A Human-is-the-Loop Approach for Semi-Automated Content Moderation

Daniel Link
European Research Center for Information Systems, Germany
daniel.link@ercis.de

Bernd Hellingrath
European Research Center for Information Systems, Germany
bernd.hellingrath@ercis.de

Jie Ling
University of Münster, Germany
j_ling04@uni-muenster.de

ABSTRACT

Online social media has been recognized as a valuable information source for disaster management whose volume, velocity and variety exceed manual processing capacity. Current machine learning systems that support the processing of such data generally follow a human-***in***-the-loop approach, which has several inherent limitations. This work applies the human-***is***-the-loop concept from visual analytics to semi-automate a manual content moderation workflow, wherein human moderators take the dominant role. The workflow is instantiated with a supervised machine learning system that supports moderators with suggestions regarding the relevance and categorization of content. The instantiated workflow has been evaluated using in-depth interviews with practitioners and serious games. which suggest that it offers good compatibility with work practices in humanitarian assessment as well as improved moderation quality and higher flexibility than common approaches.

Keywords

Disaster management, social media analysis, human-is-the-loop, content moderation, supervised machine learning

INTRODUCTION

In recent years, online social media accessed via smart mobile devices and microblogging platforms has been recognized as a valuable information source potentially contributing to situation awareness in disaster contexts (Imran, Elbassuoni, Castillo, Diaz & Meier, 2013a; Vieweg, Hughes, Starbird & Palen, 2010; Yin, Lampert, Cameron, Robinson & Power, 2012). Since the exchanged information is usually very short, noisy, highly unstructured and its usefulness and quality vary significantly, it is necessary to examine incoming messages, semantically enrich content and control dissemination. Existing workflows like content moderation (Link, Hellingrath & Groeve, 2013) attempt to handle this activity, which is difficult due to the time-critical nature of crisis-related information and the various tasks of content analysis (e.g. burst detection, geo-tagging or message classification) that easily overwhelm human analysts. Supervised machine learning (SML) techniques can be utilized to support the labor-intensive work of content analysis, e.g. for automatic content filtering and categorization. Several application systems have adopted SML techniques for the extraction or auto-categorization of relevant social media messages, especially from the microblogging platform Twitter; for example, the systems *Artificial Intelligence for Disaster Reponse* (AIDR) or *CrisisTracker* (Imran, Castillo, Diaz & Vieweg, 2015).

Unfortunately practitioners still find it difficult to incorporate information from social media into their decision-making, and the impact of relevant information on situation awareness at operational humanitarian agencies is unclear (IFRC, 2013; Tapia, Moore & Johnson, 2013). According to Endert, Hossain, Ramakrishnan, North,

Fiaux, & Andrews (2014), current machine learning and cyber-physical systems that enable interactive analytics commonly take a human-***in***-the-loop approach, wherein analytic algorithms occasionally consult human experts for feedback and course correction. We believe too that human analysts are often "*presented with results out of context, without understanding their meaning or relevance, and interactive controls are algorithm specific and difficult to understand*" (Endert *et al.*, 2014). Moreover, analysis with current systems is often data-driven and the matching of identified information to information needs comes secondary (Vieweg, Castillo & Imran, 2014). This runs contrary to best practices in humanitarian assessment, where experienced analysts are at the heart of the analysis process and the definition of information needs precedes data collection (ACAPS, 2015). As an alternative to human-in-the-loop thinking, the emerging field of visual analytics offers the human-***is***-the-loop concept, which focuses on *"recognizing analysts' work processes, and seamlessly fitting analytics into that existing interactive process*" (Endert *et al.*, 2014).

Taking a design and application-oriented research perspective, in this work we ask what a system could look like that applies the human-***is***-the-loop concept to support human analysts in examining incoming messages from smart mobile devices and online social media, semantically enriching content and controlling dissemination. Specifically, we focus on the assessment of information (e.g. by experts from the Assessment Capacities Project/ACAPS) for decision-makers at operational agencies (e.g. a logistician for the Red Cross), who would otherwise not become aware or be able to make use of relevant information. The goal is to provide relevant information with a high signal to noise ratio. The result of this design effort is a semi-automated content moderation workflow. The workflow is instantiated in a software prototype that utilizes supervised machine learning techniques to provide human analysts with suggestions regarding the relevance and categorization of collected information. The instantiated workflow has been evaluated using in-depth interviews with practitioners and serious games. The evaluation results suggest that the new workflow may indeed achieve better compatibility with work practices in the humanitarian sector, improved moderation quality and higher flexibility compared to other workflows.

The remainder of this paper is structured as follows. First, we describe our research methodology. Then we review methods and systems for content analysis, which provide a human-centered workflow and the necessary context in terms of state-of-the-art machine learning systems. Next, we present the designed semi-automated moderation workflow. The subsequent section highlights key features of the demonstrator prototype that instantiates the workflow. This is followed by a section on the evaluation of the workflow and prototype. The paper eventually concludes with a summary and an outlook.

RESEARCH METHODOLOGY

We followed the Design Science Research (DSR) paradigm, which is widely applied in various research works of Information Systems (IS) to solve relevant problems and simultaneously make contributions to the knowledge base with socio-technical artifacts (Gregor & Hevner, 2013). In particular, we have followed the DSR methodology proposed by Peffers, Tuunanen, Rothenberger & Chatterjee (2007). The newly developed, semi-automated workflow is a novel artifact on the method level, instantiated in a prototype as a proof-of-concept and for the purpose of evaluation (March & Smith, 1995).

REVIEW OF METHODS AND SYSTEMS FOR CONTENT ANALYSIS

In this section, we review several representative methods and systems supporting content analysis.

Types of Workflows

According to Rogstadius (2014), there are three types of workflows: fully-automated, manual (crowdsourcing) and semi-automated/hybrid; with different performance in terms of scalability, accuracy and flexibility, as described in the following. Using a machine-based fully-automated workflow to extract and manage information can achieve high scalability with a certain level of accuracy. However, it is often inflexible and not applicable to the evaluation of unseen problems. In contrast, a manual workflow cannot scale up as quickly and thus favors information overload, but tends to be highly flexible and accurate. This is because human analysts are generally more capable than machines when it comes to digest and evaluate ambiguous content and deal with previously unseen problems. A hybrid workflow can balance these advantages and disadvantages by fully automating repetitive steps and simple tasks while leaving more sophisticated to humans to ensure quality and flexibility. Since a fully automated workflow is not compatible with our fundamental human-is-the-loop approach, we

disregard this type and focus the further review on representative manual and hybrid workflows.

A Manual Workflow: Twitter Integrated Content Moderation in GDACSmobile

GDACSmobile is a solution developed for affected population ("public users") and disaster management professionals ("authorized users") alike, in order to acquire and disseminate real-time disaster-related information from the affected region (Link *et al.*, 2013). As illustrated in Figure 1, public users submit "observations" via Twitter or a mobile app that require to be analyzed by moderators to ensure content quality. Moderators assign the statuses like *rejected* or *accepted* to observations from public users to determine whether these should be disseminated (visible) to all client users, whether public or authorized. Moderators can also modify an observation if necessary, e.g. by re-assigning a category, semantically enriching content or sending a feedback request for further clarification. Observations from authorized users are assumed to be trustworthy based on the users' professional background and are consequently directly accepted and published. During a crisis, moderators are very likely to experience information overload and the moderation process may become a processing bottleneck of the entire information flow (data collection - moderation - publication).

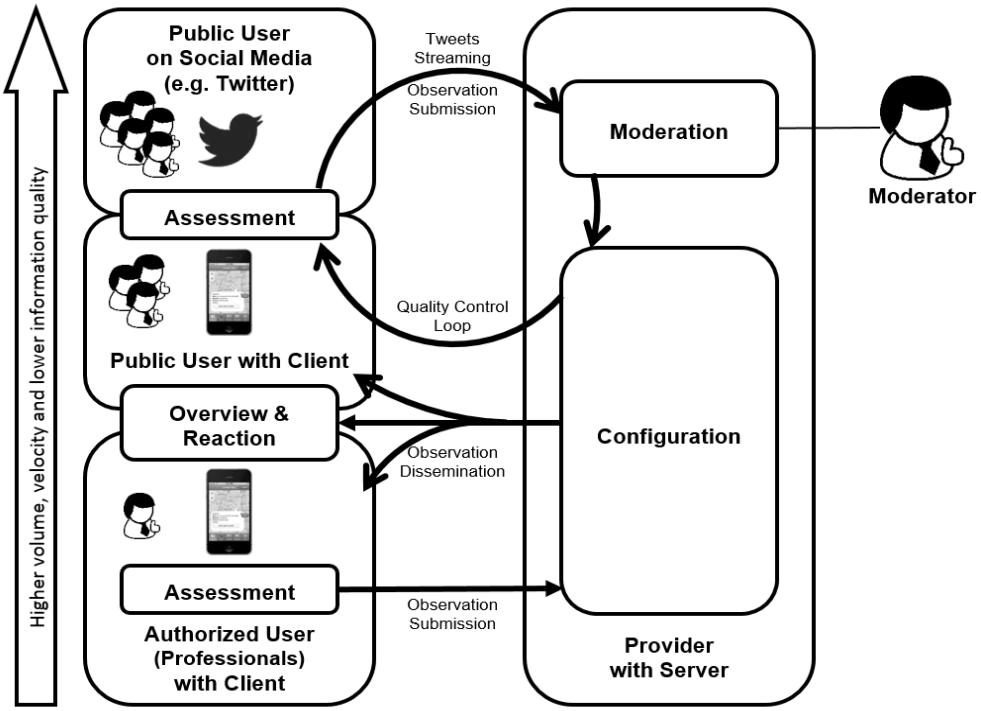

Figure 1. An Overview of GDACSmobile Information Flows

A Hybrid Workflow with Auto-Classification: the AIDR System

The AIDR (Artificial Intelligence for Disaster Response, http://aidr.qcri.org/) system implements a hybrid workflow that utilizes machine learning techniques to automatically classify disaster-related messages from Twitter and SMS into a set of pre-defined information categories (e.g. *donations* or *infrastructure damage*). As a disaster unfolds, crowdsourced human intelligence helps to improve the trained SML classifiers through *active learning*. Figure 2 display the semi-automated workflow with its two core phases: online *classification* and *active learning*. The online classification phase involves three fully-automated processing elements, including a Twitter collector, a feature extractor, and a classifier. The trained classifier can automatically classify the streamed tweets in real-time and output the confidence score of the classification. The active learning phase is semi-automated, including an automatic labeling task generator, an automatic supervised learner with feature selection, and crowdsourced human annotators. AIDR leverages the synergy of machine and human intelligence only in the active learning phase to optimize the classifiers over time by fitting crowd-sourcing annotation workflow into the fully automated classification workflow. The entire workflow is basically steered by the

system with limited flexibility.

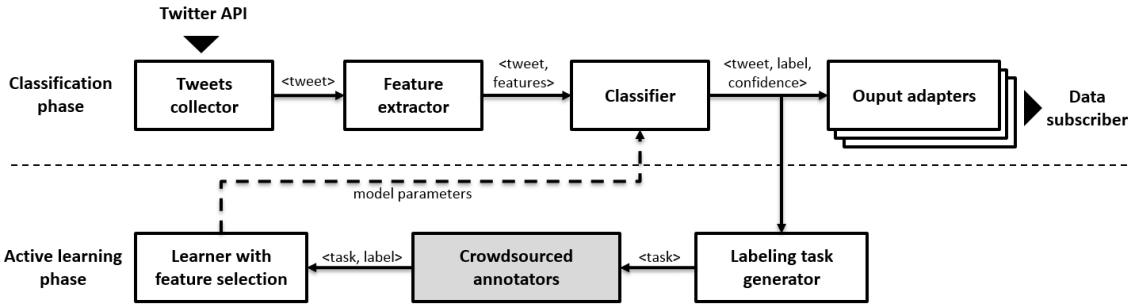

Figure 2: The hybrid model with active supervised learning in system AIDR (adapted from: Imran, Lykourentzou & Castillo, 2013b)

A Hybrid Workflow with Auto-Clustering in the System CrisisTracker

The hybrid workflow in the CrisisTracker system aims to facilitate collaborative social media analysis for disaster response (Rogstadius, Vukovic, Teixeira, Kostakos, Karapanos & Laredo, 2013). To enable a scalable, collaborative analysis workflow, CrisisTracker features a collaborative crowd-based content curation workflow for tweets analysis with automatic clustering (i.e. grouping individual tweets into *stories*). Once individual tweets are grouped into *stories*, the crowd workers only have to invest their precious time on the *top stories*, clusters of tweets posted by at least a certain number of users (e.g. 50), which makes the manual steps in the moderation workflow more scalable and helps to summarize the individual tweets. Moreover, the top stories are automatically ranked according to the number of unique users mentioning them, in order to help optimizing the allocation of the subsequent manual processing work and to ensure the timeliness of situation awareness with the given timeframe and labor budget. To further accelerate the process, CrisisTracker uses supervised classification techniques that support meta-data extraction (i.e. auto-classification of messages). As illustrated in Figure 3, the workflow uses a *classification ahead of clustering* approach. After data collection, all tweets in the stream are firstly classified independently with a label and confidence level attached for each tweet as an output. The tweets with "null" labels and low confidence levels are filtered out before clustering the individual tweets to *stories* based on their parsed textual content in the same way as in the old workflow. Analyzing only top stories implies a substantial amount of information loss of unique or unpopular stories. These unpopular stories may not contribute too much to the general situation awareness but can be essential to the decision making for professional responders who have specific information needs.

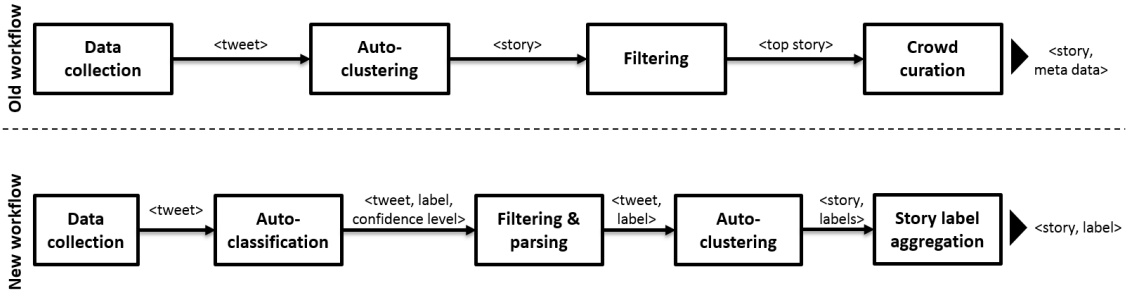

Figure 3. The workflow of CrisisTracker before and after supervised classification (adapted from: Rogstadius, 2014)

Discussion of Human-in-the-Loop Thinking in Supervised Machine Learning

In the field of supervised machine learning, the human-in-the-loop design concept is reflected by an active learning scheme. Active learning enables a feedback loop from human to machine that helps to improve the performance of algorithms to a satisfactory and usable range. The basic idea of active learning is to label only

the data that may contribute more to the enhancement of the classifiers so that a sufficient level of classifier performance can be achieved with less labor cost of labeling. The potential contribution is often determined by the prediction confidence that is estimated by the classifier itself. The most uncertain data instances are more likely to be suitable candidates for training classifiers effectively (Cuzzillo, 2015). Despite its promising aspects for interactive workflow design, the human-***in***-the-loop concept poses some inevitable usability issues, like algorithm specific and difficult to understand controls as well as presenting results out of context (Endert *et al.*, 2014). For example, the active learning scheme applied in AIDR generates individual labeling task instances for crowdsourced annotators to label based on the confidence level assigned by classifiers. However, the crowd workers typically have limited familiarity with the context (i.e. the specific, constantly evolving information needs; see Vieweg *et al.*, 2014) and the big picture of the current classification status in the system (e.g. category distribution and classification performance). If crowd workers' familiarity with the context is limited, the effectiveness of annotator agreement (e.g. 3 workers making the same judgement) as a strategy to mitigate false positives, i.e. mistakenly tagging the information as relevant, is limited too. Thus, experts should be involved in the content moderation process at same point; respecting their specific requirements in workflow and system design. Furthermore, human operators (including experts) should have more control over the entire workflow and ultimate decisions, in order to increase the quality of results and achieve a better fit with existing work practices.

A SEMI-AUTOMATED CONTENT MODERATION WORKFLOW

This section presents an interactive, semi-automated content moderation workflow based on the manual moderation workflow in GDACSmobile.

Requirements for the New Moderation Workflow of GDACSmobile

To balance the quality and scalability of the moderation process, the new workflow shall combine machine and human intelligence not only in the training phase but also in the classification phase (semi-automated moderation).

To ease the information load on moderators, the system shall perform various filtering tasks, as follows. Redundant, irrelevant or uninformative pieces of information should be filtered before human moderation. In contrast, information that likely is of a high quality should be able to bypass human moderation. Furthermore, similar content should be grouped to allow for an aggregated view during human moderation.

To support quality control, on the one hand, uncategorized observations (e.g. from Twitter users) should arrive with suggestions for fitting information categories at human moderation. On the other hand, information categories selected by app users should be checked automatically for correct categorization and, if needed, flagged for human moderation.

To detect a decrease in quality of observations from trusted users whose observations can bypass human moderation, their observations too should be automatically examined and, if suspicious, flagged for human moderation.

A High Level View on the New Workflow

Figure 4 gives an overview of the new workflow. As in the manual GDACSmobile workflow, the first phase is *data collection* from source APIs (e.g. Twitter streaming API), GDACSmobile clients or other external databases. The second phase of *pre-processing* parses incoming data and extracts domain-specific keyword features as well as domain-independent NLP and source-specific features. Subsequently, the *semi-automated content moderation* phase consists of *informative filtering*, *categorization*, *de-duplication* and *quality control*. Informative filtering determines whether an analyzed observation is useful or not, so that only informative content is passed to the following steps. Categorization enriches observations with domain-specific structural information by classifying them into a pre-specified set of categories and adding the achieved confidence level. These steps are described separately here to ease understanding although technically they may be performed simultaneously by a single multi-class classifier. To ease the load on human moderators in later steps, *de-*duplication and near-duplicates detection groups similar data entries into meaningful clusters for joint review (Feldman & Sanger, 2007). During quality control, human moderators review single and grouped observations and decide whether they should be visible to users. This may include several tasks if necessary, like checking automatic tags, aggregating content, refining keywords, interacting with authors in a feedback loop, or

modifying content. Publication eventually disseminates relevant content to the system's user base.

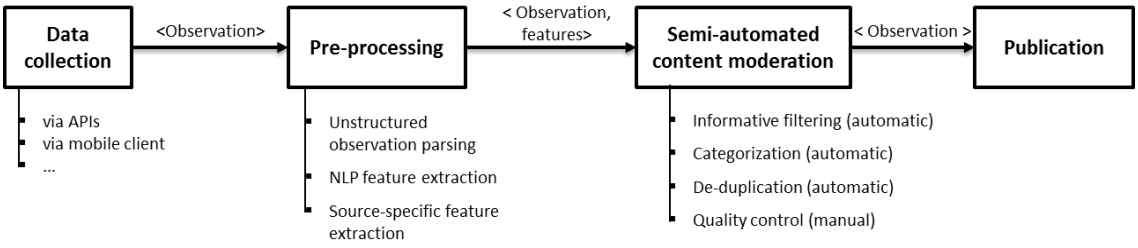

Figure 4. High Level View on the New Workflow

Differentiating Source Types by Trustworthiness

As the manual GDACSmobile workflow distinguishes between untrusted and trusted users, the high level workflow can be decomposed into two sub-flows depending on the information source: *untrusted sources* providing mostly unstructured information that likely requires content moderation and *trusted sources* providing mostly semi-structured information that may bypass content moderation. Figure 5 displays the decomposed workflow.

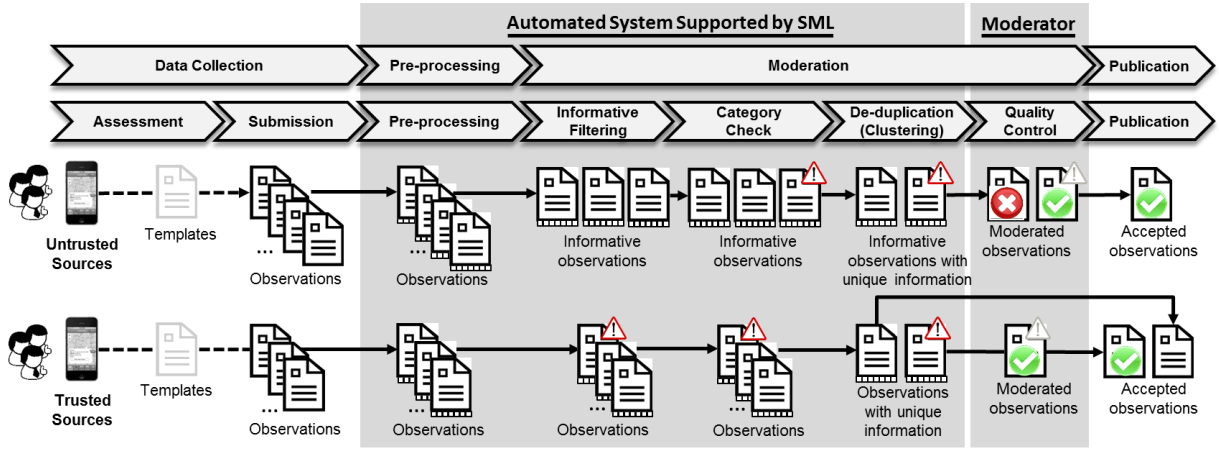

Figure 5. Decomposed Workflow Differentiating Source Types by Trustworthiness

Incoming observations that are semi-structured (via the app) or unstructured (via Twitter) are initially pre-processed. Subsequently, the system examines all observations to identify and flag informative observations and determine the best category fit, based on the content of observations[1]. If an observation's author is considered an untrusted source or if there is a conflict between the user's categorization and the classifiers' judgement, the observation is flagged for manual review during quality control. Then, for each category, an unsupervised de-duplication algorithm is performed to remove identical observations and group the observations with high similarity scores for joint human moderation. In consequence, only qualified candidates from untrusted sources and questionable candidates from trusted sources are passed to quality control for manual review.

The original GDACSmobile state transfer model (Link et al., 2013) includes various states for observations, e.g. *accepted* or *rejected*. To incorporate SML-based suggestions, we extended the original model with three new states: *auto_approved*, *auto_reviewed* and *auto_rejected*. The state of a submitted observation will be transferred to *auto_approved* only if SML classifiers decide that the observation is relevant and its information category can be correctly assigned with a high confidence level. On the contrary, the state will be transferred to

[1] The system relies on observations' content alone and doesn't also take into account the user group (i.e. trust level), in order to prevent a feedback loop between manual assignment of users to source types and the system's automatic checks of observations, which would introduce bias.

auto_rejected if SML classifiers judge that the observation is of low relevance. When SML classifiers' confidence about relevance or category assignment is too low, the corresponding observations' state is transferred to *auto_reviewed*. All the observations in the *auto_reviewed* state and all observations from untrusted sources in the *auto_approved* state need to be reviewed before publication[2].

Addressing Knowledge Transfer with a Continuous Learning Process

Most application systems merely utilize a data-driven (bottom-up) approach to analyze Twitter datasets without considering any domain-specific knowledge, which may cause misclassification of text messages especially in a transfer scenario, i.e. when reusing pre-trained classifiers of one crisis to another crisis. Even when the classifiers are trained with similar events, the performance of reusing classifiers with a data-driven approach usually still suffers from a substantial loss of accuracy (Imran, Castillo, Lucas, Meier & Rogstadius, 2014). The reason may lie in the fact that a data-driven approach focuses merely on what knowledge or information the underlying datasets can offer. To address this issue, recent research has taken a knowledge-driven (top-down) approach that pays more attention to specific information needs in order to inform analysis (Link, Horita, Albuquerque, Hellingrath & Ghasemivandhonaryar, 2015). In an attempt to improve classifier performance in transfer scenarios, we designed a continuous learning process that incorporates both domain-specific keywords as proxies for domain-specific knowledge and training sets from previous disasters; see Figure 6.

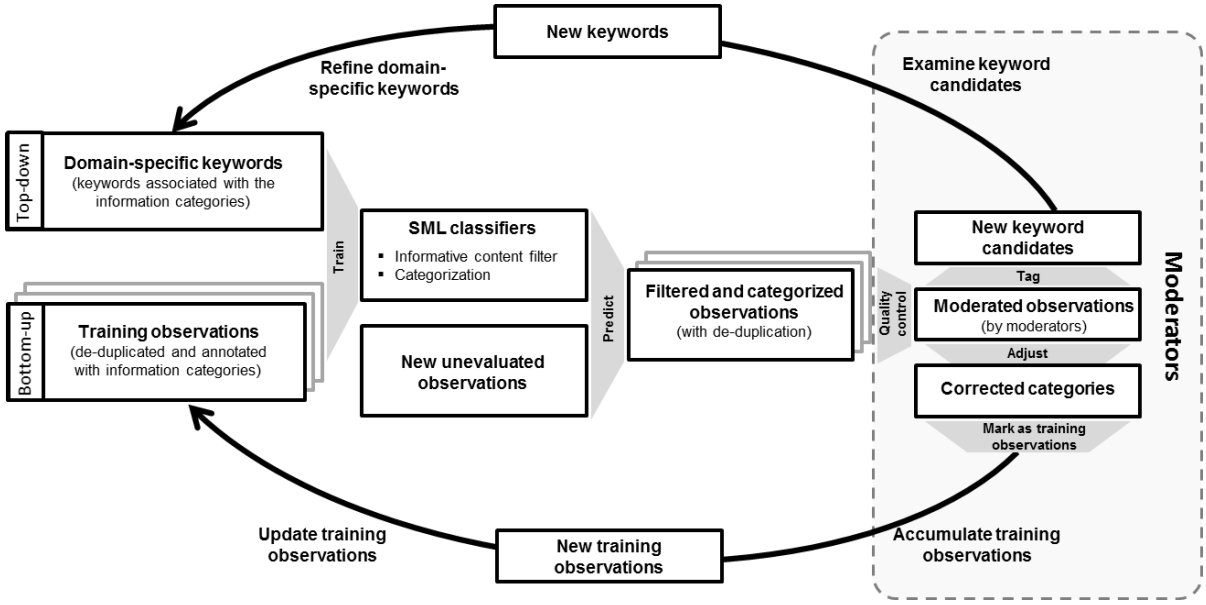

Figure 6. A Continuous Learning Process of SML Classifiers

PROTOTYPE INSTANTIATING THE WORKFLOW

In this section we describe the key features of the software prototype instantiating the semi-automated content moderation workflow.

The Moderator's Inbox

The *Inbox* is the moderators' central workbench. As shown in Figure 7, its left-hand side shows a summary with the number of observations for different states. Moderators can use these to allocate and prioritize their work. On the top-right, moderators can define information categories and source types to focus their review and distribute work among multiple moderators to better leverage specific expertise and for load balancing. The *auto_rejected* part of the inbox can be regarded as a spam box. Depending on their strategy and workload,

[2] Moderators may choose to review observations with other states as well if it fits their moderation strategy and their workload permits.

moderators may choose to check the spam box for false negatives and thus reveal new patterns. The contained observations can be ordered by their estimated degree of relevance to one of the information categories. Duplicates are displayed last, as redundant information usually doesn't require any further actions.

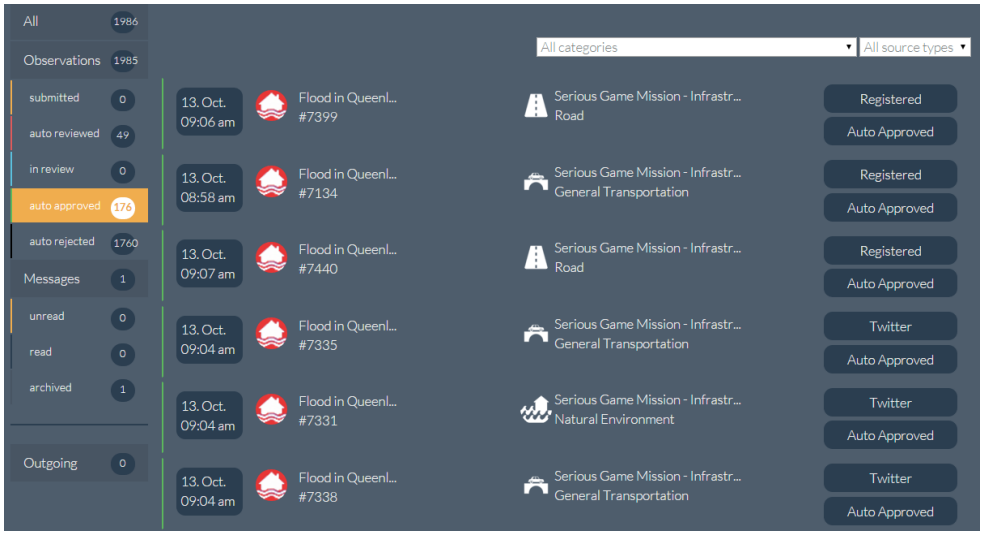

Figure 7. A View on the Moderator's Inbox

Automatically Approved Observations Bypassing Quality Control

According to the workflow, observations that successfully pass all automatic quality checks can skip manual review. Figure 8 shows a detailed view on such an automatically approved observation.

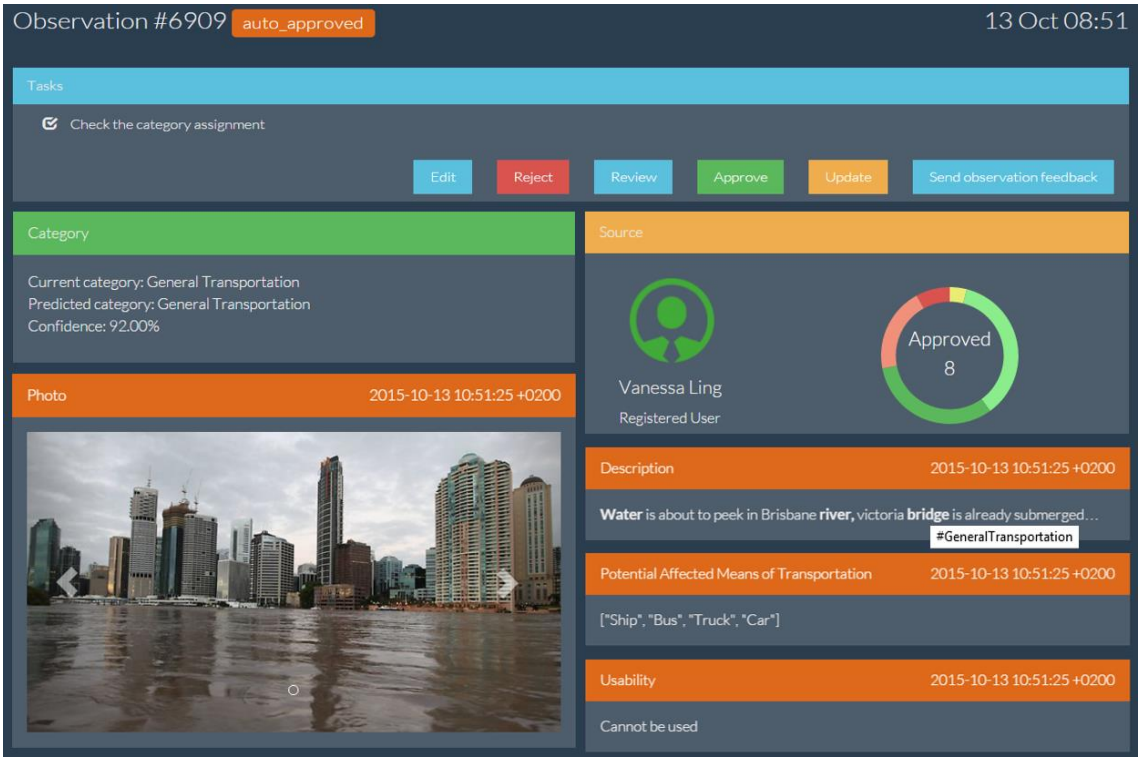

Figure 8. Details of an Automatically Approved (auto_approved) Observation

The category section indicates that SML classifiers placed the observation in the *General Transportation* category with a confidence level of 92%, which matches the observer's initial selection. The tasks section lets

the moderator modify the observation, change its status or interact with the observer via observation-centered messages. It can show a checklist of mandatory tasks, whose entries are marked us done either by the SML system or manually by moderators. The source section provides statistics that indicate the trustworthiness of the observation's author, which the moderator can use to decide whether that user should be transferred to the moderated user group. This includes a pie chart, whose colors indicate the status of the user's other observations (e.g. for *auto_rejected*), showing the exact number on mouse over. In the text description of the observation, the identified domain-specific keywords are highlighted, showing the corresponding information categories on mouse over (e.g. "bridge" is associated with *General Transportation*).

Manual Moderation

When an observation doesn't contain enough data for automatic analysis or the SML system didn't reach a satisfying confidence level, the observation is marked as auto_reviewed state for manual moderation. For example, Figure 9 shows an observation with less than 50% confidence for correct categorization. The task section consequently asks moderators to "check the category assignment". During manual review, a moderator update the observation and mark it as a training instance. Furthermore, there is a section at the bottom showing similar observations in descending order by their degree of similarity.

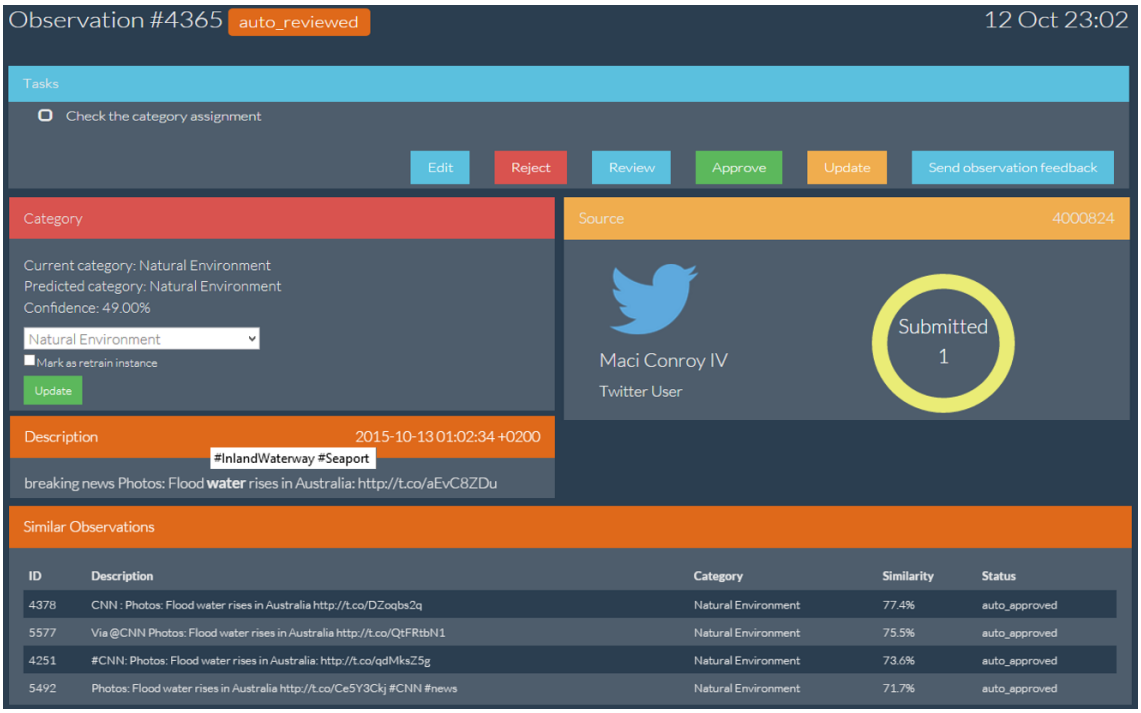

Figure 9. An (auto_reviewed) Observation Requiring Manual Moderation

EVALUATION

The evaluation aims to demonstrate the feasibility of the semi-automated moderation workflow and qualitatively assess its performance by demonstrating the prototype in in-depth interviews with practitioners and by deploying it in a serious game.

In-Depth Interviews with Practitioners

In-depth interviewing is an effective qualitative evaluation method to explore respondents' thoughts, perspectives and feedback on the ideas and outcomes of a study. The major advantage of in-depth interviews is that they enable detailed information and feedback collection via intensive discussions with a small number of respondents in comparison to other evaluation methods, such as surveys. Despite its potential pitfalls, such as being prone to bias, time-intensive and having generalization issues (Boyce & Neale, 2006), it is still a valuable method that is widely used in practice and academic research. For this work, we conducted semi-structured

interviews with four practitioners: two senior logisticians, with one of the two having extensive experience conducting field assessments, an expert on humanitarian assessments, and a senior digital volunteer.

The interviewees agreed on the need to support the processing of social media data, in order to identify relevant information (which is often lacking) and to avoid irrelevant information (which is often abundant). To this end, the developed workflow and prototype were perceived as feasible, simple and flexible solutions. However, the inflow of information may still be too high, requiring an increased focus on the most pressing information needs.

Organizations heavily favor the information sources they are used to, and increasing adoption of systems like the one presented here would require building trust towards not only the system but also its operators and the data sources utilized. In particular, social media data is perceived as an especially subjective and thus problematic source in higher need of quality control before it can be disseminated to decision-makers; to the degree where it would only be considered useful in natural disaster contexts but not conflicts. A possible way to utilize at least some parts of social media would be to enable further distinction between users according to their trustworthiness, e.g. by assigning a high weight to a key app user or to the official Twitter account of a reliable government agency.

The workflow is limited insofar as it focuses on analysis but neglects synthesis. When moderators spend time reviewing observations, they could already take note of observations that inform certain questions. The workflow and the instantiating system should be extended with features for synthesis, which help to make sense of content of sufficient quality in the light of higher level information needs; e.g. moving from individual reports of road blocks to a summary of issues of access.

Serious Games

Serious games can be regarded as an effective and resource-efficient method for evaluating IT tools in the crisis domain by balancing the involvement of non-professional players with realism and thus validity (Meesters, 2014). For this evaluation, we rely on the game design workflow proposed by Link, Meesters, Hellingrath & Van de Walle (2014) to evaluate the feasibility of the semi-automated moderation workflow.

The in-game crisis context is based on the tropical cyclone that hit Queensland, Australia, from 17th to 28th January 2013. In the game, two aid organizations respond to another cyclone hitting the east coast of Australia on 12th October 2015, causing heavy rainfalls, wide-spreading flash floods and infrastructure damages. Each organization employs a logistician, who plans delivery routes within the affected region, and an analyst, who is supposed to supply the planner with relevant information. The inflow of 229 observations and 1976 tweets from mobile client and Twitter users is high enough to certainly exceed the analysts' processing capacity. The serious game took three hours in total, including one hour of preparation (i.e. briefing and test run), 75 minutes of game execution (incl. 60 minutes of moderation and 75 minutes of logistics planning) and 45 minutes of feedback discussion.

A first round of gameplay for testing, involving different players than the final round, showed that some information categories (e.g. buildings, natural environment, country overview) are assigned with a high level of agreement among coders although the categorization was wrong. This led to a more extensive briefing with better explanations of information categories, including examples. This emphasizes the weakness of inter-coder agreement as a mitigation strategy for false positives in case of limited coder expertise, supporting the case for expert involvement in content moderation. For the final round, only categories very chosen where full agreement correlates with correct categorization.

In the final round, both moderators were incapable of keeping up with incoming observations. They chose to focus mainly on auto_approved and auto_reviewed observations. As shown in Figure 10 and Table 1, they were able to examine approximately half of the incoming observations from app users and a fraction of tweets with varying rates of approval. The results suggests that the developed system can successfully support moderators to identify useful information under information overload. The differentiation of automatic states (*auto_approved*, *auto_reviewed* and *auto_rejected*) not only helps to filter out the highly irrelevant or not useful information to reduce the workload of moderators, but also supports their prioritization decisions of content moderation.

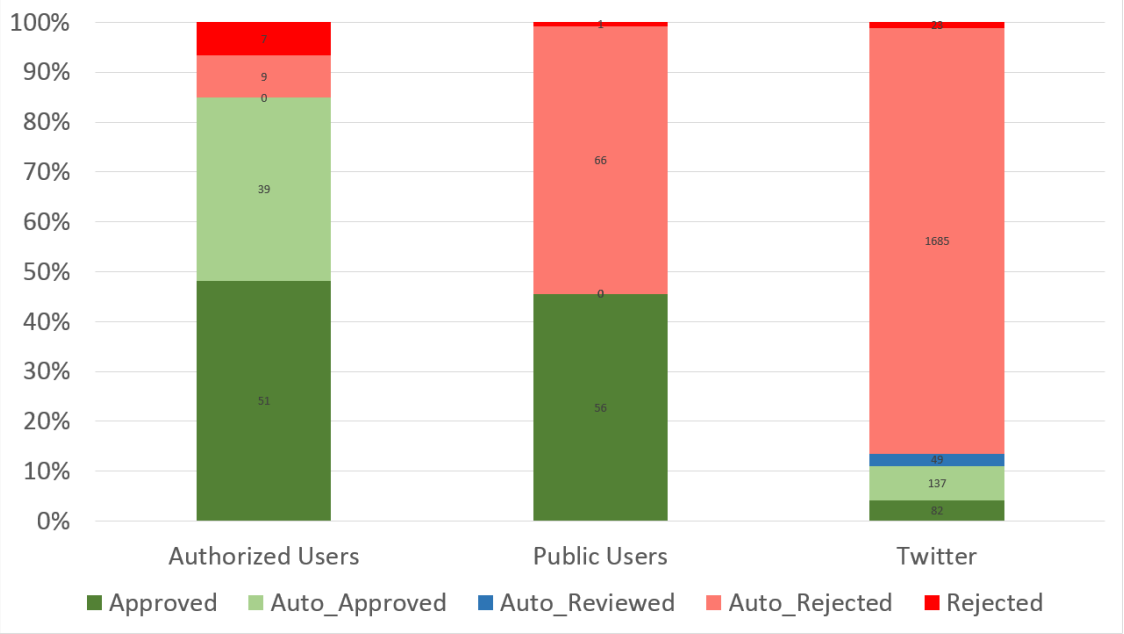

Figure 10. Serious Game Moderation Statistics

Final state / Initial state	Approved	Auto_approved	Auto_reviewed	Auto_rejected	Rejected
Auto_approved	124	176	0	45	15
Auto_reviewed	54	0	49	19	5
Auto_rejected	11	0	0	1696	11

Table 1. State Transfers in the Serious Game

The major infrastructure damages or obstructions, such as road closures and flooded bridges, were mostly identified and incorporated by the logisticians into their route planning. The logisticians agreed that the moderated information is useful to plan initial routes, check their feasibility and identify alternatives. They would like to see a built-in, automatically updated, route-based view on observations instead of using Google Maps in a separate browser window. In addition, they suggested interface improvements, like more detailed filtering options (e.g. by region, timestamp or number of views).

Another interesting result from the evaluation is that coder's perception of relevance seems to be relative. That is, categories with a higher share of useful information tend to be judged more critically, while coders seem to be more lenient when the share of relevant information is low.

CONCLUSIONS

Information from smart mobile devices and online social media can add value to decision-making in disaster management. Making valuable information available to practitioners requires to overcome the limits of human information processing capacity. To this end, supervised machine learning techniques have proven to be useful. The implementing information systems are, however, often based on human-in-the-loop thinking, which poses usability issues and other problems that lower their impact on decision-making. In this work, we relied on the human-is-the-loop approach from visual analytics to create an alternative design that revolves around human analysts and their work processes.

The resulting semi-automated content moderation workflow and the instantiating software prototype utilize supervised machine learning techniques to provide human analysts with suggestions regarding the relevance and

categorization of collected information. In-depth interviews with practitioners and a serious game suggest that the semi-automated workflow does indeed promise better compatibility with work practices in humanitarian assessment, improved moderation quality and higher flexibility. The existing filtering options and various measures, such as moderation statistics or levels of confidence for automatic classification, allow moderators to better focus their sparse capacity on most promising observations.

There are various limitations. Despite the current level of support, human analysts still have to deal with too much information. A finer distinction between information sources in terms of their trustworthiness and better filtering options would enable analysts to better focus on the sources they deem most suitable to address pressing information needs. Giving analysts the option to assign weights to individual users, such as the Twitter accounts of reliable government agencies, would be another step towards building trust in online social media as an information source. Furthermore, the use of identified, relevant information in areas of decision-making such as logistics planning, would benefit from task-oriented decision support modules. For example, there could be a module that uses route waypoints to filter relevant observations and notifies the planner of updates.

ACKNOWLEDGEMENTS

We want to thank the following students at the Research Group on Information Systems and Supply Chain Management for their contribution: Anton Becker, Carsten Bubbich, Friedrich Chasin, Jonathan Dölle, Jonas Juchim, Sven Kronimus, Ferdinand Knoll, Magdalena Lang, Stefan Laube, Marius Pilgrim, Philipp Saalmann, Mohamed Junaid Shaikh, Yannic Schencking, Martin Vanauer, and Patrick Vogel. The first design cycle owes heavily to the involvement of Adam Widera. For sharing their views in many discussions during design and development, we also want to thank practitioners Minu Limbu (UNICEF Kenya), Gintare Eidimtaite and Thomas Peter (both UNOCHA) as well as researchers Tom de Groeve, Alessandro Annunziato and Ioannis Andredakis (all JRC) and various members of the Humanitarian Logistics Association (HLA). Furthermore, we want to thank the participants of the in-depth interviews and serious games for this study. The research leading to these results has partly received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement n°607798.

REFERENCES

1. ACAPS (2015) Humanitarian Needs Assessment: The Good Enough Guide. Practical Aciton Publishing, Rugby, UK.
2. Boyce, C. & Neale, P. (2006) Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input. Pathfinder International, Watertown, MA.
3. Cuzzillo, T. (2015) Real-World Active Learning: Applications and Strategies for Human-In-the-Loop Machine Learning, 1st. O'Reilly Media, Inc, Sebastopol, CA.
4. Endert, A., Hossain, M. S., Ramakrishnan, N., North, C., Fiaux, P. & Andrews, C. (2014) The human is the loop: new directions for visual analytics. Journal of Intelligent Information Systems, 43 (3), 411–435. doi: 10.1007/s10844-014-0304-9.
5. Feldman, R. & Sanger, J. (2007) The Text Mining Hand Book - Advanced Approaches in Analysing Unstructured Data. Cambridge University Press.
6. Gregor, S. & Hevner, A. R. (2013) Positioning and Presenting Design Science Research for Maximum Impact. MIS quarterly, 37 (2), 337–355.
7. IFRC (2013) World Disasters Report 2013: Focus on technology and the future of humanitarian action, Geneva.
8. Imran, M., Castillo, C., Diaz, F. & Vieweg, S. (2015) Processing Social Media Messages in Mass Emergency. ACM Computing Surveys, 47 (4), 1–38. doi: 10.1145/2771588.
9. Imran, M., Castillo, C., Lucas, J., Meier, P. & Rogstadius, J. (2014) Coordinating Human and Machine Intelligence to Classify Microblog Communications in Crises. In: Proceedings of the 11th International ISCRAM Conference. Hiltz, S. R., Pfaff, M. S., Plotnick, L., Shih, P. C. (eds.).
10. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. & Meier, P. (2013a) Extracting information nuggets from disaster-related messages in social media. In: Proceedings of the 10th International ISCRAM Conference, Baden-Baden, Germany. Comes, T., Fiedrich, F., Fortier, S., Geldermann, J., Yang, L. (eds.), pp. 1–10, Baden-Baden, Germany.

11. Imran, M., Lykourentzou, I. & Castillo, C. (2013b) Engineering Crowdsourced Stream Processing Systems. arXiv preprint arXiv:1310.5463 (October), 1–32.
12. Link, D., Hellingrath, B. & Groeve, T. de (2013) Twitter Integration and Content Moderation in GDACSmobile. In: Proceedings of the 10th International ISCRAM Conference. Comes, T., Friedrich, F., Fortier, S., Geldermann, J., Müller, T. (eds.), pp. 67–71, Baden-Baden.
13. Link, D., Horita, F. E. A., Albuquerque, J. P. de, Hellingrath, B. & Ghasemivandhonaryar, S. (2015) A Method for Extracting Task-related Information from Social Media based on Structured Domain Knowledge. In: Proceedings of the 2015 Americas Conference on Information Systems.
14. Link, D., Meesters, K., Hellingrath, B. & Van de Walle, B. (2014) Reference Task-based Design of Crisis Management Games. In: Proceedings of the 11th International ISCRAM Conference. Hiltz, S. R., Pfaff, M. S., Plotnick, L., Shih, P. C. (eds.), pp. 592–596.
15. March, S. & Smith, G. (1995) Design and Natural Science Research on Information Technology. Decision support systems, 15, 251–266. doi: 10.1016/0167-9236(94)00041-2.
16. Meesters, K. (2014) Towards using Serious Games for realistic evaluation of disaster management IT tools. In: Actes de la 2ème Journée AIM de recherche Serious Games et innovation. Proceedings of the 2nd AIM Research and Innovation Day Serious Games. Boughzala, I., Lang, D., Said, A. (eds.), pp. 38–48.
17. Peffers, K. E. N., Tuunanen, T., Rothenberger, M. a. & Chatterjee, S. (2007) A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems, 24 (3), 45–78. doi: 10.2307/40398896.
18. Rogstadius, J. (2014) Enhancing Disaster Situational Awareness Through Scalable Curation of Social Media. Doctoral Thesis, Funchal, Portugal.
19. Rogstadius, J., Vukovic, M., Teixeira, C. a., Kostakos, V., Karapanos, E. & Laredo, J. a. (2013) CrisisTracker: Crowdsourced social media curation for disaster awareness. IBM Journal of Research and Development, 57 (5), 4:1- 4:13. doi: 10.1147/JRD.2013.2260692.
20. Tapia, A. H., Moore, K. A. & Johnson, N. J. (2013) Beyond the Trustworthy Tweet: A Deeper Understanding of Microblogged Data Use by Disaster Response and Humanitarian Relief Organizations. In: Proceedings of the 10th International ISCRAM Conference. Comes, T., Friedrich, F., Fortier, S., Geldermann, J., Müller, T. (eds.), pp. 770–779, Baden-Baden.
21. Vieweg, S., Castillo, C. & Imran, M. (2014) Integrating Social Media Communications into the Rapid Assessment of Sudden Onset Disasters. In: Proceedings of the 6th International Conference on Social informatics. Aiello, L. M., McFarland, D. (eds.), pp. 444–461.
22. Vieweg, S., Hughes, A. L., Starbird, K. & Palen, L. (2010) Microblogging during two natural hazards events: What Twitter May Contribute to Situational Awareness. In: Proceedings of the 28th international conference on Human factors in computing systems. Mynatt, E. (ed.), pp. 1079–1088. ACM Press, New York, NY.
23. Yin, J., Lampert, A., Cameron, M., Robinson, B. & Power, R. (2012) Using Social Media to Enhance Emergency Situation Awareness. IEEE Intelligent Systems, 27 (6), 52–59. doi: 10.1109/MIS.2012.6.