Social Media Analyst Responding Tool: A Visual Analytics Prototype to Identify Relevant Tweets in Emergency Events

Mahshid Marbouti
University of Calgary
mmarbout@ucalgary.ca

Irene Mayor
University of Calgary
ihmayor@ucalgary.ca

Dianna Yim
University of Calgary
dyim@ucalgary.ca

Frank Maurer
University of Calgary
fmaurer@ucalgary.ca

ABSTRACT

Public and humanitarian organizations monitor social media to extract useful information during emergencies. In this paper, we propose a new method for identifying situation awareness (SA) tweets for emergencies. We take a human centered design approach to developing a visual analytics prototype, SMA-RT ("Social Media Analyst Responding Tool"), informed by social media analysts and emergency practitioners. Our design offers insights into the main requirements of social media monitoring tools used for emergency purposes. It also highlights the role that human and technology can play together in such solutions. We embed a machine learning classifier to identify SA tweets in a visual interactive tool. Our classifier aggregates textual, social, location, and tone based features to increase precision and recall of SA tweets.

Keywords

Situation Awareness, Social Media, Emergency Management.

INTRODUCTION

When an emergency happens, social media may provide information from the public that can contribute to SA of emergency operation centers (EOCs) (Hiltz and Plotnick, 2013). 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). This on-the-ground and updated information enhance an emergency response team's SA by helping them allocate resources and coordinate rescue actions (Imran, et al., 2014). In the tsunami of data from social media, only a few gems provide information that is operationally relevant. In addition, information often arrives at a high rate, making it difficult for social media analysts in EOCs to manually monitor, analyze, and filter such texts in time-critical emergencies (Hiltz, et al., 2014).

In response to these challenges, we have developed a visual analytics tool that would help identify SA tweets in emergency situations. In the academic literature, several efforts use machine learning classifiers to develop novel filters to categorize social media posts (Imran, et al., 2015) . These automated classification approaches use a quantitative evaluation in which they measure the accuracy of a classifier according to a given test dataset of emergency events. These studies demonstrate the applicability of classifiers and text mining techniques for automated social media classification (Hughes, et al., 2014). However, Applying these automated classifiers in another emergency can drop their accuracy because they may be over-fitted with the training data set (Power, et al., 2013). Therefore, analysts need filters that can learn from historical data and yet change according to analysts' needs in real-time. To facilitate such filters, we design SMA-RT in a way so that analysts would be able to train the machine learning classifier in real time.

Our approach combines human expertise as well as intelligence with machine learning methods through a set of visual analytics techniques. We develop interactive visual emergency-related filters, and use the microblog Twitter as an example of social media streams that can provide relevant information in emergencies. This research has the following contributions: (1) Develop a machine learning classifier to identify SA related tweets using textual, social, location and tone based features to increase precision and recall of SA tweets (2) Design and develop a visual analytics prototype informed by social media analysts in an iterative manner to improve task performance of analysts in the emergency situations.

RELATED WORK

In this section, we provide an overview of how data mining and visual analysis techniques have been used for identifying relevant tweet in emergency events. Some studies apply machine learning classification to identify different tweet types. *Emergency Situation Awareness (ESA)* (Yin, et al., 2012) distinguishes those tweets that are reporting an infrastructure damage using classification learning method based on NLP and social based features. Diakopoulos et al. (Diakopoulos, et al., 2012) developed a tool known as Seriously Rapid Source Review (SRSR) to enable journalists to assess sources around breaking news. The classification based studies demonstrated the applicability of learning models with text mining techniques to automate Microblog filtering (Hughes, et al., 2014). Various textual and social based features had been extracted from the social media posts to be able to classify this information into various categories. Verma et al. (Verma, et al., 2011) designed classifier models to distinguish tweets that are contributing to SA from tweets that are not. This study used text-mining and tone based features to identify tweets that represent SA. Rudra et al. (Rudra, et al., 2015) proposed a framework to extract a summary of SA information from tweets. The proposed framework combine and summarize tweets.

Another study proposed a visual classifier tool named *Scatterblogs* (Bosch, et al., 2013). 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, and then train and update the classifier model. In Thornton et al. (Thornton, et al., 2016), the feedback-based techniques were used to improve SA 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. In our work, we extend this approach. First, we bring insight to the design of such tools by interviewing social media analysts in the emergency domain. Next, we extract and aggregate different groups of features (social based, tone based, location based, and textual) to identify SA tweets. Finally, we embed the proposed classifier in a visual analytics tool to bring the human into the classification loop and provide feedback to the classifier.

THE DESIGN PROCESS

SMA-RT was iteratively developed using an agile process focused on social media analysts. Social media analysts are those who are monitoring and analyzing social media information. Depending on the organization, the terminology of what the role can be called might be different (e.g. digital communication officer, public information officer, or social media analyst). Regardless of naming, all those 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, rumours, etc.). Initially, we **interviewed** emergency practitioners and social media analysts in emergency domain to understand their challenges and requirements when monitoring social media (Marbouti and Maurer, 2016). One of their main requirements that we gathered from these people was that there is a need for the ability to identify posts that can contribute to their SA during an emergency event. 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 analyst's knowledge and machine learning algorithms. We produced **low-fidelity prototypes** and showed these to social media analysts to gather their reaction and comments. Using this feedback, we iterated on the design and implemented the core features in a fully functional prototype. Finally, we demoed the prototype to the same analysts to gather further feedback. In the remainder of this section, we describe SA classifier, then we describe how we reflected analysts' feedback in the SMA-RT development and provide an overview of the core elements of the prototype and, finally we share the initial feedback on the prototype that we gathered from social media analysts.

Situation Awareness Classifier

We developed and implemented a classifier for identifying situation awareness related tweets. Compared to conventional document classification problems (Sebastiani, 2002), tweet classification is more challenging because less information is available as input for the classification. tweets use non-standard vocabulary, abbreviations, and are short in length (Tuarob, et al., 2014). Because of the short length, they often do not provide sufficient context information. Spelling and grammar problems reduce the performance of standard automation techniques such as named entity extraction (Rogstadius, et al., 2013). In each classification problem, we need to represent our instances (i.e. tweets) as a set of features. Choosing the right features seems crucial when it comes to tweet classification and can improve the accuracy of a classifier model. We extract several features that contribute to deriving situation awareness information (see Table 3). The features are inspired by the related work (Landwehr and Carley, 2014), interviews, and manual exploration of emergency-related tweets. We group them into four categories:

(1) Textual features extract information from the message using text-mining techniques. Considering the short length of a microblog, use of abbreviations, and the ambiguity of natural language, using only these features would not be enough for determining an effective classifier (Tuarob, et al., 2014). The common text mining features are Bag of words (BoW) and TF-IDF. **(2) Tone** Based Features measure the tone of the tweet which can contribute in determine SA (Verma, et al., 2011). Assessing the emotional and social content in a tweet can help in assessing the situation on the ground. We use IBM Watson Tone Analyzer (part of the IBM Watson Developer Cloud toolchain) to extract different tone based features from the tweet text after removing the URLs. IBM Watson utilizes NLP and machine learning techniques to analyze and extract insights from vast unstructured volumes of text. In this section, we extract social and emotional tone analyzers. **Emotional tone** infers from different types of emotions and feelings that people express in their language. For each of these emotions, the Watson service generates a score between 0 and 1, which expresses the probability that the emotion came across in the text. **Social tone** measures the social tendencies in people's writing. Similar to emotional tone, the service outputs a score that lies between 0 and 1, which indicates tendency toward each social tone. **(3) Location** based features can determine the importance of a tweet in the emergency context (Schulz, et al., 2013). For example, the location of the author indicates if the author is a potential eyewitness or not. If the tweet is not geo-tagged, we investigate if the author uses any location words to describe a specific area within the emergency region. **(4) Social Media** Based features are related to Twitter specific characteristics. Along with other features they can contribute to predicting SA (Karimi, et al., 2013). Table 1 shows the detail features that we used for each feature category.

Table 1. List of features for identifying SA Related Tweets

Feature Category	Feature Name		Feature description
Textual Features	BoW + TFIDF		BoW (Aggarwal and Zhai, 2012) represents the text as a vector of words. TF-IDF (Salton and Buckley, 1988) measures Term frequency within a tweet whereas the frequency within all tweets. Each word in the bag of words feature is represented with its TF-IDF score.
Tone based Features[1]	Emotional Tone	Fear	"*A response to impending danger. It is a survival mechanism that is a reaction to some negative stimulus. It may be a mild caution or an extreme phobia.*"
		Sadness	*"Indicates a feeling of loss and disadvantage. When a person can be observed to be quiet, less energetic and withdrawn, it may be inferred that sadness exists."*
		Disgust	"*An emotional response of revulsion to something considered offensive or unpleasant. It is a sensation that refers to something revolting.*"
	Social Tone	Extraversion	*"The tendency to seek stimulation in the company of others*."
		Openness	*"The extent a person is open to experience a variety of activities."*
		Agreeableness	*"The tendency to be compassionate and cooperative towards others."*
Location Based Features	Tweet Distance		If the tweet is geo-tagged, we measure the distance of the tweet to the center of the crisis.
	User Distance		Since according to Twitter less than 1% of the tweets are geo-tagged, we also consider the registered location of the users as a measure of how close the user is to the center of the crisis. By using geographical APIs, we extract geographical coordinates for users if they have a valid place name in their profile.
	Location mentions		Since not all users have a valid and registered place in their account, we use tweet text as another source of extracting location information. We determine number of location mentions in the text.
Social Media Based Features	Number of Mentions		Tweets with important information usually mention authorities or government related accounts (Karimi, et al., 2013). In this feature, we measure the number of times an account has been mentioned in the tweet.
	Photo		Sometimes people provide photos of the scene when posting about it. In this feature, we measure whether a post contains photo or not.
	Mobile		Eyewitnesses at the scenes usually send SA posts with a mobile phone (Kumar, 2015). In this feature, we determine if a tweet has been sent using a mobile device or not.

[1]The definitions are based on Watson tone analyzer's documentation (https://www.ibm.com/watson/developercloud/doc/tone-analyzer/understanding-tone.shtml)

Location Extraction Process

One of the features in the SA classifier was determining if a tweet text contains location information or not. In this section, we describe the process in which we extract locations from a tweet text.

Box-based Location Extraction Process: In this process, we start with query for locations based on a bounding box around the incident area. Then, with these extracted locations, we check if any of them are contained within the tweets. Figure 1 demonstrates the process.

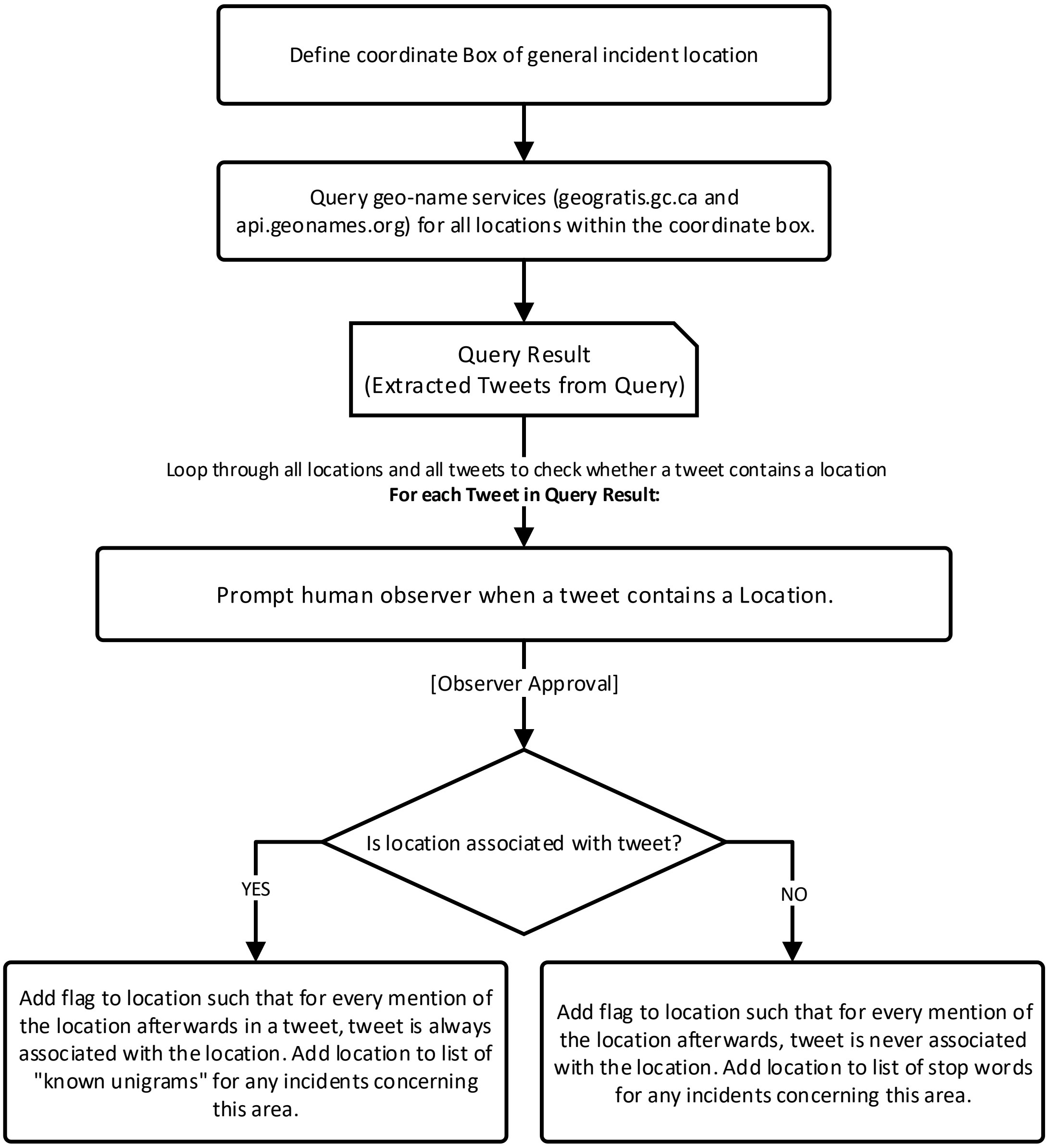

Figure 1 Box Based Location Extraction Process

Bigram Based Location Extraction Process: A second location extraction process has also been performed to support the first process in the likely cases that the first process does not cover all location mentions. This process is based on extracting bigrams for each tweet by using the NLTK python library. Figure 2 shows the process detail.

Step 1: Use stop words and "known unigrams" list from the Box-based process. Supplement "known unigrams" with additional Observer-chosen locations based upon researching the incident.

eg: "known unigrams": *{"ymm", "toronto", "fortmac", "yyc", "yea", "yed", "yec", "yeg", "highriver", "calgary", "lethbridge", "edmonton", "wetaskawi n", "ftmcmurray", "fortmacmurray", "anzac", "attawapiskat", "chipewyan", "mcmurray "}*

Step 2: Remove Links from the tweet: e.g. *"WATCH LIVE: Long says new evac centres opened up in Drayton Valley and St. Paul #ymmfire #abfire"*

Step 3: Tokenize the tweet: *"WATCH" "LIVE:" "Long" "says" "new" "evac" "centres" "opened" "up" "in" "Drayton" "Valley" "and" "St." "Paul" "#ymmfire" "#abfire"*

Step 4: Find Bigrams Tuples of each tweet: *"('WATCH', 'LIVE')","('LIVE', 'Long')", "('Long', 'says')","('says', 'new')","('new', 'evac')", "('evac', 'centres')","('centres', opened')", "('opened', 'Drayton')", ('Drayton', 'Valley')", "('Valley', 'St')", "('St', 'Paul')", "('Paul', '#ymmfire') ", "('#ymmfire', '#abfire')"*

Step 5: Remove special characters, ellipses and stop words. Set all to lower cases. *"('watch', 'live')","('live', 'long')","('long','')", ('',''), ('','evac'),"('evac', 'centres')","('centres', 'opened')", "('opened', 'drayton')", "('drayton', 'valley')", "('valley', 'st')", "('st', 'paul')", "('paul', 'ymmfire')", "('ymmfire', 'abfire')"*

Step 6: Remove all empty string items: *"('watch', 'live')", "('live', 'long')","('evac', 'centres')","('centres', 'opened')", "('opened', 'drayton')", "('drayton', 'valley')", "('valley', 'st')", "('st', 'paul')", "('paul', 'ymmfire')", "('ymmfire', 'abfire')"*

Step 7: Create all possible two word combinations: *"watch live", "live watch", "live long", "long live","evac centres","centres evac","centres opened", "opened centres","opened drayton","drayton opened", "drayton valley","valley drayton", "valley st", "st valley", "st paul", "paul st", "paul ymmfire","ymmfire paul", "ymmfire abfire", "abfire ymmfire"*

Step 8: check if any of the tokens from Step 3 are in list of "known unigrams"? None are associated with the example tweet in this case.

Step 9: Query geoname service for unigram/bigram in given Country (optional state/province). If a location is found, associate location with tweet. For example, the combination of *"drayton valley"* returned Drayton Valley Location, and *"St Paul"* returned St. Paul Location.

Figure 2. Bigram Based Location Extraction Process

We evaluate the location extraction process by applying it into three emergency events datasets and randomly sampled 1% from each. To get a more uniform distribution, the dataset that we created consisted of 50% that our location extractions process identified a location in and 50% that our process did not identified a location in. Table 2 shows the evaluation results of this process. The process shows a higher performance for centralized events. The hurricane Matthew event has the lowest precision out of our samples, as this was a more distributed event with several locations.

Table 2. Location extraction accuracy results based on 1% random sampling

Emergency event	Precision	Recall	Location List
Fort McMurray Fire	0.58	0.95	Fort McMurray
Alberta Floods	0.78	0.98	Edmonton, Calgary
Matthew Hurricane	0.50	0.96	Cuba, Haiti, South Carolina(USA), North Carolina(USA), Florida(USA), Virginia(USA), Georgia(USA)

Classifier Evaluation

In this section, we report our initial results of the SA classifier performance. We use Twitter Search API to obtain publicly available tweets during emergency events. We collect tweets for FortMacmurry Fire and Hurricane Matthew events using search terms that we chose through an initial investigation of public Twitter stream. We also purchased Alberta floods event dataset. We removed all retweets and duplicates in tweet texts as these do not provide additional information. We also removed the URLs from the tweet text. We randomly sampled 3% of the tweets from this set and performed a manual coding on whether a tweet is contributing to SA or not. Tweets were coded by 3 volunteers. For example, the following tweets coded as SA as they provide actionable insights regarding an event that can be useful for both affected public or emergency responders and help them make informed decisions.

Thickwood North under voluntary evacuation with 30 minutes notice expected for mandatory evacuation #ymmfire #fortmac

UPDATED: Forest fire has shifted away from homes. #ymm #ymmfire

On the other hand, the following tweets were code as not SA. These tweets may mention the event within their text but they did not provide any actionable information for emergency responders or affected public.

That situation in Alberta don't sound too good #ymm #ymmfire
Hope everyone in #ymm is safe this evening. Wish #yyc was a bit closer so we could offer some shelter. Thinking of you all! #ymmfire

Since we have heterogeneous features we used a Support Vector Machine (SVM) classifier with a linear kernel which is known to perform well on heterogeneous features (Kotsiantis, 2007). Performance was evaluated by measuring precision, recall and F1-score. We randomly split the data 80% for training and 20% for testing. Table 3 displays the classification results. It also displays the most effective features groups for each event. We plan to compare the performance of our classifier with the benchmark datasets.

Table 3. SA Classification results for three emergency events

Event Name	Search terms	Features used	Precision	Recall	F1-Score	Numebr of labeled tweets
FortMacmurry Fire 2016	#ymmfire, #FortMacfire, #ymm, #ymmHelps	TFIDF+Tone+location	0.83	0.80	0.81	2760
Hurricane Matthew 2016	Matthew, Hurricane, HurricaneMatthew	TFIDF+Tone+ location+ Social	0.89	0.86	0.87	2673
Alberta Floods 2013	This dataset was purchased from Gnip (www.Gnip.com)	TFIDF+Tone+ location	0.90	0.85	0.87	3171

SMA-RT development

This section describes the design of the SMA-RT application, including the various features that we have developed based on interviews with analysts. Figure 3 shows the core elements of our proposed design. The SMA-RT interface combines the ability to collect, search and sort tweets alongside the functionality to filter and organize them with machine learning techniques. In the following we describe the core parts of the design in more detail.

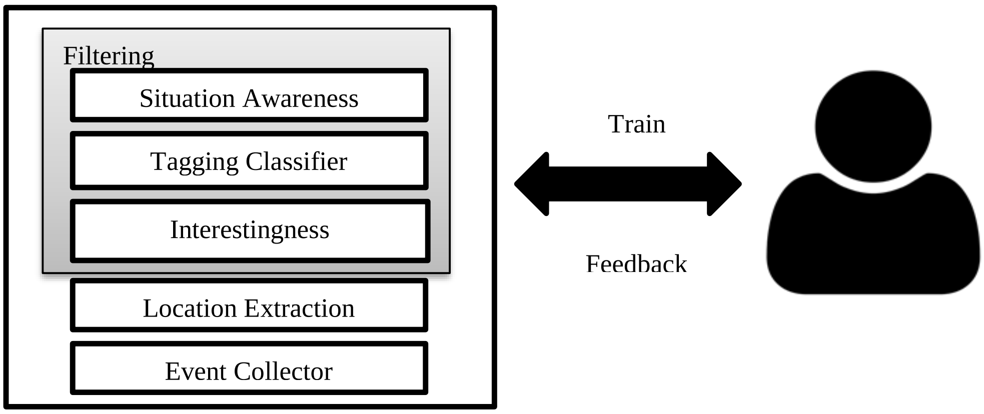

Figure 3. SMA-RT design overview

Event Collection: An essential feature of a social media monitoring tool is to be able to collect social media posts while an event is occurring. The collection component in SMA-RT partitions tweets into three types of collections: Event Collections, Filtered Event Collections and Manually Filtered Event Collections. Event Collections (Figure 4-A) contain the total tweets related to one large main event. Upon the start of using the prototype, a user must create an Event Collection with keyword-defined collectors (Figure 4-B) to collect tweets. Event Collection can then be further split into Filtered Event Collections and Manually Filtered Event Collections. As their names implies, these other two type of collections are tweet collections with further filters placed upon Event Collections. A Filtered Event (Figure 4-C) is defined by keyword filters and is used to view a small subset of the Event Collection. The SMA-RT can also extract "Topics" (Figure 4-D) using 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 certain keywords and can be added to a Filtered Event.

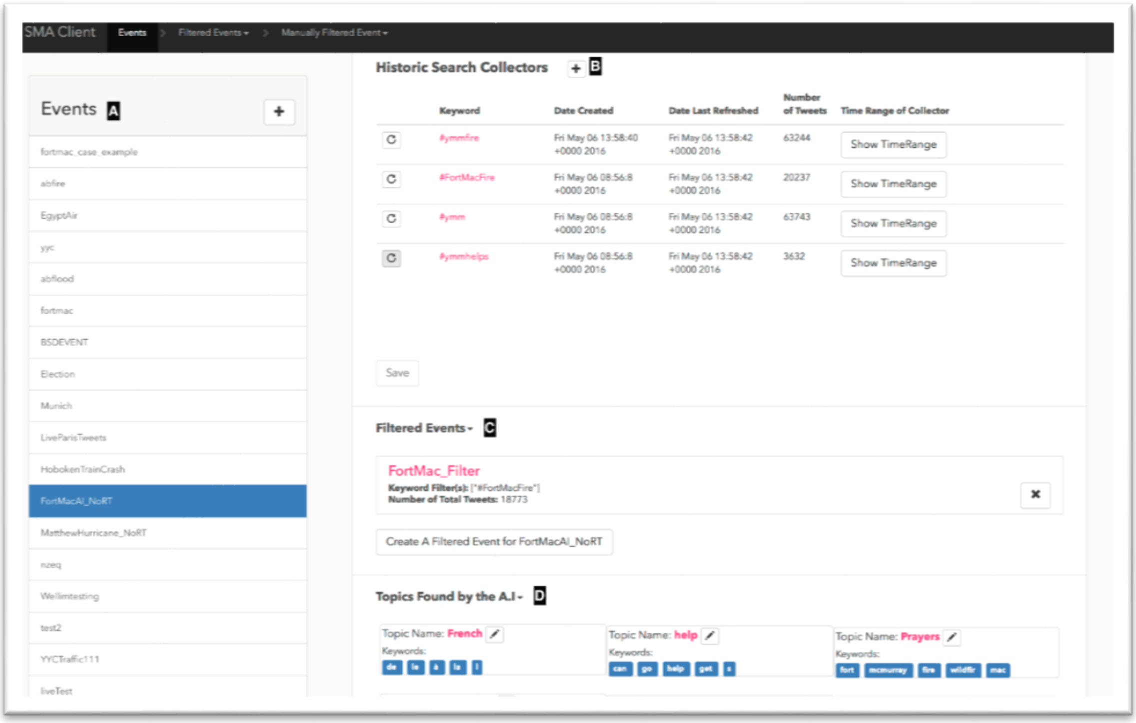

Figure 4. SMA-RT- Event Exploration. User interface for exploring events(A), create collectors to gather tweets from twitter API (B), create filtered events to filter through collected tweets (C), and extract topics from collected tweets.

Collaboration support: The pressure upon people within EOCs makes the analysis job difficult, as the responders are often local people who are also concerned for the safety of their friends and families. The participants who had experienced the Alberta floods in 2013 or the huge fire at Slave Lake in 2011 mentioned that the magnitude around those events were unprecedented. Responders may not sleep for several days and bear lots of pressure. As they stated, having a degree of separation and use the help of outsiders can be useful. For example, digital volunteers from outside the community can help in monitoring and identifying information from social media. That's why we designed web-based tool to be more accessible that supports **collaboration** between analysts and other roles in an EOC. Multiple analysts can view a filtered event and within each filtered event stream, an analyst can forward a specific post to another stream. A Manually Filtered Event allows users to manually select certain tweets from multiple events to save separately, or send off to the common operating picture of an emergency response system, such as the Emergency Operations Center of the Future, or otherwise known as EOC-F (Chan, et al., 2016). The arrow button in Figure 5-J let analyst move a tweet to either manual stream or a wall display in EOC-F.

Interactive Filtering: From interviews, we extracted that during an emergency, it is crucial that users get information quickly and accurately with easy-to-use filtering of tweets. By using machine learning classifiers to assist with this process, results can be gathered with minimal loss of time. Our design brings the human into the classification loop by letting users interact with the classifier. This solution will let users explore live streams of tweets, label incoming tweets, train the classifier and provide feedback regarding the information. It was observed that there is some resistance and mistrust for using automated filters; to address this, we ensured that our tool can work in a **manual** as well as a **semi-automated filtering mode** and allows analysts to distinguish between machine learning and human labels. Any automatically assigned label can also be changed by the analyst if necessary. We embedded three machine learning classifiers to facilitate interactive filtering.

1) Filtering by situational awareness: We embed aforementioned SA classifier in SMA-RT. In the interface, situational awareness is tracked by a lit or unlit lightbulb symbol for each tweet, allowing for people to make quicker decisions during a time critical situation. If the SA classifier predicts situational awareness, the lightbulb symbol will have a red background to allow analysts 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 (BoW+ TFIDF). As an example, an analyst manually tagging tweets will also train the classifier to tag similar tweets with the same tags, thus increasing the speed that information can be filtered. The automatically tagged tweets will be pink in color, and can be changed if the tag is deemed inaccurate by the user.

3) Filtering by interestingness: We can relate the interestingness of a tweet by its number of retweets (Webberley, et al., 2015). According to exploration of emergency and non-emergency tweets, retweeting increases in emergencies (Marina, et al., 2015). We attempt to identify an interestingness score in emergency context using this behavioral change. Interestingness compares the number of retweets to the predicted number of retweet counts as estimated by a classifier based on each user and tweet characteristics, and is displayed using either a yellow star for interesting, a half-yellow star for semi-interesting, and lastly a black star for not interesting (see Figure 5-J). By identifying interesting tweets, responders can ensure that no harmful rumors are circulating, and to quickly identify important details that many people have retweeted. In SMA-RT, we let users sort tweets based on their interestingness. We compare retweeting behavior between emergency and nonemergency Twitter datasets. First, we divide the tweets into different levels according to their retweet count. We grouped retweet counts in a way so that the number of tweets in each level is approximately equal. This is to prevent having small amounts of training data in less frequent levels. Inspired by (Webberley, et al., 2015) we predict that a tweet will fall into which level then we compare the predicted retweet count level for each tweet to its actual retweet count level. We train and test a predictive model (a decision tree classifier since it provided the best result among others) to predict retweet count level of a tweet based on randomly sampled data from Twitter API (Accuracy: 85.13%). For creating a predictive model, we used the following features: *Length of the text, Is the author account verified, Is there a mention in the text, Is the tweet a "reply", Is there a URL in the tweet, is there an exclamation mark in the tweet, Number of followers of the user, Number of user friends* The predictive model predicts that a tweet will fall into which retweet level. Then we apply the model on our emergency tweet dataset. As a result, for each tweet we have a predicted value and an actual value. Using (1) we infer an interestingness score for each tweet.

$$Interestingness = RT_{Actaul} - RT_{Predicted} \qquad (1)$$

Where RT_{actual} is the actual retweet count interval of a tweet and $RT_{Predicted}$ is the predicted retweet count interval of a tweet using a decision tree classifier

The automated filtering process heavily depends on the users playing an involved role to help the classifier learn these trends for each individual event. There are easily accessible buttons (Figure 5-F) which allow the user to train both tagging and situational awareness classifiers. However, this process is not yet continuous and the user is required to update the training of the system whenever applicable. Although inconvenient, this re-training allows the classifier to be retrained and corrected even when information becomes more important than before.

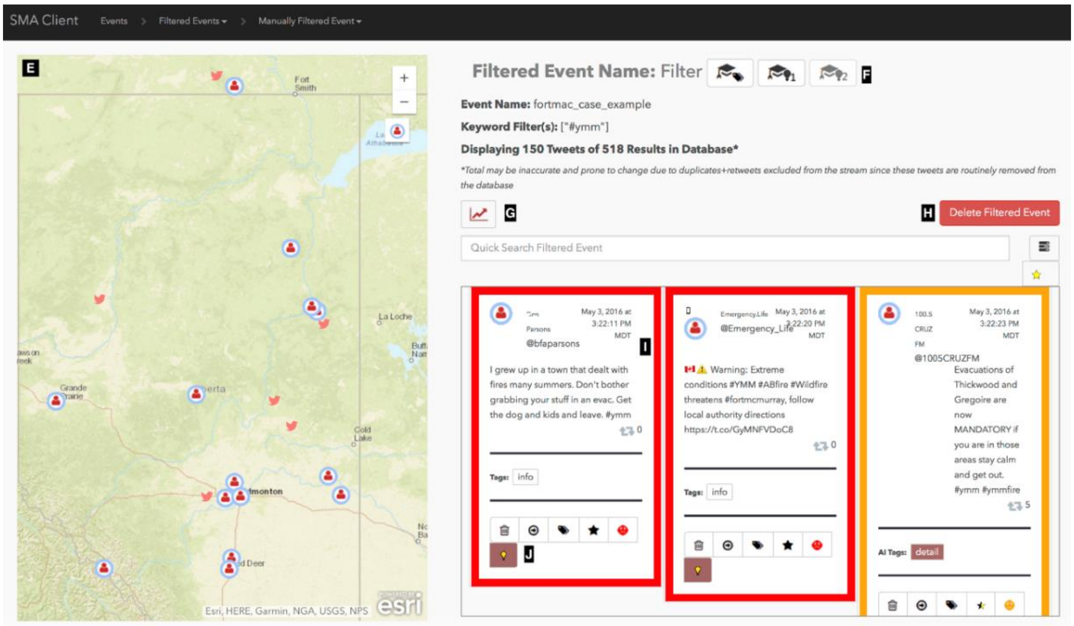

Figure 5. SMA-RT Filtered Event Page. User interface for exploration of a filtered event. Users can read and tag (J) tweets and train(F) the SA classifier iteratively.

Filtered event interface and interaction design: Simplicity in a social media tool's usage was one of the common features that the participants expected. This is especially important as depending on the emergency, personnel may decide to shift the non-trained resources to monitor social media; thus, it is important the tool would be as self-descriptive as possible. The interface was designed to be understood with minimal prior training using bold colors that standout again a clean layout. One of the participants commented that during an emergency, they usually prefer Hootsuite because of its simplicity as it displays posts in separate **stream views** that makes it easy for them to manage. Figure 5 shows a stream view, which balances displaying as much information as possible while keeping the user interface clean and easy to read. The map (Figure 5-E) shows information regarding the location of tweet, and the location of the user at the time of the tweet. If the SMA-RT extracts the location of the tweet, using the location mention extraction process, it will be red. The graduation hats (Figure 5-F) are used for training the classifier. The graph button (Figure 5-G) allows the map to be toggled off and display graphs for sentiment, interestingness, situational awareness and tags. The search bar and following buttons (Figure 5-H) can be used to change the tweets viewed within the Filtered Event. Lastly, the tweet (Figure 5-I) and the buttons below (Figure 5-J) allow quick interactions for classifying tweets and reading large amounts of tweets at once. The color of the border around each tweet displays the sentiment of individual tweets, allowing it to be easily seen.

Expert Feedback

In this section, we describe the initial feedback on SMA-RT. After developing the prototype, we conducted semi-structured interviews with 5 social media analysts we have interviewed during the design process. Our participants had at least one experience with monitoring social media in an emergency event. The goal was to understand how the system and its various elements would be helpful for social media analysts in time-critical emergency situations. At the beginning of each interview session we displayed the capabilities of SMA-RT, then proceeded to collect their feedback on the tool.

Semi-automated filters were the most controversial part of SMA-RT. There were different opinions regarding the usability of such filters. One concern that a participant mentioned was that how would junior analysts or those with limited emergency related experience know what the big picture is, such that they are able to judge what social media post is useful or not? She stated, *"Trusting people to identify information that is "important or not important" is discomforting"*. What circumstances would be worthy to set this up and use it? She stated that *"it is such a judgment call, dependent on the situation, team availability, the magnitude of the incident and whether they have the idea that the incident will "go on" long enough to benefit from the effort".* Another question was raised about how to change the filter to make previously unimportant information relevant again when the topic changes. However, participants related to public emergency organizations such as EOCs had a more positive feedback. These participants were dealing with information overload even in non-emergency situations so they realized how such filters could save them time in time-critical situations. Nevertheless, there was uncertainty towards the accuracy of such filters in the long term. One participant mentioned concerns about the possibility of the training tool misidentifying a post. He suggested that the system be "untrained" or corrected so that the same error (according to the analyst) would not appear again.

We initially designed the tool by focusing solely on the contents of the social media posts, but there was also a noticeable interest in the sources of these posts by the participants. For example, they would like to see aggregated information regarding a source with just one click or by hovering over a person icon. Such information like description, number of tweets, number of followers and followings would be helpful, they would also like to know if the person is linked to any other person in their network. They would be able to use all this information to infer reliability of a person, which is also lacking in their current tools. They also would like to pull posts based on specific users. Another addition would be to allow the ability to add a user to their "follow list" or list of reliable and unreliable users that they want to keep track of and perhaps use in future streams. Another participant suggested that they want be able to "silence" people so they don't keep populating your stream if this person is not worth "listening" to. One participant also suggested that if they could manually set the "reliability" for a user.

CONCLUSION AND FUTURE WORK

Social media has brought new opportunities and challenges in response to emergency events; the widespread usage of social media has made it difficult for analysts to extract their information needs from the data. Current commercial tools that are being used by analysts do not let users prioritize the social media posts based on their importance or the level of information. In this paper, we propose a machine learning classifier to identify SA tweets using textual, social, location, and tone based features to increase precision and recall of SA tweets. We take a human centered design approach to developing a visual analytics prototype (SMA-RT) to bring social media analysts in the classification loop and increase their task performance. The SMA-RT interface combines the ability to collect, search and sort tweets alongside the functionality to filter and organize them with machine learning techniques. SMA-RT allow analysts to filter tweets based on their situational awareness, tagging, and retweeting beyond expectation (interestingness). It also helps analysts to locate tweets if there is a location mention in the text. Initial feedback on the core parts of tool was positive.

We plan to conduct a usability study for the proposed visual analytics tool and thus evaluate the tool using social media analysts that have interviewed during the design process. The goal is to evaluate their task performance based on the amount of time these analysts will spend to gain the target information, how much information they will collect, how well they achieve the predefined goals behind each task, and lastly, how satisfied they will be with the tool.

REFERENCES

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

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

Imran, M., Castillo, C., Lucas, J., Patrick, M., Rogstadius, J. (2014) Coordinating human and machine intelligence to classify microblog communications in crises, *Proceedings of international ISCRAM conference*.

Hiltz, S.R., Kushma J. and Plotnick, L., (2014) Use of social media by US public sector emergency managers: Barriers and wish lists, *Proceedings of international ISCRAM conference*.

Imran, M., Castillo, C., Diaz, F. and Vieweg, S. (2015) Processing Social Media Messages in Mass Emergency: A Survey, *ACM Comput. Surv.*, 47, 1-38.

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

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

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

Diakopoulos, N., Choudhury, M.D. and Naaman, M. (2012) Finding and assessing social media information sources in the context of journalism, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM,* Austin, Texas, USA, pp. 2451-2460.

Verma, S., Vieweg, S., Corvey, W.J., Palen, L., Martin, J.H., Palmer, M., Schram, A. and Anderson, K.M. (2011) Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency, *Fifth International AAAI Conference on Weblogs and Social Media*, 385- 392.

Rudra, K., Ghosh, S., Ganguly, N., Goyal, P. and Ghosh, S. (2015) Extracting situational information from microblogs during disaster events: a classification-summarization approach, *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM*, pp. 583-592.

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

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

Marbouti, M. and Maurer, F. (2016) Social Media Use During Emergency Response–Insights from Emergency Professionals. 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

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

Sebastiani, F. (2002) Machine learning in automated text categorization, *ACM Comput. Surv.*, 34 1-47.

Tuarob, S., Tucker, C.S., Salathe, M. and Ram, N. (2014) An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages, *Journal of biomedical informatics*, 49 255-268.

Rogstadius, J., Vukovic, M., Teixeira, C., Kostakos, V., Karapanos, E. and Laredo, J. (2013) CrisisTracker: Crowdsourced social media curation for disaster awareness, *IBM Journal of Research and Development*, 57 4: 1-4: 13.

Landwehr, P.M. and Carley, K.M. (2014) Social Media in Disaster Relief, Data Mining and Knowledge Discovery for Big Data: Methodologies, Challenge and Opportunities, Springer Berlin Heidelberg, pp. 225-257.

Schulz, A., Ristoski, P., Paulheim, H. (2013) I See a Car Crash: Real-Time Detection of Small Scale Incidents in Microblogs, *Extended Semantic Web Conference. Springer Berlin Heidelberg*, pp. 22-33.

Karimi, S., Yin, J. and Paris, C. (2013) Classifying microblogs for disasters, *Proceedings of the 18th Australasian Document Computing Symposium, ACM,* Brisbane, Queensland, Australia, pp. 26-33.

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

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

Karimi, S., Yin, J. and Paris, C. (2013) Classifying microblogs for disasters, *Proceedings of the 18th Australasian Document Computing Symposium*, *ACM*, pp. 26-33.

Kumar, S. (2015) Social Media Analytics for Crisis Response, Doctoral dissertation, in Arizona State University.

Kotsiantis, S.B. (2007) Supervised machine learning: A review of classification techniques, Informatica 31, pp. 249-268.

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

Chan, E., Anslow, C., Seyed, T. and Maurer, F. (2016). Envisioning the Emergency Operations Centre of the Future. *Collaboration Meets Interactive Spaces* (pp. 349-372). Springer International Publishing.

Webberley, W.M., Allen, S.M., Whitaker, R.M. (2015) Retweeting beyond expectation: Inferring interestingness in Twitter, *Computer Communications* 0140-3664.