Web Data Mining and Social Media Analysis for better Communication in Food Safety Crises

Christian H. Meyer*, Martin Hamer*, Wiltrud Terlau*, Johannes Raithel[#], Patrick Pongratz[#]

** International Center for Sustainable Development, University of Applied Science Bonn-Rhein-Sieg, Grantham-Allee 20 53757 Sankt Augustin, Germany,* izne.info@h-brs.de

[#]European IT Consultancy, EITCO GmbH, Am Bonner Bogen 6, 53227 Bonn, kontakt@eitco.com

Abstract

Although much effort is made to prevent risks arising from food, food-borne diseases are an ever present-threat to the consumers' health. The consumption of fresh food that is contaminated with pathogens like fungi, viruses or bacteria can cause food poisoning that leads to severe health damages or even death. The outbreak of Shiga Toxin-producing enterohemorrhagic E. coli (EHEC) in Germany and neighbouring countries in 2011 has shown this dramatically. Nearly 4.000 people were reported of being affected and more than 50 people died during the so called EHEC-crisis. As a result the consumers' trust in the safety of fruits and vegetables decreased sharply.

In situations like that food crisis managers from public authorities as well as from privately owned companies must react quickly: They have to identify and track back contaminated products and they have to withdraw them from the market. At the same time they have to inform the stakeholders' about potential threats and recent developments. This is a particularly challenging task because when an outbreak is just detected information about the actual scope is sparse and the demand for information is high. Thus, ineffective communication among crisis managers and towards the public can result in inefficient crisis management, health damages and a major loss of trust in the food system. This is why crisis communication is a crucial part of successful crisis management, whereas the quality of crisis communication largely depends on the availability of and the access to relevant information.

In order to improve the availability of information, we have explored how information from public accessible internet sources like Twitter or Wikipedia can be harnessed for food crisis communication. In this paper we are going to report on some initial insight from a web mining and social media analysis approach to monitor health and food related issues that can develop into a potential crises. We have chosen Twitter and Wikipedia as sources for our study since it's publicly accessible and reveal what people state about certain topics and what people are looking for in order to answer their questions.

Keywords: Crisis Communication; Web Data Mining; Social Media Analysis; Issues Monitoring

Introduction

The consumption of fresh food that is contaminated with pathogens like fungi, viruses or bacteria can cause food poisoning that leads to severe health damages or even death. As a consequence actors in the agri-food system are challenged to design quality and safety assurance systems that are efficient, reliable and internationally compatible. This requires organisational structures, appropriate action plans and technical solutions that guarantee an efficient collection and distribution of information in order to improve communication procedures and to shorten the time of decision making. But barriers for information sharing still exist, not only between different regions or different countries but also between public and private organisations, especially during food crises. Although public and private data sets could provide useful knowledge and insight, there are no sufficient concepts for a routine information exchange within a short time (Hamer et al., 2014). Consequently, there's a need for alternative information gathering.

For this reason we aim to develop an online food safety issue monitoring system that integrates public accessible internet sources like Wikipedia and Twitter in order to detect potential threats as soon as possible

and to make predictions about the information needs and sentiments of the public to improve crisis management and crisis communication.

Issues monitoring in this respect is the task to keep an open eye on trends and topics that can have an impact on an organisation or on the trust in safe food in general. Issues can be seen and interpreted differently by different stakeholders. Usually they are of public interest with a high potential for conflicts (Ingenhoff and Röttger, 2008). This makes it important to have immediately available knowledge of what's actually going on and how a crisis is perceived by the public in order to tailor appropriate measures and communication plans to manage a crisis (Hamer et al., 2014).

The EHEC outbreak 2011

The ever present threat of food borne diseases is an issue that needs permanent attention. The outbreak of the *Shiga Toxin-producing Escherichia coli O104:H4 (EHEC)* in Germany and neighbouring countries in 2011 has shown this vividly. According to the central federal institution responsible for disease control and prevention in Germany the Robert Koch Institute (RKI) it was the largest outbreak that was taken to the records in Germany and with respect to *haemolytic-uremic syndrome (HUS)* it was the largest known outbreak worldwide (RKI, 2011). Based on the onset of the diseases the outbreak lasted from early May 2011 until 26 July 2011, when RKI officially declared its end.

However, although there's an official surveillance system in Germany in accordance with the Protection against Infection Act (IfSG) the RKI was not notified before 19 May, when a cluster of three paediatric HUS cases in Hamburg was reported. Before this day it wasn't apparent that there already were higher numbers of infections (Figure 1). In the evaluation report of the outbreak RKI states that in practice the period from the disease's onset until the notification of the RKI expanded from a few days to several weeks (RKI, 2011). After more infections became known, the public was informed by local authorities.

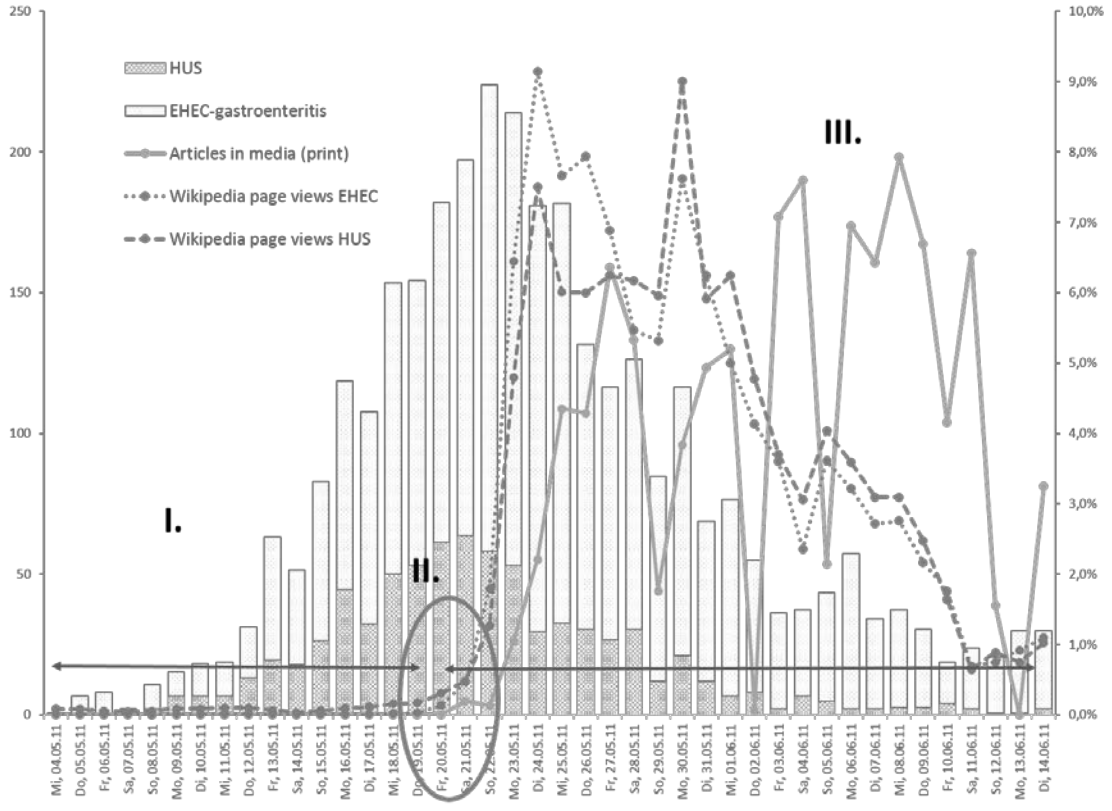

Figure 1: Epidemiological curve of HUS and EHEC adapted from RKI (2011) (right axis, daily count), Wikipedia page views related to articles about "Enterohämorrhagische Escherichia coli (EHEC)", "Hämolytisch-urämisches Syndrom (HUS)" and

articles that were printed in a selected sample of newspapers related to the EHEC-outbreak adapted from BfR (2011) and related Wikipedia page views (left axis, relative frequency)

First online articles about the EHEC outbreak that we could track back appeared on 20 May 2011 (Phase II., see Fig. 1) In the course of the day some more newspapers in Northern Germany picked up the topic and published articles about it. First articles that were printed in the newspapers appeared on 21 May according to the BfR's media monitoring (BfR, 2011). Around the same day the outbreak in terms of infections reached its peak in Northern Germany. Media articles reporting on the EHEC-topic in terms of printed articles peaked with a delay around 2 June (Linge et al. 2012). In retrospect (Figure 1) it can be shown that the demand for information rose sharply (Phase III., see Fig. 1) after the outbreak was detected and officially notified, since Wikipedia page view numbers started to soar. Surprisingly enough, Wikipedia page views started to increase from 18 May 2011, which is one days prior to the official announcement.

On 3 June 2014 a task force composed of experts from different public authorities was set up in order to deal with the outbreak and to assess its impact (BVL, 2011). In the task force's final report it is stressed that due to cooperation between governmental and local authorities and between public health and food safety authorities the source of the outbreak could be identified. In addition to this it is emphasized that information from multiple sources where collected and analysed which helped to discover sprouted seeds as the source of the infections.

It became apparent that timely, sound and robust information was crucial for the coordination task of authorities and the communication among all involved parties and towards the public (European Commission, 2011). However, improvements still can be made by harnessing the internet for communication and analysis of web data.

Web data mining and social media analysis

The internet is a constantly growing source of data that is public accessible and can be mined to discover useful information or knowledge by analysing the Web's hyperlink structure, page content and usage data. This process is referred to as web data mining (Liu, 2011). It is based on traditional data mining techniques, but it takes the particularities of the internet into account, like the various types of data that range from structured tables and semi-structured websites to unstructured texts and multimedia files. Other technical terms can found depending on the mining tasks that have to be fulfilled. Extracting knowledge from texts is referred to as text mining (Weiss et al., 2005) or if text messages that are retrieved from Twitter are analysed the term Tweet content mining can be found (Yoon, Elhadad, Bakken, 2013) as well as Twitter mining (Velardi et al., 2014).

The analysis of social media data is a growing field for research. Especially data from social media like blogs, online forums or social networks like Twitter attract a lot of attention from researchers. Much of the analysis is done in order to predict stock prices and volatility, product sales, the outcome of elections or the outbreak of diseases (Kalampokis, 2013, Schoen et.al, 2013). In this respect Twitter for example was used to predict flu trends (Achrekar et al. 2011) or the status of hay fever among people in Japan (Takahashi, 2011). Another aspect of analysis is to classify Tweets according to emotional content that they transport especially in crisis situations (Brynielsson et al., 2013, Gaspar, 2013, Torkildson, 2014)

Recent research suggests that risk and crisis communication could benefit from the integration of social media for various reasons: On the one hand the use of social media for communication opens additional ways of getting in dialogue with an organisation's stakeholders and the public (Artman et al, 2011) or with journalists as representatives of traditional media (Veil et al., 2011). On the other hand it has been observed that people in times of crises make statements about their coping strategies using social media like Twitter. This information can be harvested for further analysis, in order to detect upcoming issues (Gaspar et al., 2013) and ongoing developments (Palen et al., 2007, Nilsson et al., 2012). Another aspect of people using social media to express their opinions about certain topics is that this makes social media a valuable source for improving the early warning routines as a complement to traditional reporting mechanisms (e. g. Diaz-Aviles et al., 2012).

But not only Twitter is a valuable source of information. Wikipedia can also be harnessed for monitoring tasks (Generous, et al. 2014, McIver and Brownstein, 2014) since people use Wikipedia as a prominent source for health related information (Laurent and Vickers, 2009).

Methodology

As the EHEC outbreak has shown food borne diseases can affect the people's health badly. This is why we put the focus of our research on EHEC related web mining questions and social media analysis, in order to improve issue monitoring as an essential preliminary of successful crisis communication. Nevertheless, mining useful information and knowledge from internet data requires deep insight into to patterns and exceptions. Much of the related research has taken English language into account. We take a closer look at German by gradually digging deeper into the informational haystack. This is why our research is rather qualitative than quantitative in order to raise further research question.

For data retrieval from Twitter we used the uberMetrics Delta application. This is a fee-based media monitoring service offered by the uberMetrics Technologies GmbH (Germany) that allows to save web content for further processing. To learn about the development of Wikipedia page views we were using R which is a free software environment for statistical computing and graphics (see references). Furthermore, we used the Twitter search engine for single Tweet detection.

Core challenge: Dealing with the noise

Much of the data that can be found on the internet is noisy which means the data is highly redundant and irrelevant for particular mining tasks. This is why it is often necessary to remove the noise from the data by pre-processing it prior to analysis (e.g. Liu, 2011, Yoon, Elhadad, Bakken, 2013, Saif et al., 2014).

Taking into account that millions of Tweets are produced every day it is an impossible task to go through them by hand. This is why automated solutions must be applied. One approach to reduce the degree of noise is to apply rule and keyword based filters that bring up only material containing selected keywords. But even filtered material still can contain a lot of noise as the following two Tweets taken from Twitter illustrate:

*Tweet 1: „Aktualisierung: Im kath. Kindergarten St. Rupert in Gerolfing sind drei Fälle von **EHEC** aufgetreten. Hotline ab 23.12.14, 7:30 Uhr" (Tweeted on 22.12.2015)*
(„Update: In the catholic Kindergarten St. Rupert in Gerolfing three cases of EHEC occurred. Hotline from: 23.12.14, 7:30 o´clock)

Tweet 2: An dieser Stelle sollten wir kurz innehalten und an die Gurken denken, die zu Unrecht unter EHEC-Verdacht gestanden haben. (Tweeted on 31.12.2014)
("At this point we should pause shortly and think of the cucumbers that were wrongfully under EHEC-suspicion")

Both the Tweets contain the keyword EHEC, but whereas Tweet 1 refers to an actual case in Bavaria the author of the second Tweets is joking about cucumbers (Gurken) that came under suspicion to be the source of the EHEC-outbreak in 2011.

Wikipedia page views as an early warning signal

Inspired by the work of McIver and Brownstein (2014) and Generous et al. (2014) we took a second look at Wikipedia page views statistics of articles about *Enterohämorrhagische Escherichia coli (EHEC)* and *Hämolytisch urämisches Syndrom (HUS)* and added Wikipedia page views for the article about *"Escherichia coli (E. Coli)"* for a four month period from 1 September until 31 December 2014 (Figure 2).

Plotting the data in a time line reveals that page views drop towards the end of a week and rise at the beginning of a week. Towards the end of the year between Christmas and the New Year's Eve a lower level of page views can be observed, too. In addition to this, at some days there are exceptionally more page views than on other days. Hence, we looked for an explanation and possibly found it at Twitter. We found for some

cases that page view peaks could be matched with Tweets on the same topic that link to online news sources respectively online articles that were published around the same time (I.-V.) (Figure 2.). These reports were mainly published by local newspapers of local branches of national newspapers.

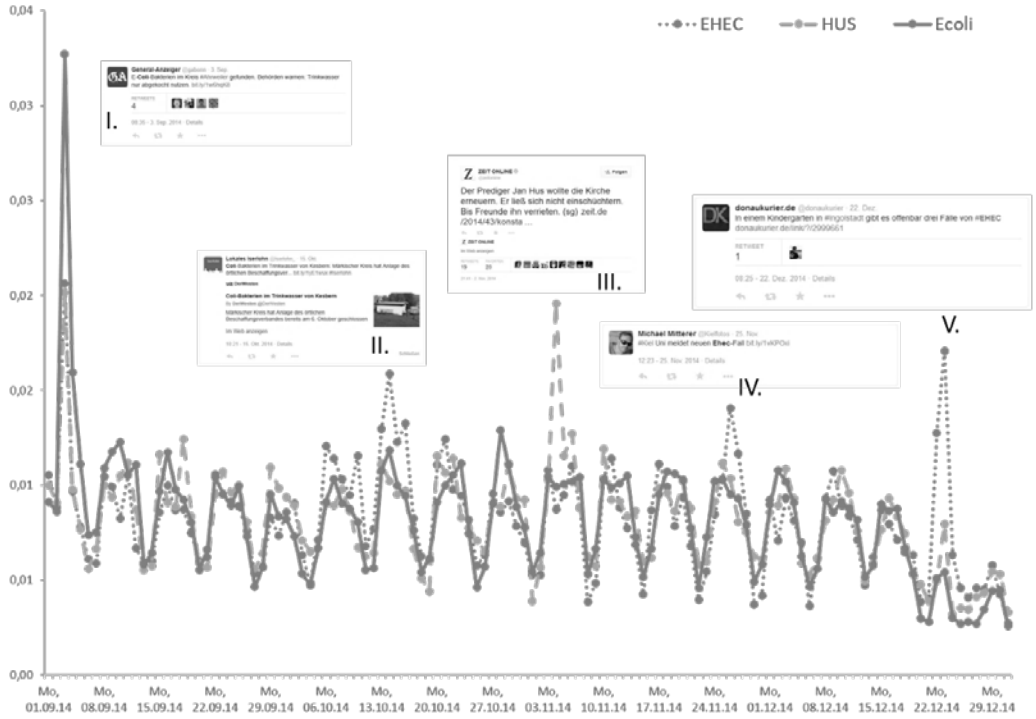

Figure 2: Wikipedia page views from September 1st to December 31st, 2014 relating to articles about "Enterohämorrhagische Escherichia coli (EHEC)", "Hämolytisch-urämisches_Syndrom (HUS)" and "Escherichia coli (E. Coli)" (relative frequency) and Tweets that report a related topic at the time of peak

Furthermore, Peak II. represents an increase in page view numbers for the term EHEC, but the related online news article is about E. Coli in drinking water found in North Rhine-Westphalia, whereas page view Peaks IV. and V. represents the term EHEC which can be related to the discovery of EHEC infections in Schleswig-Holstein and Bavaria (Table 1.)

Table 1: Wikipedia page views peaks and related Twitter content

Number	Date of peak	Event	Date and source of first related online article
I.	03.09.2014	Detection of E.Coli in public water pipes in **Rhineland-Palatinate (Germany)**	03.09.2014 (regional newspaper)
II.	14.10.2014	Detection of E. Coli in public water pipes near of **North Rhine-Westphalia**	15.09.2014 (regional news source)
III.	04.11.2014	Online Article about the Czech priest Jan Hus	02.11.2014 (national magazine)
IV.	26.11.2014	Man with EHEC was brought to hospital n **Schleswig-Holstein**	25.11.2014 (link to regional newspaper)
V.	23.12.2014	Three children infected with EHEC in an public Kindergarten in **Bavaria**	22.12.2014 (announcement via Twitter)

However, the peak of page views related to the term HUS (Peak III.) was related to the Czech priest and philosopher Jan Hus (see Table 1.). In this case Tweets relating to the HUS disease could not be found.

Tweets that count and Tweets that don't

When a particular topic or food safety issue is named it is easier to count and analyse Tweets that relate to this topic (e. g. Diaz-Aviles, 2012). Unfortunately, due to restrictions and limitations of the Twitter application programming interface it wasn't possible for us to retrieve enough Tweets that were published before the 2011 EHEC outbreak for further in depth analysis. This is why we looked ahead and focused our analysis on Tweets that were published between 1 December and 31 December 2014.

The aim of our Twitter analysis is to find patterns and co-occurrences of words in order to optimise keyword and rule based search filters for noise reduction in health related Tweets. This is necessary to reduce the amount of collected web content that has to be explored "by hand" until more sophisticated and automated sorting procedures exist. But what content do Tweets have before an ongoing food safety issue is officially notified?

Table 2: Number of frequent bigrams occurring in a set of Tweets containing the keywords "Durchfall", "Bauchschmerzen" and "Übelkeit"

Durchfall (n=1.129)		Bauchschmerzen (n=1.929))		Übelkeit (n=873)	
• durchfall und *)	99	• ich hab	248	• und übelkeit **)	229
• bei durchfall	40	• bauchschmerzen und	147	• die übelkeit	58
• wie durchfall	39	• hab bauchschmerzen	145	• gegen übelkeit	41
• durchfall hat	38	• mit bauchschmerzen	116	• kopfschmerzen und	26
• mit durchfall	31	• ich habe	110	• mit übelkeit	26
• ich habe	25	• ich bauchschmerzen	84	• übelkeit ist	21
• ich hab	23	• die bauchschmerzen	76	• bei mir	20
• gegen durchfall	21	• hab ich	68	• kopfschmerzen übelkeit	19
• in der	21	• bauchschmerzen vor	67	• ich bin	17
• hat durchfall	20	• habe bauchschmerzen	66	• der übelkeit	16

*) including the bi gram "und durchfall" **) including the bigram "übelkeit und"

To answer this question, we explored a selection of health related Tweets containing keywords respectively symptoms of food-borne diseases (Matzik, 2014) like *"Durchfall (Diarrhoea)"*, *"Bauchschmerzen (abdominal pain)"* and *"Übelkeit (sickness)"* with the intention of understanding how and what people tweet when they use these terms. The keywords were chosen on the assumption that people rather tweet about their symptoms than about the medical name of the diseases (Velardi et al., 2014).

For our analysis we prepared the data by removing duplicate Tweets and Re-tweets from the data set in order to create an unbiased list of unique Tweets to analyse word frequencies and frequently used bigrams (set of two terms that occur next to each other) (Table 2). In order to keep maximum information, we didn't remove stopwords. Stopwords are terms that don't carry none or not much information like "what", "the", "it" and the like. It is reported that removing stopwords can be problematic for further sentiment analysis (Saif et al., 2014).

We found that Twitter users that tweet about abdominal pain ("*Bauchschmerzen")* often state that they are suffering from abdominal pain themselves, whereas Tweets containing the keyword diarrhoea (*"Durchfall")* more often make statements about the symptom itself (Table 2).

Furthermore, in many cases people state that they suffer from more than one symptom as the frequent use of the word “und (and)” indicates. People who are tweeting about “Übelkeit (sickness)” apparently often suffer from *“Kopfschmerzen (headache)”*, too.

Discussion

Food crises have two major dimensions: One is to manage a crisis in terms of protecting the public from food borne threats by identifying sources of the crisis and the second is to inform stakeholders and the public about what is going on. Particularly when a food safety-issue that potentially develops into a crisis is detected, time and information are sparse resources. Since insufficient procedures for exchanging data from private organisations and public authorities still exist (Hamer, 2014), we looked at how data from public available internet sources can be harnessed to improve the availability of reliable information.

As the EHEC outbreak has shown, the numbers of infections with EHEC and HUS were already high before the outbreak was detected. Consequently, early warning and food safety issue monitoring procedures must be implemented. Diaz Aviles et al. (2012) propose Twitter for an open early warning system and showed that a total of five Tweets containing the term “EHEC” triggered the topic about the EHEC outbreak in 2011. Our findings support the proposition that a combination of online sources like Wikipedia and Twitter supports the detection of potentially threatening food safety issues also. We have found cases in which Wikipedia page views rise suddenly when articles in the media report about health related issues. This supports the study of Laurent and Vickers (2009) that Wikipedia is a prominent source for seeking health information. But our findings also suggest that the peoples’ searches are not consistent which the associated media topic. This is why a set of Wikipedia articles should be taken into account for monitoring. However, once a threat is named and announced monitoring is much easier since people can refer to this specified topic. The public’s interest in that topic can be measured, as corresponding Wikipedia page view counts show.

Furthermore, online resources from local newspapers may be also a good and quick source for early warning, since they seem to report on health safety issues first. An advantage of monitoring local newspapers is that they are usually in close contact with local authorities. Furthermore, local journalists usually have a good network of informants, they get to know what’s going on very early. On the other hand journalists are also seeking information from many sources to write their articles. Partnering with journalist via social media can contribute to a better crisis communication in terms of communication towards the public. This turns crisis communication into a two-way dialogical communication and is in line with the recommendations and findings from Veil et al. (2011) and Artman (2011).

In addition to this, monitoring food safety issues as an early warning procedure can complement official reporting channels particularly in Germany, where decentralised political organisation on local, federal state and national level are dealing with food safety issues. The same procedures can be applied in cross border information distribution between different private and public organisations as well as authorities from different countries. As a consequence, the preparedness for cross-border food safety crises as the European Commission (2011) demands could be improved. If the issue monitoring system will be integrated into a communication platform, decentralised crisis management teams with access to that platform will have an additional channel for information. This could be of importance in scenarios where those teams don’t have formal structures for communication.

Unfortunately, we were not able to retrieve Tweets related to the EHEC outbreak 2011. This is why we took a closer look on symptoms in order to learn how and what people tweet about. A closer look on the semantics of the Tweets revealed that classifying Tweets proved to be a more complex task. Since Tweets carry only a limited amount of text (max. 140 characters) they cannot be classified easily as being “relevant” or “not relevant” for the mining task - not even “by hand”. This is left to future research.

Conclusion and future work

The internet is a rich source for data that can be used for crisis management and crisis communication in particular. We looked at only a small fraction of internet data that is produced every day, but we started to learn how people in Germany search for health related information and how they tweet about particular symptoms that can be observed with food borne diseases. Nevertheless, even though finding useful knowledge and patterns by mining the web and analysing social media are complex tasks, we scratched the surface deep enough to find some starting points for future research.

Our future work will deal with more in depth analysis and automated classification of Tweets as being "relevant" or "not relevant" in order to monitor health and food related issues and to make better predictions about how a crisis develops. Furthermore, we want to approach the field of sentiment classification. In order to understand what people are concerned about during a crisis in order to tailor crisis communication plans that fit to the particular crisis. However, as our first results indicate, in the near future it still is necessary that domain and communication experts contribute their knowledge to crisis communication since much of the analytical work cannot be fully automated yet.

Nevertheless, crisis communication can benefit from web mining and social media analysis, since the monitoring of food safety issues can be improved. On the one hand we can improve keyword and rule based search filters for information retrieval. On the other hand the data that is mined from the internet can be processed and visualised online, so that crisis managing teams can have a look at the information independent from where they are. This approach helps to overcome organisational and regional borders.

Acknowledgements

This work is co-financed by the INTERREG IV A Germany-Netherlands programme through the EU funding from the European Regional Development Fund (ERDF), the Ministry for Economic Affairs, Energy, Building, Housing and Transport of the State of North-Rhine Westphalia and the Dutch Ministry of Economic Affairs and the Province of Gelderland. It is accompanied by the programme management Euregio Rhein-Waal.

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., Liu, B. (2011). Predicting Flu Trends using Twitter Data, IEEE Conference on Computer Communications Workshop, IEEE, pp 702-707.

Artman, H., Brynielsson, J., Johansson, Trnka, J. (2011). Dialogical Emergency Management and Strategic Awareness in Emergency Communication. Proceedings of the 8th International ISCRAM Conference, Lisbon, Portugal, May 2011.

BfR (2011). EHEC-Ausbruch 2011, Aufklärung des Ausbruchs entlang der Lebensmittelkette, Federal Institute for Risk Assessment (BfR), Germany 2011, Berlin Online: http://www.bfr.bund.de/cm/350/ehec-ausbruch-2011-aufklaerung-des-ausbruchs-entlang-der-lebensmittelkette.pdf [accessed 07.01.2014].

Brynielsson, J., Johansson, F., Westling, A. (2013). Learning to Classify Emotional Content in Crisis-Related Tweets. In: Proceedings of the 11th IEEE International Conference on Intelligence and Security Informatics (ISI 2013), (Seattle, Washington, 2013), pp 33–38. doi:10.1109/ISI.2013.6578782.

BVL (2011). Report on the results of the German EHEC Task Force on the EHEC O104:H4 disease outbreak investigation in Germany, Federal Office of Consumer Protection and Food Safety (BVL), Germany 2011, Berlin 2011 Online: http://www.bvl.bund.de/SharedDocs/Downloads/01_Lebensmittel/4Task_Force/Task_Force_EHEC_Ergebnisbericht_23_09_2011.pdf?__blob=publicationFile&v=2 [accessed 16.12.2014].

Diaz-Aviles, D., Avaré, S., Velasco, E., Denecke, K., Nejdl, W. (2012). Epidemic Intelligence for the Crowd, by the Crowd, (full version), online: http://arxiv.org/pdf/1203.1378v1 [accessed 16.12.20141].

European Commission (2011). Lessons learned from the 2011 outbreak of Shiga toxin-producing Escherichia coli (STEC) O104:H4 in sprouted seeds. Commission Staff Working Document SANCO/13004/2011, Brussels. Online: http://ec.europa.eu/food/food/biosafety/salmonella/docs/cswd_lessons_learned_en.pdf [accessed 16.12.2014].

Gaspar, R., Gorjao, S., Seibt, B., Lima, L., Barnett, J., Moss, A., Wills, J. (2013). Tweeting during food crises: A psychosocial analysis of threat coping expressions in Spain, during the 2011 European EHEC outbreak, *International Journal of Human-Computer Studies*, Vol. 72, pp 239-254.

Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., Priedhorsky, R. (2014). Global Disease Monitoring and Forecasting with Wikipedia, *PLOS Computational Biology*, Vol. 10(11): e1003892.

Hamer, M, Telau, W., Breuer, O., van der Roest, J., Petersen, B. (2014), The EHEC-Crisis – Impact and Lessons Learned – Sustainable Cross-Border Crisis Control and Communication, paper accepted for the 29th International Horticultural Congress, Sustaining Lives, Livelihoods and Landscapes, 17-22 August 2014, Brisbane, Australia.

Ingenhoff, Diana, Röttger, Ulrike (2008). Issues Managememt – Ein zentrales Verfahren der Unternehmenskommunikation, In: Meckel, Miriam, Schmid, Beat F.: Unternehmenskommunikation. Kommunikationsmanagement aus Sicht der Unternehmensführung. Wiesbaden: Gabler Verlag.

Kalampokis, E., Tambouris, E., Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*. Vol. 23(5), pp 544 – 559.

Laurent, M. R., Vickers, T. J. (2009). Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association*, Vol. 16(4), pp 471:479.

Linge, J.; Mantero, J.; Fuart, F.; Belyaeva, J.; Atkinson, M.; Van Der Goot, E. (2011). Tracking media reports on the shiga toxinproducing escherichia coli O104:H4 outbreak in Germany. In: P. Kostkova, M. Szomszor, D. Fowler (eds.) Electronic Healthcare - Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 91, pp 178-185.

Liu, B. (2011). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Data Centric Systems and Applications, DOI 10.1007/978-3-642-19460-3_1. Springer-Verlag.

Matzik, S. (2014). Lebensmittelvergiftung – Symptome. Online: http://www.netdoktor.de/krankheiten/lebensmittelvergiftung/symptome/ [accessed 19.12.2014].

McIver, D., Brownstein, J. S. (2014). Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLOS Computational Biology, Vol. 10(4): e1003581.

Nilsson, S., Brynielsson, J., Granåsen, M., Hellgren, C., Lindquist, S., Lundin, M., Narganes Quijano, M., Trnka, J. (2012). Making use of new media for pan-european crisis communication. In: Proceedings of the Ninth International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012), Vancouver, Canada, April, 2012. Online: http://www.diva-portal.org/smash/get/diva2:530552/FULLTEXT01.pdf. [accessed: 18.12.2014].

Palen, L., Vieweg, S., Sutton, J., Liu, S. B., Hughes, A. (2007). Crisis Informatics: Studying Crisis in a Networked World. Third Internation Conference on e-Social Science, Ann Arbor, Michigan, October 7-9, 2007, online: http://www.cs.colorado.edu/~palen/palen_papers/palen-crisisinformatics.pdf [accessed 19.12.2014]

RKI (2011). Final presentation and evaluation of epidemiological findings in the EHEC O104:H4 outbreak, Robert Koch Institute Report, Germany 2011. Berlin 2011. Online: http://www.rki.de/EN/Home/EHEC_final_report.pdf?__blob=publicationFile [accessed 16.12.2014].

Saif, H., Fernandez, M., He, Y., Alani, H. (2014). On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In: Proc. 9th Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland (2014), pp 810-817.

Schoen, H., Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., Gloor, P. (2013). The Power of Prediction with Social Media, in: Internet Research 23(5), pp 528-543

Takahashi, Tetsuro, Abe, Shuya, Igata, Nobuyuki (2011). Can Twitter be an Alternative of Real-World Sensor?. In: Jacko, J. A. (Ed.): Human-Computer Interaction, Part III, HCII 2011, LNCS 6763: 240-249, Springer-Verlag Berlin Heidelberg, pp 240-249.

Torkildson, Megan K. , Starbird, Kate , Aragon, Cecilia R. (2014). Analysis and Visualisation of Sentiment and Emotion on Crisis Tweets, In: Y. Luo (Ed.), Cooperative Design, Visualization, and Engineering, Lecture Notes in Computer Science 8683, Springer-Verlag Berlin Heidelberg, pp 64-67.

Veil S. R., Buehner T. , Palenchar M. J. (2011). A work in-progress literature review: Incorporating social media in risk and crisis communication, *Journal of Contingencies and Crisis Management* 19 (2), pp 110-122.

Velardi, P. Stilo, G., Tozzi, A. E., Gesualdo, F. (2014). Twitter mining for fine-grained syndromic surveillance. *Artificial Intelligence in Medicine*, Vol. 61, pp 153-163.

Weiss, S. M., Indurkhya, N., Zhang, T., Damerau, F. J. (2005), Text Mining – Predictive Methods For Analyzing Unstructured Information, Springer Science and Business Media, New York.

Yoon, Sunmoo, Elhadad, Noémie, Bakken, Suzanne (2013). A Practical Approach for Content Mining of Tweets, *American Journal of Preventive Medicine* 45(1), pp 122-129.

Software

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

uberMetrics Delta, Softwarelösung für Medienbeobachtung, Pressespiegel und Krisenkommunikationsmanagement, 2011-2014, uberMetrics Technologies GmbH, Berlin, Deutschland.