Challenges in media monitoring of worldwide news sources to support public health

Joao Pita Costa
Quintelligence, Ljubljana, Slovenia
University of Rijeka, Croatia

Flavio Fuart , Marko Grobelnik , Gregor Leban
Quintelligence, Ljubljana, Slovenia
Jozef Stefan Institute, Slovenia

Evgenia Belyaeva
Jozef Stefan Institute, Slovenia

ABSTRACT

Real-time global media monitoring is nowadays an essential resource to public health. Multilingual capabilities can enrich this potential allowing a worldwide overview based on online news sources, blog posts or social media. In this paper we propose research topics related with the exploration text mining tools used to provide real-time global media monitoring in the context of health. We aim to understand how media can contribute to a better overview of health related and well being matters. With it we shall also identify open research questions that motivate further technological development to better fit the needs, interests and workflow of public health professionals.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

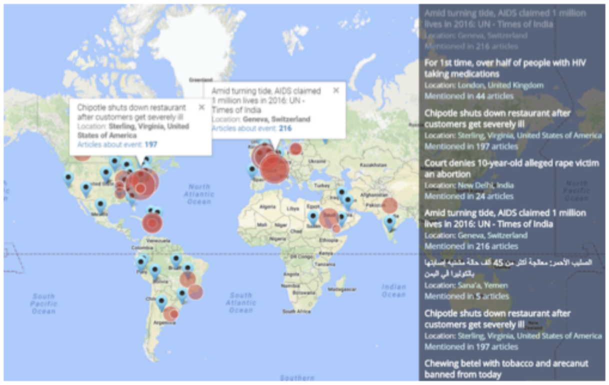
Measurement, Performance, Languages.

Keywords

Text mining, public health, news, media.

1. INTRODUCTION

Dealing with media (written online media in particular) has several issues, one of which is a lack of common publishing standards. Another issue is related to the global nature of the service: the mere fact of a system being universal, requires a system that can manage a variety of languages, possibly hundreds of them. This creates issues, as today's language technologies can deal only with words and sentences, highlighting the need to bridge the gap from simple textual representation towards semantic representation, where we would want to understand the semantic and conceptual aspect of the textual content and not just lexical (words and phrases) and syntactical (sentences). Considerable research and commercial activity in the past period has led to a development of high performance online news monitoring systems and accompanying tools, methods and models. For example, a complete cross-lingual news processing pipeline consisting of the following components has achieved good results in both research and commercial usage scenarios: NewsFeed system [16] to monitor, gather an produce clear-texts of online (HTML) news articles; Sentiment Detection module [17]; Enrycher module for text annotation [7]; Wikifier for advanced text categorization [2]; Cross-lingual document linking [11]; Document clustering [1]; and Event Registry [9], an advanced news visualization and analysis tool. There is a relatively large offer of similar online news monitoring ("clipping") systems available, each of them with a distinctive set of features. Authors, however, are not aware, of a monitoring system that would implement a rich set of cross-lingual features as the online news-processing pipeline mentioned above. According to the ECDC (European Centre for Disease Prevention and Control), the objective of epidemic intelligence is to produce timely, validated and actionable intelligence on events related to communicable diseases or of unknown origin that are of interest for public health and health authorities [4]. A similar definition is provided also by the WHO [18]. Part of this effort is to gather unofficial, unstructured and unverified information about those events, that are then later verified and analyzed by Public Health experts. Several ICT solutions to support these efforts have emerged. Most notable solutions used by authorities are GPHIN [3] (Global Public Health Intelligence Network), developed and operated by the Canadian Government, MedISys [15], developed and operated by the Joint Research Centre of the European Commission and Healthmap.org, a system developed by Boston Children's Hospital receiving external funding. All those systems are multi-lingual to some extent, i.e. they monitor news in more than one language. However, it seems they do not leverage the usefulness of cross-lingual approaches to increase the quality of detected health events. Also, they seem no to use Wikipedia, which is nowadays the biggest freely available knowledge base, to extract meaningful information from news articles. In this paper we aim to identify research questions, related to features highlighted above (cross-linguality, wikification, among others), that can contribute to the appropriate software development serving the needs of Public Health professionals. Part of the work presented in it was developed in the context of the European Union research project MIDAS, under the program Horizon 2020.



Chipotle shuts down restaurant after customers get severely ill
Location: **Sterling, Virginia, United States of America**
Articles about event: **197**

Amid turning tide, AIDS claimed 1 million lives in 2016: UN - Times of India
Location: **Geneva, Switzerland**
Articles about event: **216**

Amid turning tide, AIDS claimed 1 million lives in 2016: UN - Times of India
Location: Geneva, Switzerland
Mentioned in 216 articles

For 1st time, over half of people with HIV taking medications
Location: London, United Kingdom
Mentioned in 44 articles

Chipotle shuts down restaurant after customers get severely ill
Location: Sterling, Virginia, United States of America
Mentioned in 197 articles

Court denies 10-year-old alleged rape victim an abortion
Location: New Delhi, India
Mentioned in 24 articles

Amid turning tide, AIDS claimed 1 million lives in 2016: UN - Times of India
Location: Geneva, Switzerland
Mentioned in 216 articles

الصليب الأحمر: معالجة أكثر من 45 ألف حالة مشتبه إصابتها بالكوليرا في اليمن
Location: Sana'a, Yemen
Mentioned in 5 articles

Chipotle shuts down restaurant after customers get severely ill
Location: Sterling, Virginia, United States of America
Mentioned in 197 articles

Chewing betel with tobacco and arecanut banned from today

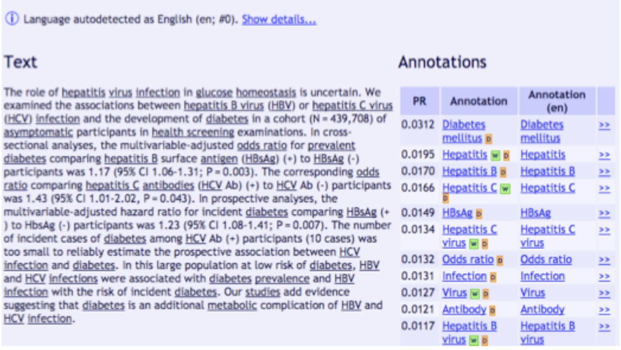
Fig. 1. The contribution of text mining efforts based on online multi-lingual news to public health raises several research questions. In the image above shows an ER visualisation module representing a real-time stream of news that permits us to explore the range of some of those questions.

2. COLLECTING THE DATA

Online news media sources represent a reliable and structured near-real time stream of a heterogeneous, multilingual text documents that describe real-word events. A range of services offers the aggregation of both social media and online news, following the uptake of news publishing through the latter channel. Several media news aggregators provide web crawlers for information extraction and media monitoring aiming for newsworthy stories. In order to extract meaningful information for a particular application domain, automatic event detection mechanisms exist allowing us to measure the media impact of public health awareness rising campaigns, health related news coverage and bias across different outlets. Those also include sentiment detection and cross-lingual linking of documents applied to news sources. In the public health domain, these approaches have been used to detect disease outbreaks and other public health threats (e.g. monitoring of international social and sports events, anti-vaccination campaigns, among others). Among the available news aggregator and analysis services that provide some level of access to online news streams we highlight the NewsFeed system (available at newsfeed.ijs.si). It is a real-time aggregated stream of semantically enriched news articles tracking over RSS-enabled global media sources worldwide. In particular, the online media monitoring system NewsFeed currently monitors around 900.000 RSS news feeds (~800.000 web sites) and collects between 350.000 and 600.000 articles per day, assuming an article archive available since May 2008. Using the current API it is possible to get annotated articles since June 2013. Currently, about 50% of all articles are in English. All languages present in Wikipedia are included, but are covered in respect of quality, volume and extent of analysis performed [8] to differing degrees. For research purposes, authors where granted access to the NewsFeed system, available at newsfeed.ijs.si. The data is accessed through a HTTP API and the result is provided in XML format. Additional metadata can be obtained through the API: named entities, concepts, categories, mentioned places and similar [14]. Furthermore, this technology could be used for additional data annotation and analysis tailored to public health applications (e.g. the annotation with *wikifier* as later discussed in Section 3). The news monitoring software EventRegistry (ER) (available at www.eventregistry.org) feeds on Newsfeed, tracking over 100,000 global media sources in near-real-time, operating across 100 languages, and aggregating global media content in a semantically meaningful way. It collects over 300,000 news articles on average per day and arranges them into events, events are further connected into story-lines which enable tracking of evolving topics [9]. Since ER covers most of the global media reporting on any topic, it can be used also to track topics like health and well-being on different levels of resolution – from small local issues, up to the higher level country issues and global trends..

3. AUTOMATIC ANNOTATION

The complexity of identification of entities makes the automatic multilingual text analysis a difficult task. Wikifier (available at wikifier.ijs.si) takes profit of Wikipedia, the biggest open, online and up-to-date knowledge repository on the internet, to annotate text and link it to relevant knowledge resources. Wikifier allows for annotating large quantities of free text in a very short time. The type of analysis provided permits us to identify trends (e.g. health related lifestyles) or ask questions like: *Provide me all texts (articles) that are about vaccination campaigns in Africa in cities with population of more than 2 million people*. Figure 2 shows the output of the automatic annotation of the abstract of a recent article on Cholera. The top ten concepts annotated include "Vibrio cholerae", "Fresh water" or "Bacteria". By being constructed over the knowledge-base of Wikipedia, it is essential that Wikipedia coverage on health topics is of high quality which is itself an open question. It is also another open question of whether an extension of Wikipedia can capture all of the concepts covered by *MeSH Headings*.



ⓘ Language autodetected as English (en; #0). Show details...

Text

The role of hepatitis virus infection in glucose homeostasis is uncertain. We examined the associations between hepatitis B virus (HBV) or hepatitis C virus (HCV) infection and the development of diabetes in a cohort (N = 439,708) of asymptomatic participants in health screening examinations. In cross-sectional analyses, the multivariable-adjusted odds ratio for prevalent diabetes comparing hepatitis B surface antigen (HBsAg) (+) to HBsAg (-) participants was 1.17 (95% CI 1.06-1.31; P = 0.003). The corresponding odds ratio comparing hepatitis C antibodies (HCV Ab) (+) to HCV Ab (-) participants was 1.43 (95% CI 1.01-2.02, P = 0.043). In prospective analyses, the multivariable-adjusted hazard ratio for incident diabetes comparing HBsAg (+) to HbsAg (-) participants was 1.23 (95% CI 1.08-1.41; P = 0.007). The number of incident cases of diabetes among HCV Ab (+) participants (10 cases) was too small to reliably estimate the prospective association between HCV infection and diabetes. In this large population at low risk of diabetes, HBV and HCV infections were associated with diabetes prevalence and HBV infection with the risk of incident diabetes. Our studies add evidence suggesting that diabetes is an additional metabolic complication of HBV and HCV infection.

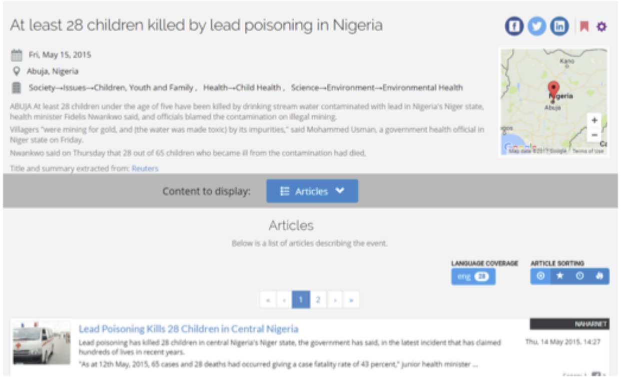
Annotations

PR	Annotation	Annotation (en)	
0.0312	Diabetes mellitus D	Diabetes mellitus	>>
0.0195	Hepatitis W D	Hepatitis	>>
0.0170	Hepatitis B D	Hepatitis B	>>
0.0166	Hepatitis C W D	Hepatitis C	>>
0.0149	HBsAg D	HBsAg	>>
0.0134	Hepatitis C virus W D	Hepatitis C virus	>>
0.0132	Odds ratio D	Odds ratio	>>
0.0131	Infection D	Infection	>>
0.0127	Virus W D	Virus	>>
0.0121	Antibody D	Antibody	>>
0.0117	Hepatitis B virus W D	Hepatitis B virus	>>

Fig. 2. The automatic annotation tool Wikifier used to enrich a WHO article on vaccination taken as an example. It is copied to the *Text* field and has several underlined words corresponding to the identified Wikipedia concepts. When hovering the emphD icon on disease name, we notice that this article is mostly about Yellow Fever, we see that yellow fever is a disease and we see that Kinshasa is a settlement. If we further explore the DBpedia entry the area, number of inhabitants, country and other information in available.

4. FROM GLOBAL TO LOCAL MEDIA MONITORING

In the context of Public Health, a graphical dashboard can provide contribution to the real-time monitoring of the global health by allowing to a continuous observation of medical and well-being issues, limiting to the presentation of articles/items related to those topics (a possible health dashboard is available at [6]). Such a dashboard presents the incoming multi-lingual health related media content published somewhere in the world. That allows the health professional to have an overview of what is happening globally at any given moment in time (cf. [5]). This is a convenient way to observe global health monitoring by using the ER system introduced above in Section 2. Its key feature is to be able to observe health issues across many languages and in temporal detail, over a variety of scales, which is what most other systems have difficulties [10]. The image in Figure 1 shows a snapshot of the dashboard with health related events on July 20th, 2017. The event of the Zika outbreak is identified immediately after the collection of news articles that report about it. With it, the health professional can explore the evolution of the news publisher's awareness of the epidemics in time by looking at the related news articles represented in a world map, as they were identified or updated during a selected period of time. ER can find articles and events related to a particular entity, topic, date, location or category, as well as measure their impact on social media (in particular, Twitter). Moreover, its cross-lingual capabilities allow considering events where the news appears only in e.g. Chinese or some unknown language where the news never came into English speaking space. This is of particular interest when considering the monitoring of rare diseases worldwide. Another perspective is the micro view of a particular health related event happening



At least 28 children killed by lead poisoning in Nigeria

Fri, May 15, 2015

Abuja, Nigeria

Society→Issues→Children, Youth and Family , Health→Child Health , Science→Environment→Environmental Health

ABUJA At least 28 children under the age of five have been killed by drinking stream water contaminated with lead in Nigeria's Niger state, health minister Fidelis Nwankwo said, and officials blamed the contamination on illegal mining.

Villagers "were mining for gold, and (the water was made toxic) by its impurities," said Mohammed Usman, a government health official in Niger state on Friday.

Nwankwo said on Thursday that 28 out of 65 children who became ill from the contamination had died.

Title and summary extracted from: Reuters

Content to display: Articles

Articles

Below is a list of articles describing the event.

LANGUAGE COVERAGE: eng 28

ARTICLE SORTING

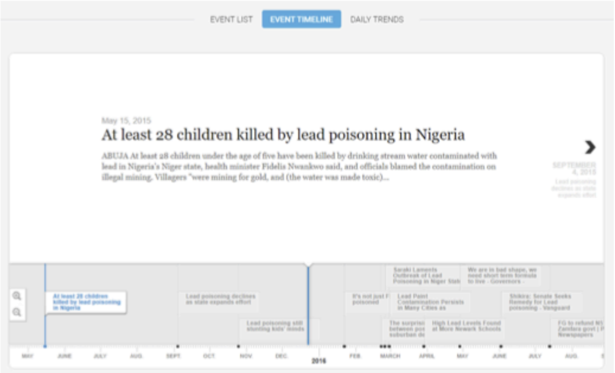
Lead Poisoning Kills 28 Children in Central Nigeria

NAHARNET — Thu, 14 May 2015, 14:27

Lead poisoning has killed 28 children in central Nigeria's Niger state, the government has said, in the latest incident that has claimed hundreds of lives in recent years.

"As at 12th May, 2015, 65 cases and 28 deaths had occurred giving a case fatality rate of 43 percent," junior health minister ...

Fig. 3. ER screenshot showing the event of “lead poisoning” in Nigeria on May 12, 2015. It includes date, location, categories, number of languages covered and social media (Twitter) count.



EVENT LIST | EVENT TIMELINE | DAILY TRENDS

May 15, 2015

At least 28 children killed by lead poisoning in Nigeria

ABUJA At least 28 children under the age of five have been killed by drinking stream water contaminated with lead in Nigeria's Niger state, health minister Fidelis Nwankwo said, and officials blamed the contamination on illegal mining. Villagers "were mining for gold, and (the water was made toxic)...

SEPTEMBER 4, 2015 — Lead poisoning declines as state expands effort

At least 28 children killed by lead poisoning in Nigeria
Lead poisoning declines as state expands effort
Lead poisoning still stunting kids' minds
It's not just F poisoned
Saraki Laments Outbreak of Lead Poisoning in Niger State
Lead Paint Contamination Persists in Many Cities as
The surprisi between po suburban d
High Lead Levels Found at More Newark Schools
We are in bad shape, we need short term formula to live - Governors -
Shikira: Senate Seeks Remedy for Lead poisoning - Vanguard
FG to refund N Zamfara govt | P Newspapers
MAY
JUNE
JULY
AUG.
SEPT.
OCT.
NOV.
DEC.
2016
FEB.
MARCH
APRIL
MAY
JUNE
JULY
AUG.

Fig. 4. ER screen-shot showing story-line developed after the event of “lead poisoning” in Nigeria on May 12, 2015. It describes below the event a timeline of related news.

somewhere in the world. The event of ``lead poisoning'' in Nigeria on May 12th, 2015, which went mostly unnoticed at a global level, is an example that can be explored through ER. With a simple query to the system, the health professional can extract all the reports related to that event, and check the event in the context of other related events. To illustrate this, the screen-shots in Figures 3 and 4 show the event itself and the story-line developed after the event, respectively.

5. Historical perspective

Another view relevant to any health related issue is a historical perspective based on an aggregation of a particular topic. The evolution of Ebola virus and reporting after 2014 until 2017 can be an example to be explored through ER. Querying ER for *Ebola virus disease* provides over 20,000 events related to Ebola appearing after 2014. The content (over 200,000 news articles) could be analysed through several visual modules: temporal intensity, geographical spread, topical spread, among others. The Figures 5 and 6 illustrate temporal (with the peak in October 2014) and geographical spread (West Africa and the US) of Ebola related events.

We highlight that ER was able to notice the outbreak of the epidemics before it was made public. Though, the question was not asked then, i.e., the query wasn't done because it was not yet an identified issue and there was a lack of continuous attention. Though a complete attention of all such epidemic related topics could be a heavy burden to the system. This rises the research problem of predicting and alerting of high density events like this. In ER the Zika outbreak event is identified immediately after the collection of the news articles that report about it. One can explore the evolution of the news publishers awareness of the epidemics in time by looking at the related news articles represented in a world map, as they were identified or updated during a selected period of time. ER can find articles and events related to a particular entity, topic, date, location or category, as well as measure their impact on social media (specifically, Twitter). This social media monitoring is complemented by TwitterObservatory (cf. [12]), leveraging in-house technology that uses data observation, enrichment and storage techniques for social media data presentation, search and analytics. Moreover, there have been several successful tests done to extract sentiment from news based on the sentiment of tweets associated with news [13]. The sentiment directly from news is still an open problem that shall be tackled.

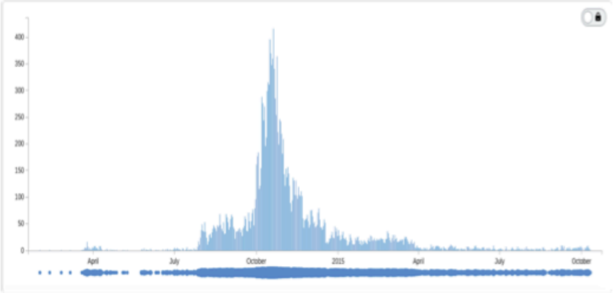
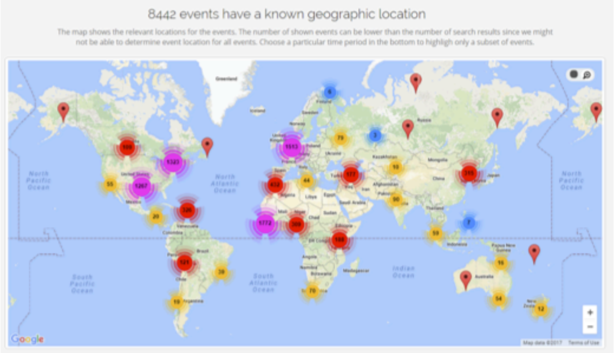


Fig. 5. A temporal intensity visualisation in ER for the query ’Ebola virus disease’ showing that the system could notice the outbreak of the epidemic before it was made public.



8442 events have a known geographic location

The map shows the relevant locations for the events. The number of shown events can be lower than the number of search results since we might not be able to determine event location for all events. Choose a particular time period in the bottom to highligh only a subset of events.

Fig. 6. ER screenshot showing a geographical spread visualisation module for the query ’Ebola virus disease’. Each mark shows the number of related news per location in the map.

6. CONCLUSIONS AND FURTHER WORK

In this paper we discussed the potential of several text mining tools dedicated to explore worlwide multilingual news, focusing matters of interest to Public Health. Further research (exploring the potential of Newsfeed, Wikifier and ER) includes: (i) the correlation of high level concepts with low level features; (ii) the showcase of hierarchies (e.g. in some disease) and how they can be drilled down to variety of sub topics (e.g. different aspects of such a disease; (iii) the analysis of the impact of health related issues on society (e.g. Ebola news impact in adherence to insurance); (iv) the presence of PubMed/Medline in global news; and (v) the prediction of consequences of a health event. Other research directions consider the dynamics of Public

Health/Healthcare institutions where activities happen affecting: (1) the decision makers that make choices based on the legislation and on the available in-house monitoring systems operated by their own data scientists, (2) those data scientists that explore the data and extract relevant information that contributes to evaluation of Public Health scenarios, and (3) the technical team that deploys and maintains the data infrastructure where data scientists are active. It could be very useful to have easy to handle data visualisation modules (like the ones offered by Kibana and shown in Figure 7) allowing decision-makers to choose with a few clicks the representation of data that makes sense to the problems they focus on. The data scientists could then manipulate the current workflows, maximising critical outputs, presenting data in a meaningful way whilst minimising resource required to drive data interrogation and presentation.

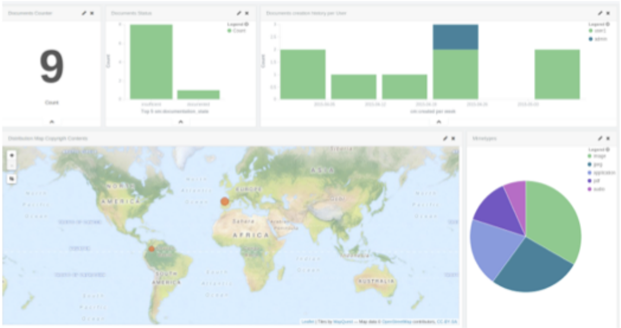


Fig. 7. Kibana screenshot showing the different visualisation modules based on queries to elasticSearch composing an interactive dashboard. This is a puppet example from venzia.es to show the practical potential of this tool.

7. ACKNOWLEDGMENTS

The authors would like to thank the EC Horizon 2020 MIDAS Project and funded by the European Union under grant agreement no. 727721, under the call SC1-PM-18-2016 - Big Data supporting Public Health policies.

8. REFERENCES

[1] J. Brank, G. Leban, and M. Grobelnik. A high-performance multi-threaded approach for clustering a stream of documents. In *Proceedings of the 17th International Multiconference Information Society*, 2014.

[2] J. Brank, G. Leban, and M. Grobelnik. Annotating documents withrelevant wikipedia concepts. In *Proceedings of SiKDD 2017,* forthcoming, 2017.

[3] M. A. Dion M, AbdelMalik P. Big data and the global public health intelligence network (gphin).*CCDR*: Vol. 41-9, September 3, 2015:Big Data, 2015.

[4] ECDC. Epidemic intelligence. ecdc.europa.eu/en/threats-and-outbreaks/epidemic-intelligence. Accessed: 2017-09-05.

[5] M. Grobelnik. Observing global health and well-being. http://www.midasproject.eu/2017/07/24/observing-global-health-and-well-being/. Accessed: 6/8/2017, 2017. MIDAS Project Blog, 2017.

[6] M. Grobelnik and G. Leban. Eventregistry's health panel. https://tinyurl.com/wits2017qnt, 2017. Accessed: 2017-07-25.

[7] M. Grobelnik and D. Mladenić. Simple classification into large topicontology of web documents. *CIT*, 13.4:279–285, 2005.

[8] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik. Cross-lingual detection of world events from news articles. In *Proceedings of the 2014 International Conference on Posters and Demonstrations Track, CEUR-WS. org*, 1272:21–24, 2014..

[9] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, ACM, 2014.

[10] J. P. Linge, J. Belyaeva, R. Steinberger, M. Gemo, F. Fuart, D. Al-Khudhairy, S. Bucci, R. Yangarber, and E. van der Goot. Medisys: medical information system. In *Advanced ICTs for disaster managementand threat detection: collaborative and distributed frameworks*, 131–142, 2010.

[11] A. Muhič, J. Rupnik, and P.Škraba. Cross-lingual document similarity. In *Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces* (ITI), IEEE, 387–392, 2012.

[12] I. Novalija, M. Papler, and D. Mladenić.. Towards social media mining: Twitterobservatory. In *Proceedings of the SiKDD 2014*, 2014.

[13] L. Rei, M. Grobelnik, and D. Mladenić. Event detection in twitter withan event knowledge base. In *Proceedings of SiKDD 2015*, 2015.

[14] J. Rupnik, A. Muhic, G. Leban, P.Škraba, B. Fortuna, and M. Gro-belnik. News across languages-cross-lingual document similarity andevent tracking. *Journal of Artificial Intelligence Research*, 55:283–316,2016.

[15] R. Steinberger, F. Fuart, B. Pouliquen, and E. van der Goot. Medisys:A multilingual media monitoring tool for medical intelligence and earlywarning. In *Proceedings of the International Disaster and Risk Conference IDRC Davos 2008,* 612-614, 2008.

[16] M. Trampus and B. Novak. The internals of an aggregated web newsfeed. In *Proceedings of 15th Multiconference on Information Society IS-2012*, 2012.

[17] T. Štajner, I. Novalija, and D. Mladenić. Informal multilingual multi-domain sentiment analysis. *Informatica*, 37.4:373–380, 2013.

[18] WHO. Epidemic intel. www.who.int/csr/alertresponse/ epidemicintelligence/en/. Accessed: 2017-09-05