# Evaluation results for a Social Media Analyst Responding Tool

**Mahshid Marbouti**
University of Calgary
mmarbout@ucalgary.ca

**Craig Anslow**
Victoria University of Wellington
Craig.Anslow@ecs.vuw.ac.nz

**Frank Maurer**
University of Calgary
fmaurer@ucalgary.ca

**ABSTRACT**

We take a human-centered design approach to develop a fully functional prototype, SMART ("Social Media Analyst Responding Tool"), informed by emergency practitioners. The prototype incorporates machine learning techniques to identify relevant information during emergencies. In this paper, we report the result of a user study to gather qualitative feedback on SMART. The evaluation results offer recommendations into the design of social media analysis tools for emergencies. The evaluation findings show the interest of emergency practitioners into designing such solutions; it reflects their need to not only identify relevant information but also to further perceive the outcome of their actions in social media. We found out there is a notable emphasis on the sentiment from these practitioners and social media analysis tools need to do a better job of handling negative sentiment within the emergency concept.

**Keywords**

Situation Awareness, Social Media, Emergency Management, User Study.

## INTRODUCTION

Effective usage of social media data in emergency management can lead to successfully restoring safety and recovering essential services. However, the widespread usage of social media data makes it hard for authorities and emergency responders to make sense of thousands of incoming messages. During emergency events people actively broadcast different kinds of emergency-related information (Hughes and Palen, 2009). Affected people no longer wait for official communications (Cameron, et al., 2012; Plotnick, et al., 2015). Since Twitter is one of the most widely used real-time microblog services (Stieglitz, et al., 2017), we use Twitter as an example of social media streams that can provide relevant information in emergencies. For example, Twitter reported that during Hurricane Sandy in 2012, people sent more than 20 million tweets about the storm within 6 days (Marina, et al., 2015). A closer inspection of these social media posts shows that in the hours after the event, people provided on-the-ground information including eyewitness descriptions, impact on the community, requests for help, and expressions of fear (Chokshi, 2015; Imran, et al., 2014). Users outside the emergency area may also contribute to social media platforms by sharing information about wishes, volunteering and donations, caution and advice, or posting about people in the affected areas. This up-to-date information may enhance the disaster response teams' situation awareness (SA) by helping them allocate resources and coordinate rescue actions. However, not all the shared information in the affected areas is related to the disaster event, and not all related information contributes to situation awareness of emergency responders (Hughes, et al., 2014). In addition, information often arrives at a high rate which makes it difficult for analysts in the Emergency Operation Centers (EOCs) to manually filter through, monitor, and analyze such texts in a time-critical emergency (Hiltz and Plotnick, 2013). Finally social media brings reliability and trust issues (Marcelo, et al., 2010). Emergency practitioners face challenges such as determining the credibility of social media sources, or dispelling rumors (Marbouti and Maurer, 2016; Plotnick, et al., 2015). On the other hand, the expectation of the crowd from official responders has changed. People expect authorities to be able to catch quickly urgency requests and

enquiries using these new communication channels but the authorities are not there yet (Flizikowski, et al., 2014; Marbouti and Maurer, 2016; Reuter, et al., 2016b). The goal of this study is to bring insight into the design of expert-informed machine learning solutions for identifying relevant information from social media.

We iteratively designed and developed a Social Media Analyst Responding Tool (SMART) using a human centered design approach (Constantine and Lockwood, 1999). The target users are those who are monitoring and analyzing social media information during emergency events. Depending on the organization, the terminology of what the role can be called can be different (e.g., digital communication officer, public information officer, or social media analyst). Regardless of naming, all of these roles monitor social media information to get a sense of a situation (e.g., what public and media are saying about an event, what are the trends, rumors). In this study, we refer to this role as ***social media analyst*** or sometimes as ***analyst***. Initially, we interviewed emergency practitioners and social media analysts in the emergency domain to understand their challenges and requirements when monitoring social media (Marbouti and Maurer, 2016). After the initial interview phase, we pursued a scenario based design approach (Carroll, 2000) to consider the variety of context in which the tool might be used and shape the feature set. We focused on designing a prototype to help develop advanced filters using the analyst's knowledge combined with machine learning classifiers. The machine learning classifier identifies SA tweets by aggregating textual, social, location, and tone based features to increase precision and recall of SA tweets (self-citation). The design process proceeded with producing low-fidelity paper prototypes. We showed these to social media analysts to gather their reaction and comments. Using this feedback, we evolved the prototype to a fully functional web based prototype (SMART). SMART involves analysts in determining the filters according to their tasks and help them identify relevant tweets.

In this paper, we focus on the result of our evaluation of SMART. The goal is to evaluate the usefulness of the proposed design and gather qualitative feedback. It was important for us to see if the proposed design matches analysts' expectations during a disaster. The evaluation findings can bring insight into the design of social media analysis tools. To incorporate machine learning classifiers into the design of social media analysis tools, the training tasks should fit within social media analysts' existing workflow. According to our findings, emergency practitioners care a lot about sentiment, but sentiment analysis tools need to do a better job of handling negative sentiment to be useful to those people.

## RELATED WORK

Many recent studies discuss the importance and rise in the usage of social media (Reuter and Kaufhold, 2018). In this section, we provide an overview of the current approaches to identify relevant tweets during emergencies.

### Supervised Classification based studies

Some studies applied machine learning models to solve the problem of separating relevant tweets from irrelevant ones in large-scale emergencies (Moon, et al., 2013). Some studies applied binary classifiers to detect small scale incidents like car crashes (Schulz, et al., 2013) or to detect fires (Power, et al., 2013). Verma et al. (Verma, et al., 2011) study designed a classifier model to distinguish tweets that are contributing to SA from tweets that are not. The classification based studies, demonstrated the applicability of learning models with text mining techniques to automate Microblog filtering (Hughes, et al., 2014) but they do not provide applicable tools. Imran et al. (Imran, et al., 2015) surveys different studies for processing social media in mass emergencies. In our study, we embed supervised machine learning classifiers into a fully functional prototype to address the needs of practitioners.

### Social Media Analysis tools

Several papers explored how to extract relevant information from social media data and more specifically from Twitter and resulted into social media analysis tools. SensePlace2 (MacEachren, et al., 2011) is a web-based tool that was built upon Twitter data to provide situation awareness and sense-making by revealing the connection between people, places, topics, and organizations. While this study provided useful tools for monitoring and filtering tweets, they provided simple filtering techniques such as keyword based filtering or filtering by location. In Rogstadius et al. (Rogstadius, et al., 2013) disaster awareness tweets were extracted using sets of keywords, constructing stories, and clustering tweets using their lexical similarity. The Twitcident system (Abel, et al., 2012) used to filter relevant information about an incident using handcrafted rules. This method enabled filtering, searching and analyzing tweets in real-time. Twitcident used semantics techniques to filter tweets. They gathered geo-tagged tweets and employed a learning model based on handcrafted rules (e.g., if a tweet contains specific attributes then it is classified in a specific category). In our proposed prototype, we

provide machine learning classifiers to identify situation awareness and user-defined tags and we let users train the classifiers in real time.

Emergency Situation Awareness (ESA) (Yin, et al., 2012) provides a Microblog mining tool to extract and visualize real-time information that can contribute to situation awareness about an incident. ESA distinguishes those tweets that are reporting an infrastructure damage using classification learning method based on NLP and social based features. The other proposed a visual classifier tool named Scatterblogs (Bosch, et al., 2013) for Microblog analysis and classification. Scatterblogs utilized a binary classifier based on text similarity to distinguish emergency-related tweets. This method allowed users to visually filter related tweets in real-time then train and update the classifier model. In Thornton et al. (Thornton, et al., 2016), the feedback-based techniques were used to improve situation awareness instead of traditional key-word based by updating the selected features in the learning model. AIDR (Artificial Intelligence for Disaster Response) (Imran, et al., 2014) study proposed a platform to apply automatic classification on emergency-related microblog posts. Their objective was to classify microblogs into a set of user-defined categories, combining machine learning classification techniques and human participation in labeling emergency-related microblogs in real-time. These tools provide automated methods for analyzing and exploring Twitter data. However, none of these approaches directly addresses the needs of social media analysts or emergency practitioners.

Our work extends these studies (Hughes and Shah, 2016; Reuter, et al., 2016a; Reuter, et al., 2017) that designed and developed an application based on the social media needs in the emergency context. However, in their applications they do not incorporate machine learning filters. In this study, we also designed and developed a prototype based on the needs of social media analysts but our evaluation brings insights into the design of applications that embeds supervised machine learning classifiers in monitoring social media during emergency scenarios. As far as we are aware this is the first tool to explore this concept in social media analysis for emergency management.

## SMART

SMART is a social media analysis prototype that allows analysts to identify relevant information during emergencies. SMART combines the ability to collect, search and sort tweets alongside the functionality to filter and organize them with machine learning classifiers. A more detail description can be found elsewhere (Marbouti, et al.).

SMART utilizes set of data **collectors** to provide historical and live data gathering (Figure 1-A). SMART can extract "**Topics**" (Figure 1-B) using Latent Dirichlet allocation (LDA) topic modeling (Blei, et al., 2003), which helps the user find sub-events or trends within a larger event. These topics are defined by a list of five top keywords.
Users can further narrow down their search by creating **Filtered Events** (Figure 2). A filtered event can be populated based on the tweets that analysts manually forwards to the filter (we refer to it as manual filtered event) or automatically based on keywords. Figure 2 shows a filtered event view. The tweet and the buttons below (Figure 2-C) allow quick interactions for labeling and reading tweets at once. The color of the border around each tweet displays the sentiment of individual tweets. We embedded two main machine-learning classifiers to facilitate interactive filtering. The solution let users label tweets to train the classifiers. Any automatically assigned label can also be changed by the analyst if necessary. Users can label tweets based on SA or based on user-defined labels:
**(1) Filtering by situational awareness (SA)**: We embedded a SA classifier in SMART. In the interface, SA is tracked by a lit or unlit lightbulb symbol for each tweet (Figure 2-C). If the SA classifier predicts situational awareness, the lightbulb symbol will have a red background to allow analysts to distinguish between classifiers' output and manually labeled tweets. **(2) Filtering by tagging:** To facilitate categorization of tweets into user defined labels we embed a multinomial Naïve Bayes classifier using basic textual features (Bag of Words (Aggarwal and Zhai, 2012) and TF-IDF (Salton and Buckley, 1988)). As an example, an analyst manually tagging tweets to categories such as Detail, Questions, Media, Misinformation or user-defined categories will also train the classifier to tag similar tweets with the same tags (using the tagging button in Figure 2-C), thus increasing the speed that information can be filtered. The automatically tagged tweets will be red in color, and can be changed if the tag is deemed inaccurate by the user.
Figure 2-A shows **the map view** in which users can see information regarding the location of the tweets, and the location of the users based on their profile information. If SMART extracts the location of the tweet from its text, using the embedded location extraction process (Marbouti, et al.), it will be red. **The graph button** (Figure 2-D) allows the map to be toggled off and display graphs for sentiment, interestingness, situation awareness and tags.

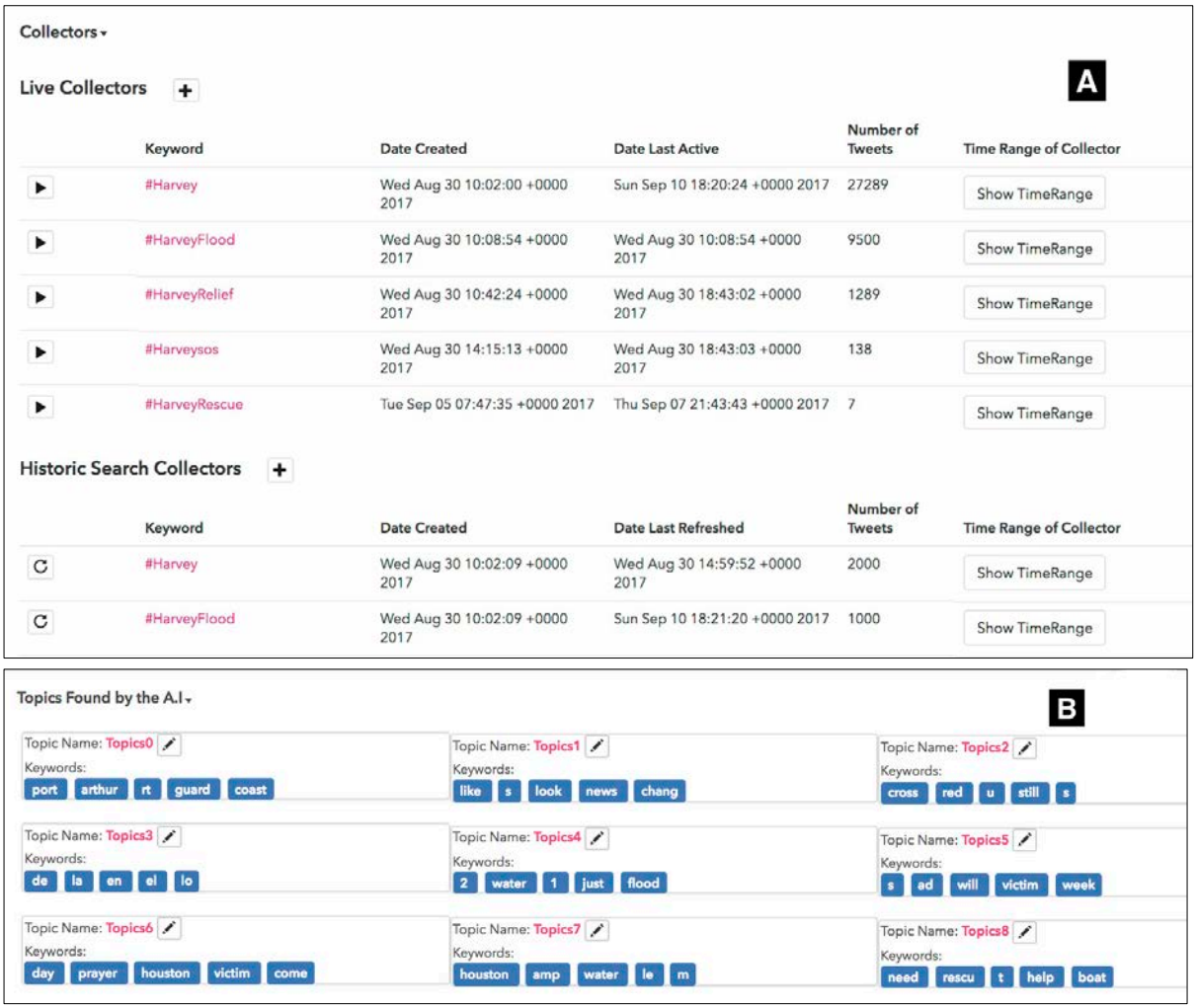


**Figure 1- shows a list of live and historical collectors for hurricane Harvey(A) and a list of topics regarding the event (B)**

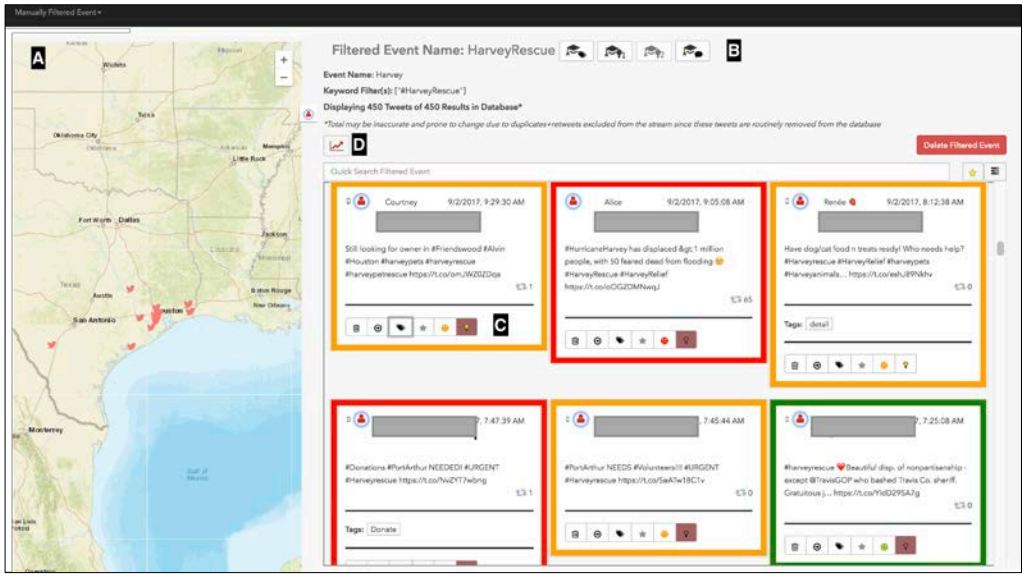


**Figure 2. SMART Filtered Event Page: This view shows the filtered event page in which users can view and label streams of tweets. The graduation hats (Figure 2-B) are used for training the tagging and SA classifiers. Users can also view the users' and tweets' locations in the map view.**

## EVALUATION STUDY

In this section, first we elaborate the evaluation procedure, participants' profiles, emergency scenarios and finally we discuss our findings.

### Evaluation Procedure and Participants

We conducted four evaluation sessions with nine participants. Although this is a small number, it is still relevant, because it is difficult to schedule meetings with emergency experts. We conducted one group session with six participants and three individual sessions. We recruited emergency practitioners and social media analysts in the emergency domain for evaluating SMART. For conducting a usability session, we contacted Calgary Emergency Management Agency(CEMA), and scheduled a time with their communication team. We preferred one on one meetings with each of their staff but it was not possible due to the nature of their work, so we conducted a group session. Six out of nine participants have direct social media responsibilities. These six participants monitor social media in their daily tasks and because the nature of their organizations they also monitor social media data extensively at the time of emergencies. Participants belonged to public sector organizations including CEMA, Calgary Police, Fire department, City of Calgary, and County of Frontenac. All participants had experienced dealing/monitoring social media data in at least one emergency event.

All study sessions were conducted on a laptop in places like the participants' workplace, in a coffeehouse near their workplace or remotely through Skype. Since we knew all participants (except one) from our previous interview studies, we did not conduct any pre-interview at this phase. Study sessions took between 90 to 120 minutes. After consenting to be in the study, we gave the participants an overview of different parts of the system. We also emphasized that participants should see the tool as proof of concept and do not expect the commercial stability. All features of SMART were demonstrated and explained for participants such as how to create events, how to create historic and live collectors. We also showed them how to create filtered events and how to train the filtered events for identifying situation awareness and user defined tags. We demoed the tool using historic events that we already had in the tool. Depending on working on a live or historic event, we either asked participants to create new events or work with the already created events. We asked participants to perform the following tasks as these are common tasks according to prior work (Marbouti and Maurer, 2016) and the fact that these tasks cover the core components of SMART:

T1. Give an overview of the situation: Participants are asked to find relevant messages and events
T2. Find the important topics that people are talking about
T3. Find the region most affected by the event
T4. Provide a report of the tweets in which people provided detail

We also asked participants to think aloud (Lewis and Rieman, 1993) during the study, we observed their use of the tool, took notes, and responded to their questions. We interviewed them after the study. To analyze the notes we followed a process inspired by grounded theory (Glaser and Strauss, 2009). Grounded theory involves iterative coding of concepts (open-coding) and finding patterns apparent in the text (axial coding) in order to form typologies.

### Emergency Scenarios

We evaluated the tool based on real emergency scenarios. In the event page of SMART, we had several emergency events. We demoed the tool usually based on familiar events for participants such as Alberta Floods and then proceeded with another one or two events upon the desire of participants. For 3 sessions with individuals, we focused on events such as the Fort McMurray fire event or Hurricane Harvey pretending these events are happening real time at the time of the study. For the group study, we found a real-time emergency that was happening in Queensland Australia; a few days before our scheduled meeting we noticed there was a tropical cyclone in Queensland Australia called Cyclone Debbie. We started gathering tweets around that event to get familiar with it. We decided to focus on that event during our study with CEMA. Cyclone Debbie started on March 23rd, 2017 and intensified to a named cyclone on March 26th. The cyclone made landfall near Airlie Beach, just north of Prosperine, on March 28th. The cyclone caused major damage. Many communities went out of power and phone reception. The group session took place on March 28th at the Calgary Emergency Management Agency (CEMA). During the study, some participants noticed a plane crash in Saskatchewan and started gathering data around that incident as well. Since both events were going on at the time of study we used real time data.

## Findings

In this section, we provide detail over analysts' feedback when performing the given tasks. The tasks were not performed in a specific order. We asked participants to assume that they are on a response team in the designated scenario and start performing tasks. We asked some participants how much these tasks match to what they are currently doing, one(P8) put it in this way: *"we do these tasks very unconsciously, first we do a situation report, participants are asked to find relevant messaging and events, and then what we do start developing messages [to push out to public]"*

### *Finding relevant keywords*

To get an overview of the situation participants started creating events and gathering tweets (T1). Participants mostly used live collectors; the historical search collectors were less used among the participants. On the utility of live and historical search collectors P7 expressed: *"I want to start following this thing , all of sudden 100 thousand hockey fans spill out of 17th avenue , so I need to monitor red mile tonight, so I start a live collector, for my keyword red mile right now, BUT maybe they were indicators yesterday or last week throughout the playoffs that this may happen, so then I would create a search [historical] collector, same keyword red mile , to go back two weeks to search for people talking about partying on 17th Ave."*

For real time events in the group session, participants often started with general trending keywords from the twitter website or news articles and then added more detailed keywords. For example, in the Cyclone Debbie event one participant started with general hashtags such as #cyclonedebbie (both live and search), #cyclone (live), #qld(live), #debbie(live) at 13:29 and then added a hashtag regarding one of the landfall locations: #Mackay(live) at 14:24. Another participant added #injuries at 13:46 to the historical collector.

One participant (P5) liked to be able to delete previously saved keywords with all their collected data. They were trying to gather data around some plane crash event in *Saskatchewan*. Initially they used *Saskatchewan* hashtag but then later they realized that the hashtag was not useful for this event and they were getting lots of unrelated data using that hashtag. Another participant (P4) also asked for the ability to edit collectors to reflect past changes. For example, they wanted to remove all previously collected data for a specific hashtag.

Many participants expressed the difficulty and challenges they are facing with finding appropriate hashtags and keywords to pull data. As mentioned by one participant, general hashtags and keywords are not enough for them to make sense of the situation specially at the early hours of an emergency event. Participants expressed excitement towards the **topic modeling** view, because it helped them find sub events and topics that people were discussing. They used topic modeling keywords as a basis to create filtered events. For example, in Cyclone Debbie event, one participant created a filtered event for collecting data regarding the evacuation with keyword "evac" which he noticed in the topic keywords detected by the topic modeling component. In the case of the Hurricane Harvey event, one participant added port Arthur as a filtered event since it was being detected by the topic component module. For the individual sessions based on historical events we added a set of collectors with general hashtags. In these sessions, all three participants usually looked at the general hashtags, then added additional collectors based on the topic modeling keywords. For example, P8 started from the Fort McMurray event page, saw a collector with hashtag #ymm with 500 tweets behind it, then moved to the keywords in topic modeling and said (as he was viewing the words in the topic view): *"OK donations, OK umm safety stay…or Hgwy 63 and then based on that I create a filter say Hgwy 63"*.

We also noticed three participants added some collectors and filtered events regardless of tweets' content that made common sense such as injuries with keyword "injuries", landfalls with keyword "landfall". Since the keywords were not based on the underlying tweets vocabulary, these filtered events were not very successful in gathering relevant data.

### *Need to facilitate reading tweets*

The filtered event view includes a tweet text box view, and they could also toggle between a map/analytics view. At the beginning of the study some participants (4) that mainly monitor social media as a part of their daily tasks had much less interaction with the map/charts and they would like to view the tweets in a more prominent way. Two participants asked if they could have a more prominent view of tweets and see more tweets in one view. One (P4) stated *"In terms of display I think it would be nice to without having to just go into one keyword group, to be able to see the tweets more prominently than all the charts and graphs on the side and to have more of a broader view of what's going on"*. However, as they were getting to know the location extraction

module their use of map view increased. Another participant made another point in this regard that labeling would be easier for them in a view with more tweets.

### *Sentiment in the Emergency Concept*

All participants expressed attention towards public sentiment during an emergency event both on their early interaction with the filtered event view and later when they were exploring to see what topics people are talking about. Four participants, sentiment was the first thing that caught their eyes when they entered a filtered event. They were also curious about the sentiment in the chart page. In cyclone Debbie, one participant t(p5) noticed the big chunk of negative and neutral sentiment in the pie chart for the first few hours. They saw that as some sort of instant feedback. Another participant(P2) in the emergency management agency added**:** "I would *see that on one of our big screens on top Iive [in the emergency operation center], continuously…anybody can check that out all the time to see what are we doing right, where are the gaps…it's like hey [Author's name], we are trending negatively on this, what's the problem, and somebody says what's killing you is the roads".* One participant suggested to click on a sentiment in the pie chart graph and see the keywords or tags changes at the bottom, specific to that negative sentiment. Regarding sentiment, another participant (P8) added that he only cares about sentiment to the degree it relates to the safety of people: *"In crisis we want to know what people thinking, yeah, but first we want to make sure that people are safe."* Another participant (P2) suggested to be able to search for sentiment under certain hashtags for things like *"fail, poor service, tax dollars…."* Mostly participants saw negative sentiment as a combination of feedback to their service and as request for help. One(P9) put it in this way when we asked how he perceived negative sentiment: *"by sentiment I meant something that I'm not happy with your response, but they are both, if somebody injured I expect that tweet to be red and if somebody says the road to my house is blocked I expect that tweet to be red"*

### *Search Box Utility*

The **search box** capability in the filtered event view had utility when exploring the filtered event (T1 &T2). One (P2) participant used it to search for injuries for the cyclone Debbie event and another used it to put a question mark in a filter to find questions and enquiries from people (See Figure 3).

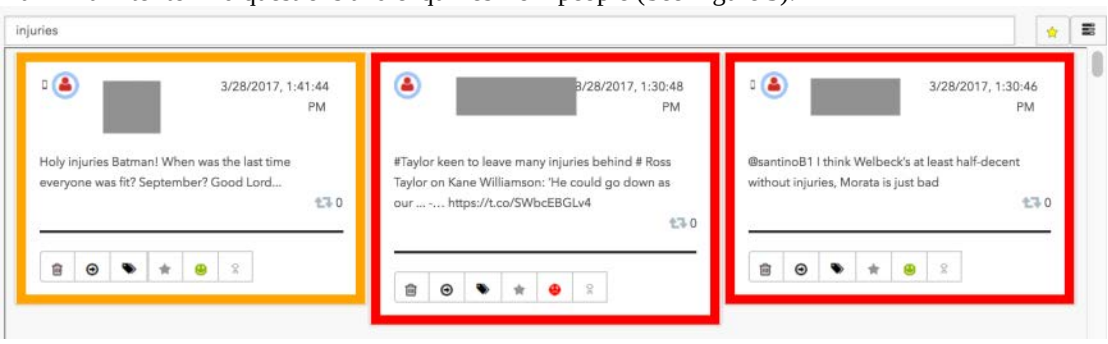


**Figure 3. This figure shows a view of search box within a filtered event**

### *Location extraction component utility*

To find the most impacted regions impacted by event (T3), participants considered **location extraction** useful. One participants mentioned the geolocation tweets on the map as the first thing they would like to view when entering a filtered event and the reason was: *"Cause this tells you where people are experiencing certain problems"* For example, in the Harvey Hurricane one participants noticed the geo-located tweets on the map about a sub region called Port Arthur (Figure 4). The location extraction component identified 5 tweets on that region in which people were asking for help (more specifically for boats). However, two participants had concerns toward the credibility of such information. When viewing the first tweets P8 was more cautious: *"It's pretty vague this tweet "boats needed in Port Arthur", I'd flag this we are getting people that need boats in Port Arthur, but where in Port Arthur, I'd respond there hey what do you need? we want to help you, how many people are you, where are you exactly, is anybody injured"* However, as he was viewing more tweets on that location on the map, he was getting a better picture of the problem. After viewing a few tweets on that location, he noticed two tweets that had addresses in their text:

> *323 Pinehurst Avenue, Port Arthur needs rescue of 40 animals & 2 women*

- *56 patients need to be evacuated from 4225 Lake Arthur Drive in Port Arthur (nursing home; nurses gone*

Afterwards, P8 mentioned *"ok we are getting six tweets from port Arthur area, …I'd respond and say ok hold tights…if there are injured people I'd say ok we need boats there maybe helicopter…"* Another participant(P9) also noticed these geo-located tweets from the Harvey Hurricane. These tweets were 2 from 34 tweets that were grouped by our location extraction component for that sub region in Port Arthur:

- *PLEASE HELP MY GREAT GRANDFATHER!!!340 West 17th St. Port Arthur,*
- *Children □□□ #sosharvey 500 Natchez Ave. PORT ARTHUR #cajunnavy @insuranceboy no phone injured adult #harveysos @CNN*

We asked him, what he would do when he sees such tweets, P9 said *"I probably want to check if that is already in the 911 system or in my call system… did they tag cnn? … next thing you know CNN is calling me and they are saying why did you go to 500 Natchez Ave?... if it turns out that is a very important call and I ignored it that's also not good!"* similar to the previous participant, he also expressed doubt toward the reliability of the information, he referred to the text in the tweet that says *"no phone injured adult"* and said *"that's pretty serious but you don't know if that's true!"* We asked P9 how he would verify such information, his response was: *"[with] basic stuff check username, check the number of previous tweets, profile photo, history, that stuff that I can do myself, I call superintendent that I communicate with, I'd say hey I hear that there's an injured adult have you heard about this?"*

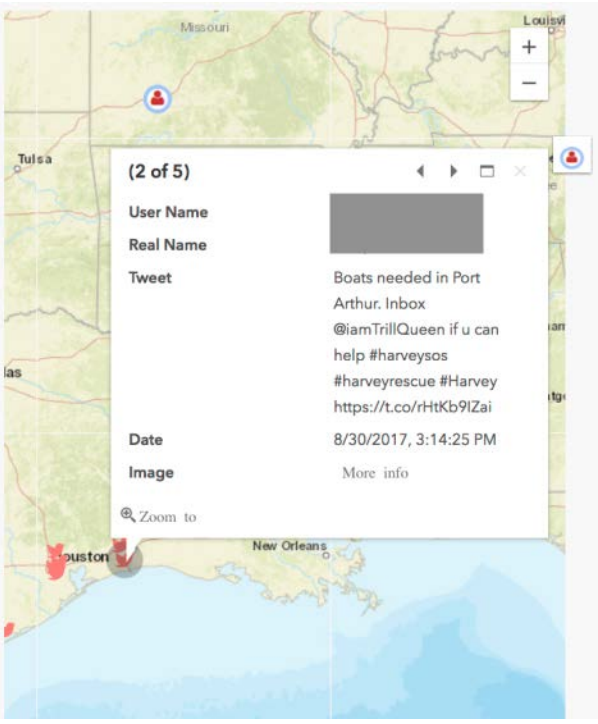


Figure 4. Geo-located Tweets in Hurricane Harvey Map View

### *Reliability of tweets' authors*

During the sessions we heard from two participants that despite having this information aggregated for a user, they still would like to go and check people's profiles because that gives them a better sense of reliability and because there still might be some flaws with the characteristics: *"For me sometime I'd like to go to the native profile cause I want to have a look to say volume of tweets, but if I go to their profile and see they tweet for every meal they had „…,then volume of tweets does not necessarily speak to the credibility of that user(P7)."* P7 also mentioned an example for fake accounts: *"for dealing with trolls and detractors that hate us, 90% of the time they just created that account in the last 24 hours to just hate the police, so I'd say ok I want to be able to search ok who's had a profile longer, but then some people may had a profile since 2007 and they never user it, so I really can't put value to that."* For these reasons, P7 still would like to go the users' profile and scroll down to say is this a normal twitter user or not.

*Manually populated filters*

To create a report of relevant messages (T4) the forwarding button to a manual populated filter has the most utility. However, due to the time limitation, few participants did that. One participant was curious about the forward button, when we explained using the forward button he could send a post to a manual stream he expressed interests toward that feature and mentioned a possible use case as P8 was thinking aloud: *"I see somebody wants help with the address, normally what I would do is just I go print it off and hand it to EM response managers, hey go check that out, … [so using forward button ]… I can create a manual stream and forward it to them? Ok I like that!"* Participants used manually populated filters to keep track of the urgency tweets and to be able to share them with the rest of the team.

*Training the classifiers*

To find more details regarding an event, participants were training the SMART classifiers**.** For simplicity, we called this training ability Artificial Intelligence (AI) in the study. The feedback towards the training was very positive. All six social media analysts' participants mentioned that they were willing to use this capability, as it would save them lots of time. *"We tag all of our incoming messages and it's 100% manual, so any starting point is an advantage…(P4)"*. One of them (P6) asked if we could add the same functionality behind the sentiment identification as well. *"I like the tool, how you filter and how you filter it over a large event, it's very positive very positive, to not have sort through those tweets manually, to be able to program and empower the AI, AI empowered around sentiment would be good too."*

One participant (P8) expressed that their emergency response tasks should be combined with training the classifiers. He went to the highway 63 filtered event page in Fort McMurray event and started labeling and viewing the tweets, After one round of training he noticed one with a red border color (which represented negative sentiment): *"So I'm gonna go here, ok this one here is red, what's going on here…[tweet text:]Lineup for gas here is at least 50 cars long. Traffic on the highway still moving slowly.#ymm,"* He also referred to another tweet that said: *"Go north if you're north of Gregoire. The fire has jumped the highway"* and mentioned that he would like to cross reference that with the information that he had. When participants saw such tweets that brought situation awareness for them usually they used the forward button to push the tweet to a manual stream so that they can refer to it later. However, there were some tweets that we might have considered as situation awareness but one participant was not interested in them. For example, on this tweet: "*Highway 63 is being blocked by police because the fire is on both sides of the highway. Please turn around & go north!"* P8 said*: "nothing there!"* Another participant had a different point of view regarding the combination of response and training the classifiers, he thought for this capability to work this should be separated from the response, and so there should be a dedicated person that monitors and trains the system. Although we did not explicitly ask the rest of participants regarding this combination, we noticed that most of them liked to perform their emergency response tasks (such as cross referencing the information, reply to the users) while they were training the classifiers.

When labeling the tweets in the filtered event view, one participant mentioned that they would automatically skip the green tweets (that means positive sentiment) and stop on the red ones (negative sentiment) that's why it's important to combine urgencies with the negative sentiment (see for Figure 2).

Two of the participants suggested simplifying the training for non-technical users. A participant (P2) that was an emergency manager expressed that he prefers a refresh button instead of a training button (The graduation hats in Figure 2-B). Another social media analyst participant mentioned that assigning tags should take the least amount of time, because what he's doing now is just scanning the tweets quickly to find the important ones so he might not have time in chaos of a disaster to click on the tagging button, assigning a tag and then click save.

On the other hand, one participant (P4) requested for more complex tags that could possibly help with their daily tasks. For example, P4 suggested tags such as identifying which business department within an organization a post belongs. *"I also don't think that tags are too complex, if we going to have more complex tags for example, see which business of city of Calgary that person is talking about …."*

We also asked participants explicitly which type of training (assigning a tag such as detail, media or labeling as SA or not SA) they would use. All of them liked to keep both options but some more comfortable with assigning tags as the tags made common sense and they could have users defined tags. Another reason was that tagging help them to get rid of irrelevant posts. An analyst in the police department gave us an example that once he was entering a keyword "Calgary COP", but he found tweets talking about Canada Olympic Park

(COP), he would have liked to tag that tweet as irrelevant and as a result the next incoming tweets about Canada Olympic Park (COP) would be tagged as irrelevant. However, one participant (P7) mentioned he would slightly prefer SA tagging as it will be faster to label and during chaos of an emergency they don't have much time: *"I use light[bulb] for the urgent [tweets]"*

One participant (P7) was interested toward the misinformation and sympathy tag types to be able to train the classifiers to identify tweets that are contributing to a rumour: *"Misinformation is pretty straightforward that might be an opportunity for us to provide education that we don't necessary have an answer for but if we do it we may have calmed the masses during an emergency and then sympathy is an interesting one, cause at first I thought I wouldn't really need that. Sympathy to me is more sentiment … it could be related to like compliments or something."*

Two of the participants mentioned that for a social media analysis tool to be useful in the times of a real emergency, the design has to be useful for their day to day tasks so that they have enough training and familiarly with the tool beforehand: *"When there's a real emergency you don't want to learn something new you have to know it works absolutely perfect, and you have to completely rely on it. If you do not practice with it a lot then you have to wonder, what tool am I gonna go with first (P9)"*

Finally, on training the classifiers P9 mentioned that he likes the control he has over training the system in our design because he did not have that control in social media platforms: *"[he was talking about Social Media platforms] what I found really annoying the tweets are not selected in the order they issued, I get to see posts that they think I want to see, I find that very annoying... and in Twitter and Instagram you can't even change it. you have no choice, they provide the posts in the order they want... but in your program, I control what AI can do"*

#### DISCUSSION

Overall the findings show the interest of participants into embedding machine learning classifiers for determining SA related and user defined tags. We found that to use such capabilities in real time scenarios the design of a monitoring tool should allow for the combination of labeling the tweets (training the classifier) and emergency response tasks. As requested by some participants (2) social media analysis tools should also support their daily tasks to be effective during a disaster. Analyst needs to be familiar with a tool before its use in an emergency. They need to be able to trust the classifiers capability in identifying relevant information before an emergency occurs. Participants suggested some interface enhancement to make the labeling faster and more intuitive; for example, one suggested to represent the urgency of tweets by a border color instead of sentiment or to combine urgency with negative sentiment. However, they saw the training capability as one piece in the whole pie. The interactive filtering allows them to get relevant information from their community faster. But getting informed is not enough, they want to influence how their responses are being perceived. They want to act based on the information that they are getting from interactive filters and they want to further explore what difference their actions are making in the social media data. One (P3) said: *"that AI piece exists but its only in its ability to further provide information …it's like now you have that data, how can you further mine it, filter based on negative, positive [sentiments] , now can you influence what is decided and fix those things and then also as we react, can [the tool] attract the difference that is making"* In other words, these analysts were not only interested to learn what has happened but to reflect what is going to happen using a social media monitoring tool. In addition to use training based filtering, analysts can still view the social media posts in the usual chronological order, so in this way they can get a sense of how well the underlying classifier is performing in terms of recall.

We designed SMART with focus on pulling and monitoring social media data. However, the participants expressed the importance of being engaged with people which means the ability to push out information and reply to people questions and concerns. They expected that a desirable tool should have both capabilities (pulling data and pushing out information) at the same time. As one stated in the first eight hours of any emergency they are pushing out information. We noticed most of the participants are going back and forth to the Twitter. They mentioned that they were doing the same with the commercial tools such as Hootsuite. This could be because Twitter website is still the easiest platform for them to work with but also because of lack of trust in the other tools that they wanted to verify the contents, and sometimes because the engagement options that Twitter can provide for them like replying, and exploring users' profiles. Some of them asked if they can have access to the native form of a tweet through a link from inside the SMART.

### LIMITATIONS AND FUTURE DIRECTIONS

Viewing sentiments of tweets both in charts and in filtered event view was a critical benefit for participants. What they perceived as negative sentiment was different from the general perception. Negative sentiment for them was a combination of urgency tweets and concerns and complains toward their response. Some (4) participants expressed interest towards having interactive filtering for identifying posts with negative sentiment. That kind of insight from the community provides them a way to see how their actions are perceived and help them coordinate their resources. This could be an interesting direction for future research. P6 imagined that by using such interactive filtering they could train the filters to identify negative sentiment.*" [P6 was referring to a tweet] 'why route to the foothills hospital is closed, come on you guys'… have we got somebody on that, oh we don't , we've got people doing xyz, but we have lots of patients coming out of this incident, we need those roads to be open, we need some way to say we're responding, and then suddenly you see a tweet come out [from their department] that says 'hey we are clearing the road, please retweet' people say it's awesome, we phone mayor and say hey can you retweet this for us, and then suddenly we shift twitter traffic from negative to positive"*

We noticed that participants try to pull the required information by querying general emergency related keywords such as injuries, landfall. One useful feature for a monitoring tool would be to find similar words with the underlying datasets and suggest them to the users.

SMART focuses on Twitter data since it is easily accessible and it is being widely used by the public in a crisis. However, two participants encouraged us to gather Facebook data as it helped them to know about public comments and concerns and quality of their service. One of them stated that the character limitation in Twitter makes it more challenging for people to express their views. Another advantage of Facebook is that people comment below a post so the conversation is more focused around a topic compared to Twitter where the conversation is very wide and diverse.

### CONCLUSION

In this paper, we discussed the user study that was conducted for a social media analysis tool. The goal was to evaluate effectiveness of the tool by gathering qualitative feedback from social media analysts based on real emergency scenarios. We designed study tasks in such a way to demonstrate if the tool would be effective during an emergency scenario, and we asked domain experts to participate in the evaluation. All our participants had several years of experience in the emergency field. Our findings revealed that identifying the most appropriate keywords, especially at the early hours of an event, is one of the challenging tasks for social media analysts. We also found that to combine human expertise with machine intelligence, the training tasks should fit practitioners' current workflow. According to our findings analysts are not only interested in finding relevant information using a social media analysis tool but further to perceive how their actions impact the public sentiment. Our results also show the importance of sentiment and the perception of emergency practitioners around negative sentiment. The findings highlight the need for identifying negative sentiment in the emergency concept for developing future applications.

### REFERENCES

F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, K. Tao, (2012) Semantics+ filtering+ search= twitcident. exploring information in social web streams, in: Proceedings of the 23rd ACM conference on Hypertext and social media, ACM, pp. 285-294.

C.C. Aggarwal, C. Zhai, (2012) Mining text data, Springer Science & Business Media.

D.M. Blei, A.Y. Ng, M.I. Jordan, (2003) Latent dirichlet allocation, the Journal of machine Learning research, 3 993-1022.

H. Bosch, D. Thom, F. Heimerl, E. Puttmann, S. Koch, R. Kruger, M. Worner, T. Ertl, (2013) Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering, Visualization and Computer Graphics, IEEE Transactions on, 19 2022-2031.

M.A. Cameron, R. Power, B. Robinson, J. Yin, (2012) Emergency situation awareness from twitter for crisis management, in: Proceedings of the 21st international conference companion on World Wide Web, ACM, Lyon, France, pp. 695-698.

J.M. Carroll, (2000) Making use: scenario-based design of human-computer interactions, MIT press.

A. Chokshi, (2015) Designing Social Media Tools for Emergency Response, Master's thesis, in: Computer Science, University Of Calgary.

L.L. Constantine, L.A. Lockwood, (1999) Software for use: a practical guide to the models and methods of usage-centered design, Pearson Education.

A. Flizikowski, W. Holubowicz, A. Stachowicz, L. Hokkanen, T.A. Kurki, N. Päivinen, T. Delavallade, (2014) Social media in crisis management-The iSAR+ project survey, In: Information Systems for Crisis Response and Management (ISCRAM)

B.G. Glaser, A.L. Strauss, (2009) The discovery of grounded theory: Strategies for qualitative research, Transaction Publishers.

S.R. Hiltz, L. Plotnick, (2013) Dealing with information overload when using social media for emergency management: emerging solutions, in: Proceedings of the 10th international ISCRAM conference, pp. 823-827.

A.L. Hughes, L. Palen, (2014) Social Media in Emergency Management: Academic Perspective, in: Critical Issues in Disaster Science and Management: A Dialogue Between Scientists and Emergency Managers., FEMA in Higher Education Program.

A.L. Hughes, L. Palen, (2009) Twitter adoption and use in mass convergence and emergency events, International Journal of Emergency Management, 6 248-260.

A.L. Hughes, R. Shah, (2016) Designing an Application for Social Media Needs in Emergency Public Information Work, in: Proceedings of the 19th International Conference on Supporting Group Work, ACM, Sanibel Island, Florida, USA, pp. 399-408.

M. Imran, C. Castillo, F. Diaz, S. Vieweg, (2015) Processing Social Media Messages in Mass Emergency: A Survey, ACM Computing Surveys (CSUR) 47:pp 67:1–67:38

M. Imran, C. Castillo, J. Lucas, M. Patrick, J. Rogstadius, (2014) Coordinating human and machine intelligence to classify microblog communications in crises, Proceedings of Information Systems for Crisis Response and Management (ISCRAM)

C. Lewis, J. Rieman, (1993) Task-Centred user Interface Design: A Practical Introduction, available on the web: ftp.cs.colorado.edu.

A.M. MacEachren, A. Jaiswal, A.C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, J. Blanford, (2011) SensePlace2: GeoTwitter analytics support for situational awareness, in: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, pp. 181-190.

M. Marbouti, F. Maurer, (2016) Social Media Use During Emergency Response–Insights from Emergency Professionals, In: Dwivedi Y. et al. (eds) Social Media: The Good, the Bad, and the Ugly. I3E 2016. Lecture Notes in Computer Science, vol 9844. Springer, Cham.

M. Marbouti, I. Mayor, D. Yim, F. Maurer, Social Media Analyst Responding Tool: A Visual Analytics Prototype to Identify Relevant Tweets in Emergency Events, In: Information Systems for Crisis Response and Management (ISCRAM)

M. Marcelo, P. Barbara, C. Carlos, (2010) Twitter under crisis: can we trust what we RT?, in: Proceedings of the First Workshop on Social Media Analytics, ACM, Washington D.C., District of Columbia, pp. 71-79.

K. Marina, P. Leysia, M.A. Kenneth, (2015) Think Local, Retweet Global: Retweeting by the Geographically-Vulnerable during Hurricane Sandy, in: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM, Vancouver, BC, Canada, pp. 981-993.

S.P. Moon, Y. Liu, S. Entezari, A. Pirzadeh, A. Pappas, M.S. Pfaff, (2013) Top health trends: An information visualization tool for awareness of local health trends, in: 10th International Conference on Information Systems for Crisis Response and Management, pp. 177-187.

L. Plotnick, S.R. Hiltz, J.A. Kushma, A. Tapia, (2015) Red Tape: Attitudes and Issues Related to Use of Social Media by US County-Level Emergency Managers, In: Information Systems for Crisis Response and Management (ISCRAM)

R. Power, B. Robinson, D. Ratcliffe, (2013) Finding fires with twitter, in: Proceedings of the Australasian Language Technology Association (ALTA) Workshop, Brisbane, Australia, pp. 80-89.

C. Reuter, C. Amelunxen, M. Moi, (2016a) Semi-automatic alerts and notifications for emergency services based on cross-platform social media data-evaluation of a prototype, Informatik 2016.

C. Reuter, M.-A. Kaufhold, T. Ludwig, (2017) End-User Development and Social Big Data – Towards Tailorable Situation Assessment with Social Media, in: F. Paternò, V. Wulf (Eds.) New Perspectives in End-User Development, Springer International Publishing, Cham, pp. 307-332.

C. Reuter, M.A. Kaufhold, (2018) Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics, Journal of Contingencies and Crisis Management, 26 41-57.

C. Reuter, T. Ludwig, M.-A. Kaufhold, T. Spielhofer, (2016b) Emergency services′ attitudes towards social media: A quantitative and qualitative survey across Europe, International Journal of Human-Computer Studies, 95 96-111.

J. Rogstadius, M. Vukovic, C.A. Teixeira, V. Kostakos, E. Karapanos, J.A. Laredo, (2013) CrisisTracker: crowdsourced social media curation for disaster awareness, *IBM Journal of Research and Development*, 57 4-4.

G. Salton, C. Buckley, (1988) Term-weighting approaches in automatic text retrieval, Information processing & management, 24 513-523.

A. Schulz, P. Ristoski, H. Paulheim, (2013) I See a Car Crash: Real-Time Detection of Small Scale Incidents in Microblogs, in: P. Cimiano, M. Fernández, V. Lopez, S. Schlobach, J. Völker (Eds.) The Semantic Web: ESWC 2013 Satellite Events, Springer Berlin Heidelberg, pp. 22-33.

S. Stieglitz, D. Bunker, M. Mirbabaie, C. Ehnis, (2017) Sense-making in social media during extreme events, Journal of Contingencies and Crisis Management.

J. Thornton, M. DeAngelus, B.A. Miller, (2016) Feedback-based social media filtering tool for improved situational awareness, in: 2016 IEEE Symposium on Technologies for Homeland Security (HST), pp. 1-6.

S. Verma, S. Vieweg, W.J. Corvey, L. Palen, J.H. Martin, M. Palmer, A. Schram, K.M. Anderson, (2011) Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency, In International AAAI Conference on Weblogs and Social Media

J. Yin, S. Karimi, B. Robinson, M. Cameron, (2012) ESA: emergency situation awareness via microbloggers, in: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, pp. 2701-2703.