

# A New Age of Public Health: Identifying Disease Outbreaks by Analyzing Tweets

**Manuela Kriek**

Governmental Institute of Public Health of Lower Saxony  
Hannover, Germany  
Roesebeckstraße 4 - 6  
Manuela.Kriek@nlga.niedersachsen.de

**Johannes Dreesman**

Governmental Institute of Public Health of Lower Saxony,  
Hannover, Germany  
Roesebeckstraße 4 - 6  
Johannes.Dreesman@nlga.niedersachsen.de

**Lubomir Otrusina**

Brno University of Technology  
Brno, Czech Republic  
Bozetechova 2  
iotrosina@fit.vutbr.cz

**Kerstin Denecke**

L3S Research Center  
Hannover, Germany  
Appelstraße 9a  
denecke@L3S.de

## ABSTRACT

Traditional disease surveillance is a very time consuming reporting process. Cases of notifiable diseases are reported to the different levels in the national health care system before actions can be taken. But, early detection of disease activity followed by a rapid response is crucial to reduce the impact of epidemics. To address this challenge, alternative sources of information are investigated for disease surveillance. In this paper, the relevance of twitter messages outbreak detection is investigated from two directions. First, Twitter messages potentially related to disease outbreaks are retrospectively searched and analyzed. Second, incoming twitter messages are assessed with respect to their relevance for outbreak detection. The studies show that twitter messages can be – to a certain extent – highly relevant for early detecting hints to public health threats.

## Categories and Subject Descriptors

J.3 [Life and Medical Sciences]:

### General Terms

Experimentation

### Keywords

Twitter Analysis, Epidemic Intelligence, Content Analysis, User Study

## 1. INTRODUCTION

Increasing numbers of people are becoming digitally connected. In particular, younger generations use the World Wide Web as their primary source of information and communication. Twitter, Facebook or StudiVZ are commonly used tools to stay connected and informed. One of the new fast growing medium for communication and updates is twitter.com. Twitter Incorporation offers a social networking and micro blogging service which allows user to write messages up to 140 characters. The so-called tweets are posted on the users profile or users blog. These tweets are publicly accessible and searchable on Twitter search.

According to a recent report on twitter users [1] approximately two-third of all twitter users are located in the USA. Germany is with 1,51 % ranking sixth in the list of twitter user after UK,

Canada, Australia and Brazil. Further, studies from Epidemiology showed that SMS messaging is very fast in communicating information on outbreaks of infectious diseases [2]. It is still an open question to what extent disease outbreaks are reflected in Twitter messages and how to make use of them.

According to the law on German Protection against Infection Act (Infektionsschutzgesetz (IfSG), 2001) the traditional disease surveillance relies on data from mandatory reporting of cases by physicians and laboratories. They inform local county health departments (Landkreis) which in turn report to state health departments (Land). At the end of the reporting pipeline, the national surveillance institute (Robert Koch Institute) is informed about the outbreak. It is clear that these different stages of reporting take time and delay a timely reaction.

Since new media and communication tools are available and, as stated before, are in use by the population also for communicating information on diseases, the objective of this work is to assess the relevance of Twitter messages for outbreak detection. Further, it will be studied how and to what extent real disease outbreaks are reported in Twitter.

On the one hand, twitter messages have been collected regularly for some time period and explored retrospectively for information on known disease outbreaks. We investigate the time delay between identification of disease information in twitter messages and their notification at health departments. On the other hand, twitter messages containing descriptions of symptoms of diseases are analyzed with respect to their relevance for disease reporting.

The overall goal of this work is to identify hints to infectious diseases and in this way, contribute to an early detection of disease outbreaks and to the reduction of the spread of diseases.

The main contributions of this work are:

- Analysis of tweets with respect to their relevance for disease surveillance,
- Content analysis and content classification of tweets,
- Linguistic analysis of disease-reporting twitter messages,
- Recommendations on search patterns for tweet search in the context of disease surveillance.

The remainder of the paper is structured as follows. Section 2 provides an overview on related work. Twitter with its

characteristics and our observations on linguistic peculiarities are presented in Section 3. The methods for collecting Twitter data and for analysing the data are summarized in Section 4. Results from our assessments are described in Section 5 and are discussed in Section 6. The paper finishes with conclusions and remarks on future work in Section 7.

## 2. Related Work

Due to the increased availability of information in the Web, in the last years, a new research area has been developed, namely Infodemiology. It can be defined as the “science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy” [8]. As part of this research area, several kinds of data have been studied for their applicability in the context of disease surveillance. Google flu trends exploits the search behavior to monitor the current flu-related disease activity [9]. It could be shown by Carneiro and Mylonakis [10] that Google Flu Trends can detect regional outbreaks of influenza 7–10 days before conventional Centers for Disease Control and Prevention surveillance systems.

The focus in our work is on Twitter messages and their relevance for disease outbreak detection. It has been reported already that especially tweets are useful to predict outbreaks such as a Norovirus outbreak at a university [3]. Chew et al. [5] analysed twitter news during the influenza epidemic 2009. They compared the use of the term “H1N1” and “swine flu” over the time. Furthermore, they analysed the content of the tweets (ten content concepts) and validated twitter as a the real time content. They analysed the data via Infovigil an infosurveillance system by using an automated coding. To find out if there is a relationship between automated and manual coding, the tweets were evaluated by a Pearson’s correlation. Chew et al. found a significant correlation between both coding in seven content concept.

Culotta [6] and Lampos et al. [7] perform a similar analysis of Tweets. Both studies track flu related keywords to estimate influenza rates with a high accuracy by using an automated method to filter spurious twitter messages. The obtained results correlate with the national influenza rate.

In this paper, we want to present a similar procedure as Chew et al. [5]. The main differences to our work are the following: The focus of our study is Germany. Since according to studies the twitter behaviour in Germany differs from the one in English-speaking countries in general and the U.S. in particular [1], it needs to be investigated whether this source might be of relevance for detecting disease outbreaks in Germany. Therefore, only German keywords are exploited to identify Twitter messages. Further, we are not only interested in influenza-like illnesses as the studies available so far, but also in other infectious diseases (e.g. Norovirus and Salmonella). Also differing from existing work is the study of linguistic peculiarities of the disease-or symptom-reporting tweets as an important prerequisite for developing methods for an automatic analysis of Twitter messages in future work.

## 3. Tweet Characteristics

Twitter messages have a common format: [username] [text] [date time client]. The length is restricted to 140 characters. In terms of linguistics, each twitter user can write as he or she likes. Thus, the variety reaches from complete sentences to

listing of keywords. Hashtags, i.e. terms that are combined with a hash (e.g. #flu) denote topics and are primarily utilized by experienced users.

Referring to the study from Chew et al. 2010 [5] we categorise tweets according to their contents (Table 1). In more details, Twitter messages can

- Provide information,
- Express opinions or
- Report personal issues.

If information is provided, the authority of that information cannot normally not be determined, so it might be unverified information. Opinions are often expressed with humor or sarcasm and may be highly contradictive in the emotions that are expressed. Consider for example the tweet: “I feel so sick. I have Bieber fever. ☹”. On the one hand it reports about the sickness which is rather negative. On the other hand, there is the smiley which denotes that there is no serious illness, but only “Bieber fever”, which is not really a disease or even fever. It is related to the young pop star Justin Bieber.

Content	Description	Example Tweets
<b>Resources</b>	Resource tweets contain news, updates, or information about diseases or outbreaks. The title of the linked article might be mentioned.	#schweinegrippe Neue Schweinegrippefälle in Europa: Kulmbacher Gesundheitsamt warnt vor P... <a href="http://tinyurl.com/36o4nqh">#http://tinyurl.com/36o4nqh</a> #influenza #h1n1
<b>Personal opinion with linking</b>	Twitter users post their opinion on a disease, virus, symptom.	Schweinegrippe :D <a href="http://yfrog.com/h3zuhtj">#http://yfrog.com/h3zuhtj</a>
<b>Personal opinion and information</b>	Twitter users post only their personal feelings or health status.	Hallo Freitag - Hallo Erkältung
<b>Marketing</b>	Tweet contains an advertisement for an H1N1-related product or service.	It’s National Influenza Vaccination Week. Get vaccinated to fight flu!
<b>Spam</b>	Tweet is unrelated to diseases or symptoms, but contains mentions of diseases or symptoms	"Das soll gegen Erkältung helfen!" <a href="http://twitpic.com/3fl9q0">http://twitpic.com/3fl9q0</a> (via @haraldmeyer)

**Table 1: Categorization of Tweets**

Tweets that contain mentions of symptoms or diseases can be further distinguished based on their content in informing about the health status of the (1) author of the tweet, (2) a friend of the author or (3) a prominent person. Rarely, they are reporting about health status of animals. Further, tweets are reporting about general health information or health education, official information or advices from travel medicine. Characteristically, tweeters are using short sentences (e.g. *I have fever*), or just keywords (e.g. *fever, cough, headache*). Abbreviations are wide spread and sometimes difficult to understand due to a lack of context.

Twitter messages can have a certain value for syndromic surveillance. It can be a source of public sentiment, but also health officials are using it to inform the public of current outbreaks or vaccination campaigns. The objective of this work is to analyze this value in more depth by monitoring the content of Tweets from various perspectives. The concrete methods applied are described in the next section.

## 4. METHODS

In this section, the method underlying our study is presented. First, we describe the data collection process. Second, the analysis criteria are introduced.

### 4.1 Data Collection

The data underlying our studies has been collected using the social web search engine Topsy (<http://www.topsy.com>) and Twitter API (<http://apiwiki.twitter.com>). For selecting tweets, a list of symptoms and disease names has been created manually by domain experts (epidemiologist of the state health department of Lower Saxony). This list contains symptoms of infectious diseases such as *fever*, *headache* and disease names such as *swine flu* or *H1N1* (their German correspondents). For the relevance assessment of Twitter messages (Study 1), only this list has been used for filtering and selecting the Twitter messages for the evaluation.

For the retrospective analysis (Study 2), additional search terms and their combinations with disease names and symptoms have been exploited. Additional search terms include therefore keywords (e.g., Kindergarten, Elderly house, canteen), person groups (e.g., children, students) or locations (e.g. Hannover, Berlin). Since study 2 is a retrospective analysis, more information about the disease outbreaks is already known. From this additional knowledge the additional keywords have been selected. Our queries were limited to search terms which were found in the weekly national epidemiology records (Epidemiologische Lagekonferenz) reporting about concrete outbreaks and SurvNet@RKI, the national health surveillance system used in Germany. The twitter search then exploits a combination of one keyword or symptom/disease and one place name. By including the location into the search, it is ensured that outbreaks or related information taking place at the right location are considered.

The search was limited to the particular month, for which the outbreak was associated with. In particular, the tweet searches included data from September 2010 till February 2011.

### 4.2 Data Analysis

Two analyses are been performed. As part of **study 1 (Relevance Assessment of Tweets)**, tweets matching a disease name or symptom are manually classified as relevant or irrelevant for disease surveillance. Any tweet should be labeled as positive or case regardless whether it is a confirmed, putative or probable case, if:

1. it confirms that the user is infected with a disease or symptom, e.g. *I am sick now... I got inuenza and I need medicine*,
2. it confirms that another subject (e.g., animal, ) has a disease or symptom,
3. a test result is mentioned which confirms an infection, e.g. *Tyler is inuenza positive!!!!*, or if

4. a suspicion is mentioned, e.g., *my son is suspected to has swine u* , or
5. another outbreak or danger is described.

To consider the reduced length of Twitter messages and the language peculiarities, tweets are also labelled *positive* when only a disease name or symptom is mentioned.

Any tweet which confirms that there is no case or which contains text that is unrelated to a case is labelled negative.

A negative tweet is any tweet that:

1. is a question, e.g. What is this Bieber Fever Thing?
2. contains a condition, e.g. *If I have the flu again I will kill someone*.
3. offers advices like *#Kids health:you should prevent your child from getting #dengue fever*
4. negates an infection, e.g. *I don't have measles*,
5. contains a disease definition, statistics, or describes past outbreaks, jokes about diseases or outbreaks.
6. is outside of the disease outbreak domain.

The number of relevant and irrelevant tweets will be quantified as part of the analysis.

In the second study (Retrospective Analysis of Tweets), tweets returned by the search engine were classified into relevant and irrelevant tweets with respect to the disease outbreak for which the search terms have been selected. Furthermore the relevant tweets were divided into their content: resources, personal experience and personal note and misinformation (marketing and spam). It is assessed how many relevant tweets could be identified using the adapted search term list and which content is provided.

## 5. RESULTS

In this section, we report the results of our studies.

### 5.1 Study 1: Relevance Assessment

Table 2 lists the numbers of relevant and irrelevant marked tweets per search term. It can be seen that depending on the keyword, the number of relevant and irrelevant tweets differs.

Keyword	#Relevant	#Irrelevant
Malaria	62	199
Dengue	24	119
Yellow Fever	10	55
Influenza	33	73
Measles	36	91
Poisoning	130	93
Cholera	17	77
Typhoid	40	52
Hepatitis	50	116
Smallpox	11	77
Headache	174	39
Fever	28	68

**Table 2: number of relevant and irrelevant labelled tweets**

For the term *headache* around 80% of the labelled tweets are relevant for disease reporting and outbreak detection. In contrast, only 30% of the labelled tweets matching the keyword *fever* are relevant. We conclude that there are keywords referring to symptoms that are specific enough to provide a significant number of relevant tweets. Other symptoms are used in non-outbreak related combinations (e.g. Bieber-Fever) so that

lots or irrelevant tweets are received when using these keywords.

For twitter messages matching disease names we can recognize that most of the tweets are irrelevant. A closer look to these tweets show that tweets with mentions of disease names are often extracts from news, contain information about vaccinations or medications. This means that the general public is not using these terms in their messages. This is understandable since normally, people are unaware of the disease they have before they go to see a doctor. But they observe certain symptoms which are reported in tweets as this study showed. In conclusion, it is crucial to search for tweets matching symptom terms instead of disease terms to identify tweets relevant for disease outbreak detection.

## 5.2 Study 2: Retrospective Analysis of Tweets

During six months (September 2010 to February 2011) we collected 920 tweets containing at least one place name and one keyword or one symptoms/disease name for 46 real world outbreaks. The manual classification as relevant or irrelevant resulted in 102 relevant tweets for nine outbreaks:

- Salmonellosis in Taunusstein (Hessen)
- Influenza in Göttingen (Lower Saxony)
- Norovirus at a high school in Freital (Saxony)
- Measles in Lübeck (Schleswig-Holstein)
- Q-Fever (North Rhine Westphalia)
- Hepatitis C (Mallorca Inca clinic)
- Norovirus in Austria Hospital (Austria)
- Influenza in all federal states
- Norovirus in all federal states

Analyzing the different outbreak we found that Influenza had the largest number of relevant tweets.

Table 3 shows the frequency of matched search terms of the three categories (additional keywords, place names, diseases / symptoms) in all tweets and in relevant tweets. According to this, keywords seem to be less relevant for identifying outbreak-related tweets since only 6,12% of the relevant tweets matched one of these keywords. "Schüler" (pupils, n=17) was the most frequently used keyword. The highest percentage of matches with tweets (66,31%) was found with symptoms and disease names. Thus, we conclude that disease and symptom-describing terms are most relevant for identifying outbreak-related twitter messages.

Search terms	Total number of tweets	relevant number Tweets	% of relevant Tweets
keywords	637	39	6,12%
place name	681	85	12,48%
symptoms/ disease	187	124	66,31%

**Table 3: number of mentioned search terms in tweets**

Table 4 shows the frequency of tweets matching certain disease names or symptoms. It can be seen that the most popular disease name was "Schweinegrippe" (swine flu) followed by "H1N1"(n=23) and "Norovirus" (n=16) (Table 4).

word	frequency
Schweinegrippe	42
H1N1	23
Noroviren	16
Influenza	10
Erkältung	7
Durchfall	5
Grippewelle	4
Hepatitis	4
Masern	4
Übelkeit	2
Dialyse	2
Grippe	2
Q-Fieber	2
Erbrechen	1

**Table 4: most popular symptoms and diseases found in relevant tweets**

Only a few tweets contained symptoms or disease names and were considered relevant. Only tweets about strategic marketing or those in foreign languages where the terms have the same spelling (e.g. "Grippe" (German, French) or "Influenza" (German, English) or universal language such as "H1N1" were irrelevant.

Approximately 51% of the tweets contained title and headlines which were linked to a news website. Tweets expressing personal opinion without a links had the with almost 8% the lowest percentage of relevant tweets.

Resources	Personal opinion with linking	Personal opinion
50,98% (n=52)	41,18% (n=42)	7,84% (n=8)

**Table 5: Descriptions and Examples of Content Categories**

## 6. DISCUSSION

Widespread use of social media also involves several important challenges for disease surveillance. Although twitter is growing rapidly, it remains less widespread and accessible than traditional media. It is possible that not all relevant tweets were represented in our study since it is likely that not all tweets match our keywords or had a different spellings. Not considered by our data collection process were common misspellings, slang, abbreviations or plural of terms.

Further, the conjunction of one keyword and the name of the town/city has benefits, but also shortcomings. The benefit is that more relevant tweets can be identified when the query does not make the inclusion of a certain location mandatory. On the other hand, it is then difficult to correctly locate the tweet, i.e. whether it is really referring to the outbreak under consideration. Our studies showed that by dropping this criteria of connecting disease / symptom or keyword with a location helps to increase the number of relevant tweets.

We found that news and information are the most common tweets (50,98%) and not the personal description of symptoms. News information normally repeats official information and is therefore not helpful in early detection of disease outbreaks. So, the number of personal tweets referring to the disease outbreaks that have been monitored is limited. Anyway, the relevant personal tweets identified in our search provided indications to the outbreaks and could therefore be very relevant for early detection of disease activity.

For the first study, only a snapshot of data from one day has been considered. In future work, such analysis should be repeated for some longer time period. Further, only two symptoms have been considered. To draw more reliable conclusions on the usefulness of symptoms in such process, more symptoms need to be considered in the future studies.

Further work is planned to analyze more tweets retrospectively and prospectively. In the future, we plan to compare our results to the traditional surveillance system to verify whether social media is faster than traditional disease reporting.

## 7. CONCLUSIONS

In this paper, we analysed the relevance of Twitter for disease surveillance. The results give us a first impression of the potential for twitter for the purpose of disease surveillance. Tweets can be used for real-time content analysis and sentiment analysis since relevant information is contained to a certain extent. It could support a quicker respond to disease activity and helps health officials to be better informed about public health threats and outbreaks or disease activity as reflected by the population. The studies also showed that sophisticated filtering needs to be applied to exclude irrelevant messages automatically. Further, it is insufficient to only use terms referring to symptoms and diseases as selection criteria. Instead, additional terms need to be added to the queries. Future research will assess which additional terms are of relevance.

## ACKNOWLEDGMENTS

The research is partly funded by the European Commission under 247829.

## REFERENCES

- [1] Sysomos Inc. *Exploring the use of twitter around the world online report*: <http://blog.sysomos.com/2010/01/14/exploring-the-use-of-twitter-around-the-world/>; 2010 January
- [2] M. Keller et al. *Use of unstructured event-based reports for global infectious disease surveillance*. *Emerg Infect Dis*. 2009 May;15(5):689-95
- [3] Velasco, E., Otrusina, L., Linge. J., Dreesman, J., Eckmanns, T., Kriek, M., *Social media and epidemiology: Tweets indicate Norovirus outbreak at a university*, International Meeting on Emerging Diseases and Surveillance (IMED 2011).
- [4] <http://topsy.com/>
- [5] Chew, C., Eysenbach G., *Pandemic in the Age of Twitter: Content Analysis of tweets during the 2009 H1N1 Outbreak* Plos one 2010 Nov
- [6] Culotta A. *Detection influenza outbreaks by analyzing Twitter messages* (Submitted on 24 Jul 2010 Department of Computer Science, Southeastern Louisiana University Hammond, LA 7402)
- [7] Lampos V., Cristianini N., *Tracking the flu pandemic by monitoring the social web*, In 2<sup>nd</sup> IAPR Workshop on Cognitive Information Processing (CIP 2010), pages 411-416, 2010.
- [8] Eysenbach G. *Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet*. *J Med Internet Res* 2009;11(1):e11
- [9] Eysenbach G: *Infodemiology: Tracking Flu-related Searches on the Web for Syndromic Surveillance*. *AMIA Annu Symp Proc*. 2006 2006: 244-48
- [10] Carneiro HA, Mylonakis E: *Google Trends. A Web-based Tool for Real-time Surveillance of Disease Outbreaks*. *Clinical Infectious Diseases*, 2009, 49(10), 15557-64