baseline. In essence, both re-ranking methods discover new relationships that help identify more relevant tweets in the whole collection.

Additional medical conditions, identified using LDA or hash-tag co-occurrences, lead to better improvements in terms of precision. Followed by the locations, and finally, the complementary context.

All entities extracted from the latent topics tend to boost the ranking performance higher than the ones identified using hash-tags. But this improvement comes at a price, since keep tracking of recurring hash-tags is less expensive than computing LDA periodically.

## CONCLUSIONS AND FUTURE DIRECTIONS

To show the potential of Twitter for early warning, we focused on the recent EHEC/HUS outbreak in Germany, and monitored the Twitter social stream. We showed that early Tweets were detected and nine of them were enough to generate an alarm on May 20th, 2011, a day ahead of well established early warning systems, such as MedISys.

We learned that selecting features from a low dimensional space constructed using LDA and from the hash-tag co-occurrence helped in supporting our analysis of outbreak.

We are currently working closely with German and global public health institutions to help them integrate social media monitoring into their existing surveillance systems.

As future work, we plan to address the information overload faced by the domain experts during the document analysis stage, as a personalized document ranking problem. We have learned that in order to support personalization, we are faced with a limited context for building user profiles, moreover, the tweets themselves are sparse so this type of sparse text may contain no explicit information, which makes profile-document matching challenging. Initials results, derived from the analysis presented here, can be found in [3] and [4].

We have shown the potential of Twitter for epidemic intelligence in the presence of a particular sudden outbreak. 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.

**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

1. Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*, 2nd ed. Addison Wesley, 2011.
2. Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res. 3* (Mar. 2003), 993–1022.
3. Diaz-Aviles, E., Stewart, A., Velasco, E., Denecke, K., and Nejdl, W. Epidemic Intelligence for the Crowd, by the Crowd. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, June 4-7, 2012* (2012).
4. Diaz-Aviles, E., Stewart, A., Velasco, E., Denecke, K., and Nejdl, W. Towards Personalized Learning to Rank for Epidemic Intelligence Based on Social Media Streams. In *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012* (2012), 495–496.
5. Khan, S. A. Handbook of biosurveillance, m.m. wagner, a.w. moore, r.m. aryel (eds.). elsevier inc. isbn-13: 978-0-12-369378-5. *Journal of Biomedical Informatics* (2007).
6. Linge, J., Mantero, J., Fuart, F., Belyaeva, J., Atkinson, M., and Van Der Goot, E. Tracking media reports on the shiga toxin-producing escherichia coli o104:h4 outbreak in germany. In *ICST Conference on eHealth, 2011* (2011).
7. Paul, M., and Dredze, M. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)* (2011).
8. Robert Koch Institute (RKI). Final presentation and evaluation of epidemiological findings in the EHEC O104:H4 outbreak, Germany 2011. Tech. rep., Robert Koch Institute (RKI), September 2011. `http://goo.gl/9tciB`.
9. Semiocast. Countries on Twitter. http://goo.gl/RfxZw, 2012.