University of Calgary

PRISM Repository **https://prism.ucalgary.ca**

The Vault Open Theses and Dissertations

2018-05-23

Analyzing Twitter Data for Emergency Managemer

Marbouti, Mahshid

Marbouti, M. (2018). Analyzing Twitter Data for Emergency Management (Doctoral thesis, University of Calgary, Calgary, Canada). Retrieved from https://prism.ucalgary.ca. doi:10.11575/PRISM/31944 http://hdl.handle.net/1880/106677

Downloaded from PRISM Repository, University of Calgary

UNIVERSITY OF CALGARY

Analyzing Twitter Data for Emergency Management

by

Mahshid Marbouti

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMPUTER SCIENCE

CALGARY, ALBERTA

MAY, 2018

© Mahshid Marbouti 2018

Abstract

Social media is an important part of our lives. It is hard to ignore the role of social media in our everyday lives and during disastrous events. During emergencies, emergency personnel need to make strategic decisions in a short amount of time, coordinate actions and prioritize tasks. Social media can be a powerful source of information that comes directly from the public; it can reflect public sentiment, needs, and questions.

In this research, I performed an interview study to find the use cases and challenges that emergency-related organizations encounter when dealing with social media. The findings reveal the needs of practitioners for designing social media monitoring tools to help them find the information they need. One of the main challenges for practitioners is that commercial tools are not designed for emergency response, and academic approaches do not consider their requirements. This dissertation brings insight into the design of expert-informed machine learning solutions for identifying relevant information from social media by following a human-centered design approach. By actively being involved with emergency practitioners throughout three years, I designed, developed, and evaluated a social media monitoring tool for emergency response. The evaluation results show the effectiveness of bringing analysts into the classification loop to train and get feedback to machine learning classifiers. It also shows that analysts would like to combine the training tasks with their response tasks. Another aspect of this research is exploring the significance of various categories of features and machine learning algorithms and automatically identifying situational awareness information in different emergency event datasets. Results show that significant features vary across different events which indicates that training should happen during the event.

Keywords: Situational Awareness, Social Media, Machine Learning, User Study

Acknowledgements

I would like to deeply thank my supervisor, Dr. Frank Maurer for his great mentoring, guidance and support through my PhD journey. It has been a pleasure and an honor to pursue my PhD under his supervision. I learned a lot from his, knowledge and professionalism.

I would like to thank my other supervisory committee members, Dr. Behrouz Far and Dr. Wesley Willet for all their helpful feedback on my work. Their thoughtful comments and detailed feedback helped me a lot. I would like to thank my examiner committee, Dr. John Braun, and Dr. Mario Costa Sousa for their great feedback that increased the quality of my work.

This work would not have been possible without the help of my co-authors, Dianna Yim, and Irene Mayor who have helped me in the development of the proposed tool. I would also like to thank another co-author, Dr. Craig Anslow, for all his great help and timely feedback. I also want to thank Dr. Theodore D. Hellmann, who guided me especially when I was at the beginning of my PhD journey.

Last but not least, I would like to express my deepest love and gratitude to my family members. My husband, Majid, with his invaluable emotional and technical support. This could not have been possible without his help, kindness, and encouragements. I want to thank my mother, Malekeh, who has been there for me on each and every step of my life and who has always put me ahead of herself. I would also like to thank my brothers, Farshid and Farid, for all they have done for me and for being great brothers.

Dedication

I dedicate this thesis to my mother and to the memory of my father.

Table of Contents

Abstract ii

Acknowledgements iii

Dedication iv

Table of Contents v

List of Figures and Illustrations viii

List of Tables xi

List of Symbols, Abbreviations and Nomenclature xii

Chapter 1: Introduction 1
- **1.1 Research Scope** 4
- **1.2 Research Questions** 6
- **1.3 Research method** 8
- **1.4 Contributions** 9
- **1.5 Overview of Dissertation Chapters** 10

Chapter 2: Background and Related Works 11
- **2.1 Emergency Response: Main Concepts** 11
- **2.2 Emergency Events** 11
- **2.3 Emergency Organizations** 12
- **2.4 Social Media Analysts** 13
- **2.5 Situational Awareness** 13
- **2.6 Using Social Media for Emergency Management** 14
- **2.7 Requirement Elicitation Studies** 17
- **2.8 Microblog Classification in Emergency Domain** 18
- **2.9 Social Media Analysis tools** 27

Chapter 3: Requirement Elicitation 31
- **3.1 Method** 31
- **3.2 Results** 33
  - 3.2.1 Patterns of Use 33
  - 3.2.2 Types of Information 35
  - 3.2.3 Information Overload – The Need for Filtering and Categorization 36
  - 3.2.4 Degree of Separation 38
  - 3.2.5 Collaboration and Communication 38
  - 3.2.6 Reliability of Social Media 39
  - 3.2.7 Current Tools 40
- **3.3 Chapter Summary** 42

Chapter 4: Situational Awareness Classifier 44
- **4.1 Microblog Classification for Emergency Management** 44
- **4.2 Situational Awareness (SA) Classifier Features** 47
  - 4.2.1 Textual Features 48
  - 4.2.2 Tone based Features 48
  - 4.2.3 Location Based Features 49

4.2.4 Social Media Based Features 53
4.3 Test/Training Datasets 56
4.4 Feature Selection 57
4.5 Classifier Models 58
4.6 Analysis of Variance (AONVA) 60
4.7 Results 61
4.8 ROC Curve 67
4.9 Chapter Summary 69

Chapter 5: Design Process 70
5.1 Example Scenario 70
5.2 Feedback on Low Fidelity Prototype 77
5.3 Findings 78
5.3.1 Selecting Keywords vs Geographic Boxes 78
5.3.2 Posts View 80
5.3.3 Manual Mode 80
5.3.4 Ability to Make Internal Comments 80
5.3.5 Advanced Filtering 81
5.3.6 Source Reliability 83
5.4 Design Considerations 84
5.5 Chapter Summary 85

Chapter 6: Social Media Analyst Responding Tool (SMART) 86
6.1 SMART Prototype Design and UI 86
6.1.1 Event Collection 86
6.1.2 Interactive Filtering 87
6.1.3 Collaboration Support 90
6.1.4 Easy to Use Design 91
6.1.5 Location Extractor component 92
6.2 SMART Architecture 94
6.2.1 Data Storage 94
6.2.2 Tweet Collection 94
6.2.3 Web Application Components 96
6.2.4 Analytical Components 97
6.3 Case Study 97
6.3.1 Fort McMurray Fire 2016 98
6.3.2 Fort McMurray Event Creation 99
6.3.3 Extracted Topics and Filtered Event Creation 99
6.3.4 Classifier Training 100
6.4 Chapter Summary 104

Chapter 7: Evaluation Study 105
7.1 Evaluation Procedure 105
7.2 Participants 106
7.3 Emergency Scenarios 107
7.4 Findings 108
7.4.1 Finding Relevant Keywords 108
7.4.2 Need to Facilitate Reading Tweets 110
7.4.3 Sentiment in the Emergency Concept 111
7.4.4 High Use of Search Box 112
7.4.5 High Use of Location Extraction Component 112

7.4.6 Reliability of Tweets' Authors 114
7.4.7 Manually Populated Filters 115
7.4.8 Training the Classifiers 115
7.5 Discussion 119
7.6 Limitations and Future Directions 121
7.7 Chapter Summary 122
Chapter 8: Conclusion and Future Work 124
8.1 Dissertation Summary 124
8.2 Discussion and Future Directions 126
8.2.1 Bringing Human in the Loop 127
8.2.2 Public vs Private Sector 127
8.2.3 Strategies to Mitigate the Challenges 128
8.2.4 More Feature Categories 129
8.2.5 Long Term Evaluation Study 129
8.2.6 Various Geographical Regions and Influence of Demographic Factors 129
8.3 Chapter Summary 130
References 131
Appendix A- List of publications 138
Appendix B- Requirement Elicitation Interview Questions 140
Appendix C- Evaluation Study Tasks 141
Appendix D- Permission of Copy Right 142
Appendix E- Consent Form for Requirement Elicitation Study Participants 143
Appendix F- Consent Forms for Evaluation Study Participants 146

List of Figures and Illustrations

Figure 1. Emergency Management Cycle: This research targets Response phase. 5

Figure 2. Target users when using social media for Emergency Response: This research targets Social Media Analysts. 5

Figure 3. Different types of use of social media in emergencies: This research targets identifying situational awareness. 5

Figure 4. Research Method 9

Figure 5. Categorizing Tweets (A) Manual Mode (B) Automated Mode 46

Figure 6. Bigram Based Location Extraction Process. 52

Figure 7. Box Based Location Extraction Process 53

Figure 8. Classifier Process. 61

Figure 9. Retweet count group. 63

Figure 10. Cross Validation accuracy results for each event and each feature category. 64

Figure 11. Cross Validation F1 results for each event and each feature category. 64

Figure 12. Box plot for 10-fold cross validation accuracy scores for Fort McMurray dataset. Each box plot shows the results per feature category and per algorithm. 65

Figure 13. Box plot for 10-fold cross validation accuracy scores for Hurricane Matthew dataset. Each box plot shows the results per feature category and per algorithm. 65

Figure 14. Box plot for 10 fold cross validation accuracy scores for Alberta Flood dataset. Each box plot shows the results per feature category and per algorithm. 66

Figure 15. Box plot for 10 fold cross validation accuracy scores for Colorado Wildfire dataset. Each box plot shows the results per feature category and per algorithm. 66

Figure 16. ROC Curve for Fort McMurray using SVM. 68

Figure 17. ROC Curve for Fort McMurray using LR. 68

Figure 18. ROC Curve for Hurricane Matthew using SVM. 68

Figure 19. ROC Curve for Hurricane Matthew using LR. 68

Figure 20. ROC Curve for Alberta Flood using SVM. 68

Figure 21. ROC Curve for Alberta Flood using LR. 68

Figure 22. ROC Curve for Colorado Wildfire using SVM. 69

Figure 23. ROC Curve for Colorado Wildfire using LR. 69

Figure 24. Basic filtering view- By selecting a set of keywords, list of locations, a time range and clicking on the search button, John can create the main view. 71

Figure 25. Main view- Based on the basic filtering menu, a query will be sent to the Twitter API and will result in a stream of tweets in the Main View. 72

Figure 26. User characteristics for determining reliability and proximity. 73

Figure 27. Create Water rumor stream. 73

Figure 28. Water rumor advanced filter. 74

Figure 29. Feature list. 74

Figure 30. Water rumor stream view. 75

Figure 31. Create a detail stream. 76

Figure 32. Detail stream advanced filter. 76

Figure 33. Detail stream view. 77

Figure 34. A-shows a list of live and historical collectors for hurricane Harvey (the top image). B- a list of topics regarding the event (the bottom image). 87

Figure 35. Time range for each collector 87

Figure 36. SMART Filtered Event. 90

Figure 37. SMART provides three ways of filtering. (1) by situational awareness (light-bulb), (2) by tagging (tagging button) and (3) by interestingness (star button) 90

Figure 38. Advanced filtering menu 93

Figure 39. Geo-located Tweets in Hurricane Harvey Map View 93

Figure 40. SMART Architecture Overview. 95

Figure 41. The result of topic extractor module for Fort MacMurray fire event 100

Figure 42. A sample tweet that was automatically identified as situational awareness (the light bulb), and automatically tagged as detail by SMART. 101

Figure 43. Red bird icons on the map shows the locations that smoke has been reported in Tweets. 103

Figure 44. Number of tweets over time for smoke filter. 103

Figure 45. Number of tweets over time for Airport Filter. 103

Figure 46. This figure shows a view of search box within a filtered event 112

Figure 47. Geo-located Tweets in Hurricane Harvey Map View. 114

Figure 48. 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 47-B) are used for training the tagging and situational awareness. 117

List of Tables

Table 1. Summarize feature categories and labels in microblog classification. 26

Table 2. Requirement Elicitations user profiles. 32

Table 3. Summary of the requirement elicitation study findings. 43

Table 4. Sample tweets for emotional tone features. 49

Table 5. Sample tweets for social tone features. 49

Table 6. List of features for identifying SA Related Tweets. 55

Table 7. Labeled set size with number of SA vs not SA tweets along with the percentage of SA and not SA tweets in the sample set. 57

Table 8. Location extraction accuracy results based on 1% random sampling. 62

Table 9. 10-fold cross validation accuracy results for predicting retweet count interval. 63

Table 10. Sample tweets for each interestingness value. 63

Table 11. Pairwise comparison of statistical difference between feature categories and classifier algorithms. 67

Table 12. A summary of interview study findings. 85

Table 13. Sample of tweets identified with the location extraction module for the evacuation filter 100

Table 14. Summary of feedback on final prototypes. 123

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
EM	Emergency Management
SA	Situational awareness
SMA	Social Media Monitoring
CEMA	Calgary Emergency Management Agency
EOC	Emergency Operations Center
COP	Common Operating Picture
SVM	Support Vector Machines
KNN	K-Nearest Neighbour
LR	Logistic Regression

Chapter 1: Introduction

When an emergency occurs, responders need to gain situational awareness to work effectively (Alexander, 2014). Social media is considered a critical resource for an effective response because of its large user base, fast communication, and ease of use (Pipek, Liu, & Kerne, 2014). Social media has facilitated multi-channel and bi-directional conversations (Tapia & Moore, 2014). Data from social media can be used as one of the major sources from which emergency managers can draw information to help identify and prioritize new issues or provide more detail about known issues. However, extracting meaningful information from social media is challenging due to the sheer volume of data and its high noise to signal ratio.

Emergency responders use social media platforms to communicate with the public. Affected people also use social media to share information regarding the event. This shared information from the public can contribute to situational awareness of the Emergency Operations Centers (EOCs) team and impact their decision-making (Huiji, Barbier, & Goolsby, 2011). Affected people no longer only rely on traditional channels to get their information. For example, as one of the most used social media platforms, Twitter reported that during Hurricane Sandy in 2012, people sent more than 20 million tweets about the storm within six days (Marina, Leysia, & Kenneth, 2015). During the event, people actively broadcast different kinds of emergency-related information. A closer inspection of these tweets reveals that, in the hours after the event, local users provided on-the-ground information, including eyewitness descriptions, impact on the communities, requests for help, and expressions of fear. Users outside the emergency area may also contribute to social media platforms by sharing information about people's wishes, volunteering and donations, providing caution and advice, or post about people in the affected

areas. This up-to-date information enhances the disaster response team's situational awareness by helping them allocate resources and coordinate rescue actions. On the other hand, officials use social media to push out information and to be a part of the communication during a disaster. They provide updated information regarding an event and answer people's questions and concerns. They also use social media to dispel the spread of rumors and misinformation (Chokshi, 2015).

This adoption of social media in the emergency management process brings its own complications and challenges (A. L. Hughes & Palen, 2009). Not all of the shared information in the affected areas is related to the disaster event, and not all shared information is informative. In addition, information often arrives at a high rate which makes it difficult for social media analysts in the EOCs to manually filter through, monitor, and analyze such texts during a time-critical emergency (Hiltz & Plotnick, 2013). According to interviews with emergency practitioners, they are usually able to respond to a small portion of social media posts during an emergency event. However, the expectation of the public concerning official responders has changed. People expect authorities to be able to quickly catch urgent requests and inquiries using these new communication channels. However the authorities usually are not able to respond in a timely fashion and meet the expectation (Reuter, Ludwig, Kaufhold, & Spielhofer, 2016). Current tools used by emergency practitioners do not prioritize information based on their level of importance or situational awareness, instead only based on the chronological order of the information. Another drawback of current tools is that they have not been developed specifically for emergencies nor based on the requirements of real practitioners (Marbouti & Maurer, 2016; Reuter & Kaufhold, 2018). Another complication that social media has brought for emergency practitioners is reliability and trust issues (Marcelo, Barbara, & Carlos, 2010). They cannot act solely based on

social media information before assessing the reliability of the sources or cross-checking information.

In this research, I focused on the information overload challenge in the adoption of social media during emergencies. Numerous academic studies have been done to filter and categorize Microblog posts using natural language processing and machine learning techniques [8]. Supervised classification is one of the most widely used machine learning approaches that has been used for filtering and identifying relevant information from textual data [9]. While these studies demonstrate the applicability of classifiers and text mining techniques for automated Microblog classification, they are application-specific using traditional Natural Language Processing (NLP) features such as bag of words (BOW) (Aggarwal & Zhai, 2012) and TF_IDF (Salton & Buckley, 1988). However, short text classification is challenging due to non-standard vocabulary, abbreviations, and are short length (Tuarob, Tucker, Salathe, & Ram, 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 (Jakob Rogstadius et al., 2013). Another problem with current academic approaches that use machine learning for identifying relevant information from social media is that they do not consider the requirements of emergency practitioners (Hughes, Amanda, & Palen, 2014) and focus solely on the quantitative performance. My dissertation *brings insight into the design of expert-informed machine learning solutions for identifying relevant information from social media* by following a human-centered design approach (Constantine & Lockwood, 1999). I designed and developed a social media monitoring tool based on the requirements of emergency practitioners. In my research, I also explored the *significance of various categories of features* for automatically

identifying situational awareness information using feature selection and supervised machine learning techniques.

1.1 Research Scope

This research lies at the intersection of social media and emergency management. Emergency management cycle has four basic phases (Figure 1) (Walle et al., 2009): Mitigation, Preparedness, Response, and Recovery. *Mitigation* refers to identifying and reducing risks which happen before an emergency occurs. *Preparedness* involves preparation actions to help authorities and the public to adequately respond to an incident. *Preparedness* also happens before an emergency event. The next two phases happen after an event occurs. The *Response* phase includes actions taken immediately after an event unfolds, and the *Recovery* phase includes actions necessary to get back to the normal situation.

Different users might use social media during emergencies. These users can be citizens or authorities. Citizen includes journalists, digital volunteers, affected people, or outsiders. Authorities' roles include roles such as operational, liaison officers, or communication roles (Figure 2) (Reuter & Kaufhold, 2018). Use of social media during emergencies brings several opportunities (Figure 3). For example, social media data can be used to identify events, data from social media can be summarized to help get an overview of the situation, emergency personnel can use social media data to help them decide how to coordinate resources, and social media can help authorities communicate important messages out to the public. Although it does not cover all members of a population, social media platforms have a considerable number of users. Another use of social media is to identify relevant and possibly important messages from the sheer volume of social media data. This relevant information varies from small details regarding the event or urgency requests for help (Reuter & Kaufhold, 2018).

This research focuses on the *Response* phase of the emergency management cycle and targets *Communication* people who monitor social media (i.e., *Social Media Analysts*) to support them *identify relevant information* or more specifically *situational awareness* information.

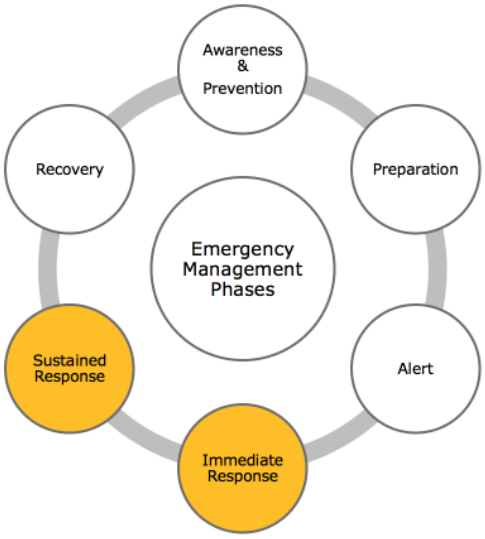


Figure 1. Emergency Management Cycle: This research targets Response phase.

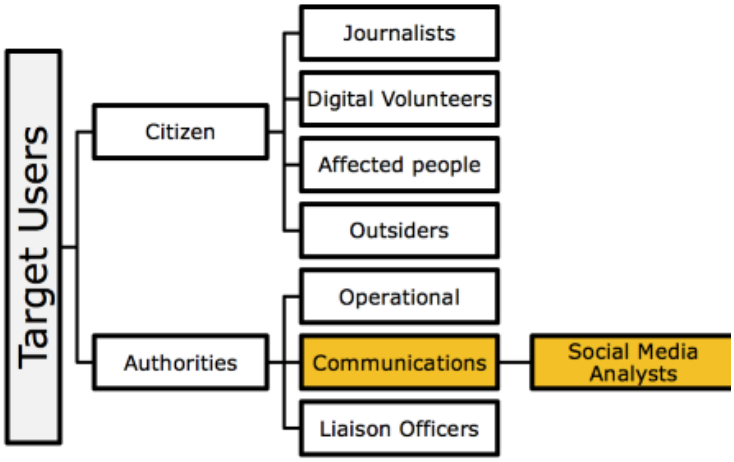


Figure 2. Target users when using social media for Emergency Response: This research targets Social Media Analysts.

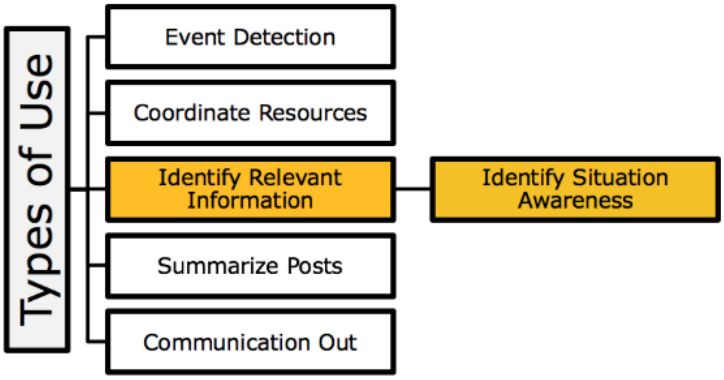


Figure 3. Different types of use of social media in emergencies: This research targets identifying situational awareness.

1.2 Research Questions

An objective of this research is to investigate the characteristics of emergency-related social media posts and how a combination of human expertise and machine learning techniques can improve the situational awareness of emergency responders based on social media.

As Twitter is one of the most widely used real-time microblog services with free accessible data (Stieglitz, Bunker, Mirbabaie, & Ehnis, 2017), I used Twitter as an example of social media streams that can provide relevant information during emergencies. Twitter facilitates the spread of user-generated content. The messages on Twitter, called tweets, can contain user mentions, hashtags, and links to external sources, photos or videos. My approach can likely be applied to any other microblog data as well. With this research, I want to help emergency responders effectively use streaming Twitter data to improve their situational awareness during an emergency event.

One of the big challenges of working with social media during a time-critical emergency is information overload (Hiltz & Plotnick, 2013). Emergency personnel need to act quickly and they need as much information as possible. However, thousands of tweets spreading in a short time can be quite overwhelming. For emergency communication teams, a primary challenge is to be able to find relevant information and filter out the noise. Relevant information could be about people, properties, or the environment that the event includes. Being able to find relevant tweets accurately in a short amount of time is a challenge that I am going to address in this research. In this research, I seek to answer the following research questions:

RQ1: What are the requirements of practitioners and emergency responders for using social media during emergencies?

I conducted sixteen interviews with emergency practitioners to identify their main requirements. The participants were from a variety of roles within the emergency management field. I discussed

what uses social media has during an emergency, how social media was useful for them, and what challenges they encountered regarding social media, and how they engage with social media using current commercial tools. The interview results confirmed that in the sheer volume of data on social media, only a small amount of the shared information in the affected areas is related to the emergency event. Even in the related ones, a small percentage provide situational awareness information for the emergency response teams. According to the interviews, some emergency professionals encounter thousands of messages every day and this quantity could be much larger during an emergency. It takes a substantial amount of time and effort for emergency professionals to sort through the enormous incoming data, and they are currently able to respond to only a small portion of the messages (Marbouti & Maurer, 2016).

RQ2: How can we identify relevant tweets that help analysts achieve their tasks?

Developing techniques to extract useful information from social media is beneficial. To identify relevant tweets, I applied supervised machine learning techniques. Supervised classification is one of the most common machine learning approaches that has been used for filtering and identifying relevant information from textual data (Hughes et al., 2014). The analysis of tweets is hard due to the short, unstructured feature of the texts (Hiltz & Plotnick, 2013). In this research, I developed innovative filters using supervised machine learning techniques to help analysts view tweets based on information relevancy other than of the chronological order of the tweets. I explored various features such as textual, social, location and tone-based features and developed a machine learning filter. The goal is to find the significant features and classifier algorithms that lead to the best accuracy.

RQ3: What are the design considerations of social media monitoring tools for emergencies?

As the current tools being used by emergency practitioners are not designed based on their requirements, one of the important objectives of this research is bringing insight into the design of social media monitoring tools based on practitioners' feedback. To help analysts find relevant information using automated filters and at the same time find the interest of analyst according to the contexts of the real-time emergency event, I iteratively designed, developed, and evaluated an interactive visual analytics tool. The design process started with low-fidelity paper prototypes and evolved into a fully functional prototype that can be tested in real emergency scenarios. The tool interactively involves social media analysts in determining the filters according to their tasks. The dynamic adaptation of filters allows social media analysts to find their informational needs more efficiently.

1.3 Research method

In this research, I followed the process represented in Figure 4. Initially, I identified and interviewed emergency practitioners and social media analysts in the emergency domain to understand their challenges and requirements when dealing with social media. After the initial interview phase, I pursued a scenario-based design approach (Carroll, 2000) to consider the variety of contexts in which the tool might be used and shape the feature set. I 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 tweets that provide situational awareness by combining textual, social, location, and tone-based features to increase precision and recall of situational awareness tweets (Marbouti, Mayor, Yim, & Maurer). The design process started with producing low-fidelity paper prototypes. I showed these prototypes to social media analysts to capture their reaction and comments. Using the feedback, I evolved the prototype to a

fully functional web-based prototype. The prototype involves analysts in determining the filters according to their tasks and help them identify relevant tweets. Finally, to bring insight into the design process of social media monitoring tools, I conducted a user study with emergency practitioners on the final prototype.

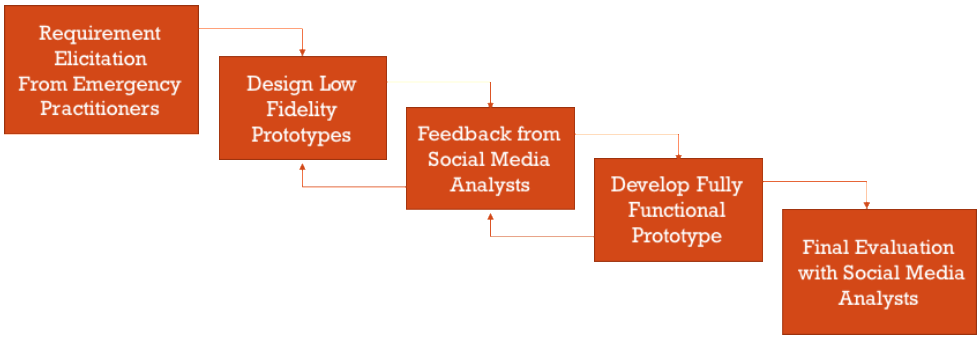


Figure 4. Research Method

1.4 Contributions

This dissertation research provides the following contributions:

(1) Presents the result of a requirement elicitation from emergency practitioners in regard to using social media. I identified challenges that emergency practitioners are facing when trying to extract information from social media, and what kinds of information types they need.

(2) Creates machine-learning filters that combine three classes of features in addition to text features. These classes are social-based, tone-based and location-based features. Social based features are based on Twitter specific characteristics such as how much a tweet is getting retweeted more beyond expectation. Tone-based features measure emotional and social tone of a tweet. Location-based features determine the author and tweet geolocation if it exists, their distance with the emergency area, and if there is a location mentioned in the tweet text. This research compares

different features and classification algorithms to determine their significance in identifying situational awareness.

(3) Designs and develops a social media monitoring tool that combines human expertise with supervised machine learning techniques based on the requirements of real practitioners. The tool design includes unique features to bring human expertise in the classification loop.

(4) Conducts an evaluation study on the proposed tool with domain experts to bring knowledge into the development of social media analysis tools that embeds machine learning. To the best of my knowledge, this is the first evaluation study to explore the concept of bringing human in the loop in the design of social media analysis tools for emergency response.

1.5 Overview of Dissertation Chapters

This dissertation is organized as follows: Chapter 2 covers background and related work for this research in the areas of emergency response and machine learning. Chapter 3 reports a requirement analysis study that presents the findings of interviews with emergency practitioners to understand their uses and challenges regarding social media during emergency events. Chapter 4 describes the supervised machine learning techniques to identify relevant information during emergency events. Chapter 5 covers the design process, and the evaluation findings based on the practitioners' feedback on the early social media analysis prototypes. Chapter 5 explains the final prototype developed for analyzing and monitoring social media data. It also provides detail regarding the architecture and implementation of the proposed tool. Chapter 7 evaluates the prototype by gathering qualitative feedback from emergency practitioners. Finally, Chapter 8 summarizes research findings and sheds light on the lessons learned and future directions.

Chapter 2: Background and Related Works

In this chapter, I provide a background of the essential concepts of emergency response and a review of related works that use social media for emergency management. I review existing literature that discusses the use of social media for emergency management, conduct requirement elicitation studies for using social media in emergency management, apply microblog classification for information filtering in the emergency domain, and propose social media analysis tools for emergency response.

2.1 Emergency Response: Main Concepts

The aim of this research is to increase situational awareness of social media analysts during emergencies. Therefore, it is important to define what I mean by emergencies, what are different phases in the emergency management cycle, what are different emergency organizations, who are social media analysts, and what is situational awareness in the emergency context.

2.2 Emergency Events

To have a common understanding of emergency response, I provide a short description of what an ***emergency event*** means. These include everyday emergencies such as traffic accidents and house fires. A disaster is a serious disruption to the functioning of a society. A catastrophe is a disaster with a much more severe and widespread level of disruption; in a catastrophe most houses are unusable, schools are shut down, infrastructure is disrupted, local organizations cannot operate normally and need the help of outsiders (Walle et al., 2009). Often when authors refer to emergency management in literature, they do not mean small-scale emergencies, but rather they refer to disasters and catastrophes. During disasters and catastrophes, the scope of the event dramatically increases the amount of information in social media and causes an overload for the

emergency responders. In this dissertation, an emergency event means disaster and catastrophes that create crisis situations and need emergency managers to save human lives, infrastructure, and environment. Such kinds of emergency events dramatically increase information overload.

Turoff et al. (Turoff et al., 2009) identify four basic phases for emergency management cycle: Mitigation, Preparedness, Response, and Recovery. *Mitigation* refers to identifying and reducing risks which happen before an emergency. *Preparedness* phase involves preparation actions to help authorities and the public to respond to an emergency event adequately. Preparedness also occurs before an emergency event. The next two phases (*Response* and *Recovery*) happen after an emergency event. The *Response* phase happens immediately before a foreseen event, and during and after the event. *Response* phase includes actions that help saving people, properties, and environment. Finally, *Recovery* phase is about helping affected people return to their normal lives.

2.3 Emergency Organizations

Van De Walle et al. (Walle et al., 2009) views four types of organizations related to emergencies. *(1) Established Organizations* that are engaged in response activities and their organizational structure is the same during, before and after the event. Emergency Operations Centers (EOCs) are an example of these types of organizations. An EOC is a command and control operation center that is triggered when an emergency is happening. EOCs are responsible for carrying out the principles of emergency preparedness and emergency management, and they usually act as an integration hub for related involved organizations (Chokshi, 2015). *(2) Expanding Organizations* that are routinely engaged in response, but they must expand to do so and normally recruit volunteers; Red Cross is an example of such an expanding organization. *(3) Extending Organizations* that do not normally get involved in emergency response. However, depending on the emergency event scale, extending organizations can participate using their existing

organizational structure. For example, a church community providing meal services during a flood would be an example of an extending organization. *(4) Emergent Organizations* are the ones that are not usually participating in response and need a new organizational structure. A community group formed to collect donations would be an emergent organization (Walle et al., 2009).

2.4 Social Media Analysts

As part of this research, I designed a tool for those who are monitoring and analyzing social media information during emergencies. Depending on the emergency organization, the terminology to of the role and position can be different. Digital communication officer, public information officer, or social media analyst are examples of social media analyst roles. Regardless of the name, 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, and rumors). In this study, I refer to this role as *social media analyst* or sometimes as *analyst*.

2.5 Situational Awareness

Situational awareness is "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995). This definition suggests that gaining situational awareness requires humans to be able to identify and comprehend an appropriate set of aspects of an event. More simply, situational awareness is knowing what is going on around you. Gaining situational awareness is critical for dealing with emergencies as it helps emergency managers make better decisions during a time-critical situation (Hiltz & Plotnick, 2013). In the context of this research, I define situational awareness as any actionable information that helps emergency responders make better-informed decisions.

2.6 Using Social Media for Emergency Management

Many recent studies discuss the importance and rise in the usage of social media. Imran et al. (Imran, Castillo, Diaz, & Vieweg, 2015) survey different studies for processing social media in mass emergencies. A group of studies discuss social media usage by exploring datasets of social media posts across various emergencies (Dashti et al., 2014; P. M. Landwehr & K. M. Carley, 2014; St Denis, Palen, & Anderson, 2014). Exploring emergencies shows the importance of social media in extracting on the ground information and the social media adoption by people during the emergencies (A. L. Hughes & Palen, 2009). Takahashi et al. (Takahashi, Tandoc, & Carmichael, 2015) explore how ordinary people and organizations use social media and what factors impact this use by analyzing tweets during Typhoon Haiyan in the Philippines. An analysis of tweets from Hurricane Sandy demonstrates that retweet activity increased during the event (Marina et al., 2015).

Stieglitz et al. (Stieglitz et al., 2017) analyzed three emergency related datasets to find patterns regarding commentary related information such as expression of feelings, reporting, and trying to make sense of an event. In their analysis, they considered factors such as sentiment, number of retweets, mentions, and URLs. They also conduct a social network analysis of the datasets colour coded on groups of users. Their goal was to compare the similarity of events in regard to these sentiment and network based aspects.

Another group of studies analyze rumors in Twitter datasets to extract and analyze propagation patterns and the content of the rumor related tweets (Marcelo et al., 2010; Oh, Kwon, & Rao, 2010; Starbird, Maddock, Orand, Achterman, & Mason, 2014).

Visual Analytics techniques have been used for designing tools for emergency response (Mazumdar, Lanfranchi, Ireson, & Ciravegna, 2014). More specifically, Visual analysis of Social

media can provide a rich mechanism for extracting useful information before, during, and after emergencies. Before an emergency, social media data can be monitored and analysed to identify risks (Anson, Watson, Wadhwa, & Metz, 2017; Dou, Wang, Skau, Ribarsky, & Zhou, 2012) and help emergency organizations consider establishing and creating their social media connections in the community and be able to create trust, know the influencers in the long run and thus be able to act better in the short amount of time they have in response phase (Anson et al., 2017). During an emergency, small-scale incidents and help requests can be identified (Yin, Karimi, Robinson, & Cameron, 2012). The visual analytics field uses interactive visual interfaces to help users search and filter, detect anomalies, events, or topics from microblogs in a real-time manner. Another application is to extract public sentiment about an issue, which is useful for decision makers in EOCs (Caragea, Squicciarini, Stehle, Neppalli, & Tapia, 2014; Dewan, Gupta, Goyal, & Kumaraguru, 2013).

Another group of studies cover different technological aspects of gathering, integrating, and visualizing emergency related information from various sources including social media, other online sources, maps, or satellite images. For example, Albert-Green et al. (Albert‐Green, John Braun, Martell, & Woolford, 2014), present exploratory data analysis tools to model the Fire Weather Index which can lead to decision support tools for fire management agencies. Crisis mapping studies allow the public to contribute real-time information during emergencies. For example, images can be used as supplementary information by responders for a more effective response. One of the major examples of using emergency mapping is the 2010 Haiti earthquake in which many volunteers gathered to help in mapping the basic infrastructure, to translate text messages, add them to live web-based mapping services like Crisis Camp Haiti, OpenStreetMap[1],

[1] https://www.openstreetmap.org

Ushahidi[2], and GeoCommons[3](MacEachren et al., 2011; Middleton, Middleton, & Modafferi, 2014; Zook, Graham, Shelton, & Gorman, 2010). In my proposed tool, I implemented a map view to let analysts explore tweets on the map. My proposed approach also provides location information extracted from social media posts and users' profiles over different emergency areas using Twitter data.

Some researchers applied analytics techniques for event exploration. Fuchs et al. (Fuchs, Andrienko, Andrienko, Bothe, & Stange, 2013) investigated two key assumptions regarding social media visual analytics by applying visual analysis on German flood tweets over the course of eight months. The results showed how spatiotemporal clustering of emergency related tweets might allow detection of both -time and space- related events. However, sometimes tweets are referring to remote events and sometimes tweet clusters may refer to sub-events (e.g., a gas problem in an earthquake). This study also relates extracted clusters with other sources (e.g., news). This study applied initial filtering based on a predefined vocabulary for known events.

Twitinfo (Marcus et al., 2011) is a real-time visual analytics platform for Twitter exploration. Twitinfo automatically detects and labels unusual bursts in real-time Twitter streams. It receives a query and then visualizes and summarizes the events with an aggregated view of user sentiment. Twitinfo enables users to identify sub-events. It provides geographic distribution of tweets representing public sentiment in different regions. However, Twitinfo does not focus on emergency related events so it does not consider emergency related features in the subevent or sentiment identification.

[2] http://www.ushahidi.com
[3] http://geocommons.com

Another application of social media visual analytics is information diffusion, which deals with the propagation of information through social media. Visualization of information diffusion allows tracking when and where information is spreading, how different topics compete for public attention, and what the public's reaction is to a particular event. Being able to track information diffusion can be useful in the Response phase for understanding how and when information spreads. Cao et al. (Cao et al., 2012) provided a novel visualization for tracing information diffusion on Twitter. Opinion Flow (Wu, Liu, Yan, Liu, & Wu, 2014) allows analysts to explore and trace public opinions and diffusion of these opinions around a topic. Analysts can explore different topics based on time. The visualization also displays the sentiment of the public around a topic using different color ranges. Furthermore, one can explore users' influence on each other. Whisper (Cao et al., 2012) provides a novel visualization based on a sunflower, to visualize the spread of ideas (when, where) using Twitter data. Xu et al. (Xu et al., 2013) demonstrate topics competition for public attention and influence of opinion leaders on topics over time. They visualize how different topics compete, the influence of leaders' opinions on the topics, and the trend of topic transition over time. In their study (Xu et al., 2013), authors assume that users define opinion leaders and topics. Information diffusion can be applied in counteracting rumors. In this way, analysts can assess how much the rumor has spread amongst a group of users and counteract it by sending messages to those groups.

2.7 Requirement Elicitation Studies

Another group of studies explore social media usage and the related issues by gathering requirements from practitioners. Tapia et al. (Tapia, Moore, & Johnson, 2013) conducted semi-structured interviews to understand the usage of microblogged data by emergency related organizations. They showed the landscape of the microblogged data usage among different

organizations is varied. They found microblogged data useful when responders lack information. However, there are concerns with quality and reliability of microblogged data. Hiltz et al. (Hiltz, Kushma, & Plotnick, 2014) discussed the results of interviews with US emergency managers about their barriers and wish lists when using social media. They identified three main barriers to using social media: lack of resources, lack of policies and guidelines for social media use, and concern about the trustworthiness of posts. A survey (Plotnick, Hiltz, Kushma, & Tapia, 2015) of 241 U.S. emergency managers at the county level explored if organizations used social media and discussed how they could improve this usage. The study stated that only about half of these organizations use social media in any way.

The use of social media is evolving rapidly and the type of use is different from one event to another and from one organization to another (Houston et al., 2015; P. Landwehr & K. Carley, 2014; P. M. Landwehr & K. M. Carley, 2014). The use and requirements for social media can vary by geopolitical region. As previous studies are mostly US centered, I intend to reveal the usage of social media within Canadian organizations and to understand if the challenges are the same and what specific strategies Canadian organizations are performing to tackle those challenges. I targeted both emergency managers and social media analysts in the emergency field.

2.8 Microblog Classification in Emergency Domain

Most of the existing literature is about event detection in general and not determining the situational awareness or the amount of valuable information within each post (Laylavi, Rajabifard, & Kalantari, 2017). In this section, I focus on using microblog classification for emergency response. In the emergency domain, several efforts use machine-learning classifiers to filter and categorize microblogs. In microblog classification, we are dealing with short, noisy, unstructured

and informal text data. Compared to conventional document classification problems, microblog classification is more challenging because less information is available as input for the classification. Having an appropriate set of features is one of the important factors in improving the accuracy of a classifier. From the literature review, I identified five groups of features being used for microblog classification:

1. *Text mining features:* This group of features targets the content of the text and tries to 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 following is a set of common text mining features in the literature:
    - n-gram features are a contiguous sequence of n grams of items from a given text, an n-gram of size 1 is referred as a unigram and an size 2 is referred to as a bigram (Aggarwal & Zhai, 2012).
    - Bag of Words (BoW) features represent the text as a vector of words; sometimes words are represented with their frequency in the text (Aggarwal & Zhai, 2012).
    - Part of Speech Tagging (POS) assigns each word in the text to the corresponding part of speech (e.g., Noun, Verbs, Prefixes).
    - Term Frequency-Inverse Document Frequency (TF-IDF) measures term frequency within a tweet whereas the frequency within all tweets. BoW can be combined with TF-IDF so that each word in the BoW feature can be represented with its TF-IDF score (Salton & Buckley, 1988).

- Latent Dirichlet Allocation (LDA) is a topic modeling technique that discovers clusters of semantically related words that co-occur in a collection of documents (Blei, Ng, & Jordan, 2003).

2. *Space and time related features:* This category considers temporal and spatial characteristics of a tweet. For the tweets that are not geo-tagged, mentioning a location in the tweet can contribute to situational awareness and thus these features can be important (Karimi, Yin, & Paris, 2013a; Schulz, Ristoski, & Paulheim, 2013).
3. *Linked open data features:* Semantic features try to compensate the limitations of short-texts by using external sources. In this category, short text is annotated to the external sources such as DBpedia or Freebase. Another approach to enrich the semantics of short message is Named Entity Recognition (NER), which identify predefined categories in text such as name, organization, location, and time (Amparo E. Cano, Andrea Varga, Matthew Rowe, Fabio Ciravegna, & Yulan He, 2013).
4. *Social network based features:* Social media features of a tweet along with other features can contribute to determine its relevance to the target emergency. Some of these features are "if the tweet is replied / retweeted", "if it contains hashtags", or "if it contains mentions"(Diakopoulos, Choudhury, & Naaman, 2012).
5. *Source features:* These features target the characteristics of the tweet author. This category is important in determining the level of trust for the author. Some of these features are "Number of messages of the author", "Registration date of author's account", "If the account is verified", "Ratio of followers to followings", and "Length of screen name" (Diakopoulos et al., 2012; Xia, Yang, Wu, Li, & Bao, 2012).

Other than defining a set of appropriate features, different types of classification algorithms can be used according to the type of problem. Some studies applied **binary classifiers** to solve the problem of separating relevant tweets from irrelevant ones in large-scale emergencies (Sakaki, Okazaki, & Matsuo, 2010). Some studies applied binary classifiers to detect small scale incidents like car crashes or to detect fires (Power, Robinson, & Ratcliffe, 2013). Verma et al. (Verma et al., 2011) designed classifiers to distinguish tweets that are contributing to situational awareness from tweets that are not. Their system could achieve over 80% accuracy on categorizing tweets that contribute to situational awareness (Verma et al., 2011). Their study used text-mining features to identify tweets that contribute in situational awareness, while I explored a wide variety of characteristics to extract spatial, social network and source-based related features.

Karimi et al. (Karimi et al., 2013a) focused on one of the preliminarily problems in classifying tweets during an emergency: Separating relevant tweets from irrelevant ones and also distinguishing different types of emergencies like earthquakes, flooding, fire, and storms. They focused on building a classification model using past incidents to classify tweets for current incidents. They used a combination of generic features with incident specific features. They concluded that for previously unseen events, generic features work better than incident-specific ones because with incident related features, the classifier would be biased and cannot generalize the results based on other emergency types. In contrast, if specific information about an incident were already revealed to the classifier, incident specific features would be useful and contribute to increased accuracy. Sakaki et al. (Sakaki et al., 2010) interpreted Twitter users as social sensors to detect earthquakes or typhoons in real-time. In their method, they applied a binary classifier for identifying tweets referring to an earthquake. Their next step was to get related tweets as input and detect event and location of events using probabilistic models. The aim was to detect earthquakes

(or typhoons) in real-time to notify the users that are close to the event. Sakaki et al. (Sakaki et al., 2010) used the Global Positioning System (GPS) and known information about a user for a location estimate. Schulz et al. (Schulz et al., 2013) proposes a classifier to detect real-time small-scale incidents. The study employed semantic web technologies along with different combinations of Natural Language Processing techniques. Before applying the classifier on tweets, they applied temporal and spatial filtering. The study assumed that the timestamp of a tweet is not always enough for temporal filtering (since some tweets talk about past and some talk about future events), so a mechanism to extract the temporal information about a tweet has been adapted (Schulz et al., 2013). Similar to temporal extraction, the study also extracted location information (as only 1% of all tweets are geo-tagged) using the name entity recognition method. Schulz et al. (Schulz et al., 2013) results show that using n-gram features without PoS, using TF-IDF score and syntactic features (the number of "!" and "?" in a tweet) provide the best classification results. However, re-evaluating their result on a test dataset from different timelines dropped the accuracy, which shows the model has been biased to the training dataset. Finally, adding linked open data features helped to overcome over-fitting of the training dataset and improved classification on the test dataset.

Microblog classification has also been used on domains such as sentiment analysis (Liu & Chen, 2015) or traffic detection (D'Andrea, Ducange, Lazzerini, & Marcelloni, 2015). Tuarob et al. (Tuarob et al., 2014) applies classification techniques on social media posts (Facebook and Twitter) for identifying health related tweets. Alonso et al. (Alonso, Marshall, & Najork, 2013) studied classifying tweets based on interestingness. From an informal survey, they defined what features make a tweet interesting. They realized tweet authors, originality, and timeliness can be contributing factors to what makes a tweet interesting. Moreover, some factors like genres, form, and style can also contribute to interestingness. For creating training data, datasets of tweets were

labeled into the following categories: news, current, meme, commemorative, adult, and thinking-out-loud. Webberley et al. (Webberley, Allen, & Whitaker, 2015) also focused on the personalized interestingness of the tweets. For measuring interestingness, they measured the ratio of actual retweets to expected retweets. For measuring the expected retweet count, they trained a classifier using social media and source based features. In my study, I measure interestingness for the emergency context. I use the interestingness factor as one of the features in my proposed social media monitoring tool. I train a classifier based on random sampled Twitter data to identify an expected retweet count. Then I measure the difference between the actual emergency related tweets retweet count to the excepted retweet count.

Some approaches categorized emergency-related tweets into different information types and sources using *Multi-Class Classifiers* (Amparo E Cano, Andrea Varga, Matthew Rowe, Fabio Ciravegna, & Yulan He, 2013; Caragea et al., 2011; Kreiner, Immonen, & Suominen, 2013) Krenier et al. (Kreiner et al., 2013) proposed a multi-class topic classifier on the following categories; food, health, economic, personal, community, and environment as well as a source classifier for identifying media, official authorities, and non-profit organizations. The study by Leylavi et al. et al. (Laylavi et al., 2017) uses classification techniques to determine the level of informativeness of a tweet. The study uses a term class frequency technique to assign relatedness scores to tweets and reports 87% accuracy result. They evaluated their proposed method on one dataset.

The reliability of social media information is another problem that can be addressed using a classifier. Not all content published on Twitter is trustworthy. Rumours can spread quickly and lead to incorrect decisions and actions (Starbird et al., 2014). Researchers have proposed automatic ways to distinguish misinformation or unreliable sources. Features related to the content of the

message, the characteristics of users, and the way the message propagates can contribute to determining the reliability of the message (Xia et al., 2012).

Most of the microblog classification approaches assign a single label to a tweet text. *Multi-label classification* categorizes instances into multiple classes, while the associated classes are not exclusive. There are different approaches to solve the multi-label classification problem:

- Use a combination of binary classifiers where each classifier predicts if an instance belongs to a specific class or not.
- Transform the multi-label problem into a multi-class problem, where each combination of label set can be considered as a class.
- Construct a chain of n binary classifiers where each binary classifier depends on the prediction of previous ones and n is the number of labels (Kong, Shi, & Philip, 2011).

Schulz et al. (Schulz, Mencía, Dang, & Schmidt, 2014) propose a multi-label classification approach, which considers label dependencies on tweets while Caragea et al. (Caragea et al., 2011) applies multi-label classification based on ten labels by transforming the problem to ten binary classifiers.

An important question to consider when determining a classifier for emergency related tweets is, what are the types of information within the tweets? Olteanu et al. (Olteanu, Vieweg, & Castillo, 2015) constructed a comprehensive categorization of different information types. They manually categorized 26 datasets regarding crisis events by crowdsourcing to different information types. Olteanu et al. (Olteanu et al., 2015) also categorized each tweet to two major categories of informative or not informative. They also categorized informative tweets to affected individuals, infrastructure and utilities, donations and volunteer, caution and advice, sympathy and emotional support, and other useful information.

The automated classification approaches introduced in this section, use a quantitative evaluation in which they measure the accuracy of a classifier according to a given test dataset of emergency event. These studies demonstrate the applicability of classifiers and text mining techniques for automated microblog classification (Hughes et al., 2014). These classifiers are being trained based on a few experiments. Applying these automated classifiers in other conditions can decrease their accuracy as they may be over-fitted with the training data set. For example, in a study for detecting fire incidents using microblog classification, accuracy dropped from one season (winter) to another (summer) [22]. Although generic features can be used broadly, sometimes incident-specific features can have positive impact on the accuracy of a classifier (Karimi et al., 2013a). To tackle these problems, I explored different characteristics in order to choose the right features. I then embedded a classifier in a visual analytics tool to bring the human into the classification loop and provide feedback to the classifier to add or remove features. Table 1 summarizes studies that used microblog classification for the emergency domain along with the list of different features they used.

Table 1. Summarize feature categories and labels in microblog classification.

Study	Applied Feature	Classification Labels
Abel et. al (Abel, Hauff, Houben, Stronkman, & Tao, 2012)	NER, number of followers, number of tweets during incident, location of the user.	Unknown Labels.
Karimi et. al(Karimi et al., 2013a)	n-grams, hashtag set, hashtag count, mention, mention count, URL, tweet length.	Related or Not, Disaster Type.
Power et. al(Power et al., 2013)	Number of words, mention count, hashtag count, and hyperlink count, unigram, bi-grams.	Identify fires.
Schulz et. al(Schulz et al., 2013)	Word unigrams, character n-grams, TF-IDF, the number of "!" and "?" in a tweet, Spatial and Temporal unigram features, Linked Open Data features.	Identify Small Scale Incident.
Cano et. al (Amparo E Cano et al., 2013)	Bag of unigram weighted by TF-IDF, POS tagging, Linked Open Data based on DBpedia and Freebase.	WarConflict , LawCrime and DisasterAccident.
Sakaki et. al (Sakaki et al., 2010)	The number of words in a tweet, and the position of the query word within a tweet, the words in a tweet, and the words before and after the query word.	Related or Not.
Verma et. al (Verma et al., 2011)	Unigrams and their raw frequency, bi-grams and their raw frequency, Part-of-speech tags.	Situational awareness or Not.
Imran et. al (Imran, Castillo, Lucas, Patrick, & Rogstadius, 2014)	Unigram, bi-grams.	Unknown Labels.
Yin et. al (Yin et al., 2012)	Tweet length, number of person names, number of location names, number of time words, number of organizations, number of verbs, has Web links or not, number of other nouns, number of stop words, number of words after segmentation, number of total words, frequency of stop words, containing "" or not, containing ''#'' or not, number of punctuations, number of all the nouns, number of followers.	Related or not
Schulz et. al (Schulz et al., 2014)	3-grams, Sum of TF-IDF scores, the number of '!' and '?, spatial features.	Shooting, Fire, Crash, Injury.
Caragea et. al (Caragea et al., 2011)	BoW, a bag of m topic words using LDA.	Medical emergency, people trapped, food shortage, water shortage, water sanitation, shelter needed, collapsed structure, food distribution, hospital service.
Diakopoulos et. al (Diakopoulos et al., 2012)	The number of Twitter followers, frequency of posts and presence of URLs, use of Twitter re- tweets, mention and reply functions that indicate interaction with other users, named entities mentioned by the user, and the topic distribution of user posts.	Eyewitness or Not/ Source Category.

2.9 Social Media Analysis tools

In this section, I will focus on studies that discuss designing social media monitoring tools for emergency context. Several studies explored how to extract relevant information from social media data, more specifically from Twitter, and designed and developed social media monitoring tools.

TweetTracker (Kumar, Barbier, Abbasi, & Liu, 2011) is a tool to track, analyze, and monitor tweets (Kumar et al., 2011). It enables users to monitor and analyze time series data of tweets based on specific keywords in different frequencies (day, week, and month). It enables users to monitor top Uniform Resource Locators (URLs), users, and hashtags. It also allows filtering based on keywords, geo-location of users, and languages. SensePlace2 (MacEachren et al., 2011) is a web-based tool built upon Twitter data. Its aim is to provide situational awareness and sense-making by revealing the connection between people, places, topics, and organizations. If tweets from one location mention other places, SensePlace2 visualizes this by pointing to those locations on maps. It also provides temporal controls in different granularities. SensePlace2 involves users by allowing them to promote or demote tweets and report errors in the geo-locations. While these studies provide useful tools for monitoring and filtering microblogs, they provide simple filtering techniques such as keyword-based filtering or filtering by location.

In my work, I aim to seek for more advanced filters and give analysts the opportunity to add or remove features that can impact the accuracy of the filters. The work of Rogstadius et al. (J. Rogstadius et al., 2013) disaster awareness tweets were extracted using sets of keywords, constructing stories, and clustering tweets using their lexical similarities.

Some studies utilized microblog classification within visual analytics systems to help analysts identify relevant information using more advanced filters**.** The study by Diakopoulos et al. (Diakopoulos et al., 2012) took a further step in Twitter classification. They developed and

evaluated a tool called Seriously Rapid Source Review (SRSR) to improve the experience of journalists when exploring social media posts and assessing social media sources around breaking news. They implemented two classifiers; the first classifier was to confirm eyewitness sources and the second to distinguish the type of each source. The classifier categorized sources into organizations, journalists/bloggers, and ordinary people. Their proposed tool, SRSR, shows a source's network and their friend's locations to help journalists assess the potential credibility of the source's information about the event. The tool also shows the top three most frequent locations in the corresponding network. SRSR additionally includes a network sketch indicating who has shared relevant event content among the source's followers. These features can help inform journalists of whether this source knows many people who are involved in the target event. While this study was not directly related to the disaster domain, the features and information cues they proposed about the reliability of a source can be applied into disaster-related scenarios to inform social media analysts about the tweet authors.

The Twitcident system (Abel et al., 2012) filters relevant information about an incident using handcrafted rules. It enables filtering, searching, and analyzing tweets in real-time. Twitcident uses semantics techniques to filter tweets. However, it does not involve users to improve filtering accuracy. It only gathers geo-tagged tweets and employs classification based on handcrafted rules (e.g., if a tweet contains specific attributes then it is classified in a specific category).

Emergency Situational awareness (ESA) (Yin et al., 2012) provides a microblog mining tool to extract and visualize real-time information that can contribute to situational awareness about an incident. The analysis package of this tool provides alert monitoring through clustering and burst detection, distinguishing tweets that report infrastructure damages using the

classification method, and allow for exploration of geo-tagged tweets. In this study, I will seek to find characteristics that make an emergency tweet informative. Mazumdar et al. (Mazumdar et al., 2014) proposes a tool that was designed for Emergency Response to increase situational awareness through real-time visualizations and a highly interactive design. However, this tool does not combine machine learning intelligence in the design.

Other studies have incorporated machine-learning classifiers into the visual analytics process. These studies engage the user in the classification process. In a supervised classification approach, a classifier gets a set of features and a labeled training set to build a model that can be used for prediction of new instances. However, preparing a training set is time-consuming and using training sets from previous events may not be efficient according to the specific characteristics of each event. To solve this problem, some studies engage users to label a set of instances online (Hoferlin, Netzel, Hoferlin, Weiskopf, & Heidemann, 2012). In online training, either the system or the user can select instances for the training set. Studies have applied this learning approach in several domains such as video analysis (Hoferlin et al., 2012), document retrieval (Heimerl, Koch, Bosch, & Ertl, 2012), and microblog analysis (Bosch et al., 2013).

Bosch et al. (Bosch et al., 2013) proposed a visual classifier tool named Scatterblogs 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. Thornton et al. (Thornton, DeAngelus, & Miller, 2016), use the feedback-based techniques to improve situational awareness instead of traditional key-word based techniques by updating the selected features in the learning model. The Artificial Intelligence for Disaster Response (AIDR) (Imran, Castillo, Lucas, Meier, & Vieweg, 2014; Imran, Castillo, Lucas, Patrick, 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. The authors proposed to consider the following factors in tweet classification in future studies: temporal effects and their impact on the performance of a classifier, and removing duplicates to avoid labeling items that are closely similar to those that have already been labeled. 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. My work extends studies (Flizikowski et al., 2014; A. L. Hughes & Shah, 2016; Reuter, Amelunxen, & Moi, 2016) that designed and developed an application based on the social media needs of public information officers in the emergency context. However, in their application, they do not incorporate machine learning filters. In this dissertation, I designed and developed a prototype based on the needs of social media analysts in the emergency domain and the prototype evaluation brings insights into the design of applications that embed supervised machine learning classifiers in monitoring social media during emergency scenarios.

Chapter 3: Requirement Elicitation[4]

This chapter reports the result of interviews conducted with emergency practitioners. The goal was to analyze issues in organizational needs for extracting information from social media during emergency response. By interviewing 16 Canadian emergency professionals, I gained insights into such things as how they currently monitor social media, what types of information they are interested in, what challenges they encounter, and what strategies they use to overcome those challenges. The most frequent requirements mentioned by participants were the needs for prioritization and categorization of social media data to mitigate information overload. These professionals are also concerned about the reliability of information and counteracting the rumors.

3.1 Method

I conducted semi-structured interviews with 16 participants across established or extended emergency related organizations (see Table 2). The organizations can be divided into two main types: End users and Service providers. End users belong to public organizations (e.g., EOC and police department) or private sectors (e.g., Oil and Gas sector). I targeted participants who had the experience in managing and responding to at least one emergency event (e.g., Slave lake Fire, Alberta Flood 2013, and Calgary Snow Storm 2014). The interviews were performed between March 2015 and December 2015. After receiving the participants consent to be in the study (for the consent form, please see Appendix E), I asked participants about their use of social media, how they monitor and extract information from public posts, and what challenges they are facing when using social media (For a full list of interview questions used, please see Appendix B). Each

[4] This chapter is based on: Marbouti Mahshid and Frank Maurer. "Social Media Use During Emergency Response-insights form Emergency Professionals." Conference on e-Business, e-Service and e-Society. Springer International Publishing, 2016.

interview lasted between 60-90 minutes and was audiotaped and transcribed. To analyze the notes, I followed a process inspired by grounded theory (Glaser & 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. The goal was to determine current usages of social media across different organizations. I used the saturate[5] application to analyze the interview notes.

Table 2. Requirement Elicitations user profiles.

Participant	Job Title	Years of Experience in Emergency Field	Type of Organization
P1	Researcher at Center of child, family, and community research.	4-5	Service Providers
P2	Police officer.	19	End Users- Public Org.
P3	Director, national security and strategic foresight.	6	Service Provider
P4	Director of emergency communication.	30	End Users- Public Org.
P5	Fire chief.	24	End Users - Public Org.
P6	Senior safety coordinator.	20	End Users- Private sector
P7	Emergency manager.	16+	End Users – Public org.
P8	Emergency manager.	12	End Users – Public org.
P9	Police officer.	20	End Users – Public Org.
P10	Executive director of operations.	28	End Users – Public Org.
P11	Police officer.	unknown	End Users Public Org.
P12	Emergency center communications officer.	unknown	End Users – Public Org.
P13	Strategic communications specialist.	unknown	End Users- Private sector
P14	Team lead communications.	10	End Users –Public Org.
P15	Digital communications officer.	11	End Users –Public Org.
P16	Vice President.	unknown	Service Provider for Private Sector

[5] http://www.saturateapp.com/notebooks

3.2 Results

In presenting the results, I will not use any percentages or statistics since this would not be suitable for the size and nature of the sample studied. Instead, I follow a qualitative research approach and describe the main themes, based on counts of coding categories, and then give examples of descriptions of these themes in the words of the interviewees themselves.

3.2.1 Patterns of Use

Most of the participants agreed that social media had become a central element in emergency response planning. *"It's not a part you can ignore or a piece that can wait."* They were familiar with the potential and importance of using social media during an emergency. *"Manage public relation - if you do not manage it, it manages you."* Twitter, Facebook, and Instagram were the most used social media services mentioned by participants. Depending on the resources and the size of the organization they were managing, the degree of their involvement with social media was different. Some mentioned that they monitor social media "very intensively" and others mentioned that they only communicate out and do not monitor social media, mainly because of "lack of resources" due to the sheer volume of information.

Public sector end-users mostly monitor social media for *"watching the conversation, communication and identify trends, reporting urgencies, … [seeing] impact on community."* They care about people and their safety, properties, and environment. It is important for these organizations to know the topics that people talk about and how they feel about those topics. *"We push information but we are also very interested and keen on what people are talking about."* However, the usage differs between everyday monitoring and during an emergency. As many participants stated, during an emergency, the volume of messages increases but one of the participants also pointed out another interesting difference: *"during a crisis event that*

conversation is very elevated and very focused … in non-crisis in 5 minutes I see 20 different topics that people are talking… maybe there is a higher need for filtering during non-crisis event since the conversation is much wider… [while during the crisis] it comes about how to filter out the noise." On the other hand, emergency events have more challenges as they usually happen fast and it takes time for authorities to get a grip on the online story. For example, one participant stated about a shooting, *"Social communication was huge and it took authorities three hours to become part of that conversation. Speculations and accusations were occurring; investigators were required to follow up noise to track the efforts. It's a mess and we don't know what the formula is".*

The police social media analysts described some additional usages in regard to social media. Their tasks include general enquiries such as *"less formal questions, [and] legal advice type questions",* proactive community policing such as *"sharing infographics about the latest crime trends",* reactive to crimes for example *"ask public to help find and identify suspects",* and investigative support for example: *"the officers are reaching out to me asking for assistance on social media… They found a wallet they have a name but they can't find the contact detail of the owner... Can we find this person on Facebook?"* Another police officer stated that when there is a risk of further incidents, they monitor social media intensively to determine where and what is being said to help mitigate the risks.

I also interviewed three private sector end-users. They pointed out that their companies usually do minimal monitoring on a daily basis. One of them stated that during an emergency they are looking for information about what people are saying regarding the emergency, what negative comments or inaccuracies are being reported about the company then craft a response to the posted information. She stated, *"[The company] hired a public relations company to monitor the social*

media feeds in the recent oil spill because traffic was so high. Otherwise [the company does] very minimal monitoring on a day-by-day basis, really only during an emergency."

3.2.2 Types of Information

I asked participants what kind of information social media provides that they are interested in. I extracted the following categories from their responses:

Detail: As a communication officer mentioned during an emergency they try to extract the impact of the event on people and the environment. A social media post can contain information that shows this impact. People can provide details in messages, photos, and videos as a participant commented: *"People post about things they're in the middle of or post pictures, videos, comments".* According to the participants, details can be regarding properties, their conditions, and damages during an event. Details can also be regarding affected people, their injuries, their evacuation, their needs for resources such as food, blankets, water, etc.

Questions: A post can contain a question that the author seeks an answer for. As an emergency manager stated they *"Try to answer questions quickly – before someone non-authoritative can answer."* They may ask questions regarding different subjects such as evacuation orders, volunteering and donation, or rescue processes: *"For flood people asking us with evacuation areas. There were 3 communities that they want to evacuate and then 4,5 were going to be and then they started asking … people started asking questions where should we go, where are the community help centers,…"*

Misinformation: A post can contain a rumor or false information. According to a recent analysis of emergency-related social media posts (Starbird et al., 2014), a considerable portion of emergency-related microblogs can be incorrect. Rumors or false information can

spread quickly, either purposefully or accidentally, and lead to incorrect decisions and actions by the public. Hence, identifying microblogs that contain false information can help analysts to counteract them. A participant mentioned: *"Another critical piece, [we] need to be able to dispel rumors. In a tornado, there were two rumors: I heard that there [were] two separate touchdown points..."*

Urgencies: An emergency related post can contain a request for help. For example, somebody needs immediate help or rescue. These kinds of posts are highly prioritized by emergency personnel. They try to find these posts by scrolling through lists of incoming messages but are challenged by the high volume of incoming posts, e.g. *"We had posts like I can't find my mom, can somebody go and check the house OR my neighbor has been handicapped can someone go and check."*

Sentiment: The public perception regarding an event was another category that the participants mentioned. While these posts often do not have any operational information, understanding the public sentiment can help the authorities prioritize resources. One participant mentioned the importance of sentiment, which helped discover problems that are about how people feel and react to the situations the authorities were already familiar with.

3.2.3 Information Overload – The Need for Filtering and Categorization

One of the problems that all participants mentioned was the volume of information coming from the public. According to the interviews, analysts receive thousands of messages every day and the situation is worse during an emergency. The Calgary flood infographic (Yablonski, 2013) shows there were 857,000 tweets for the most popular hashtags during the first two weeks of flooding. A considerable percentage of incoming messages are operationally irrelevant which makes it hard to

extract urgent posts: *"Got 10s of thousands of messages of support from around the world. People would report people who might need to be checked on, but this could get overwhelming fast." "If somebody needs help, we should make sure to catch that by scrolling through and its challenging and things get messed up."*

Another participant stated that sometimes businesses try to promote themselves by attaching their advertisement to the trending hashtags on Twitter. As an analyst stated one drawback of current tools is that they represent posts in a reverse-chronological order that makes it challenging for them to make sense of thousands of incoming posts with a short amount of time in an emergency situation. Another participant mentioned: *"It takes a significant amount of time and effort to sort through the incoming posts, we get thousands of incoming messages on a daily basis, we try from a corporation perspective to acknowledge or respond messages, right now we are performing about 3%!"* On the other hand, the use of social media by the public has enormously increased during the last decade and community members now expect help when they send a social media message, *"There is an increase in the expectation from the public that if they tweet asking for help they will generate help or a response, we are not there yet!"*

Some organizations have strategies to mitigate the information overload. For example, one participant mentioned that they were creating a role called "Audience Intelligence". The role monitors news and streams to determine what the trends are and understand what people are saying. This can help the rest of the social media monitoring team to catch stories online. The digital communication officer in the police department also mentioned that since they lack resources to monitor social media posts during the night shift, they put the geo-tagged ones on the map view so that everybody can see what is going on in social media.

3.2.4 Degree of Separation

The pressure that people within EOCs face makes the analyst job difficult as the responders are often local people who see their friends and families in danger. 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 was unprecedented. Responders may not sleep for several days and bear lots of pressure. As one participant mentioned, *"you may start crying in a corner after working several hours".* As they stated, having a degree of separation and use the help of outsiders for monitoring social media can be useful, *"during the flood we had one person that worked remotely, I think that was one of the keys to the success, having that degree of separation... since that person was not influenced by the crazy energy that was present in the Calgary Emergency Operation Center".*

3.2.5 Collaboration and Communication

I frequently heard from participants that communication and collaboration among people are more important than technology. During emergencies, integration between different organizations happens through people. Some participants mentioned that emergency managers might satisfy their informational needs through their trustworthy sources of information. Hence, it is important to establish those links before an emergency occurs. This could also be applied in the social media space by helping analysts identify active users within a community. It is important to be involved in social media before an emergency occurs and establish the communication links. As a participant explained, *"One of the big messages that we've been getting again and again in emergency management is that relationships in terms of people and organizations sits above technology and capability, even if you do have technology unless you have personal relationships build and trusted ahead of time still you are going to struggle!"* Another participant mentioned

that during the Alberta Flood 2013, Twitter shut down the Calgary police Twitter account as it was treated as spam due to exceeding the daily limit of tweets.

3.2.6 Reliability of Social Media

Although social media contains information that can contribute to situational awareness, one of the concerns is the reliability of the information. Almost all participants had concerns regarding the reliability of a source when talking about social media. *"Can you verify the information, who is the source of the info?"*

Source Reliability: One aspect of reliability was regarding how much the source of information is reliable. When asked about what characteristics they consider to assess the reliability of a source, the answers were: is the Twitter account valid? Do they follow their organization? Is there a known/verified follower? What past posts does this user have? How recently was the account created? What is the number of followers/friends? How active is the user? How many times is the user being reposted? Is the user an eyewitness or inside the emergency region? What is the profile picture of the user? As one participant mentioned, going through all these characteristics takes time.

Dispel Rumors: Another concern was the spread of rumors over social media. As mentioned by many participants, one of the tasks of a social media analyst is to counteract rumors. A rumor is a piece of information which the analyst knows is incorrect. According to the participants, rumors might spread intentionally by malicious people, but they can also spread unintentionally by lack of understanding or technical information, accidental misinformation or simply people's frustrations. For example, a communication officer stated that during the Calgary flood, there was a rumor that the drinking water is not safe in the city. One person found a boil water advisory from

the same day but a different year on the city website and tweeted that the water was not safe. This incorrect information spread rapidly. The city responded by dispelling the rumor but needed to repeat this many times. Some participants mentioned this concern that sometimes by counteracting a rumor, the number of people that hear about it increases – which might undesirably spread the rumor. Authorities need to target a select audience when counteracting a rumor. Sometimes by counteracting a rumor, the number of people that hear about it increases – which might undesirably spread the rumor. Thus, there is a need to target a select audience when counteracting a rumor. One participant wished that they could filter social media recipients based on different factors like geographical location or demographics so that they could target communities of interest. *"During a crisis who is your audience? ... We are less interested to counteract the rumor to a global audience. Need to know that and ensure the messaging and efforts are directed to those audience members and not spend energy dispelling rumors that are unrelated to the event and that audience."*

3.2.7 Current Tools

In this section, I discuss what the practitioners expect from social media monitoring tools and describe the current tools they use to manage social media. Most of the participants expected two common features from a social media monitoring tool: simplicity and the ability to manage multiple social media sources. They were using commercial tools: geo-feedia[6], Radian 6[7], Hootsuite[8], and SproutSocial[9]. HootSuite and Sprout Social are broadcasting tools. They let users schedule updates to communicate with the public. Radian 6 lets users listen and engage with

[6] https://geofeedia.com
[7] https://login.radian6.com
[8] https://hootsuite.com
[9] http://sproutsocial.com

multiple social media sources. In the listening part, users can track specific keywords. Finally, Geo-feedia's main advantage is to provide location-based filtering of geo-tagged posts and keyword-based filtering. One of the participants commented that in an emergency situation they usually prefer Hootsuite because of its simplicity. However, none of these tools help users automatically prioritize the social media posts based on their importance or the level of information. The chronological order makes it challenging for analysts to make sense of thousands of incoming posts within a short period of time in an emergency situation. End users currently lack advanced filtering tools to help identify and categorize relevant information. The tools should let analysts identify and track key influential users within their community. Future tools should be able to view the sequence of posts not just based on time but also based on the amount of information the posts are carrying. I noticed that none of the tools that are currently being used by these big emergency-related organizations had been developed specifically for emergencies but they had been built for a broader market. This reveals that there might be a need to design and develop practical tools specifically for emergency response and management.

Another issue is regarding the reliability of sources. Although participants knew some characteristics that can determine how much a source is reliable, they did not have enough time to check those characteristics in emergency chaos. Exploring how reliability scores can be assigned to social media sources is a current research direction to address this issue. Researchers are exploring different features (e.g., user profile characteristics and the content of a message) to be able to derive reliability scores for social media users (Diakopoulos et al., 2012; Xia et al., 2012).

3.3 Chapter Summary

This chapter addresses the organizational requirements on social media analysis tools in the context of emergency events. I conducted semi-structured interviews and gathered information about the usage and challenges of social media in emergencies. Table 3 shows a summary of the findings of requirement elicitation study. As the table shows, I discussed patterns of social media use within the emergency related organization. Based on the interviews, I realized that social media information is considered both essential and high risk. It was essential as it helped people communicate easier, faster and opened a two-way communication road. It is considered high risk, as there were a lot of speculations regarding what has happened. I discussed what types of information analysts are interested in. I also discussed what challenges they are currently facing when using social media, what strategies the practitioners are using to mitigate the current challenges, and what features are desired of future tools to help them in identifying situational awareness information. Social media analysts require tools that help them automatically prioritize and categorize social media posts according to their informational needs. They also require tools to facilitate quickly assessing the reliability of sources. Finally, this study's findings show that another area of concern for emergency practitioners is to dispel rumors. One of their challenges is that sometimes by trying to dispel a rumor they might spread it more to the people that have not heard about it. They need to be able to dispel the rumors by only targeting communities of interest.

Table 3. Summary of the requirement elicitation study findings.

Aspect	Summary
PATTERNS OF USE	• Watching the conversation • Communication and identify trends • Reporting urgencies • Monitoring impact on the community • Push out information • Inquiries
TYPES OF INFORMATION	• Detail • Questions • Misinformation • Urgencies • Sentiment
CHALLENGES	• Volume • Trust • Lack of proper tools • Dispel rumors
HOW TO MITIGATE POSSIBLE RISKS	• Using Common Operating Picture • Audience Intelligence • Degree of Separation • Collaboration and Communication

Chapter 4: Situational Awareness Classifier

In this chapter, I address the overload problem that was discussed in Chapter 3: Requirement Elicitation, from a computational perspective. I describe how I used supervised machine learning to identify situational awareness tweets. First, I provide a short description of machine learning classification as one of the most used methods for information filtering (Hughes et al., 2014). Second, I propose a Situational Awareness (SA) classifier to separate SA tweets from non SA. I explore how different categories of features contribute to identifying SA. Next, I evaluate the proposed classifier based on quantitative performance metrics. Finally, I conduct a stepwise regression to find the most contributing features for each dataset and improve performance metrics.

4.1 Microblog Classification for Emergency Management

I provide advanced filters to help analysts identify relevant information from Twitter data. Microblog classification is a widely used machine learning approach that has been used for filtering and identifying relevant information from textual data (Hughes et al., 2014). In this section, I provide a brief introduction of machine learning classification. Some of the popular areas where text classification is applied are, if email is read as spam or not and sorting movie reviews as good, bad, or neutral. In the first example, emails are input instances and being spam or not are the output labels for each email. In the second example, input instances are movie reviews and the output labels are good, bad, or neutral for each review. A classifier needs a set of features to automatically assign labels to each instance. In the movie reviews case, a set of features should be extracted from each review and be given to the classifier as input. Features could be length of review, negative emotions, positive emotions, and presence of smileys.

Machine learning defines classification, as the problem of assigning a set of unclassified instances to a set of labels given a training set whose label membership is known. Let $X =$

$\{x_1, \dots x_n\}$ be a set of instances, where $Y = \{y_1, \dots, y_m\}$ be a set of labels (which are referred to as categories or classes). The goal is to assign a label y, to an instance $x \in X$. If $m = 2$ the classifier is binary. If $m > 2$, the classifier is multiclass. Furthermore, if instances can belong to more than one label it would be a multi-label classifier (Kong et al., 2011; Sriram, Fuhry, Demir, Ferhatosmanoglu, & Demirbas, 2010). The classification problem could be formalized as a function approximation where: $y = \hat{f}(x)$ (the hat symbol represents an estimate (Kevin, 2012)).

The classification goal is to estimate the function f given a labeled training set. In text classification, instances should be represented in a structured form. I refer to these structured forms as features (certain researchers refer to these as explanatory variables), then the input of function f would be a feature vector (Kevin, 2012). In Figure 5. (A) a group of crowd-sourcing workers annotate a set of tweets manually while in Figure 5. (B) a classifier annotates labels to each tweet using a predictive model (certain researchers refer to the output labels as response variables). A classifier needs a subset of annotated tweets as a training set to build the predictive model. A feature extraction module is used to convert the tweets to a set of features. The most common technique to represent text in a structured format is Bag of Words (BOW). In this technique, the text is broken down into words and each word represents a feature. In this research, tweets along with their metadata information are instances and tweet information types are the labels. In the next section, I describe different features that are being used for short-text social media data classification.

The typical evaluation metrics for classification methods are *accuracy, precision, recall, and f-1 scores.* Assuming a binary classifier, the output labels of the classifier are either positive or negative. In fact, in the case of a multi-class classification, the metrics are computed per class and the overall measure is simply the average of the per-class measures. Positive instances can be

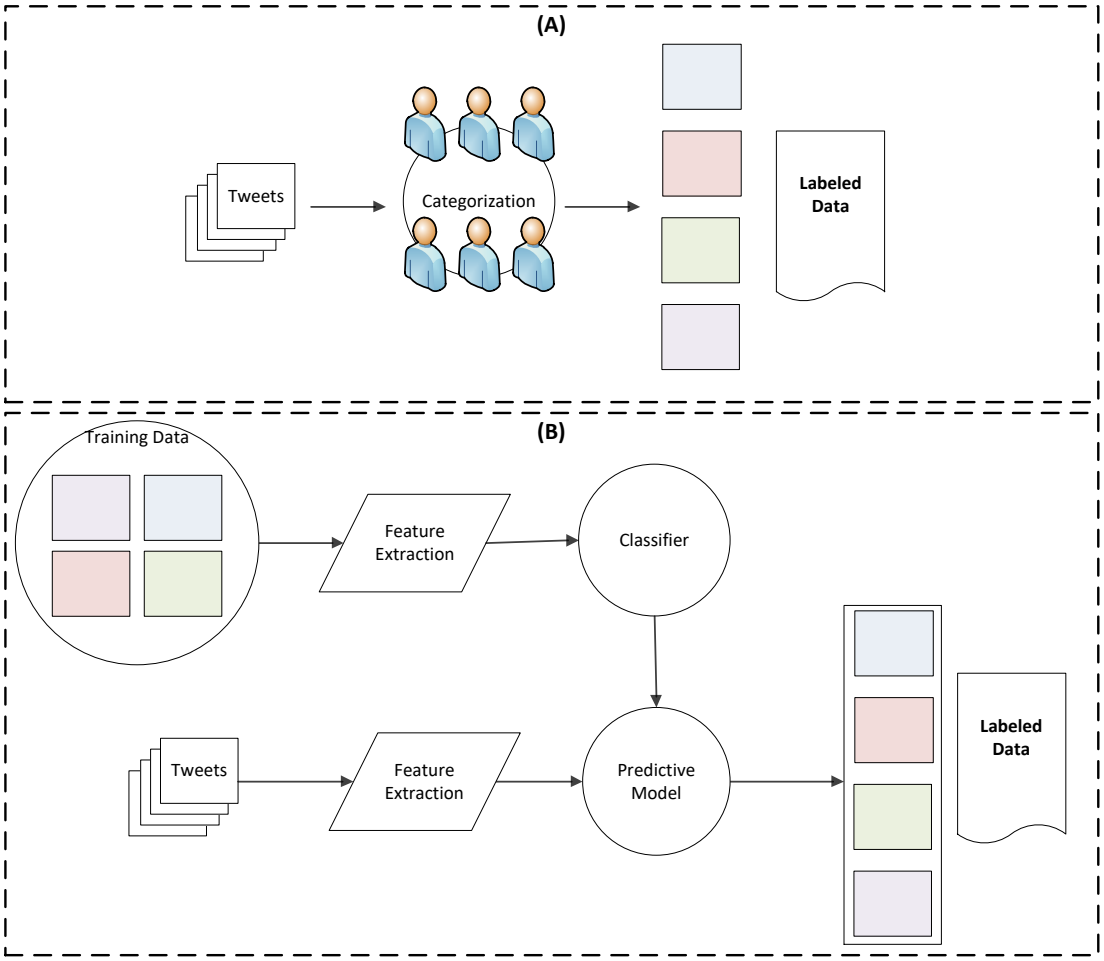


Figure 5. Categorizing Tweets (A) Manual Mode (B) Automated Mode

classified correctly (True Positives or TP) or incorrectly (False Positives or FP). Similarly, negative instances can be classified correctly (True Negatives or TN) or incorrectly (False Negatives or FN).

Based on these definitions, accuracy represents the number of correctly classified samples over the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is the number of correctly classified samples of a class, i.e., true positive class, over the number of samples classified as belonging to that class.

$$Precision = \frac{TP}{TP + FP}$$

Recall is the number of correctly classified samples of a class, i.e., positive class, over the real number of instances of that class; it represents the effectiveness of the classifier to identify positive samples.

$$Recall = \frac{TP}{TP + FN}$$

The *f-1* is the weighted geometric mean of precision and recall. It trades of precision and recall and it is used to compare different classifiers:

$$\mathrm{f} - 1 = (1 + \beta^2).\frac{Precision * Recall}{(\beta^2 * Precision) + Recall}$$

The default balanced *f-1* measure equally weighs precision and recall, which means considering β=1 (to evenly weight precision and recall).

4.2 Situational Awareness (SA) Classifier Features

Compared to conventional document classification problems (Sebastiani, 2002), tweet classification is more challenging because less information is available as input for the classification. In this section, I identify SA tweets in the emergency context. I explore possible contributing features to SA. These include social, location, tone, and textual based features. I create a classifier that combines these features. I also explore the significance of each feature for identifying SA. In each classification problem, it is needed to represent instances (i.e. tweets) as a set of features. Having an appropriate set of features is one of the important factors in improving the accuracy of a classifier model. I define several features that contribute to deriving SA

information. The features are inspired by the related work (P. M. Landwehr & K. M. Carley, 2014), interviews, and manual exploration of emergency-related tweets.

4.2.1 Textual Features

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 Term Frequency- Inverse Document Frequency (TF-IDF) which were discussed in Chapter 2.

4.2.2 Tone based Features

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. I used 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. I used the social and emotional tone analyzers. *Emotional tone* infers 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 is an indication of that emotion in the text. I used three different emotional tones: Fear, Sadness and Disgust. Table 4 provides example tweets for each of these emotions.

Table 4. Sample tweets for emotional tone features.

Text	Fear	Sadness	Disgust
#FortMacFire My thoughts are with all of you and your families. I am so sorry for all of the animals who are no longer with us. #SoSad	0.037284	0.72824	0.079801
Wildfire near Colorado Springs quickly becoming dangerous threat as thousands of residents evacuate homes	0.798841	0.07825	0.196947
Not good: 'Air drops like spitting on a campfire': Only weather chang… #ymmfire #FortMacFire	0.071805	0.139998	0.738147

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. I used three different social tones: Openness, Extraversion and Agreeableness. Table 5 provides example tweets for each of these emotions.

Table 5. Sample tweets for social tone features.

Text	Extraversion	Openness	Agreeableness
Please Help Those Affected by #YMMFIRE by Donating to @redcrosscanada!	0.149731	0.17418	0.906256
@RMWoodBuffalo Commending you on your amazing #socialmedia efforts keeping everyone up to date on the #ymmfire #FortMacFire appreciate it	0.99613	0.104105	0.793981
@CarneyGlass don't use the #ymmfire hashtag	0.244539	0.039508	0.13375
Well done @StLouisBlues, proceeds from 50/50 and silent auction tonight to help #ymmfire victims #AlbertaStrong	0.244539	0.039508,	0.13375
This is terrifying. And amidst all of it, my utmost respect to police officers and firefighters. Thank you	0.95091	0.91308	0.973212
Thinking of friends and first responders out in #FortMacMurray during this difficult time. Stay safe. #AlbertaStrong #FortMacFire	0.277439	0.055263	0.732265

4.2.3 Location Based Features

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, I investigated if the author uses any location words to describe a specific area within the emergency region. I used box-based and bigram-based processes to determine if a text of a tweet contains a location mention or not. For the creation and development of box-based and bi-gram based processes, I had the help of two undergraduate developer interns[10].

Bigram Based Process: To perform the extraction of location from tweets, we identified bigrams for each tweet by using the NLTK python library. For each event, we constructed a set of known unigram locations of interest to look for within the tweets. Next, we fetched the bigrams of each tweet. Every word within each bigram was checked against the list of stop words. This was to filter out all bigrams that are clearly not locations. The list of stop words was also increased in a later part of this process to increase relevancy. We ignored any bigrams containing stop words.

We iterated over the tweets and aggregate the cleared bigrams into a dictionary, associating each bigram with its corresponding tweet's ID or list of tweet IDs, in the case of multiple tweets mentioning the same location. These tweet IDs was used later to update the database. We considered both possible combinations of the bigram, such as "Buffalo Wood" and "Wood Buffalo", to ensure that when they are sent to the geographical database, they are matched with a location. After constructing this dictionary of bigrams and the list of associated tweets, we ran through the tweets a second time to search for the known unigram locations of interest. Associated tweets with unigrams were stored as a dictionary in the same manner as the bigrams.

After the extraction of a set of possible bigram locations and all relevant known unigram locations, these were sent to geographical databases. For an event in Alberta, Canada, we iterated

[10] The two undergraduate developers are my co-authors (the second and third authors) in this paper: *"Mahshid Marbouti, Irene Mayor, Dianna Yim, Frank Maurer (2017). Social Media Analyst Responding Tool: A Visual Analytics Prototype to Identify Relevant Tweets in Emergency Events. In Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Albi, France."* I refer to them as undergraduate interns throughout rest of the thesis.

through the locations and sent them to the Canadian GeoNames Search Service[11]. We ensured that the queries made using this search service were restricted to the populated and administrative places in the Alberta province to aid in finding the most relevant locations. This search service was placed in its strictest settings to prevent locations like "Three Brothers" to be associated with a mere mention of someone's brother.

There were, however, limitations with the strict setting of the Search Service; shortened, unofficial names or alternate location names were not matched. In order to account for missed location mentions, we combined queries to the GeoNames[12] geographical database and manually monitor returned locations. GeoNames' queries checked across a wide range of alternate names which allowed for more matches. Once more, we ran through the bigrams and unigrams, ignoring any duplicate locations that were found. we also limited the location attributes for the GeoNames queries similar to the Canadian GeoNames search service. we set the query to sort the result by the most relevant first as that is the most likely match for a location. For the first few minutes of running these queries, every location is checked before it is saved into the database. This is to increase the list of stop words and fine-tune querying for locations. A final check against the database, which is where these locations are stored, is performed to remove any obvious unintended location mention matches and increase the list of stop words for future tweets surrounding this event. Figure 6 presents the procedure of bigram process using an example. The original tweet is: "WATCH LIVE: Long says new evac centres opened up in Drayton Valley and St. Paul #ymmfire #abfire https://t.co/SF0llWqjGR"

[11] http://www.nrcan.gc.ca/earth-sciences/geography/place-names/tools-applications/9249
[12] http://www.geonames.org/

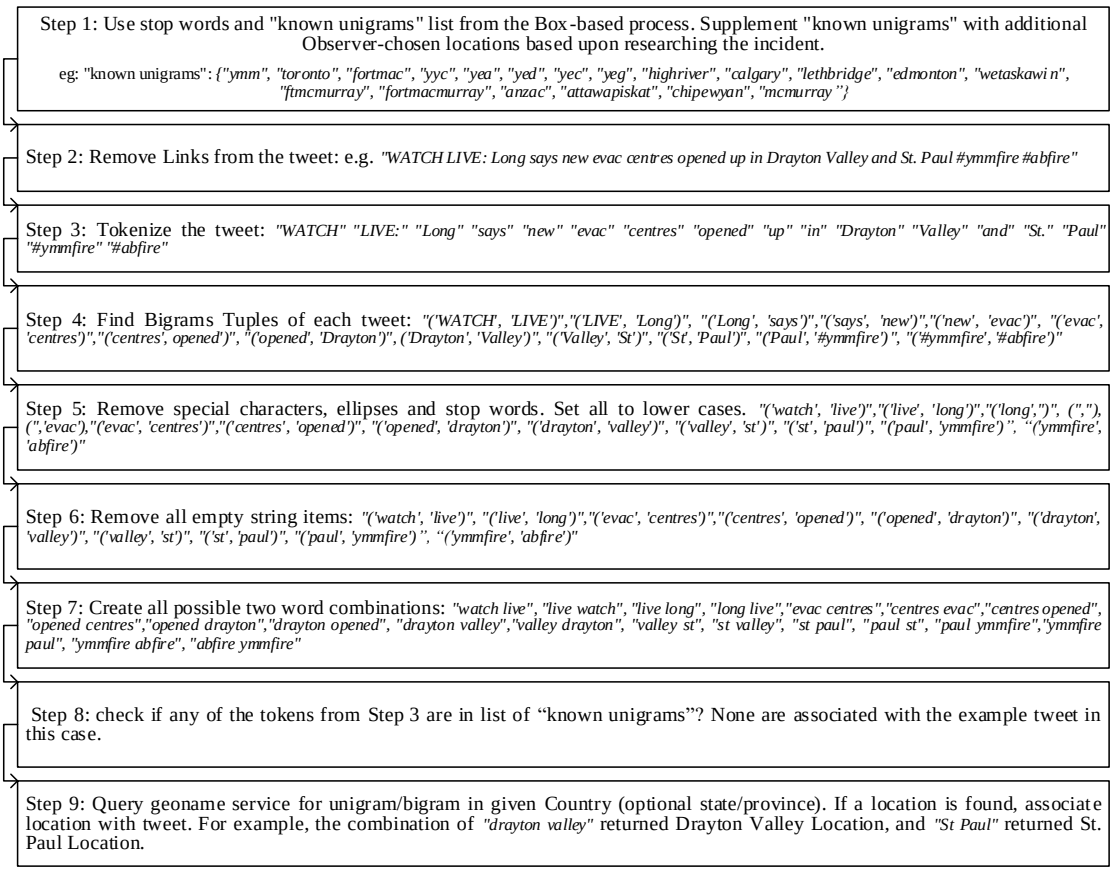


Figure 6. Bigram Based Location Extraction Process.

Box Based Process: A second location extraction process has also been performed to support the first process (see Figure 7), in the likely case that known unigrams and found bigrams do not cover all location mentions, especially those crucial to the situation. Instead of extracting possible location tweets, we started with a database query for locations based on a bounding box around the incident area. Next, with these extracted locations, we checked if any of them are contained in the tweets. A tweet contains a location if and only if the entire name is found in the tweet. Both tweet content and location names have been set to their lowercases to ensure case sensitivity does not exclude any matches. Since matching the entire name of locations restricts many invalid location names, as mentioned in the previous process, the query sent to the Canadian GeoNames Search Service (For events in Canada) allowed for all possible geographical features, instead of

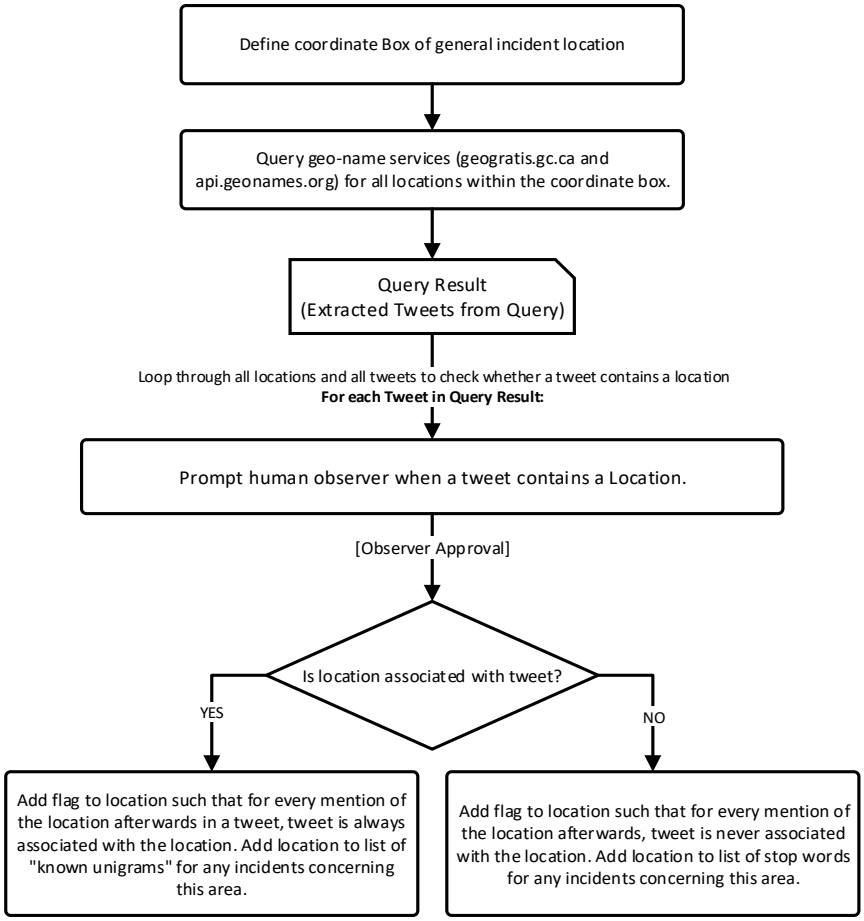


Figure 7. Box Based Location Extraction Process

solely populated places. This was done to account for nearby forest areas, as any tweets mentioning information about them, which in certain cases, would probably be important for situational awareness. All tweets that appeared to contain location names were recorded in dictionaries, in the same manner as the first process.

4.2.4 Social Media Based Features

Social Media Based Features are related to Twitter specific characteristics. Along with other features, they can contribute to predicting SA (Karimi et al., 2013a). I used four social based features: (1) **Number of Mentions:** Tweets with important information usually mention authorities or government related accounts (Karimi, Yin, & Paris, 2013b). In this attribute, I

measured the number of times an account has been mentioned in the tweet. (2) **Photo:** Sometimes people provide photos of the scene when posting about it. In this attribute, I measured if a tweet contains a photo or not. Identifying whether a photo is related or not is out of the scope of this study. (3) **Mobile:** Eyewitnesses at the scene usually post situational awareness posts with a mobile phone (Kumar, 2015). In this attribute, I measured if a tweet has been posted by a mobile phone or not. (4) **Retweeting Beyond Expectation:** For this feature, I compared retweeting behavior between emergency and nonemergency Twitter datasets. First, I divided the tweets into different groups according to their retweet count. I grouped retweet counts in a way so that the number of tweets in each group is approximately equal. This is to prevent having small amounts of training data in less frequent retweet count groups. Inspired by (Webberley et al., 2015), I created a predictive model to identify that a tweet will fall into which level, then I compared the predicted retweet count level for each tweet to its actual retweet count level. I trained and tested a labeling model (a KNN classifier since it provided the best result among others) to predict the retweet count level of a tweet based on randomly sampled data from the Twitter API. For creating a labeling model, I 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 model predicts the retweet count group a tweet will fall into. The training was based on randomly sampled data from Twitter API without any specific keywords so that the same trained model can be used for different emergency events. Then I applied the model to each emergency tweet dataset. As a result, for each tweet, there is a predicted value and an actual value. Using (1) I inferred an interestingness score for each tweet.

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

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. Table 6 summarizes the detailed features that I used for each feature category.

Table 6. List of features for identifying SA Related Tweets.

Feature Category	**Feature Name**		**Feature description**
Textual Features	BoW + TFIDF		BoW represents the text as a vector of words (Aggarwal & Zhai, 2012). TF-IDF measures the ratio of Term frequency within a tweet to the frequency within all tweets (Salton & Buckley, 1988). Each word in the bag of words feature is represented with its TF-IDF score.
Tone based Features[13]	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 from their profile page 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 relevant location mentions in the text.
Social Media Based Features	Number of Mentions		In this feature, we measure the number of times an account has been mentioned in the tweet.
	Photo		In this feature, we measure whether a post contains photo or not.
	Mobile		In this feature, we measure if the source of the tweet used a mobile device (e.g. android or iphone)
	Interestingness		Measure if a tweets get retweeted beyond expectation.

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

4.3 Test/Training Datasets

To create supervised based labeling models, I needed test/training datasets. I evaluated the performance of the classifier using four emergency tweet datasets. I collected tweets for the **Fort MacMurray Fire** and **Hurricane Matthew** events using search terms that were chosen through an initial investigation of the public Twitter stream. **Alberta floods** event dataset was also purchased. Finally, I downloaded one benchmark dataset for **Colorado wildfire** event. I removed all retweets and duplicates in the datasets as these do not provide additional information. This cleaning process is not biasing the results as it can also be automatically done in a real event. I also removed the URLs from the tweet text. I randomly sampled ~1-3% of the tweets from each set and performed a manual coding on whether a tweet is contributing to SA or not. Three volunteers (including myself and two undergraduate interns) manually labeled a random sampled set of tweets. I asked all three volunteers to label 200 random sample tweets and measured inter-rate reliability (~87% agreement). Afterwards, the three volunteers coded separate subsets. Any tweet that provided actionable insights was coded as SA. For example, the following tweets coded as SA as they provide information 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 coded as not SA. These tweets may mention the event within their text but they did not provide any valuable 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*

After manual labeling, I represented each tweet as a set of candidate features and a manually assigned SA tag that shows if the tweet contributes to SA or not. Table 7 shows the training set for each dataset and the percentage of SA vs. non SA tweets.

Table 7. Labeled set size with number of SA vs non SA tweets along with the percentage of SA and non SA tweets in the sample set.

Dataset	Sample Size	SA (%)	Not SA (%)
Fort McMurray	2911	972 (33%)	1939 (67%)
Alberta Flood	2914	567 (19%)	2347 (81%)
Hurricane Matthew	3001	869 (29%)	2132 (71%)
Colorado Wildfires	643	348 (54%)	295 (46%)

4.4 Feature Selection

I explored various categories of features to determine their impact on the classifier performance. I compared textual features, all proposed features and a subset of proposed features, which have been selected based on feature selection techniques. Feature selection is the process of reducing the number of features to remove redundant features, reduce overfitting and possibly improving the performance of the classifier. To perform feature selection, I used stepwise regression (Montgomery, Peck, & Vining, 2012) and Least Absolute Shrinkage and Selection Operator (LASSO) (Fan, 1997; Fan & Li, 2001) .

Stepwise regression (Montgomery et al., 2012) is a common technique that has been widely used for feature selection in the linear regression model and is a process of building a model by

sequentially adding or removing features based on the t-statistics of their estimated coefficients. Forward selection stepwise regression is a step by step process that starts with no feature at the beginning, and then gradually (at each step) adds a feature that makes the most statistically significant improvement in the model. The process continues until no new addition of features makes a significant improvement on the model. LASSO is another feature selection technique, LASSO performs regularization and feature selection. LASSO applies a regularization process in which it penalizes the coefficient of the regression features. Some of the features get shrunk to zero. Afterwards the features that are not assigned a zero coefficient are selected to be part of the model (Fan & Li, 2001). LASSO is generally considered a better feature selection approach than stepwise regression.

4.5 Classifier Models

The most outstanding research for extracting relevant information (e.g. events and subevents) from Twitter data is based on Natural Language Processing (NLP) and text classification (Hollander, Wolfe, & Chicken, 2013) that use classifier models such as Support Vector Machines (SVM) (Boser, Guyon, & Vapnik, 1992), K-Nearest Neighbour (KNN) (Ripley, 2002) and Logistic Regression (LR) (Ripley, 2002). In this work, I compared these three classifier models for determining SA. A comprehensive overview of techniques can be found in (Kotsiantis, 2007).

The first model is a SVM classifier with a linear kernel which is known to perform well on heterogeneous features (Kotsiantis, 2007). A SVM classifier model is a non-probabilistic binary linear classifier that was first presented in 1992 (Boser et al., 1992). SVM uses a “margin” to separate classes. Having a margin as wide as possible has proven to reduce an upper bound on the generalization error (Kotsiantis, 2007). The SVM algorithm starts by loading the training dataset, and divides it into 10-folds for the cross-validation process. The cross-validation error is calculated

for every training set of the features for 10 times and the average cross-validation error is reported. Then the optimization algorithm searches the parameter space for global minima that maximize the classification accuracy and minimize the over-fitting. Finally, the SVM model is trained with the best-optimized parameters.

Next, the KNN classifier model is an instance-based classifier which assumes that the instances within a dataset generally exist in close proximity to other instances that have similar features. The value of an unlabeled instance will be determined by observing the K-nearest labeled instances that are close to the unlabeled one. The proximity of the instances is determined by a standard Euclidean metric. We use the KNN (Ripley, 2002) classifier to find similarity between classes (relevant to situational awareness or not) based on the extracted feature set. The K similar instances are used to estimate the class. We need to choose the best K parameter that minimizes the false positive (FP) and false negative (FN) of the model (Ripley, 2002). The extracted features are divided into a training set (75%) and a test set (25%). To estimate the class of an instance in the testing dataset, the KNN algorithm compares the features to the K-nearest (in this work experiment K = 5 is the best to maximize the accuracy) records in the training set and assigns the majority class of these K neighbors as the estimated class to the unknown instance (Ripley, 2002).

The LR classifier model (Ripley, 2002) is used to model a categorical outcome variable that is modeled as a linear combination of the predictor variables. After computing a linear combination of inputs, we pass the results through a hypothesis function to predict the probability that a given sample belongs to class 0 or class 1.

For creating the classifier, I randomly sampled test and training for each dataset. The training set consisted of a set of tweets along with a manually assigned SA label. Each tweet included the tweet text and other social media and source related attributes. I transformed the test and training set to

a set of features including textual, social, location and tone-based features that described in 4.2 . I trained SVM, KNN and LR classifier models on each dataset to identify the SA tweets.

4.6 Analysis of Variance (AONVA)

Analysis of Variance (ANOVA) includes statistical models to analyze the differences between a group of means, developed by Ronald Fisher. ANOVA provides a way to perform a statistical significance test of whether the means of two or more groups are equal or not. ANOVA is similar to multiple two-sample t-tests but since it results in less type I error (Maxwell & Delaney, 2003) is suited to a wide range of practical problems. A repeated measure ANOVA is similar to ANOVA but for related, not independent groups which applies to this research as I needed to compare the means of same event datasets using different classifier algorithms and with different selected features. ANOVA uses F-distribution. The larger the F value is, it's less likely that the means between different groups are equal. For implementing ANOVA, I used R[14].

Figure 8 provides an overview of the whole process I followed to create the SA classifier.

[14] https://www.r-project.org/

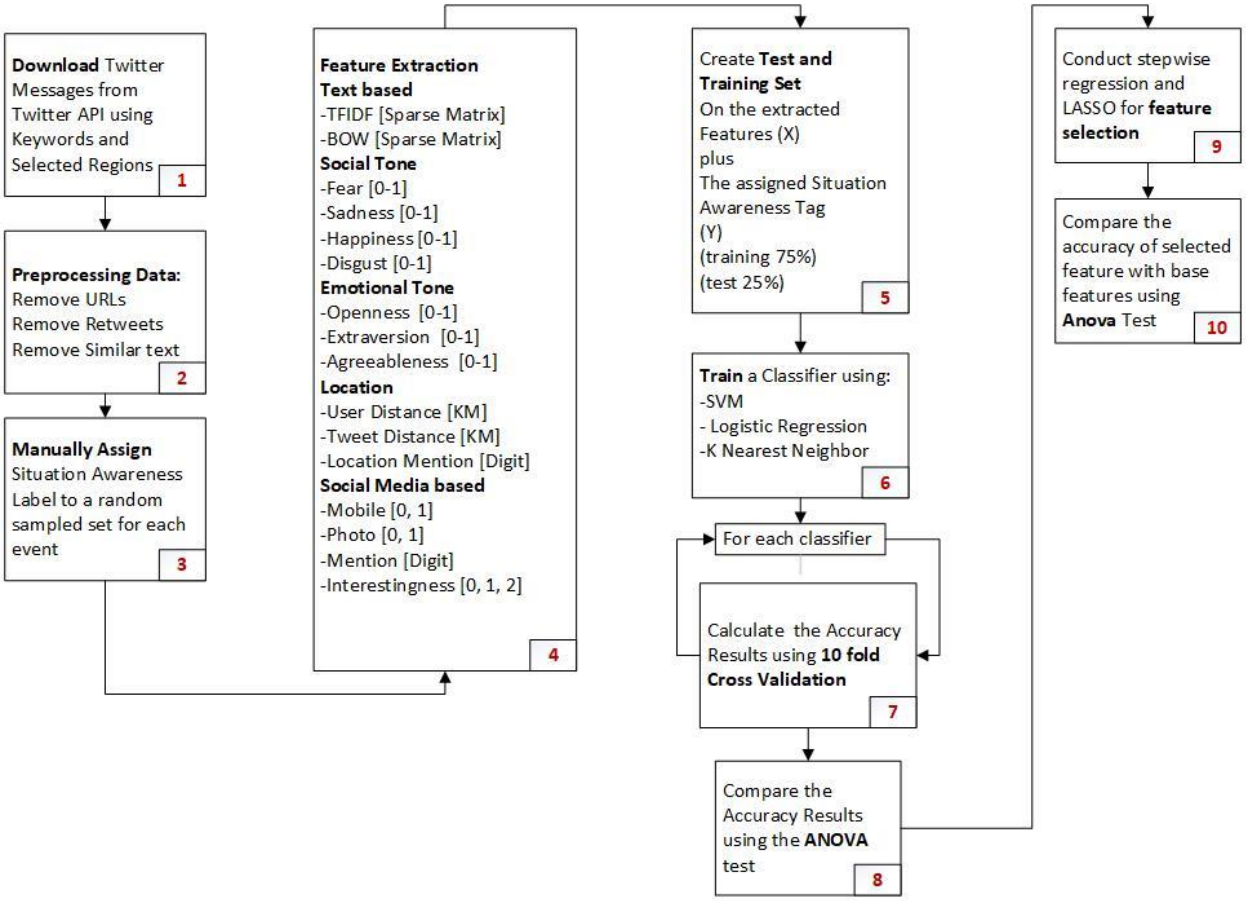


Figure 8. Classifier Process.

4.7 Results

I report the performance of the proposed features using three different classifiers. Before presenting the results for the SA classifier, I describe the accuracy results regarding extracting location based features and the predicted retweet count level (interestingness).

Location Based Features: I evaluated the location extraction process by applying it into three emergency events datasets and randomly sampled 1% of tweets from each. Table 8 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 all samples, as this was a more distributed event with several locations.

Table 8. 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)
ColoradoFires	0.76	0.92	Colorado, United States.

Retweet Count Group Prediction for Deriving Retweeting Beyond Expectation: For determining the interestingness of a tweet –as one of the input features of the SA classifier- I grouped tweets based on their retweet count (see Figure 9) and trained a classifier to predict retweet count group. I grouped tweets into 5 intervals. For example, the first bar in the Figure 9 shows the number of tweets that has been retweeted 0, 1 or 2 times. The next bar shows number of tweets that has been retweeted 3 to 10 times. The interestingness classifier determines that a tweet falls within which group.

Table 9 shows the results for SVM, decision tree and KNN classifiers. As Table 9 shows KNN provides the best accuracy results. In the next step, I applied this classifier on each emergency related dataset and based on the difference between actual retweet group and the predicted retweet group I identified three levels of interestingness for each group 0, 1, 2. Table 10 provides sample tweets for each interestingness value.

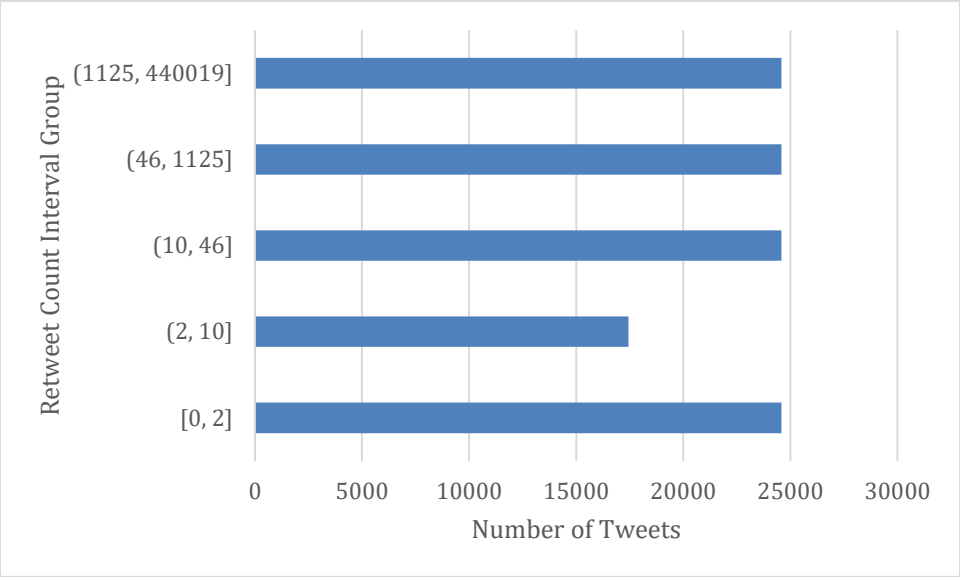


Figure 9. Retweet count group.

Table 9. 10-fold cross validation accuracy results for predicting retweet count interval.

	SVM	Decision Tree	KNN
Accuracy	0.66	0.68	**0.89**

Table 10. Sample tweets for each interestingness value.

Tweet	Interestingness
• Thank you @etalkCTV for your thoughts and mention of #ymmfire ♡ • I will not complain about the rain after seeing the devastation in Alberta #ymmfire	0
• @RMWoodBuffalo @redcrosscanada I'm trying to register with Red Cross 1-888 number. Registration available online? • If you are heading to Cold Lake my home is open #ymmhelps #ymmfire #ymm	1
• looking for families staying at @Northlands who would prefer to be in a home. 2 bedrooms available in my house in Sherwood Park #ymmfire • If you are an evacuee and require relocation for your horses, please contact us at 877-463-6233. #FortMacFire	2

Classifier Results: I trained three classifiers by extracting all the features. In the following I compare the results on the performance of four categories of features (1) Text: using only textual features, (2) All: using all proposed features discussed in section 4.2 (3) Stepwise: using features

selected by stepwise regression. (4) LASSO features based on LASSO feature selection. I trained three classifiers for each category of features. Figure 10 and Figure 11show the cross-validation accuracy and f1 results for each classifier and each emergency event dataset.

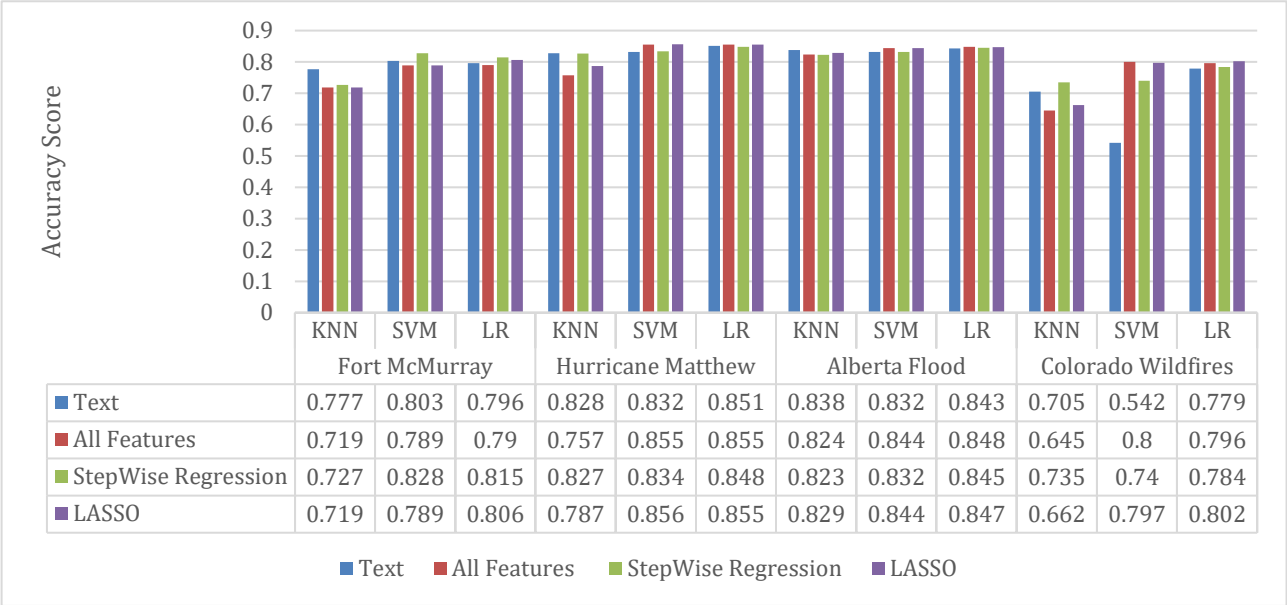


Figure 10. Cross Validation accuracy results for each event and each feature category.

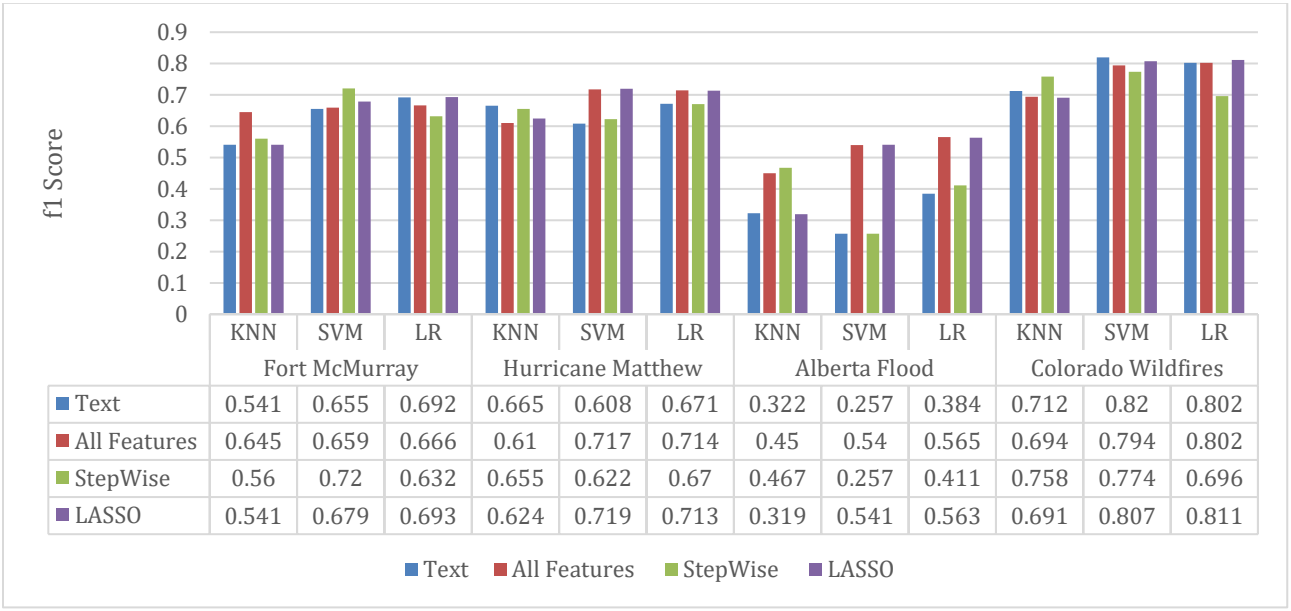


Figure 11. Cross Validation F1 results for each event and each feature category.

Figure 12 - Figure 15 shows box plot charts for 10-fold cross validation accuracy for different feature categories and for each dataset. X axis shows the result for each algorithms and y axis shows the accuracy results.

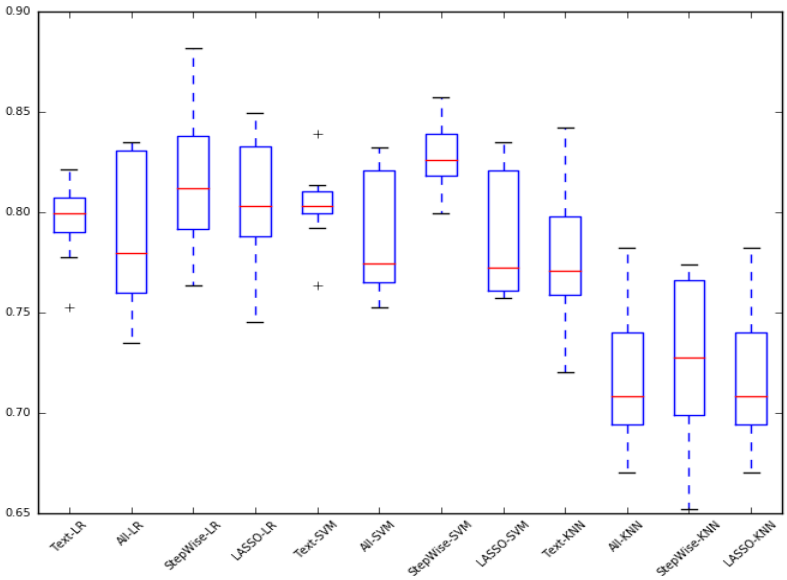


Figure 12. Box plot for 10-fold cross validation accuracy scores for Fort McMurray dataset. Each box plot shows the results per feature category and per algorithm.

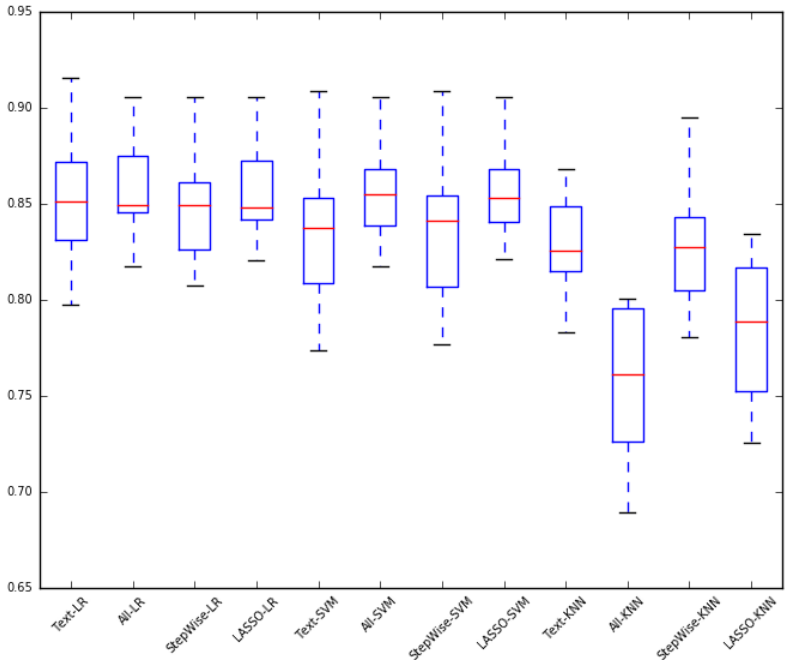


Figure 13. Box plot for 10-fold cross validation accuracy scores for Hurricane Matthew dataset. Each box plot shows the results per feature category and per algorithm.

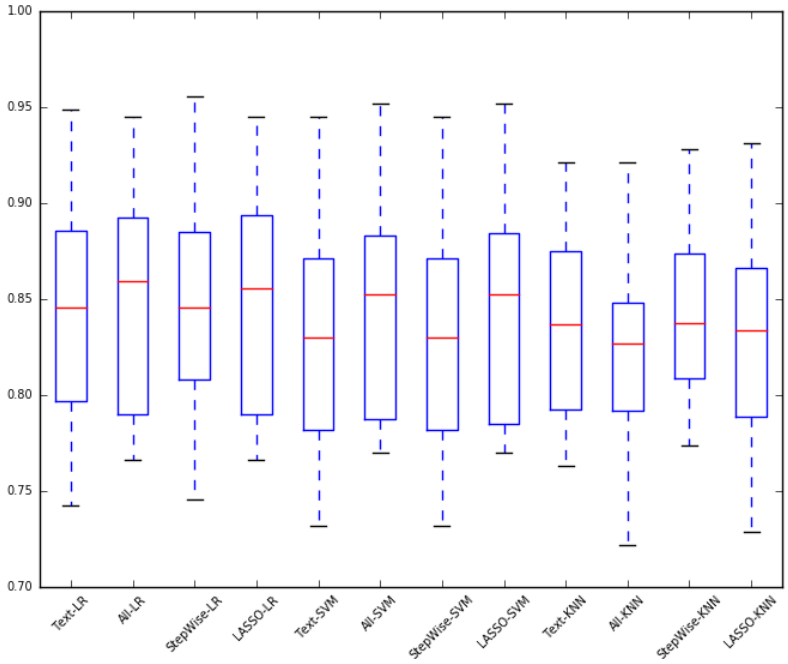


Figure 14. Box plot for 10 fold cross validation accuracy scores for Alberta Flood dataset. Each box plot shows the results per feature category and per algorithm.

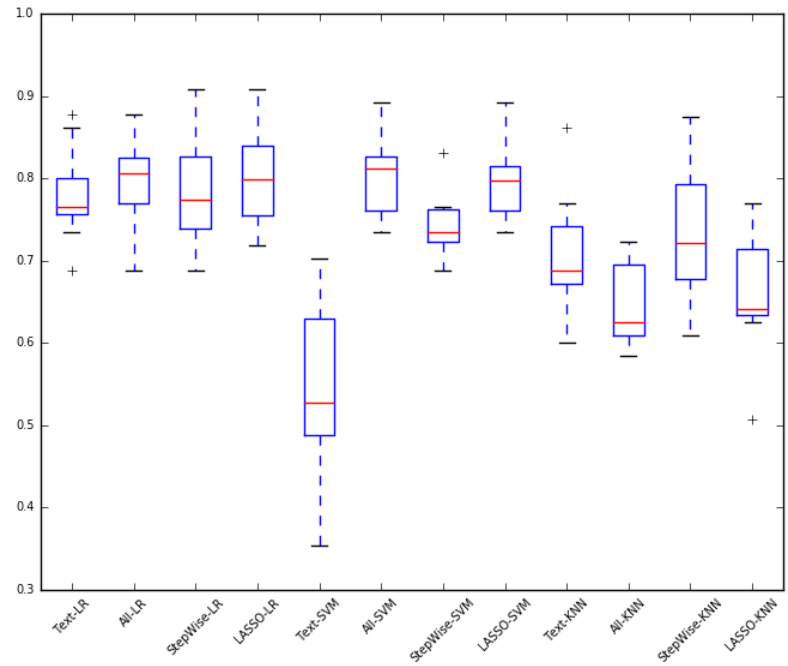


Figure 15. Box plot for 10 fold cross validation accuracy scores for Colorado Wildfire dataset. Each box plot shows the results per feature category and per algorithm.

I conducted repeated measure ANOVA test (Armstrong, Slade, & Eperjesi, 2000) for each dataset to determine if there is a main effect of selected features (Text, All, StepWise or LASSO) or selected classifier algorithm (SVM, KNN or LR). Table 11 shows a summary of results.

Results show that For Hurricane Matthew, and Colorado Wildfires there is a main effect of either the choice of features or the choice of algorithm on accuracy. Likewise, there is a significant interaction effect between the choice of features and algorithms for these two events. For Fort-MacMurray, there is a main effect of the choice of Algorithm on the accuracy and also there is a main effect on the interaction of the choice of features and choice of algorithm for this event.

Table 11. Pairwise comparison of statistical difference between feature categories and classifier algorithms.

	Algorithm effect	*Feature effect*	*Algorithm-Feature effect*
Fort McMurray	***F (2, 108) = 28.671 p-value<0.001***	*F (3, 108) =1.177 p-value =0.322*	***F (6, 108) =5.509 p-value <0.001***
Hurricane Matthew	***F (2, 108) = 49.426 p-value < 0.001***	***F (3, 108) = 2.741 p-value= 0.04***	***F (6, 108) = 3.027 p-value= 0.008***
Alberta Flood	*F (2, 108) = 0.454 p-value = 0.636*	*F (3, 108) =0.012 p-value= 0.998*	*F (6, 108) = 0.161 p-value=0.986*
Colorado Wildfires	***F (2, 108) = 24.792 p-value <0.001***	***F (3, 108) = 9.709 P-value <0.001***	***F (6, 108) =13.760 P-value <0.001***

4.8 ROC Curve

To compare the performance of various features and algorithms we cannot only rely on the accuracy score. To know that how much discrimination power the proposed features have I generated Receiver Operating Characteristic (ROC) curve. ROC curves plot false positive rate on the X axis and true positive rate on the Y axis. The closer the ROC curve is to the upper left corner, the higher the overall accuracy is. Figure 16- Figure 23 shows the ROC curve for four events for SVM and LR classifiers. Each chart shows the curve for four categories of features. We can see that in the Fort MacMurray and Hurricane Matthew events eithers features selected by stepwise or

LASSO produce better results. In Alberta Flood, using all proposed features produce better results and in Colorado Wildfire event, using textual features produce the best results.

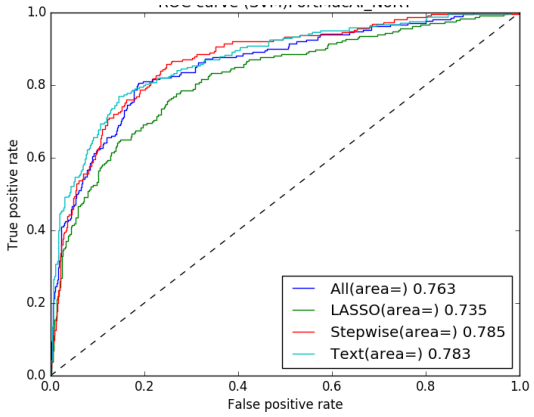


Figure 16. ROC Curve for Fort McMurray using SVM.

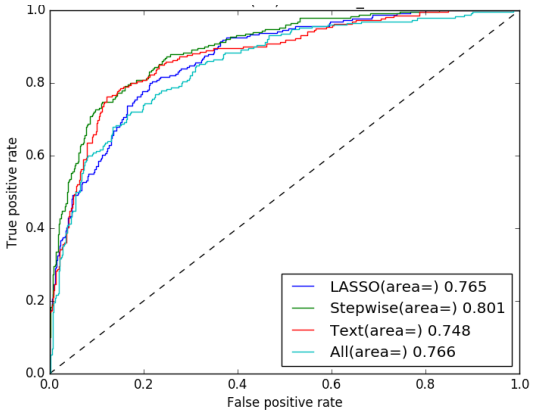


Figure 17. ROC Curve for Fort McMurray using LR.

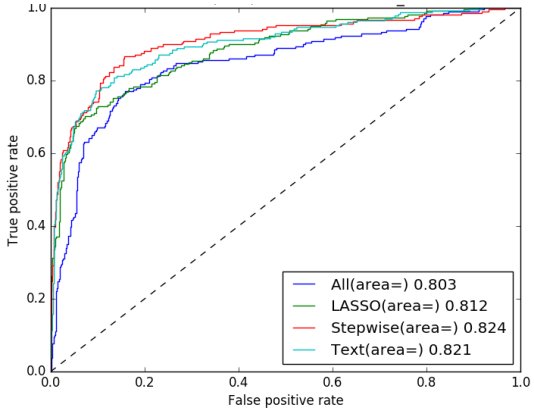


Figure 18. ROC Curve for Hurricane Matthew using SVM.

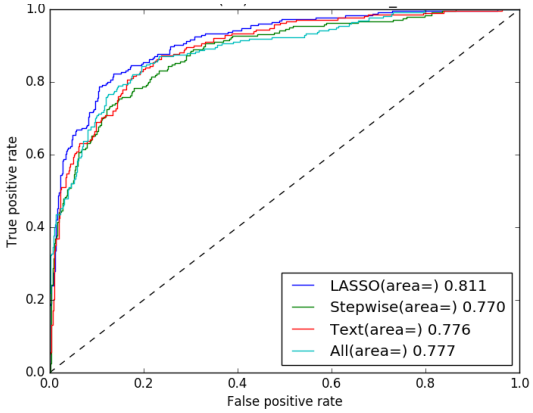


Figure 19. ROC Curve for Hurricane Matthew using LR.

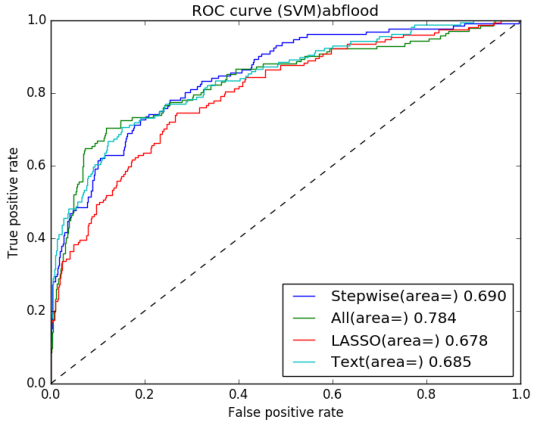


Figure 20. ROC Curve for Alberta Flood using SVM.

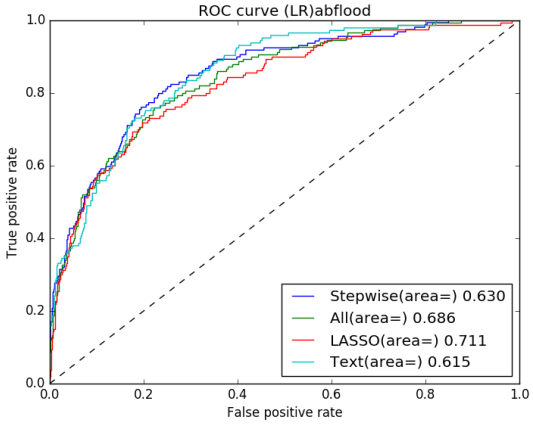


Figure 21. ROC Curve for Alberta Flood using LR.

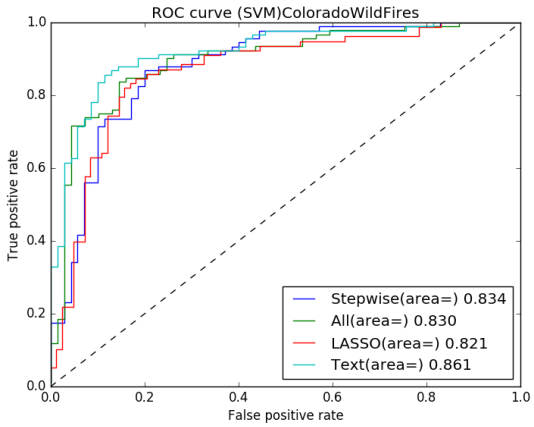


Figure 22. ROC Curve for Colorado Wildfire using SVM.

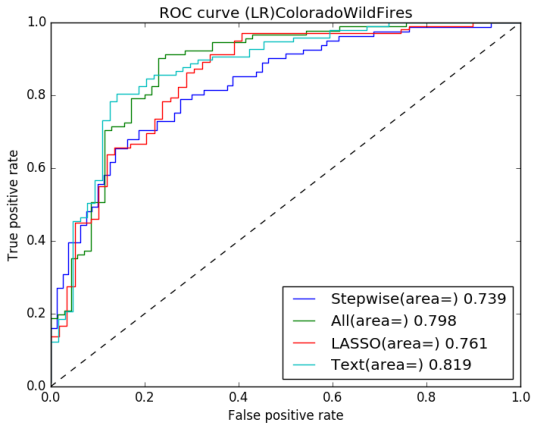


Figure 23. ROC Curve for Colorado Wildfire using LR.

4.9 Chapter Summary

In this chapter, I discussed the creation process of a machine learning classifier to identify situational awareness tweets using textual, social, location, and tone-based features. I compared performance when selecting textual based features, all proposed features or a subset of features using two feature selection techniques (Stepwise and LASSO). Results show performance varies in different events and different classifier algorithms. The results show that features contributing to situational awareness can vary based on the type of event, specific conditions for each event and geographical regions. For example, people in different geographic regions have different attitudes when using social media. For example, based on the Alberta Flood 2013 dataset, people actively posted details of water level in different regions so the social media posts contributing as small pieces to create a big picture of what is happening in the area. On the other hand, in an event like Philippines Flood people were trapped in buildings and social media posts were full of urgent help requests including the locations of people in need. This further indicates that training should happen in real time and we can mitigate these differences when bringing human in the training and classification loop.

Chapter 5: Design Process

After the initial interview phase, I 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. A scenario-based design approach also helps with better illustration and communication of the design goals to practitioners and thus receive constructive feedback. I focused on designing a prototype to help develop interactive filters using the analyst's knowledge and machine learning algorithms. The design process proceeded with producing low-fidelity paper prototypes. I presented these to social media analysts to gather their reaction and comments. In this section, I describe the low-fidelity prototypes, feedbacks from analysts, and the design considerations.

5.1 Example Scenario

I designed the early prototypes based on a flood scenario in Calgary. The flood scenario is inspired by the actual flood that happened in Calgary in 2013. In late June 2013, there was a flood in Calgary, Alberta that caused an estimated $5 billion worth of damage. In the following, I describe the scenario assuming a hypothetical social media analyst (John), from the city of Calgary, has been tasked with monitoring tweets that are related to the flood. John has been tasked with identifying detail information that people provide regarding themselves or other people (e.g. help requests and needs), properties (e.g., damages), and the environment (e.g., level of water). Another objective for John is to gather information regarding a rumor that says the water in the city is not safe to drink.

At the start of the flood, John, wants to get a sense of what is going on in social media. He logs into the system to monitor and analyze the social media posts. In the main view, John sees a stream view of the social media posts and a map view. In the stream view, John defines a new stream and names it: "Calgary Flood". He defines a filter for the stream by dropping down a basic

filtering view. In the filtering drop-down menu, John defines a set of keywords related to the flood, sets the location (Calgary), and time (last 6 hours) to pull information from Twitter (Figure 24). Hitting on the search button in the basic filtering view will result to the main view.

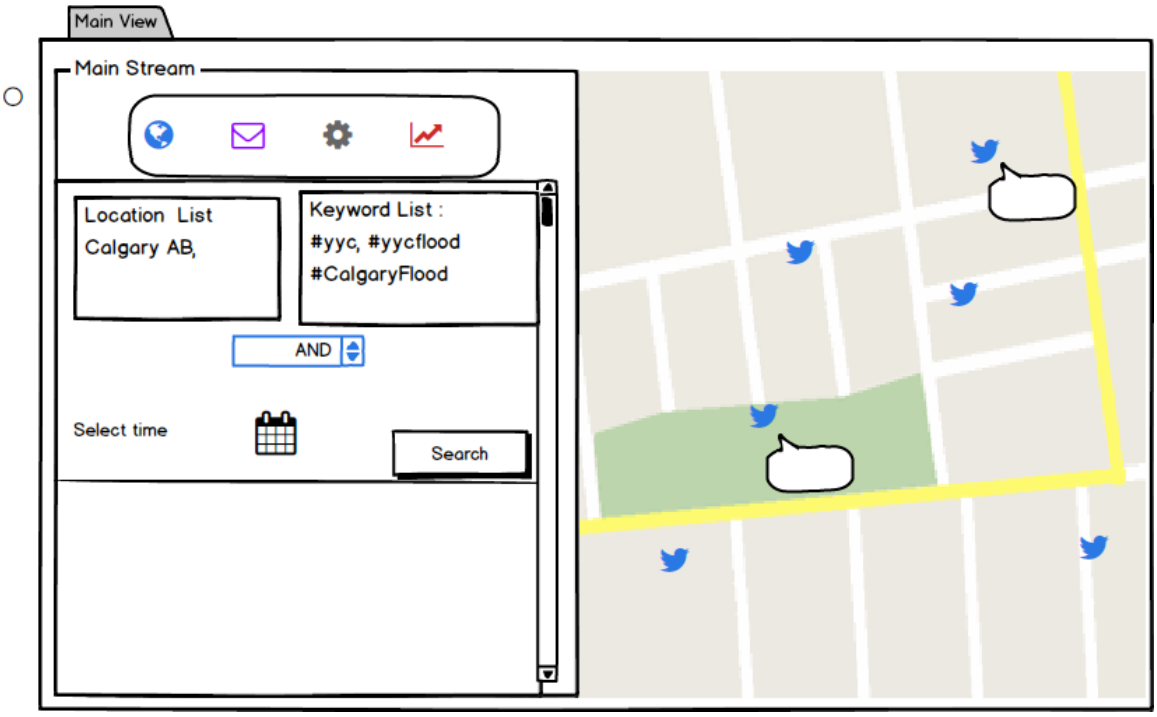


Figure 24. Basic filtering view- By selecting a set of keywords, list of locations, a time range and clicking on the search button, John can create the main view.

The resulting main stream (Figure 25) in the main view displays all the posts that have passed through the filtering menu by querying the Twitter API. John sees different information about the posts including user profile ID, SM source, text, and time. John sees the pictures and videos of posts in a small size beside the text. The main view also includes a map visualization to show geo-located tweets.

John sees a tweet that mentions that floodwater is almost at the top of one bridge in an area that he knows is vulnerable to the flood. John taps on the user id and is presented with a popup showing the ratio of followers to followings for the user, how long the account has been registered, the user's location, and the fact that the user is not verified. Based on this information, John decides

to send the tweet to the wall display. John identifies a new account, which is related to the City of Calgary and is being followed by the mayor and Calgary Police. John adds the account to the list of verified users.

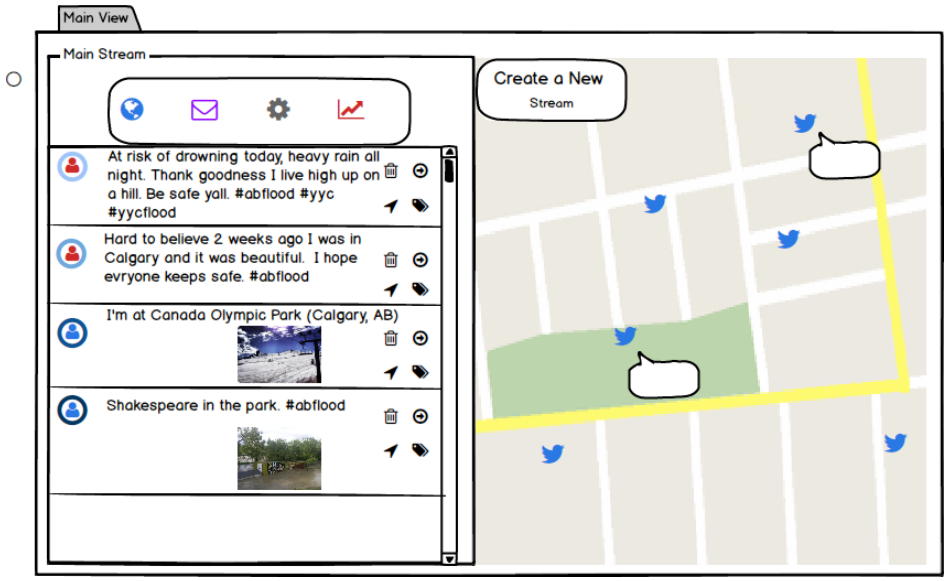


Figure 25. Main view- Based on the basic filtering menu, a query will be sent to the Twitter API and will result in a stream of tweets in the Main View.

Each user icon in the stream view has a color that determines the reliability level of the user. The circle around the user icon is color-coded to determine the proximity level of the user to the crisis region. Figure 26 shows the characteristics that are being used for determining reliability and proximity.

John knows there is a rumor that says Calgary's water is not safe to drink. In the main view, John creates a new stream and names it *WaterRumor*. In create new stream window (Figure 27), he selects filtering type (misinformation) and adds a list of keywords (Water, Drink, Flood) and spatiotemporal information (Calgary, Starting from June 19th, 2013). If he wants to follow specific users for this stream, he can add them to the user list.

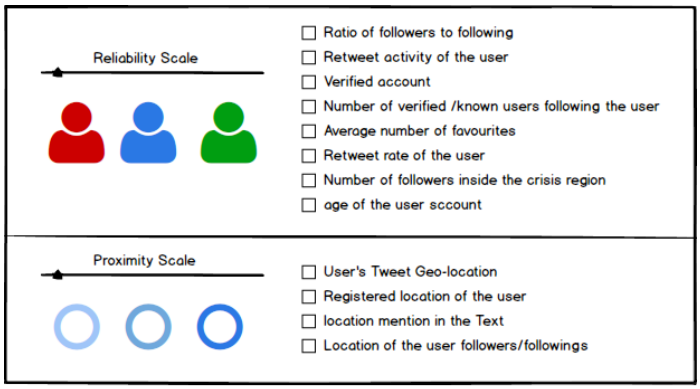


Figure 26. User characteristics for determining reliability and proximity.

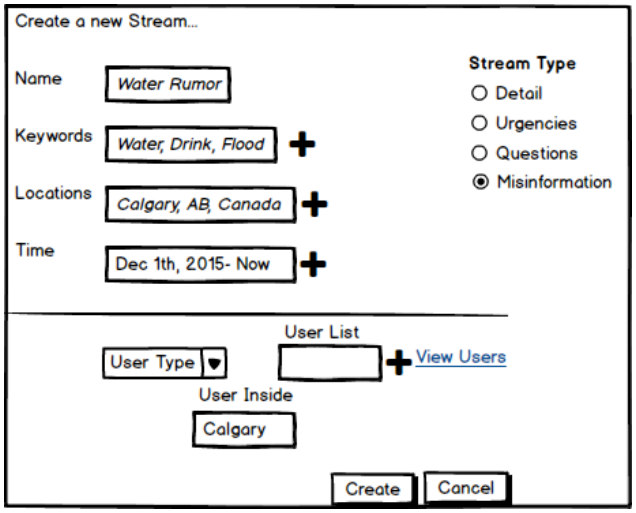


Figure 27. Create Water rumor stream.

After creating the stream, John sees the *WaterRumor* stream (Figure 28). John has some insights about what are the characteristics that are common in the messages of that filter type, so he selects those characteristics (I refer to them as features). For example, are the messages mentioning a location? (Not in this case) Do they contain links? (Not in this case). John selects the features in the advanced filtering menu. A complete list of features is presented in Figure 29. After determining advanced filtering options, John manually selects the initial incoming posts that are related to the water rumor and assigns them to the *WaterRumor* stream. Afterwards, the system decides for the rest of the incoming messages.

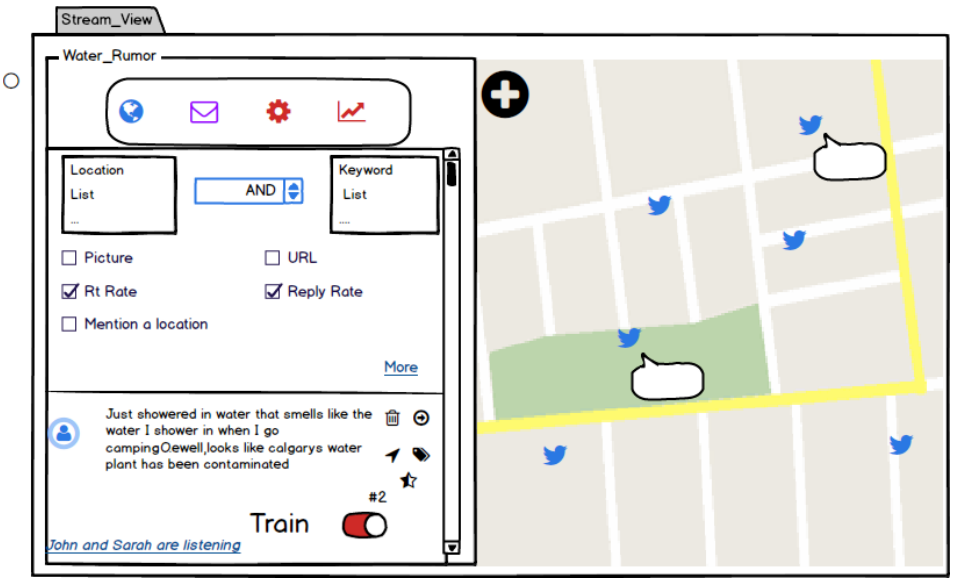


Figure 28. Water rumor advanced filter.

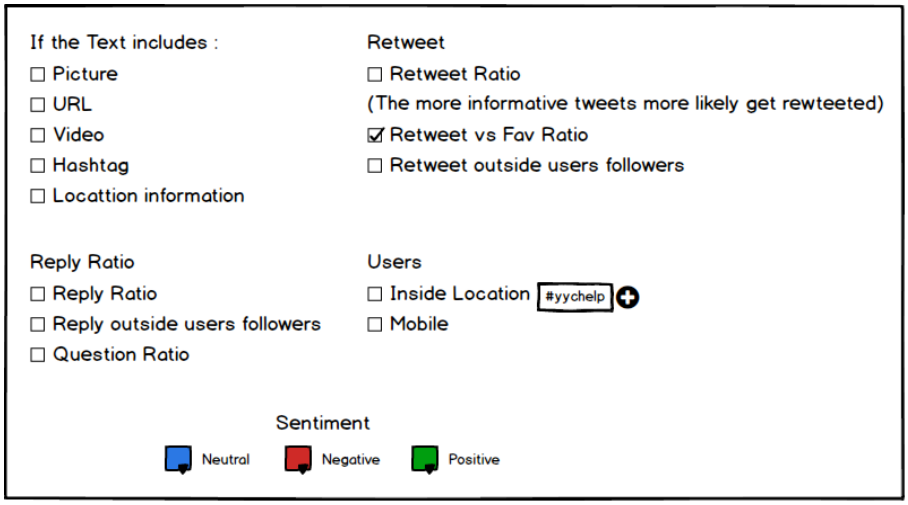


Figure 29. Feature list.

Whenever John sees any wrong label for the tweets, he selects them and provides feedback to the system. John will repeat this manual process until he is satisfied with the results (Figure 30).

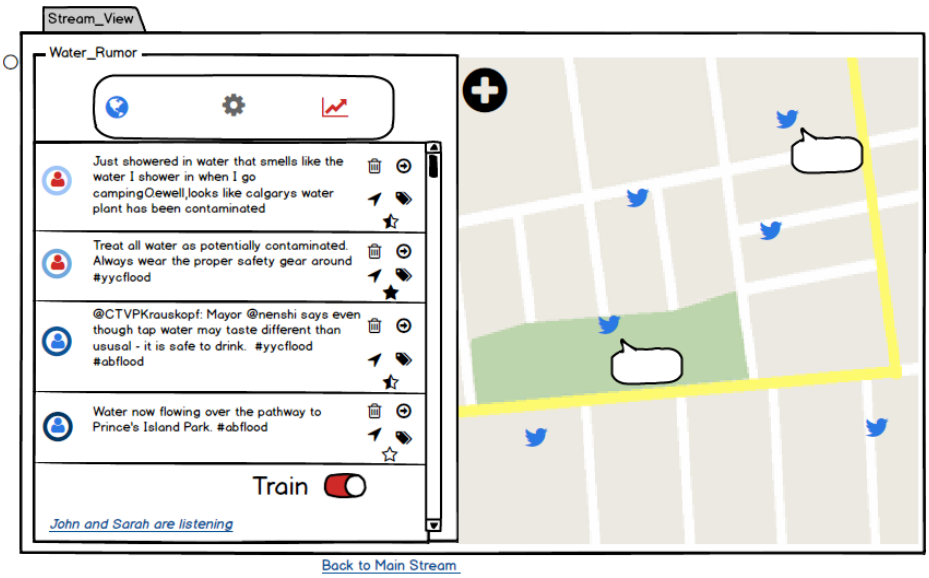


Figure 30. Water rumor stream view.

By manually exploring messages, John realizes people are sending information about flood-affected areas (e.g., pictures of different locations that are flooded in order to show affected people, damages, and level of water.). John defines a stream to identify these informative messages so that he can provide additional information to the EOC common operating picture (Figure 31). He names the stream "Detail". In the drop-down filtering menu, he selects the filter type (Informative Messages), a list of keywords (flood, yycfloods, yycflood, yycflooding) and spatiotemporal information (e.g., Calgary, Starting from June 19th, 2013).

John selects the features that he thinks are common among informative messages in the advanced filtering menu (Figure 32). For example, are the messages usually mentioning a location? (In this case, yes), Are they usually containing a link? (In this case, yes).

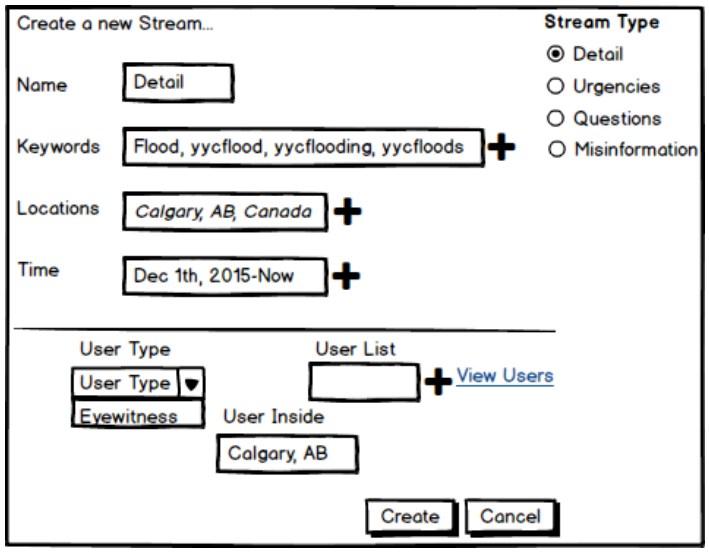


Figure 31. Create a detail stream.

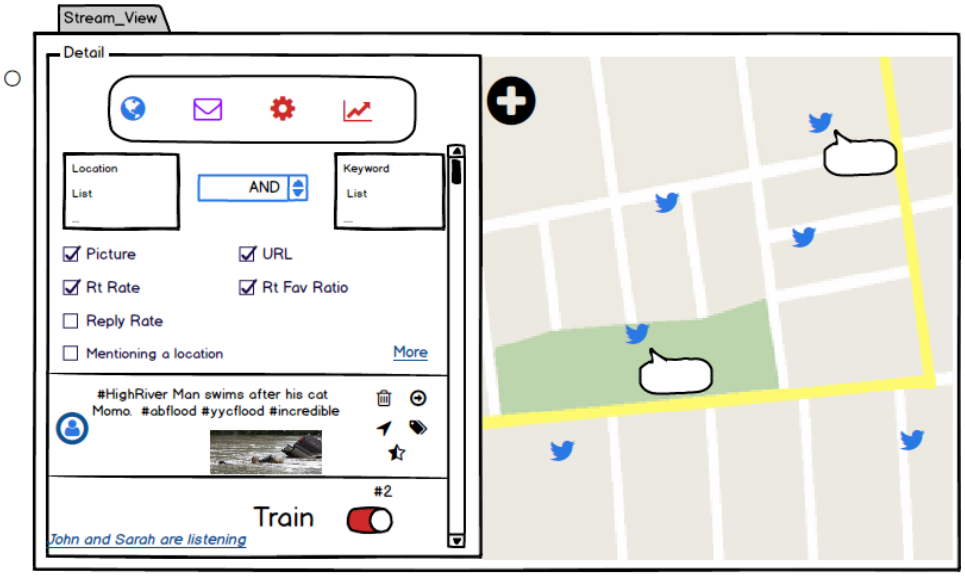


Figure 32. Detail stream advanced filter.

John manually selects the initial informative incoming posts, and the system then decides for the rest of the incoming messages. Whenever John sees any wrong filtering, he selects them and provides feedback to the system. John will repeat this manual process until he is satisfied with the results. Figure 33 shows a view of the detail stream while tagging button is selected.

Figure 33. Detail stream view.

John realizes that there are many tweets that are not geo-tagged but are mentioning a location within their text. He enables a location determination mode. The system will identify if the messages are mentioning a location and will display them on the map. John can distinguish the geo-tagged messages from the messages that are mentioning a location on the map view by their icon on the map view.

5.2 Feedback on Low Fidelity Prototype

I conducted feedback sessions with social media analysts from different emergency-related organizations to gather their feedback on the design ideas and prototypes. I conducted semi-structured interviews with six participants. I contacted different emergency organization and recruited three analysts from City of Calgary and Calgary police, a communication manager from the private sector, and two more analysts from the private sector. 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, I gave the participants an overview of the prototypes, then proceeded based on the design scenarios to collect their feedback

regarding the tool. I walked through the design with them. I wanted to understand their thought process so I asked them to think out loud (Lewis & Rieman, 1993) as they examined the prototype. I also audiotaped the sessions.

It was important to see if the proposed design matched their workflow and expectation, especially during a disaster. I transcribed the recorded audios. The transcribed notes formed the qualitative data. To analyze the notes, I followed a process inspired by grounded theory (Glaser & Strauss, 2009). Grounded theory is a systematic method for analysis of qualitative data. Grounded theory involves iterative coding of concepts (open-coding) and finding patterns apparent in the text (axial coding) to form typologies. Coding means labeling or categorizing collected data. Open coding involves reading through the data several times and creating preliminary labels. In axial coding, we create conceptual groupings from the labels.

5.3 Findings

In this following, I present findings of the feedback received on the early prototypes based on the extracted coding categories. I give examples of descriptions of these categories in the words of the interviewees themselves.

5.3.1 Selecting Keywords vs Geographic Boxes

One of the first parts of the prototypes was regarding data collection from Twitter. Posts can be collected using a combination of keywords, geographical areas, and list of users. However, most of the participants considered keywords as a primary source of data collection, because as one of them stated, by relying on geotagged based content they may lose information. One participant, initially was thinking about having a circle or box on map for pulling geotagged tweets but then he thought that some people might see an incident and then move outside of the incident area but

they still might tweet about it: *"Sometimes people are tweeting once they've left that box, so maybe they were there [when] the flood is coming through elbow park, they saw it but they were at the bus on that time now they are home, they are tweeting pictures of the flood in elbow park, but they are outside of the area I've searched in, so in some aspects for a location I don't wanna be too specific, I'm looking for flood and elbow park."*

When navigating through the prototype, participants mentioned some challenges they are currently facing when gathering data for an emergency event. For example, using general keywords to pull data could overwhelm them with a lot of *garbage* content. That's why they need more specific keywords, *"I feel like the keywords are usually the most important thing for me but the keywords that are pulled [automatically] are never the ones that we care about. It's always like yyc, of course they gonna tag their stuff Calgary but that's gonna provide zero value for us,"* she added, "*we had tweets about an underpass that is being flooded and so although keywords about that underpass are not gonna make in the list of top keywords just by sheer volume, it was among the biggest topics that was discussed.*" One participant wished that if they could get the messages based on the keywords that had been indicated by user. *"I tag this message 1st street underpass and then it pulls based on my tag."*

However, on the benefits of using general keywords, one participant mentioned that they have to use general keywords when they don't have any idea about what topics they are looking for. For example, when an eyewitness reports an event in social media it does not have any specific hashtag, they may only add some general hashtags regarding the city name (e.g., #yyc) thus having these general hashtags is useful for them when pulling information. One participant mentioned a case that an incident happens in the train and the first people on the scene do not use specific hashtags but they use general hashtags, such as #yyc, to report more details and information.

5.3.2 Posts View

The current practice of social media analysts includes a lot of manual handling and reading social media posts. They expected that the proposed design for each stream would help them increase their task performance. For example, they wanted to have the ability to differentiate between read and unread posts and have the search option for read/unread tweets in the advanced filtering menu. The delete icon for each post confused some participants. Two of them asked is this going to delete it forever or just archive it? They liked the archive option because *"I know that I read and I've dealt with it but I still could get it back from somewhere"* One participant suggested that a when tweet is forwarded to another stream, it should be removed from the original stream *"then I know as I know going through main stream I don't see it anymore and it's being assigned to another stream that a colleague maybe responsible for."*

5.3.3 Manual Mode

Participants liked to have a social media stream that is populated manually by the analysts. All social media analysts among participants considered having a manual stream very useful for their workflow. One of them also saw it as a To-do list and further suggested, *"Then maybe you should gray those setting and filters for that manual stream."*

5.3.4 Ability to Make Internal Comments

Public organizations like city municipalities and police departments have a social media team for handling their social media communication. As mentioned in the Chapter 3: analysts need to collaborate using the tool. I had further discussions with analysts in this regard. According to the prototypes, users can add notes to each post. Three out of five participants found these internal notes very useful. One mentioned he would like to create internal notes for future references. For

example, he mentioned a case in which people might ask what is a helicopter doing in their neighbourhood. He might not be able to update the public right away but he can create an internal note attached to those inquiries. He added, "*If all of sudden 100 people saying what the helicopter is doing right now, I've already saved that note, attached it for myself and my colleagues to see should we choose to engage for that post.*"

In the stream views, each user could subscribe to each stream. The idea sounded challenging for two participants. They were worried what if they tagged the same posts, *"Does system support collision avoidance?"*.

5.3.5 Advanced Filtering

The training capability was the most controversial part of the tool. By training, I mean both on the relevancy part and on the tagging. There were different opinions regarding the usability of such filters on the early prototypes. There was a different perspective between participants from private organizations vs. participants from public ones. As one participant mentioned, the private sector does not have enough resources to assign to social media tasks. *"In most nonpublic organizations if you got one person monitoring social media, that's great."*

Another concern that a participant from private sector 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."*. She asked what circumstances would be worthy to set this up and use it? She explained, "*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 problem towards advanced filters was that it could be labor intensive, especially for private sector,

as they do not have enough resources and they usually cannot project early on in the disaster scenario if the situation is going to persist long enough. *"What concerns me is that I'm in a crisis now I don't have the time to do that… this should have been done weeks ago…It would be easier to ask two persons to look at tweets. You are asking me to make a judgment call on a situation that I have no ability to that."*

However, participants related to public emergency organizations such as EOCs had 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 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 trained filters to misidentify a post. He suggested that the system be "untrained" or corrected, "retrain" so that the same error (according to the analyst) would not appear again. Another suggestion based on lack of trust to automated filters was made so that they could differentiate what is being tagged/labeled by the human intelligence or machine intelligence. A similar suggestion was made to differentiate between the geotagged posts and the one that has been geotagged based on their textual content.

Another question was raised about how to change the filter to make previously unimportant information relevant again when the topic changes. In other words, what if the interest of the user changes through the event. What if a topic that was previously unimportant starts to grow, how can the training be undone when that happens? One stated, *"If I flag something as not interesting and now it's starting to grow you need to tell me."*

In the proposed design users could rate information of a post by assigning a star to it or tag posts based on different information types. During the interviews, participants mentioned types of tags they would like to assign to social media posts. Participants from the public organizations

mentioned types such as mainstream media, public comments, inquiries, reporting, reporting that they should report to 911, reporting that they should report to someone else, complaints, compliments, and not relevant. One mentioned that tagging should be standard between all users when it comes to custom tagging, *"What if I put irrelevant and my colleague put not relevant. It should be for all users and all streams."*

5.3.6 Source Reliability

Obviously, reliability and trust issues are hot topics when it comes to social media. During the requirement elicitation phase, I discussed the requirement of practitioners in regards to the reliability of social media information. In the interviews, I received detailed feedback for the participants needs regarding assessing source reliability. Participants usually relied to the authors' accounts to gain reliability of the content. That is why *a user management piece* was missing according to some 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 number of followings would be helpful. They would also like to know if the person is linked to any other person in their network, "*Is there an influencer that I know following this person?"* They would like to 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 to 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.

One analyst with a lot of experience in emergency management was more caution towards using credibility levels for users. While he expressed its usefulness he wanted to follow a whitelist approach because he could not completely trust it, "*I would think that yeah certainly if I was going through a long list of data anything that would show up as green then at least kind of give me a place to start"* However, He was not willing to discard authors that have been identified as unreliable, *"Cause maybe they didn't hit high enough on some of the criteria but they have very relevant info for me."*

5.4 Design Considerations

Based on the feedback on the low fidelity prototypes I made changes to the initial design. I noticed that participants had difficulty with selecting the right keywords for stream views and they could not only rely on general keywords. Hence, I made substantial changes in the data collection part. A summary of interview study findings is represented in Table 12.

It was observed that there are some resistance and mistrust of using automated filters. To address this, I ensured that the tool could work in a manual as well as an interactive filtering mode and allow analysts to distinguish between machine learning and human labels. Some participants mentioned about the overload of training the advanced filters. I changed the design so that the training would be as simple as possible. More flexibility in assigning tags to social media post was provided. The design changed to allow assigning a situational awareness tag to each post. In the next chapter, I discuss the detail of the final prototype.

Table 12. A summary of interview study findings.

Aspect	Summary
ADVANCED FILTERS	• Need for manual mode • Retraining the classifier • Identifying relevant keywords
DESIGN CONSIDERATIONS	• Facilitate reading of posts • Support Collaboration and Communication
PUBLIC VS. PRIVATE ORGANIZATIONS	• Public org. were interested more in individual messages • Private org. were interested in an overview and the volume

TRUST ISSUES	• Aggregated information about users
	  • Verified account
	  • Followed by known users
	  • Number of followers / following
	  • Number of posts
	  • Number of re-shares
	  • Age of account
	  • Profile pictures

5.5 Chapter Summary

In this chapter, I elaborated on the design process, the low fidelity prototypes, and the feedback received from the emergency practitioners and social media analysts. I noticed participants had difficulty finding sub-events during an emergency, and they were confused with having multiple streams to gather data. They also had doubts regarding training advanced filters as they do not know whether the initial effort worth it or not. Based on their feedback I made changes in early prototypes and developed a fully functional prototype.

Chapter 6: Social Media Analyst Responding Tool (SMART)

Using the analysts' and practitioners' feedback, I evolved the early prototypes to a fully functional web-based tool. In this chapter, I describe the final developed version, along with details regarding the implementation and architecture of the proposed tool. The development of a web-application component of the tool and a conducted case study to demonstrate the feasibility and applicability of the tool are done collaboratively with the help of two undergraduate interns.

6.1 SMART Prototype Design and UI

I developed Social Media Analyst Responding Tool (SMART), a fully functioning web-based prototype. This section describes the design of the SMART application. The SMART interface combines the ability to collect, search and sort tweets alongside the functionality to filter and organize them with machine learning techniques. In this chapter, I describe the core parts of the design in more detail.

6.1.1 Event Collection

I made substantial changes to the early low fidelity prototypes in regard to data collection. The changes were both on the terminology and the User Interface to make the data collection part more intuitive and provide views for each collector. An essential feature of a social media monitoring tool is to be able to collect social media posts while an event is occurring. SMART utilizes a set of data collectors to provide historical and live data gathering (Figure 34- A).

The collectors contain all tweets related to one main event. Upon the start of using the tool, a user must create an Event Collection with keyword-defined collectors to gather tweets. Each event collector has a time range window to track when the collector has been active. Figure 35 shows a historic collector with *wildfire* keyword that gathered tweets from August 22nd to

September 5th. SMART extracts *Topics* for each event collection (Figure 34-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.

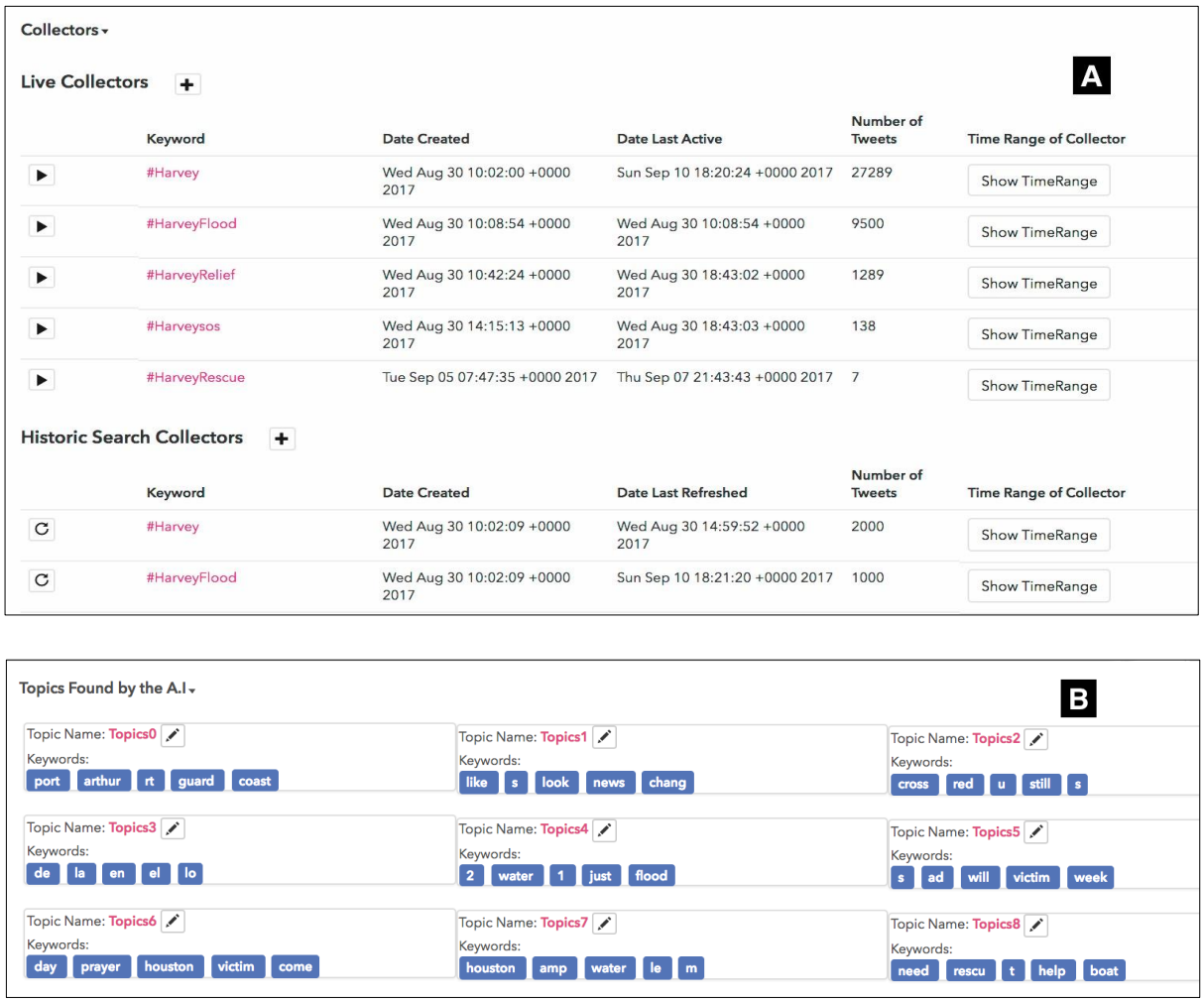


Figure 34. A-shows a list of live and historical collectors for hurricane Harvey (the top image). B- a list of topics regarding the event (the bottom image).

Figure 35. Time range for each collector

6.1.2 Interactive Filtering

Users can further narrow down their search by creating Filtered Events (Figure 36). A filtered

event can be populated based on the tweets that analysts manually forward to the filter (I refer to it as manual filtered event) or automatically based on keywords. Figure 36 shows a filtered event view. The tweet and the buttons below (Figure 36-J) allow quick interactions for labeling and reading tweets at once. The color of the border around each tweet displays the sentiment of individual tweets. Figure 37 provides a closer view of the tweets in the filtered event page.

From interviews, I 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. My 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. Any automatically assigned label can also be changed by the analyst if necessary. I embedded the following three machine learning classifiers to facilitate interactive filtering.

1) Filtering by Situational Awareness: I embed a situational awareness classifier in SMART. In the interface, situational awareness is tracked by a lit or unlit lightbulb symbol for each tweet, allowing people to make quicker decisions during a time critical situation. If the classifier predicts situational awareness, the lightbulb symbol will have a red background (Figure 37) to allow analysts to distinguish between the classifiers' output and manually labeled tweets. Details about the situational awareness filter described in the Chapter 4.

2) Filtering by Tagging: To facilitate categorization of tweets into user-defined labels, I embedded a multinomial Naïve Bayes (Murphy, 2006) classifier using basic textual features (BoW+ TFIDF). A Naïve Bayes classifier is a probabilistic classifier that applies Bayes' theorem. More detail can be found in (Kevin, 2012). Bayesian networks are known for their short

computational time for training (Kotsiantis, 2007). In the SMART design, 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 red in color (Figure 37) and can be changed if the tag is deemed inaccurate by the user.

3) Filtering by Retweeting Beyond Expectation: To help analysts order the tweets based on their level of importance and not just based on the chronological order, I assigned an interestingness level to each post. I relate the interestingness of a tweet by its number of retweets which means the more a tweet gets retweeted there's a higher possibility that the content of a tweet is interesting to authors (Webberley et al., 2015). According to the exploration of emergency and non-emergency tweets, retweeting increases during emergencies (Marina et al., 2015). I attempted to identify an interestingness score in emergency context using this behavioral change. The interestingness rate is calculated based on if a tweet getting retweeted beyond an expected level. Identifying interesting tweets can help responders quickly identify pieces of information that many people have retweeted and thus be able to identify misinformation. SMART lets users sort tweets based on their interestingness. More details regarding identifying interestingness of a tweet provided in the Chapter 4.

The interactive filtering target users should train the classifiers by labeling tweets. There are easily accessible buttons which allow the user to train both tagging and situational awareness classifiers (See buttons below each tweet in Figure 37). To add new labels, the tool can be trained again whenever applicable (Figure 36- F). This re-training allows the classifier to be retrained and corrected when the interest of analyst change over time.

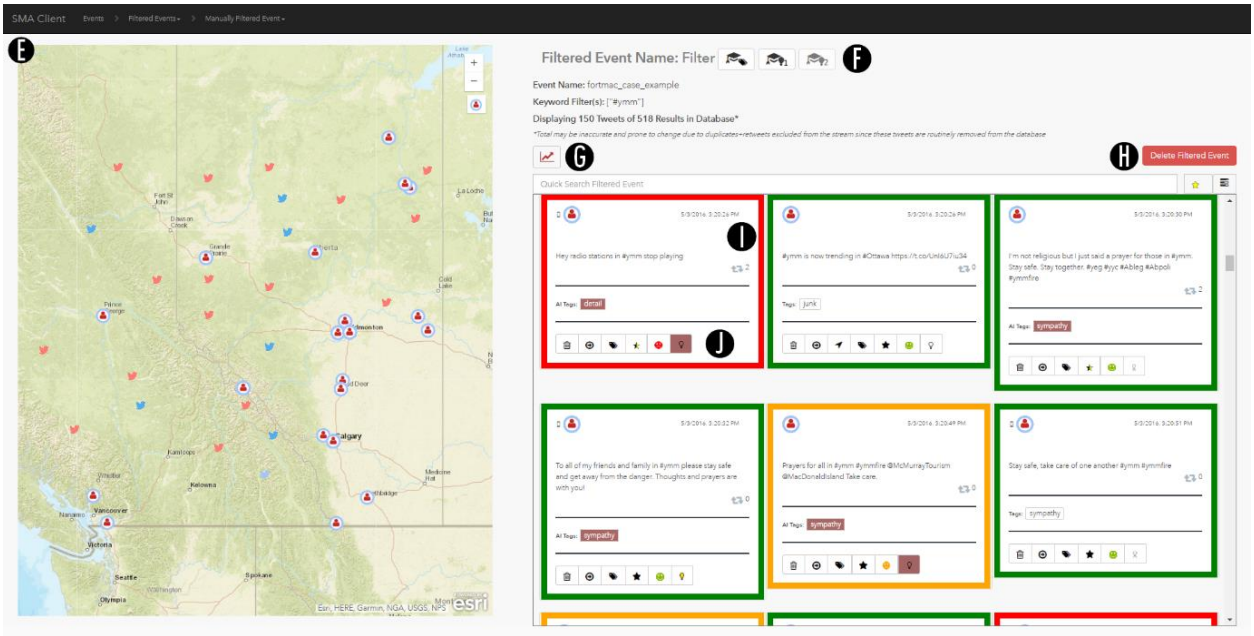


Figure 36. SMART Filtered Event.

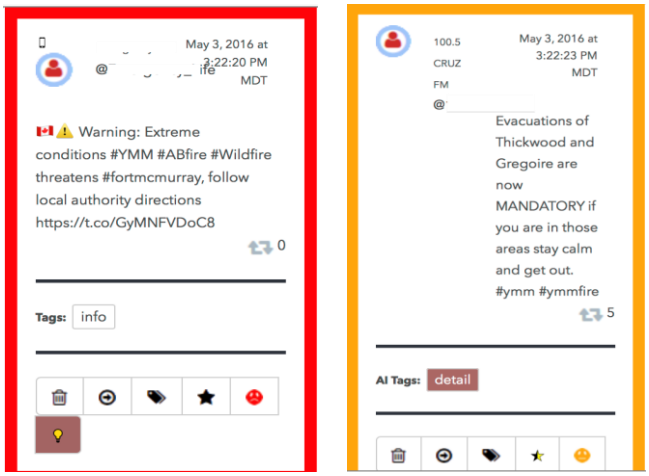


Figure 37. SMART provides three ways of filtering. (1) by situational awareness (light-bulb), (2) by tagging (tagging button) and (3) by interestingness (star button)

6.1.3 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 of those events was unprecedented. Responders may not sleep for several days

and bear a lot of pressure. As they stated, having a degree of separation and the help of outsiders, such as digital volunteers outside the crisis region, can be useful. Digital volunteers from outside the community can help in monitoring and identifying information from social media. That is why I designed a 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 and thus create manually populated filters. Users can also send certain tweets to a wall display or the common operating picture of an emergency response system (Edwin Chan, 2016). The analyst can move a post to either manual stream or a wall display using the arrow button in Figure 36-J.

6.1.4 Easy to Use 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 add new 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 stand out against 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. As Figure 36 shows, a filtered event view balances displaying as much information as possible while keeping the user interface clean and easy to read. The map (Figure 36-E) shows information regarding the location the tweet is about, and the location of the user at the time of the tweet. If SMART extracts the location of the tweet, using a location mention extraction process, it will be red. The graduation hats (Figure 36-F) are used for machine learning as mentioned before. Tags, sentiment, interestingness, and

situational awareness can all be viewed in graph form when toggled with the map. The search bar and following buttons (Figure 36-H) can be used to change the tweets order, viewed within the Filtered Event. Lastly, the tweet and the buttons below 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. The graph button (Figure 36-G) allows the map to be toggled off and display graphs for sentiment, interestingness, situational awareness, and tags.

Upon clicking on a menu icon beside the search bar, analysts can view an advanced filtering window (Figure 38). The filtering window allows analysts to filter posts based on their time, tags, sentiments, topics they are related to, and their situational awareness.

6.1.5 Location Extractor component

One of the important features in SMART was determining if a tweet text contains location information or not. I use bigram based and a box based process to determine if a tweet text contains one or more location mentions or not. I describe the process behind this component in the Chapter 4. Figure 39 shows the map view with geo-located tweets (in red color).

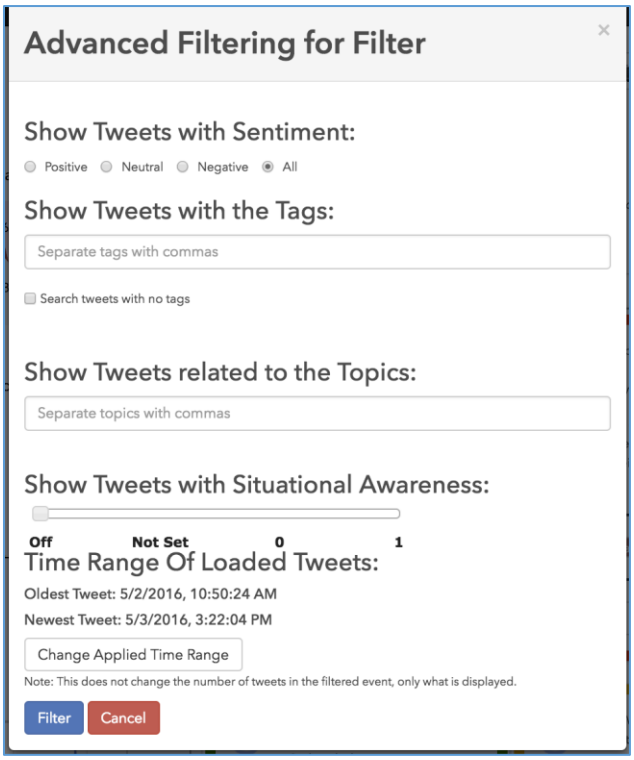


Figure 38. Advanced filtering menu

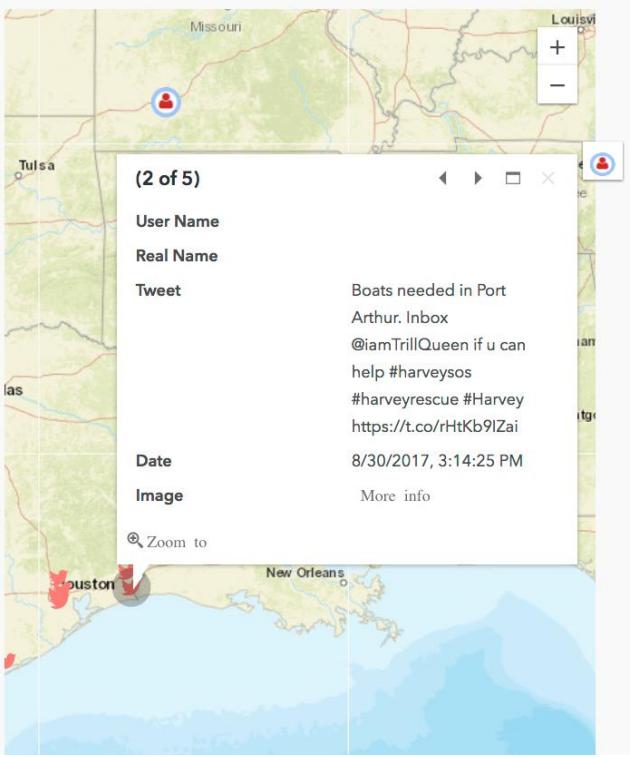


Figure 39. Geo-located Tweets in Hurricane Harvey Map View

6.2 SMART Architecture

This section provides a high-level overview over software components and technologies that create SMART. The application consists of four major components, a data storage, a web application interface, analytical components which include classifiers and the text processing modules, and the tweet collectors which retrieves the raw data from Twitter for an event. Figure 40 provides an overview of the SMART architecture.

6.2.1 Data Storage

Twitter data includes several types of metadata and is dynamic in nature. It includes arrays with varying length for some elements. Its data schema keeps changing over time as Twitter modifies features. For example, tweets can consist of multiple hashtags, URLs, user mentions, and photographs. As a result, the data schema is not consistent over time. Usually, the number of posted tweets for an event is enormous (for example in the Fort MacMurray wildfire event, the Twitter API downloaded 72,147 unique tweets after removing the retweets and duplicates). Considering these characteristics, I decided to use a document-based data storage over a traditional relational database. I chose the MongoDb[15] database to store tweets in this study. MongoDB is a NoSQL database and provides good performance for working with high volume of records.

6.2.2 Tweet Collection

To gather the tweets related to each event, I used the Twitter APIs. Twitter[16] provides a set of REST APIs to retrieve tweets from Twitter. These APIs are search based and work for historical or real-time data revival. There are Standard, Premium and Enterprises licenses and based on the license type, the amount of data that user can retrieve varies. In this study, I used the basic license

[15] https://www.mongodb.com
[16] https://developer.Twitter.com/en/docs

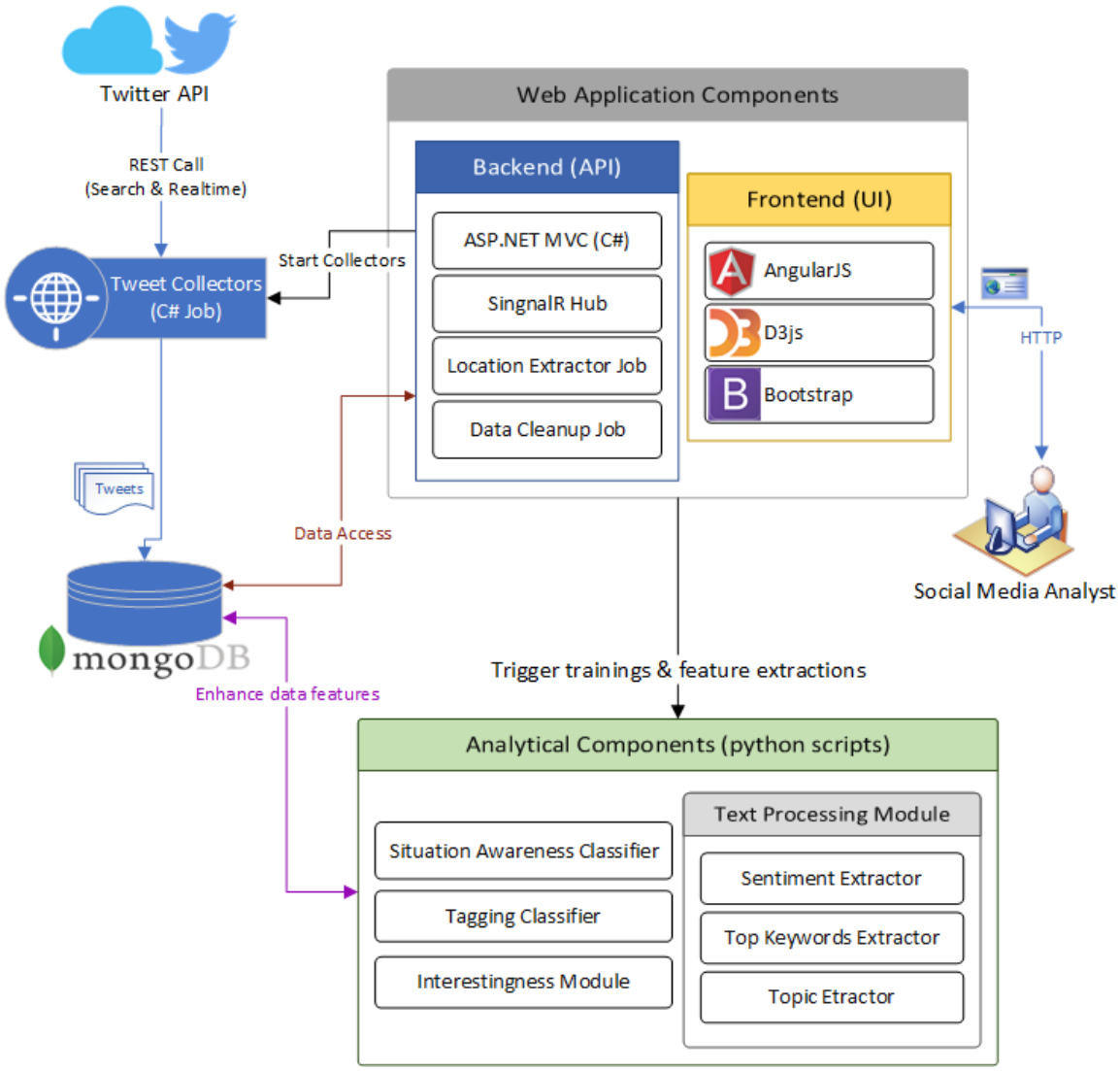


Figure 40. SMART Architecture Overview.

to use these APIs. There are some limits on using standard license. For example, the search API supports the last seven days as the time window for the search, or the real-time API supports a maximum of 400 keywords as the search term. The proposed tool includes a module written in Microsoft C#.NET to retrieve tweets and store the raw data in a MongoDB database. Later the analytics components will process this raw data.

6.2.3 Web Application Components

Social media analysts are the target users of the tool. As the analysts need to have interaction with the system and the need for dynamicity of the tool, I have designed a Single Page Application (SPA) web application. SPA allows dynamic interaction with the web page by modifying the current page rather than loading new pages. The development of web application components is done collaboratively with the help of two undergraduate interns. Our team used AngularJS[17] to implement a SPA web application for the user.

Another important feature of today's web application is to have a responsive and unique user interface across different browsers and devices. Social media analyst may access the system from a large screen monitor or small size tablet. We employed the Bootstrap[18] framework to design the user interface to provide a stable and responsive user experience.

The tool provides many charts and visualization components. These visualizations of the statistical information help the analyst to gain faster and more accurate awareness of the situation. D3[19] is a powerful JavaScript framework to enable rich visualizations that I used in the application. So far, I explained the frontend part of the web application. The backend is implemented in Model-View-Controller (MVC) architectural pattern using Microsoft ASP.NET MVC[20]. This layer provides data access layer and retrieves information from MongoDB to the frontend.

In several scenarios, the application provides real-time data push from server to the connected client. For example, while the analyst enables a tweet collector, the statistical data about

[17] https://angularjs.org
[18] https://getbootstrap.com
[19] https://d3js.org
[20] https://www.asp.net/mvc

captured tweets will be sent to the UI layer in real-time. SignalR[21] framework empowers the web application to do this real-time service.

Location extractor component is another module written in C# in the backend to extract tweet location based on its content.

The backend of the system also includes a data cleanup module to delete duplicate and similar text tweets from the database as multiple entries for the same tweet can be stored in the database by different simultaneous collectors.

6.2.4 Analytical Components

The main engine of the system is implemented using python scripts. These scripts are triggered by the web application. Analytics components include text processing modules for extracting sentiment, topics, and top keywords. It also includes components for extracting features and creating classifier models for interactive filtering. Development of analytical components was done solely by me.

6.3 Case Study

This section presents the feasibility and applicability of the approach with a case study. The case study was conducted collaboratively with the help of two undergraduate interns. The study shows how Twitter data regarding the Fort McMurray wildfire was collected and analyzed using SMART. The goal was to simulate how an analyst would use the tool to get an overview of an emergency. We conducted the study in a lab environment.

[21] https://www.asp.net/signalr

6.3.1 Fort McMurray Fire 2016

Fort McMurray is an urban service area in the Regional Municipality of Wood Buffalo, Alberta, Canada, and is surrounded by a boreal forest, which is often prone to wildfires that can get out of control as seen during May 2016. On May 1st, due to a wildfire southwest of Fort McMurray, a warning went out to the residents of Gregoire to prepare for evacuation (Kohut, 2016). On May 4th, the Alberta government declared a provincial state of emergency and approximately 80,000 people were given a mandatory evacuation order. By the time the fire had been contained, an estimated 589,995 hectares had been affected and the entire Fort McMurray urban area including nearby areas were evacuated ("Final Update 39: 2016 Wildfires (June 10 at 4:30 p.m.)," 2016). Low humidity and strong winds had played a major role in the difficulty of containing the fire. Such weather factors caused many evacuation orders to appear contradictory, as the fire jumped from place to place. During this event, social media aided in getting live announcements to the population, especially those who lost all other forms of communication except for their cellphones. On the other hand, evacuees were also able to use social media to request for help as well as give live updates on the situation. Once local radio stations were forced to evacuate, social media became even more vital. Some tweets such as the following from the Fort McMurray provided awareness about the situation which helped the emergency responders:

> *"Still stuck on the road in Abasand and the flames are getting closer. Please get us all out of here. Haven't moved in 30 mins #ymmfire #ymm".*

Obviously, there were also tweets either expressing sympathy or otherwise strategically useless information within the streams of watched hashtags. These are our target to filter out, since they are not providing information about the event or they are completely irrelevant, such as the following tweets:

- *"Get Free Twitter Followers Fast - Easy #ymmfire https://t.co/WbZ08FtBaV"*
- *"My heart and prayers are with those fleeing the #ymmfire and for those who are fighting the blaze. Stay safe everyone. <3"*

Within the timeline of 2016-05-01, 11:27:46 PM to 2016-05-06, 7:58:13 PM, we gathered 72,147 tweets after removing duplicates and retweets.

6.3.2 Fort McMurray Event Creation

For our first step, we created an event named FortMacFire to gather potentially relevant messages by supplying an initial set of keywords. Second, we used the trending hashtags regarding the event: #ymm , #ymmfire, #FortMacFire, and #ymmHelps.

6.3.3 Extracted Topics and Filtered Event Creation

The next step was to let the topic modeling component extract sub-events and topics from the gathered collection of tweets (Figure 41). We identified two themes that we wanted to explore futher: (1) *Evacuation:* we created filtered events with keywords "*mandatory*" and "*evac*" to get a sense of the evacuation status. (2*) Airport:* Fort MacMurray airport was among the hot topic discussed, we created a filtered event using "*airport*" to find the relevant information regarding the airport.

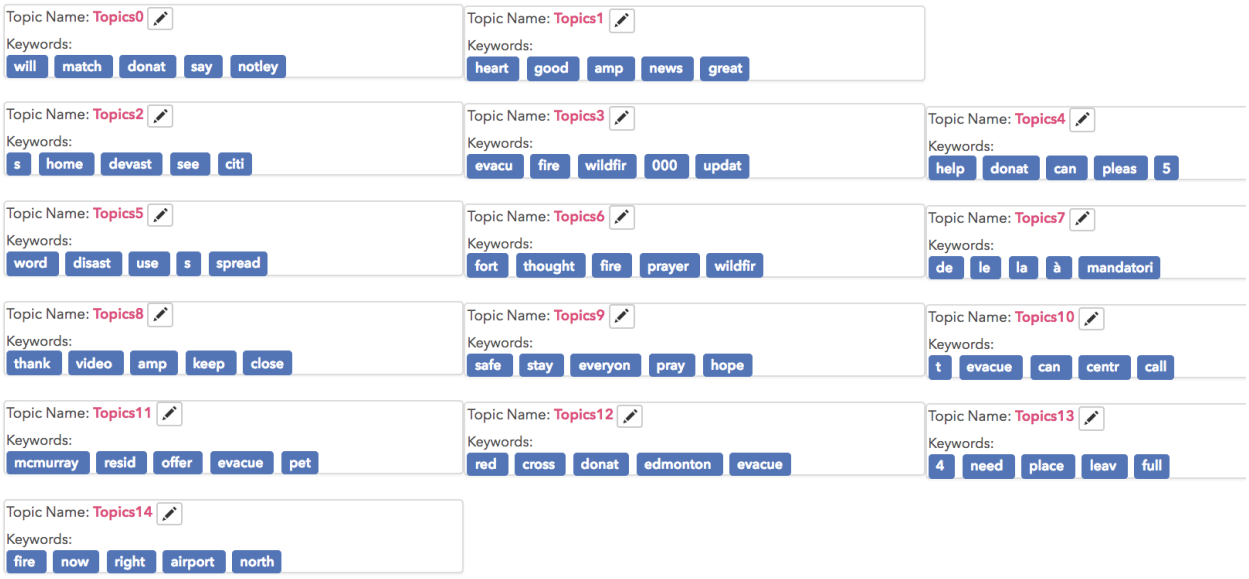


Figure 41. The result of topic extractor module for Fort MacMurray fire event

6.3.4 Classifier Training

The *Evacuation Filtered Event* resulted in 8,144 tweets. We started training by labeling the tweets. We also took advantage of the tweets that the geolocation component extracted. The geolocated extraction component identified tweets that were about mandatory evacuations in different locations on the map. Table 13 shows some samples of these tweets:

Table 13. Sample of tweets identified with the location extraction module for the evacuation filter

Tweet Text	**Location**
MANDATORY evacuation Anzac NOW COMPLETE! REOC IN Lac La Biche SETTING UP NOW	Anzac
Members of Ironworkers union from Edmonton holding a free barbeque for #ymmfire evacuees in Lac la Biche	lac la Biche
Residents from Kikino, a Metis colony of about 2,000 that's opened its doors to evacuees	Kikino
Fort McMurray wildfire update: Mandatory evacuation at Prairie Creek & Centennial trailer park	Prairie Creek

Using the search box in the filtered event, we searched for *"stay safe*"and *"pray"* to quickly find tweets that could be relevant to sympathy. On average, it took us 20 minutes to label and filter 150 tweets in the evacuation stream.

After training the evacuation filtered event, we noticed some tweets that were being identified by either situational awareness or tagging classifier and were talking about smoke in the air. We decided to create a filtered event for the smoke to gather more information about that topic.

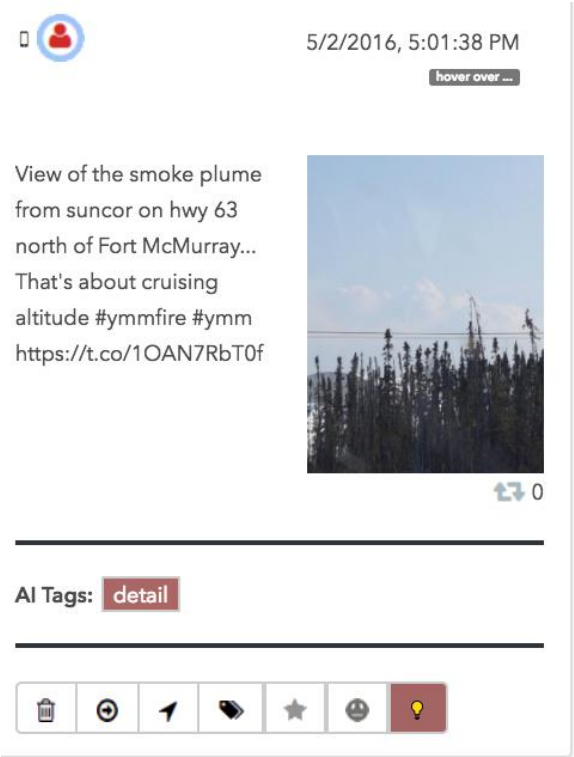


Figure 42. A sample tweet that was automatically identified as situational awareness (the light bulb), and automatically tagged as detail by SMART.

In the smoked filtered event, we discovered more informative tweets (out of 600) that were either identified as situational awareness or as detail by the interactive filters that we might have missed otherwise. Figure 42 shows a sample tweet that has been identified as situational awareness and been tagged as detail. In the following, we can see more examples of tweets that have been automatically identified as situational awareness or detail.

Tweet	Identified as
Heavy smoke is seen near the Global News crew at the crossroads of Highway 63 and Highway 881	SA
Smoke view hwy 63&881 north view Pic 2/3	SA
#nofilter #ymmfire and smoke plum as seen from Beaconhill	SA
The view from our office at King St and Manning Ave. The promised smoke plumes have arrived. #ymm #ymmfire	detail

In the following, we listed tweets in the smoke filtered event that were filtered out since SA classifier identified them as not situational awareness.

Tweet	Identified as
The fire and smoke bring such an eerie glow over YMM View from Waterways. Thanks to all who keep us safe #ymmfire	Sympathy
Everyone in my house woke up with sore throats & headaches thanks to all the forrest fire smoke in the air!	Not SA
Smoke heavy and rolling in! Fort McMurray	Not SA

The map view of the smoke filtered event (Figure 43) also showed the geo-located tweets (with the red bird icon) from locations that smoke has been reported.

The Airport filtered event found 613 tweets. Some of the posts were speculations over the possibility of fire threatening the international airport. Comparison of a number of tweets for smoke filter vs. Airport filter (see Figure 44 and Figure 45) shows that the airport conversation mostly happened around May 4th, while reports of smoke were spread across all event days.

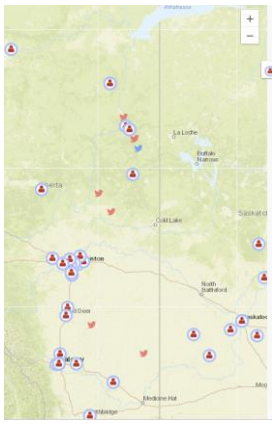


Figure 43. Red bird icons on the map shows the locations that smoke has been reported in Tweets.

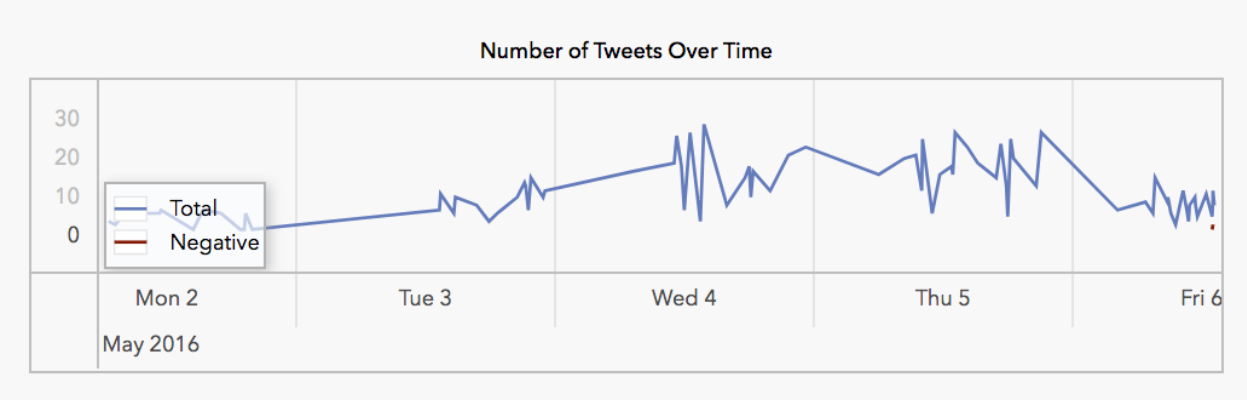


Figure 44. Number of tweets over time for smoke filter.

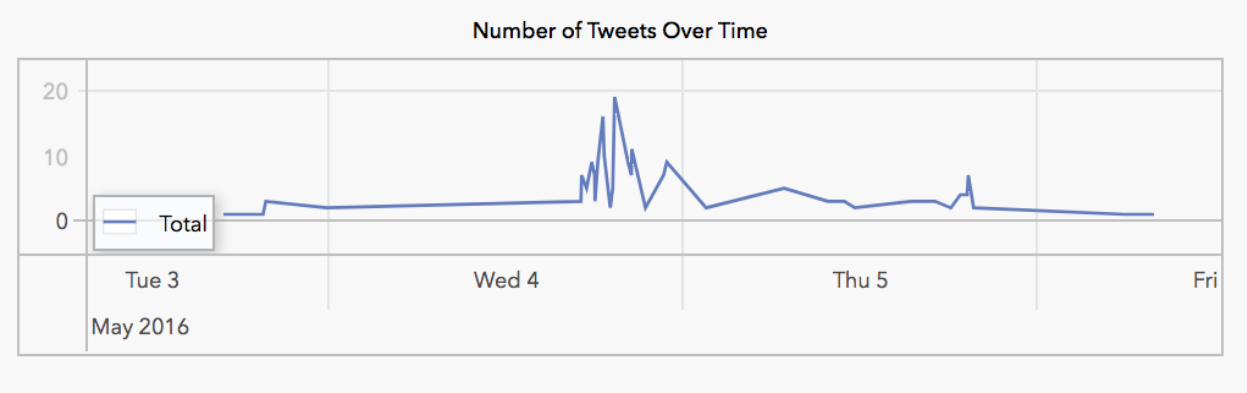


Figure 45. Number of tweets over time for Airport Filter.

By exploring the previous filtered events, we found tweets that were talking about lines of gas and the need for fuel in the evacuees in Highway 61. We named the filter fuel and defined it based on “fuel” and “gas” keywords. By training this filtered event for about another 10 minutes

(to add more training samples), the classifier could assign label and tag to new posts regarding people in the fuel lines. For example, the classifier identified the following tweet as situational awareness: "*50,000L of fuel about 30km north of #FortMcMurray on the southbound ramp of Suncor Voyageur",* but this one as not*: "You wouldn't think about how important a full gas of tank is when being evacuated from your city."* Hence, the interactive filtering saved time in separating relevant information from irrelevant ones with similar keywords.

6.4 Chapter Summary

In this chapter, I elaborated on the final prototype, architecture, and implementation details regarding SMART. I ensured that the final version of the tool is intuitive and provides interactive filters to let analysts identify relevant information based on their interest. Analysts' interest could be related to situational awareness information or any tags that analyst need to assign to a post. I demonstrated the applicability of the tool based on a fire event by iteratively train the embedded classifiers.

Chapter 7: Evaluation Study[22]

In this chapter, I report the result of a second evaluation study on SMART. I conducted a user study to gather qualitative feedback. The evaluation results offer recommendations into the design of social media analysis tools for emergencies. The evaluation findings show the interest of emergency practitioners in 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. According to the 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.

7.1 Evaluation Procedure

I conducted four evaluation sessions with nine participants. Three of these sessions were individual sessions and the fourth session consisted of six participants. All study sessions were conducted on a laptop in places like the participants' workplace, in a coffeehouse near their workplace, or remotely through Skype (through screen sharing). Since I knew all participants (except one) from previous interview studies, I did not conduct any pre-interviews at this phase. Study sessions took between 90 to 120 minutes. After receiving their consent to be in the study (For the consent form, please see Appendix F), I gave the participants an overview of different parts of the system. I 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, and how to create historic and live collectors. I also showed them

[22] This chapter is based on Mahshid Marbouti, Craig Anslow, Frank Maurer (2018). Evaluation results for a Social Media Analyst Responding Tool. To appear in Proceedings of the 15th International Conference on Information Systems for Crisis Response And Management (ISCRAM)

how to create filtered events and how to train the filtered events for identifying situational awareness and user defined tags. I demoed the tool using historic events. Depending on working on a live or historic event, I either asked participants to create new events or work with the already created events. I asked participants to perform the following tasks as these are common tasks according to prior work (Marbouti & 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.

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

7.2 Participants

To have more promising results (external validity), I intended to use domain experts instead of students for the evaluation study. However, it is challenging to schedule a time with domain experts due to the nature of their work and their busy schedules. By contacting different emergency organizations, I recruited emergency practitioners and social media analysts in the emergency domain for evaluating SMART. I selected the participants from public related organizations that

extensively monitor social media data to address public needs and concerns during emergencies. For conducting a usability session, I contacted Calgary Emergency Management Agency (CEMA), and scheduled time with their communication team. I preferred one on one meetings with each of their staff, but it was not possible due to the nature of their work. Thus, I conducted a group session. In each group session, there were nine participants, with six of them having direct social media responsibilities. These six participants monitor social media in their daily tasks, and because of the nature of their organizations, they also monitor social media data extensively at the time of emergencies. Although this is a small number, it is still relevant because it is difficult to schedule meetings with emergency experts. All nine participants belonged to public sector organizations including CEMA, Calgary Police, Fire department, City of Calgary, and County of Frontenac. All participants had experienced dealing with/monitoring social media data in at least one emergency. The participants had experienced emergency events such as Alberta Floods (2013), Calgary snow storm (2014), or Fort McMurray fire (2016).

7.3 Emergency Scenarios

I evaluated the tool based on real emergency scenarios. In the event page of SMART, there were several emergency events. I demoed the tool usually based on familiar events for participants, such as the Alberta Floods, and then proceeded with another one or two events upon the desire of participants. In the three sessions with individuals, I focused on events, such as the Fort McMurray fire or Hurricane Harvey, simulating these events as they are happening in real time at the time of the study. For the group study, I serendipitously found a real-time emergency that was happening in Queensland Australia; a few days before the scheduled meeting I noticed there was a tropical cyclone in Queensland Australia called Cyclone Debbie. I started gathering tweets around that event to get familiar with it. I decided to focus on that event during the user 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 lost power and phone reception. The group session took place on March 28th at the CEMA office. 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, I used real time data.

7.4 Findings

In this section, I provide details of analysts' feedback when performing the given tasks. The tasks were not performed in a specific order. I asked participants to assume that they are on a response team in the designated scenario and start performing the tasks. I asked some participants how much these tasks matched to what they are currently doing, 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[is] start developing messages [to push out to public]."*

7.4.1 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 hundred 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 there 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.

P5 liked to be able to delete previously saved keywords with all their collected data. They were trying to gather data around a plane crash event in *Saskatchewan.* Initially, they used #Saskatchewan but then, later on, they realized that the hashtag was not useful for this event and they were getting a lot of unrelated data. 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, especially at the early hours of the 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, for Cyclone Debbie, 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, 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, I added a set of collectors with general hashtags. In these sessions, all three participants

usually took a look at the general hashtags, then added additional collectors based on the topic modeling keywords. For example, P8 started from the Fort McMurray Wildfire 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."*

I also noticed three participants added some collectors and filtered events regardless of a 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 had a considerable portion of unrelated messages.

7.4.2 Need to Facilitate Reading Tweets

The filtered event view includes a tweet text box view, and the users can also toggle between a map/analytics view. At the beginning of the study, four participants who mainly monitor social media as a part of their daily tasks had much less interaction with the map/charts and they liked 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. 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 the map view increased. Another participant made another point in this regard that labeling would be easier for them in a view with more tweets.

7.4.3 Sentiment in the Emergency Concept

All participants were interested in public sentiment during an emergency, both on their early interaction with the filtered event view and later on when they were exploring to see what topics people were talking about. For four participants, the sentiment was the first thing that caught their eyes when they entered a filtered event. They were also curious about the sentiment on the chart page. In Cyclone Debbie, 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. 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 each sentiment in the pie chart graph and see the keywords or tags changes at the bottom, specific to that negative sentiment. Regarding sentiment, 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."* P2 suggested being 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. P9 put it in this way when I 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 is 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."*

7.4.4 High Use of Search Box

The **search box** capability in the filtered event view had utility when exploring the filtered event (T1 & T2). P2 used it to search for injuries for the Cyclone Debbie event and another participant used it to put a question mark in a filter to find questions and enquiries from people (Figure 46)

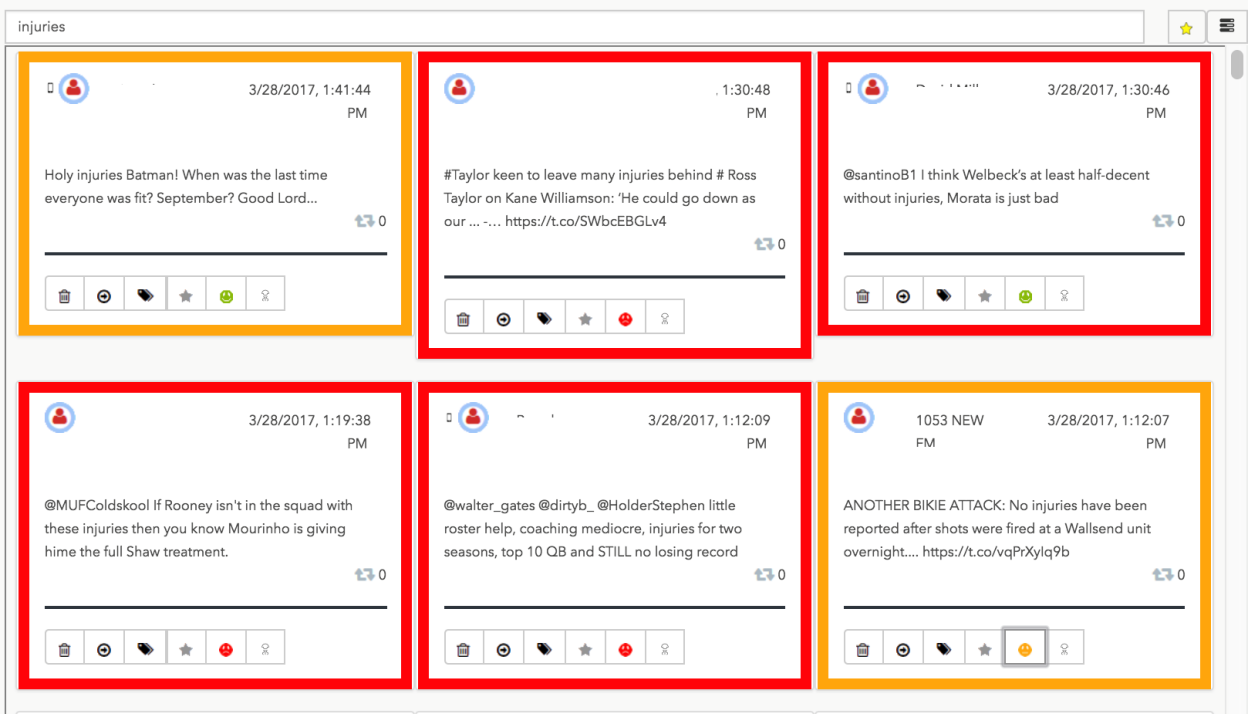


Figure 46. This figure shows a view of search box within a filtered event

7.4.5 High Use of Location Extraction Component

To find regions impacted by flood (T3), participants considered **location extraction** useful. One participant mentioned the geo-location 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 Harvey Hurricane, one participant noticed the geo-located tweets on the map about a sub region called Port Arthur (Figure 47). The location extraction component identified five tweets in 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 tight… if there are injured people I'd say ok we need boats there maybe helicopter…"* P9 also noticed these geo-located tweets from Harvey Hurricane. These tweets were two from 34 tweets that were grouped by the 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

When asked "what he would do when he sees such tweets?" P9 replied, *"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 didn't 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!"* I 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?"

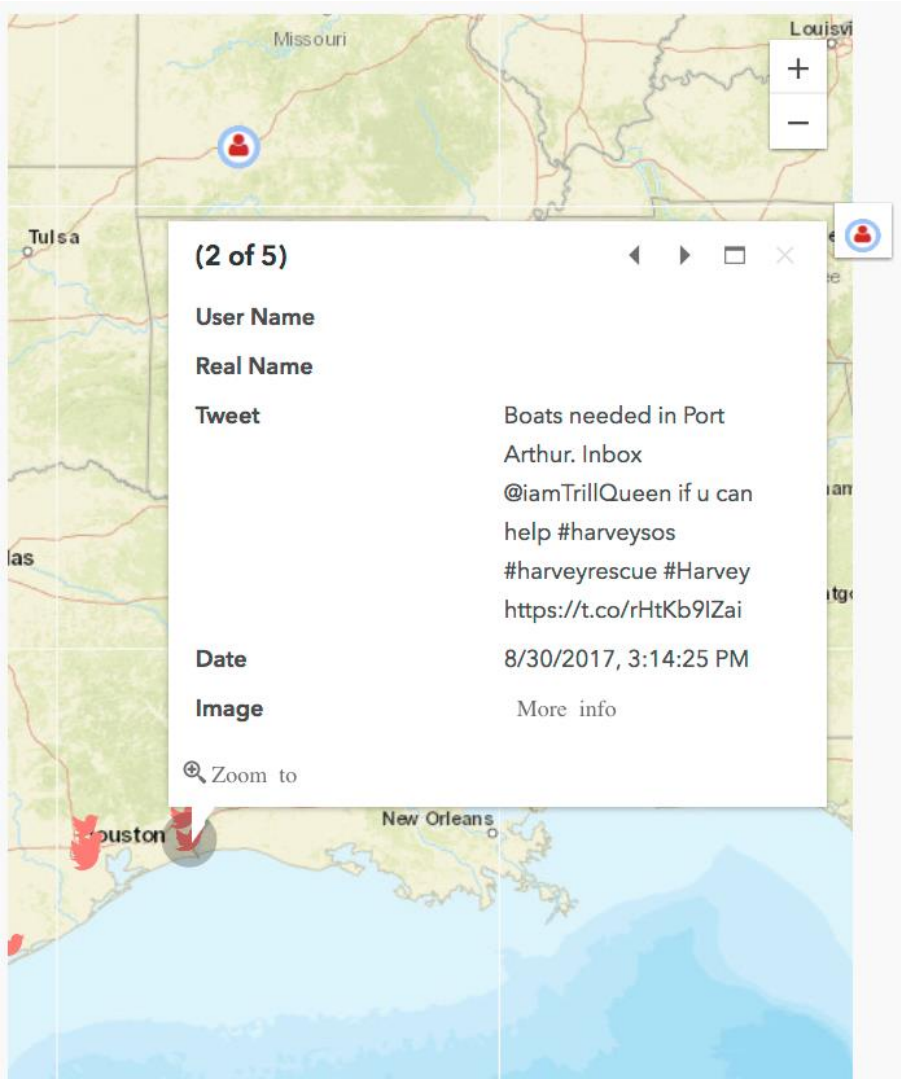


Figure 47. Geo-located Tweets in Hurricane Harvey Map View.

7.4.6 Reliability of Tweets' Authors

During the sessions I 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 used 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.

7.4.7 Manually Populated Filters

To create a report of relevant messages (T4), the Forward 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 I 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.

7.4.8 Training the Classifiers

To find more details regarding an event, participants were training the SMART classifiers**.** For simplicity, I 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 a lot of time. "*We tag all of our*

incoming messages and it's 100% manual, so any starting point is an advantage," (P4). P6 asked if I could add the same functionality to 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."*

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 a tweet 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 situational 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 I might have considered as situational awareness, but a particular participant was not interested in them. For example, regarding the following 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 I did not explicitly ask the rest of participants regarding this combination, I 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 is why it is important to combine urgencies with the negative sentiment (Figure 48).

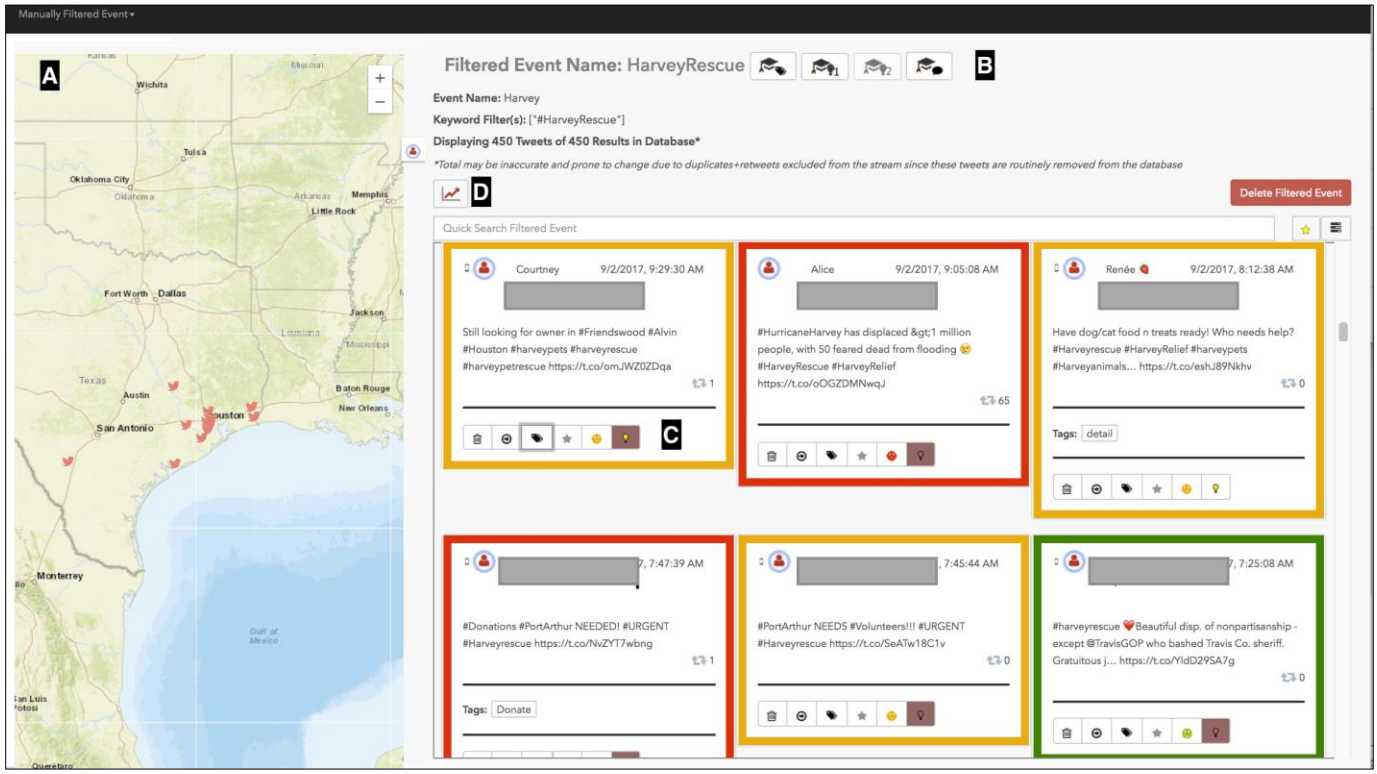


Figure 48. SMART Filtered Event Page: This view shows the filtered event page in which users can view and label streams of tweets. The graduation hats (Figure 48-B) are used for training the tagging and situational awareness.

Two of the participants suggested simplifying the training for non-technical users. P2 who was an emergency manager expressed that he prefers a refresh button instead of a training button (the graduation hats in Figure 48-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 the chaos of a disaster to click on the tagging button, assigning tags and then clicking save.

On the other hand, P4 requested more complex tags that could possibly help with their daily tasks. For example, she 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 ….*"

I also asked the participants explicitly which type of training (assigning a tag such as detail, media or labeling as situational awareness or not situational awareness) they would use. All of them liked to keep both options but some were more comfortable with assigning tags as the tags made more sense and they could have user 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, P7 mentioned he would slightly prefer situational awareness tagging as it will be faster to label and during the chaos of an emergency they don't have much time, *"I use the light [bulb] for the urgent [tweets]."*

P7 was interested towards 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 SMART 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're 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 the AI can do."*

7.5 Discussion

Overall the findings show the interest of participants into embedding machine learning classifiers for determining situational awareness and user defined tags. Some of the analysts in the public organizations were already manually labeling social media posts as a part of their daily tasks, so the real time training of situational awareness and user-defined tags' classifiers would add a minimum overload to their current tasks. I found that to use such capabilities in real time emergency 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 two participants, a social media analysis tool should also support their daily tasks to be effective during a disaster. Analyst needs to be familiar with a tool before using it during 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 colour 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. However, 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. 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 in learning what has happened, but also to reflect what is going to happen using a social media monitoring tool.

I 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. I noticed most of the participants are going back and forth to Twitter. They mentioned that they were doing the same with the commercial tools such as Hootsuite. This could be because the Twitter website is still the easiest platform for them to work with but also because of a 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 SMART.

7.6 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 about their response. Four 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."*

I noticed that participants try to pull the required information by querying general emergency related keywords such as injuries, and 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.

7.7 Chapter Summary

In this Chapter, I discussed the user study that was conducted for the SMART. The goal was to evaluate effectiveness of the tool by gathering qualitative feedback from social media analysts based on real emergency scenarios. I designed study tasks to demonstrate if the tool would be effective during an emergency scenario, and I asked domain experts to participate in the evaluation. All study participants had several years of experience in emergency management. A summary of the findings is presented in Table 14. The findings revealed that identifying the most appropriate keywords, especially during the early hours of an event, is one of the challenging tasks for social media analysts. I also found that to combine human expertise with machine intelligence, the training tasks should fit practitioners' current workflow. According to the findings, analysts are not only interested in finding relevant information using a social media analysis tool but further they used the proposed tool to perceive how their actions impact the public sentiment. The results also show the importance of sentiment and the perception of emergency practitioners around the negative sentiment. The findings highlight the need for identifying negative sentiment in the emergency concept when designing social media monitoring tools.

Table 14. Summary of feedback on final prototypes.

Aspect	Summary
MOST USED FUNCTIONALITIES	• Sentiment • Search box • Topics for extracting keywords • Location extraction • Make internal comments • Manually populated filters
ADVANCED FILTERS	• Integrate with response tasks • Cross-reference info/ Create the trust in the long-term
DESIGN CONSIDERATIONS	• Ability to integrate with daily work
TRUST ISSUES	• Aggregated information is not enough
VISION	• After being aware of what is going on around, they need to project the future

Chapter 8: Conclusion and Future Work

In this final chapter, I summarize research results, contributions, elaborate lessons learned and suggest possible future directions for analyzing social media data for emergency management.

8.1 Dissertation Summary

The goal of this dissertation is to help social media analysts overcome the information overload problem when dealing with social media data during an emergency. To get a better understanding of what are the main social media related requirements of people in the emergency domain, I conducted semi-structured interviews with people in the emergency domain. The participants were a combination of emergency practitioners that had the social media experience and social media analysts that had experience of monitoring at least one emergency event. The participants were mostly from public organizations such as CEMA, police and fire departments. In this phase, I also selected a few participants from the private sector such as oil and gas related organizations. I gathered information about how the practitioners use social media and what are their challenges when monitoring social media. Based on the interviews, I realized that social media is a source of data that these practitioners cannot ignore anymore. They had mixed feelings towards social media as they considered it both essential and high risk. On the positive side, social media made the communication easier, faster, and opened a two-way communication road. However, social media was considered high risk, as there were reliability and trust issues associated with information coming from social media. Current technologies that emergency practitioners use do not help them prioritize information based on their level of importance or situational awareness but only based on chronological order. On the other hand, academic approaches mostly ignore the requirement of emergency practitioners when designing monitoring tools (A. L. Hughes & Shah, 2016). Social Media Analysts need analysis tools that help them identify relevant information according to their

informational needs. They also require tools to facilitate quickly assessing the reliability of sources (Marbouti & Maurer, 2016).

For identifying situational awareness, I developed a machine learning classifier to identify tweets that contribute to situational awareness regarding an emergency event. I used textual, social, location, and tone based features to increase precision and recall of classification. I compared the impact of using different features, and different classifier algorithms on the classifier performance. Results show that the impact of features and algorithms varies across different events.

In the next step, I took a human-centered design approach (Constantine & Lockwood, 1999) to develop Social Media Analyst Responding Tool (SMART), with unique features over existing approaches (Marbouti et al.). During requirement elicitation, I found that analysts from public organizations are usually doing manual labeling of social media posts as a part of their internal tasks. However, there is no intelligence in the backend of current commercial tools to learn the patterns and thus be able to automatically label the incoming posts and reduce the overload of incoming posts. Thus I designed a tool to bring analysts in the classification loop and let them perform real-time training of classifiers. I implemented supervised machine learning into SMART. The tool combines the ability to collect, search and sort tweets alongside the functionality to filter and organize them with machine learning techniques. SMART allows 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 within the text. Initial feedback on the core parts of the tool was positive.

For evaluating the tool, I conducted a user study using SMART. The goal was to evaluate the effectiveness of the tool by gathering qualitative feedback from social media analysts based on real emergency scenarios. The goal was to evaluate if the tool would be effective during an

emergency scenario. All participants had several years of experience in emergencies. Findings show that to bring human in the machine learning loop for training a classifier, and to combine human expertise with machine intelligence, the training tasks should fit practitioners' current workflow. I observed that when practitioners were finding relevant information using interactive filters, they wanted to perceive how their actions impacted the public sentiment. Emergency practitioners emphasize the importance of negative sentiment as it could help them in prioritize and allocate resources during the chaos of an emergency event. Finally, 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.

The contribution of this dissertation is to *bring insight on the requirements of Social Media Analysts in the emergency domain* and *provide design recommendations on social media monitoring tools* based on my engagement with emergency practitioners throughout three years. The proposed design combines supervised learning techniques with human expertise. The qualitative evaluation with domain experts brings knowledge into the development of social media analysis tools that embeds supervised machine learning. As far as I know, this is the first study to explore this concept in social media analysis for emergency management. Another contribution aspect of this dissertation is on exploring the *significance of various categories of features* into automatically *identifying situational awareness* information using feature selection and supervised machine learning techniques.

8.2 Discussion and Future Directions

In this section, I share interesting discussion posts and possible future directions that can be pursued to continue this research.

8.2.1 Bringing Human in the Loop

The challenges with monitoring social media for emergency organizations become more visible when responders lack time and resources to deal with loads of information coming from social media. The problem has been addressed in academia by using text classification techniques. Choosing the right features seems crucial when it comes to microblog classification. Studies (Imran, Elbassuoni, Castillo, Diaz, & Meier, 2013) also show that the accuracy of classifiers drops significantly when trained on one disaster event but reused in another. To overcome this issue I designed and developed a tool with the aim of bringing the target users - in this case, social media analysts- in the classification loop. To target short-term classification problem, I used various categories of features in the embedded classifier. I received feedback from domain analysts both on early and fully functional prototypes to find if the proposed design is effective in real emergency scenarios. I provided two options for analysts. Label tweets based on their situational awareness or label based on user-defined tags. Participants liked to keep both options. The final user study showed that by bringing analyst in the loop, they could get a better sense of disaster scenario. Participants said they would use such a system if they can trust it's accuracy and for that, they need to be able to use it in the long term for their everyday tasks. Some participants were concerned towards shifting in the interest of analyst when they are training the system and they would also like to retrain the system and distinguish between human assigned labels and automatic assigned labels.

8.2.2 Public vs Private Sector

For the requirement elicitation interviews and early design feedback over low fidelity prototypes, I recruited three participants from the private sector. However, I noticed different perspectives and requirements among private and public sector participants. Private sector participants had more

limitations in assigning resources to social media and therefore they cared more about the volume of social media posts. They wanted to know if the issue is big enough that it requires action. They were more interested in the volume of social media posts to get a sense of how big the problem is. A participant that worked for the private sector stated, *"My client [a private sector organization] wants to know the volume because that tells me how important it is, who are the influencers, the degree to which I'm getting a share of voice, the degree to which it's being spoken about. Volume to talk ratio. Volume that's negative. Is there something that needs reaction."* However the public sector cared more about the importance of each individual message, an analyst from public sector mentioned, *"Whenever we compile these hourly reports during a crisis, usually what everyone wants to know is they don't even care about the volume, it's just what are people saying at these hour, one person said that she's in a wheelchair the elevator not working and she can't get down… So I think it should be very topical in terms of the information being reported."* For the final evaluation I only focused on the public sector organizations.

8.2.3 Strategies to Mitigate the Challenges

I learned that emergency practitioners are using some strategies to mitigate the overload problem. Some of these strategies can inspire the design of monitoring tools. For example, they created a role call audience intelligence, in which they briefly get an overview of the current topics by reviewing news streams having a staff member in the 311 or 911 call center so that they could quickly become part of the conversation in the social media and overcome their lack of resources. Another strategy was to put the geo-tagged messages on the common operating picture during the night shift in organizations with no resources to focus only on social media so that everybody can see what is going on in social media platforms.

8.2.4 More Feature Categories

In this work, for determining situational awareness in the feature set, I did not consider contents such as photo, video, and URLs within the tweets. One possible direction for this work is to add features related to these types of content by processing images and videos that are in tweets as possible predictors of situational awareness. Another possibly contributing feature could be related to processing the content of URLs that are embedded in tweets.

8.2.5 Long Term Evaluation Study

For using a social media analysis tool, there is a need for designing monitoring tools that address all requirements of emergency practitioners and support their response tasks. Emergency practitioners need tools that help them monitor and interact with social media posts both during their day to day tasks as well as during an emergency event. As a natural extension to this research, one can fully implement SMART with all the commercial features and make it publicly available. Afterwards, a long-term study of SMART use in real world emergency scenarios could take place to assess the effectiveness of bringing human into the classification loop. A long-term study helps to possibly create a trust between analysts and machine learning classifiers and help them make training portion as a part of their daily tasks.

8.2.6 Various Geographical Regions and Influence of Demographic Factors

The evaluation conducted in this dissertation focused on Canadian social media analysts mostly based in Calgary, Alberta. Future studies could expand this scope to include emergency practitioners and social media analysts across different regions ensuring findings from this research generalize. While I hypothesize that the dissertation findings will apply to other analysts,

there will likely be slight differences in analysts and expectations across the various geographical regions.

8.3 Chapter Summary

This research designed and developed a novel tool by involving emergency practitioners to help them monitor and analyze social media data during emergency events. In the tool, I embed machine learning classifiers to identify relevant information from sheer volume of social media posts and help analysts increase their task performance. In the classifier, I used various categories of features that can contribute to identifying situational awareness. The proposed classifier achieved a high precision and recall rate in determining relevant and situational awareness posts. One of the main challenges for practitioners is that commercial tools are not designed for emergency response and academic approaches do not consider their requirements. This research brings insight into the design of social media analysis tool for emergency response by evaluating a proposed social media monitoring tool. The evaluation demonstrates the effectiveness of bringing analysts into the classification loop to train and get feedback to classifiers.

References

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

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

Albert‐Green, A., John Braun, W., Martell, D. L., & Woolford, D. G. (2014). Visualization tools for assessing the Markov property: sojourn times in the forest Fire Weather Index in Ontario. *Environmetrics, 25*(6), 417-430.

Alexander, D. E. (2014). Social Media in Disaster Risk Reduction and Crisis Management. *Science and Engineering Ethics, 20*(3), 717-733.

Alonso, O., Marshall, C. C., & Najork, M. (2013). *Are some tweets more interesting than others?# hardquestion.* Paper presented at the Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval.

Anson, S., Watson, H., Wadhwa, K., & Metz, K. (2017). Analysing social media data for disaster preparedness: Understanding the opportunities and barriers faced by humanitarian actors. *International Journal of Disaster Risk Reduction, 21*, 131-139.

Armstrong, R. A., Slade, S., & Eperjesi, F. (2000). An introduction to analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry. *Ophthalmic and Physiological Optics, 20*(3), 235-241.

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

Bosch, H., Thom, D., Heimerl, F., Puttmann, E., Koch, S., Kruger, R., . . . Ertl, T. (2013). Scatterblogs2: Real-time monitoring of microblog messages through user-guided filtering. *Visualization and Computer Graphics, IEEE Transactions on, 19*(12), 2022-2031.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers*. Paper presented at the Proceedings of the fifth annual workshop on Computational learning theory, Pittsburgh, Pennsylvania, USA.

Cano, A. E., Varga, A., Rowe, M., Ciravegna, F., & He, Y. (2013). *Harnessing linked knowledge sources for topic classification in social media*. Paper presented at the Proceedings of the 24th ACM Conference on Hypertext and Social Media, Paris, France.

Cao, N., Lin, Y.-R., Sun, X., Lazer, D., Liu, S., & Qu, H. (2012). Whisper: Tracing the spatiotemporal process of information diffusion in real time. *Visualization and Computer Graphics, IEEE Transactions on, 18*(12), 2649-2658.

Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H.-W., Mitra, P., . . . Jansen, B. J. (2011). *Classifying text messages for the haiti earthquake.* Paper presented at the Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011).

Caragea, C., Squicciarini, A., Stehle, S., Neppalli, K., & Tapia, A. (2014). Mapping moods: geo-mapped sentiment analysis during hurricane Sandy. *Proc. of ISCRAM*.

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

Chokshi, A. (2015). *Designing Social Media Tools for Emergency Response.* (MSc), University Of Calgary.

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

D'Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-Time Detection of Traffic From Twitter Stream Analysis. IEEE Transactions on Intelligent Transportation Systems, 16(4), 2269-2283.

Dashti, S., Palen, L., Heris, M. P., Anderson, K. M., Anderson, S., & Anderson, S. (2014). Supporting disaster reconnaissance with social media data: a design-oriented case study of the 2013 colorado floods. *In Proceedings of the international ISCRAM conference*.

Dewan, P., Gupta, M., Goyal, K., & Kumaraguru, P. (2013). *Multiosn: realtime monitoring of real world events on multiple online social media.* Paper presented at the Proceedings of the 5th IBM Collaborative Academia Research Exchange Workshop.

Diakopoulos, N., Choudhury, M. D., & Naaman, M. (2012). *Finding and assessing social media information sources in the context of journalism*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, Texas, USA.

Dou, W., Wang, X., Skau, D., Ribarsky, W., & Zhou, M. X. (2012). *Leadline: Interactive visual analysis of text data through event identification and exploration.* Paper presented at the Visual Analytics Science and Technology (VAST) IEEE Conference.

Edwin Chan, D. G., Mahshid Marbouti, Irene Mayor, Sydney Pratte, Teddy Seyed, Yuxi Wang, Jade White, Frank Maurer (2016). *Multi-Surface Systems for the Emergency Operations Centre of the Future*. Paper presented at the CSCW 2016 Workshop on Collaboration and Decision Making in Crisis Situations (CADMICS), San Francisco, USA.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*(1), 32-64.

Fan, J. (1997). Comments on «Wavelets in statistics: A review» by A. Antoniadis. *Journal of the Italian Statistical Society, 6*(2), 131. doi:10.1007/bf03178906

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association, 96*(456), 1348-1360.

Final Update 39: 2016 Wildfires (June 10 at 4:30 p.m.). (2016, June 10). Retrieved from http://www.alberta.ca/release.cfm?xID=41701E7ECBE35-AD48-5793-1642C499FF0DE4CF

Flizikowski, A., Holubowicz, W., Stachowicz, A., Hokkanen, L., Kurki, T. A., Päivinen, N., & Delavallade, T. (2014). *Social media in crisis management-The iSAR+ project survey.* Paper presented at the ISCRAM.

Fuchs, G., Andrienko, N., Andrienko, G., Bothe, S., & Stange, H. (2013). *Tracing the German centennial flood in the stream of tweets: first lessons learned.* Paper presented at the Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information.

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

Heimerl, F., Koch, S., Bosch, H., & Ertl, T. (2012). Visual classifier training for text document retrieval. *Visualization and Computer Graphics, IEEE Transactions on, 18*(12), 2839-2848.

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

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

Hoferlin, B., Netzel, R., Hoferlin, M., Weiskopf, D., & Heidemann, G. (2012). *Inter-active learning of ad-hoc classifiers for video visual analytics.* Paper presented at the Visual Analytics Science and Technology (VAST), IEEE Conference.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods* (Vol. 751): John Wiley & Sons.

Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R., . . . Griffith, S. A. (2015). Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters, 39*(1), 1-22. doi:10.1111/disa.12092

Hughes, Amanda, & 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.

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management, 6*(3-4), 248-260.

Hughes, A. L., & Shah, R. (2016). *Designing an Application for Social Media Needs in Emergency Public Information Work*. Paper presented at the Proceedings of the 19th International Conference on Supporting Group Work, Sanibel Island, Florida, USA.

Huiji, G., Barbier, G., & Goolsby, R. (2011). Harnessing the Crowdsourcing Power of Social Media for Disaster Relief. *Intelligent Systems, IEEE, 26*(3), 10-14.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys (CSUR), 47*(4), 1-38.

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). *AIDR: artificial intelligence for disaster response*. Paper presented at the Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea.

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

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). *Practical extraction of disaster-relevant information from social media*. Paper presented at the Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil.

Karimi, S., Yin, J., & Paris, C. (2013a). *Classifying microblogs for disasters*. Paper presented at the Proceedings of the 18th Australasian Document Computing Symposium, Brisbane, Queensland, Australia.

Karimi, S., Yin, J., & Paris, C. (2013b). *Classifying microblogs for disasters.* Paper presented at the Proceedings of the 18th Australasian Document Computing Symposium.

Kevin, P. M. (2012). *Machine Learning: A Probabilistic Perspective*: The MIT Press.

Kohut, T. (2016, May 4). Fort McMurray fire: Timeline of events. Retrieved from http://globalnews.ca/news/2681249/fort-mcmurray-wildfire-timeline-of-events/

Kong, X., Shi, X., & Philip, S. Y. (2011). *Multi-Label Collective Classification.* Paper presented at the SDM.

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

Kreiner, K., Immonen, A., & Suominen, H. (2013, December 05 - 06 2013). *Crisis management knowledge from social media.* Paper presented at the ADCS '13,, Brisbane, QLD, Australia.

Kumar, S. (2015). *Social Media Analytics for Crisis Response.* ARIZONA STATE UNIVERSITY.

Kumar, S., Barbier, G., Abbasi, M. A., & Liu, H. (2011). *TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief.* Paper presented at the ICWSM.

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

Laylavi, F., Rajabifard, A., & Kalantari, M. (2017). Event relatedness assessment of Twitter messages for emergency response. *Information processing & management, 53*(1), 266-280.

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

Liu, S. M., & Chen, J.-H. (2015). A multi-label classification based approach for sentiment classification. *Expert Systems with Applications, 42*(3), 1083-1093.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., . . . Blanford, J. (2011, 23-28 Oct. 2011). *SensePlace2: GeoTwitter analytics support for situational awareness.* Paper presented at the Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on.

Marbouti, M., & Maurer, F. (2016). *Social Media Use During Emergency Response–Insights from Emergency Professionals.* Paper presented at the Conference on e-Business, e-Services and e-Society.

Marbouti, M., Mayor, I., Yim, D., & Maurer, F. Social Media Analyst Responding Tool: A Visual Analytics Prototype to Identify Relevant Tweets in Emergency Events.

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

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). *Twitinfo: aggregating and visualizing microblogs for event exploration.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Marina, K., Leysia, P., & 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* (pp. 981-993). Vancouver, BC, Canada: ACM.

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data: A model comparison perspective*: Routledge.

Mazumdar, S., Lanfranchi, V., Ireson, N., & Ciravegna, F. (2014). *Visual Analysis of Real-time Social Media for Emergency Response.* Paper presented at the SSA-SMILE@ ESWC.

Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems, 29*(2), 9-17.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821): John Wiley & Sons.

Murphy, K. P. (2006). Naive bayes classifiers. *University of British Columbia, (vol. 18)*.

Oh, O., Kwon, K. H., & Rao, H. R. (2010). *An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010*.

Olteanu, A., Vieweg, S., & Castillo, C. (2015). *What to Expect When the Unexpected Happens: Social Media Communications Across Crises*. Paper presented at the Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work Social Computing, Vancouver, BC, Canada.

Pipek, V., Liu, S. B., & Kerne, A. (2014). Crisis Informatics and Collaboration: A Brief Introduction. *Computer Supported Cooperative Work (CSCW), 23*(4), 339-345.

Plotnick, L., Hiltz, S. R., Kushma, J. A., & Tapia, A. (2015). Red Tape: Attitudes and Issues Related to Use of Social Media by US County-Level Emergency Managers.

Power, R., Robinson, B., & Ratcliffe, D. (2013). *Finding fires with twitter.* Paper presented at the Proceedings of the Australasian Language Technology Association (ALTA) Workshop, Brisbane, Australia.

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

Reuter, C., & Kaufhold, M. A. (2018). Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics. *Journal of Contingencies and Crisis Management, 26*(1), 41-57.

Reuter, C., Ludwig, T., Kaufhold, M.-A., & Spielhofer, T. (2016). Emergency services' attitudes towards social media: A quantitative and qualitative survey across Europe. *International Journal of Human-Computer Studies, 95*, 96-111.

Ripley, B. (2002). Statistical Data Mining. *Springer-Verlag, New York.*

Rogstadius, J., Vukovic, M., Teixeira, C., Kostakos, V., Karapanos, E., & Laredo, J. (2013). CrisisTracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development, 57*(5), 4: 1-4: 13.

Rogstadius, J., Vukovic, M., Teixeira, C. A., Kostakos, V., Karapanos, E., & Laredo, J. A. (2013). CrisisTracker: crowdsourced social media curation for disaster awareness. *IBM J. Res. Dev., 57*(5), 4-4.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors.* Paper presented at the Proceedings of the 19th international conference on World wide web (pp. 851-860). ACM.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management, 24*(5), 513-523.

Schulz, A., Mencía, E. L., Dang, T. T., & Schmidt, B. (2014). Evaluating Multi-label Classification of Incident-related Tweets. *Making Sense of Microposts (# Microposts2014),* (pp. 26-33).

Schulz, A., Ristoski, P., & Paulheim, H. (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*, (pp. 22-33). Springer, Berlin, Heidelberg.

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

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). *Short text classification in twitter to improve information filtering.* Paper presented at the Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.

St Denis, L., Palen, L., & Anderson, K. M. (2014). *Mastering Social Media: An Analysis of Jefferson County's Communications during the 2013 Colorado Floods.* Paper presented at the Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM 20014).

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *iConference 2014 Proceedings*.

Stieglitz, S., Bunker, D., Mirbabaie, M., & Ehnis, C. (2017). Sense‐making in social media during extreme events. *Journal of Contingencies and Crisis Management*, 26(1), 4-15.

Takahashi, B., Tandoc, E. C., & Carmichael, C. (2015). Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines. *Computers in Human Behavior, 50*, 392-398.

Tapia, A. H., & Moore, K. (2014). Good Enough is Good Enough: Overcoming Disaster Response Organizations' Slow Social Media Data Adoption. *Computer Supported Cooperative Work (CSCW), 23*(4-6), 483-512.

Tapia, A. H., Moore, K. A., & Johnson, N. J. (2013). *Beyond the trustworthy tweet: A deeper understanding of microblogged data use by disaster response and humanitarian relief organizations.* Paper presented at the Proceedings of the 10th International ISCRAM Conference.

Thornton, J., DeAngelus, M., & Miller, B. A. (2016, 10-11 May 2016). *Feedback-based social media filtering tool for improved situational awareness.* Paper presented at the 2016 IEEE Symposium on Technologies for Homeland Security (HST).

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

Turoff, M., Hiltz, S. R., White, C., Plotnick, L., Hendela, A., & Yoa, X. (2009). The past as the future of emergency preparedness and management. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM), 1*(1), 12-28.

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., . . . Anderson, K. M. (2011). *Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency.* Paper presented at the ICWSM (pp. 385-392).

Walle, B. V. D., Turoff, M., & Hiltz, S. R. (2009). *Information Systems for Emergency Management*: M.E.Sharpe.

Webberley, W. M., Allen, S. M., & Whitaker, R. M. (2015). Retweeting beyond expectation: Inferring interestingness in Twitter. *Computer Communications* 73:229–235.

Wu, Y., Liu, S., Yan, K., Liu, M., & Wu, F. (2014). OpinionFlow: Visual analysis of opinion diffusion on social media. *Visualization and Computer Graphics, IEEE Transactions on, 20*(12), 1763-1772.

Xia, X., Yang, X., Wu, C., Li, S., & Bao, L. (2012). *Information credibility on twitter in emergency situation*. Paper presented at the Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics, Kuala Lumpur, Malaysia.

Xu, P., Wu, Y., Wei, E., Peng, T.-Q., Liu, S., Zhu, J. J., & Qu, H. (2013). Visual analysis of topic competition on social media. *Visualization and Computer Graphics, IEEE Transactions on, 19*(12), 2012-2021.

Yablonski, C. (2013). The Impact of Social Media on the Calgary Flood.

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

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. *World Medical & Health Policy, 2*(2), 7-33.

Appendix A- List of publications

The progress of this research has been presented in a number of publications as follows:

Mahshid Marbouti, Craig Anslow, Frank Maurer (2018). Evaluation results for a Social Media Analyst Responding Tool. *To Appear* In Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM), Rochester, New York, USA.

Mahshid Marbouti, Irene Mayor, Dianna Yim, Frank Maurer (2017). Social Media Analyst Responding Tool: A Visual Analytics Prototype to Identify Relevant Tweets in Emergency Events. In Proceedings of the 14th International Conference on Information Systems for Crisis Response and Management (ISCRAM), Albi, France.

Mahshid Marbouti and Frank Maurer (2016). Social Media Use During Emergency Response-insights form 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.

Mahshid Marbouti, Theodore D. Hellmann, Frank Maurer (2016). Social Media Visual Analytics for Emergency Management: A Systematic Mapping. In CSCW 2016 Workshop on Collaboration and Decision Making in Crisis Situations (CADMICS), San Francisco, USA.

The other papers that I have published during my PhD studies are:

Maleknaz Nayebi, **Mahshid Marbouti**, Rachel Quapp, Frank Maurer, Guenther Ruhe (2017). Crowdsourced Exploration of Mobile App Features: A Case Study of the Fort McMurray Wildfire. *In Proceedings of the 39th International Conference on Software Engineering (ICSE*), pp 57-66. ACM.

Zahra Sahaf**, Mahshid Marbouti,** Haleh Alemasoom, Roberta Cabral Mota, Mario Costa Sousa and Frank Maurer (2017). PipeVis: Interactive Visual Exploration of Pipeline Incident Data. In International EuroVis Workshop on Visual Analytics (EuroVA). Barcelona, Spain.

Mahshid Marbouti, Rahul Bhaskar, Zahra Shakeri, Craig Anslow, Leland Jackson, Frank Maurer (2016). . In: Moshirpour M., Far B., Alhajj R. (eds) Highlighting the Importance of Big Data Management and Analysis for Various Applications. Studies in Big Data, vol 27. Springer, Cham

Edwin Chan, Dan Gonzalez, **Mahshid Marbouti**, Irene Mayor, Sydney Pratte, Teddy Seyed, Yuxi Wang, Jade White, Frank Maurer (2016). Multi-Surface Systems for the Emergency Operations Centre of the Future. In CSCW 2016 Workshop on Collaboration and Decision Making in Crisis Situations (CADMICS), San Francisco, USA

Rahul Kamal Bhaskar, **Mahshid Marbouti**, Julia Paredes, Waliur Rahmann, Ostap Orobets, Amir Karbalaei, Xuezhang Xu, Craig Anslow, and Frank Maurer (2014). LibraryVis: Towards Understanding Library Resource Usage. In Proceedings of the IEEE VIS Workshop on BusinessVisualization (BusinessVis), Paris, France.

Appendix B- Requirement Elicitation Interview Questions

Interview Overview

Target Participants

Those who are working with Emergency Operation Centers / Operation Centers/ Emergency Planning and Response, Disaster Recovery.

Study Type

Semi-structured interviews will be used. In semi-structured interviews, an initial set of pre-defined interview questions are used for each participant, and follow-up questions may be used to gather additional details based on participants' responses.

Interview Questions

1. During an emergency situation are you typically stationed **within the EOC** or physically **outside the EOC** in another location? Which sectors do you most communicate with?
2. What types of **tasks** and **decisions** this job actively participates in and which ones you are responsible for? What are the three most **labour intensive tasks** for you in an emergency situation or exercise? Could they be automated?
3. Talk about **flow of information** between and within organizations, could it `be **improved**?
    a. **Ex.** Please describe the most recent medium- or large-scale emergency in which you were involved.
    b. What **sources of information** did you use in dealing with this situation?
    c. What **type of information** did you need, and did these sources supply it?
    d. What **tools** did you use for analyzing and sharing this information?
        i. What did you like about [tools mentioned]?
        ii. What didn't you like about [tools mentioned]?
4. What kind of **support system, tools or technology** could have helped you?
5. What **information** is most important to support **decision making**
6. How do you make **critical decisions** (Collaborative Vs. Individual)
    a. **Ex.** Please describe the last time you made a critical decision as part of the response for an emergency.
    b. How much **time** did you have to make this decision?
        i. How much do you feel that time pressure impacted your decision?
    c. What **resources** did you use to help you make this decision?
    d. Did you **collaborate** with anyone else on this decision?
        i. Were these people working in the same role as you were?
        ii. Were these people working with the same agency as you were?
        iii. Are there issues when working with people from a different agency?
    ~~e. Do you know if you made a good decision? How?~~
    f. What kind of **support system** could have helped you?
    g. *[What would **optimize** or facilitate the process by which you **make critical decisions**?]*
7. Have you used any information from **social media** or the general public?
    a. What information could you get from the public, but not get from other sources?
    b. Were there any special issues you encountered when working with social media?
    c. Have you used social media to communicate with the public?

Appendix C- Evaluation Study Tasks

1

Purpose	Tasks
The purpose of these Tasks is to let user explore online social media, find appropriate keywords or geo-location for a target online event and ask them use SMA client to define collectors and gather data.	1. Create Event
	1. Create live collectors (as much as needed.)
	1. Create Historic Collectors (as much as needed.)
	4. Create Filtered events using the collectors/extracted topics, as much as needed.

2

Purpose	Task
The purpose of the following task is based on an after-action review on historical twitter data to understand what data was available via twitter. We use this as a training task to familiarize the participants with the tool. The goal is to evaluate their understanding of the environment, visual data analysis and reasoning, and user experience	5. Explore filtered events for a historical event, read and tag tweets and trains the machine learning if applicable. User can iterate through this task until the result is "good enough". Good enough is when user can get a sense of what's going on and the AI tags seems reliable. They can perform tagging each tweet as (Sympathy, Detail, Media.) or (SA or Not SA) and expect that AI learn these labels.

While performing this task, please answer the following question:

a. Give an overview of the situation: participants are asked to find relevant messages and events.
b. Find the important topics that people are talking about
c. Find the region most affected by the event
d. Provide a report of the tweets in which people provided detail
e. What are public questions and concerns

3

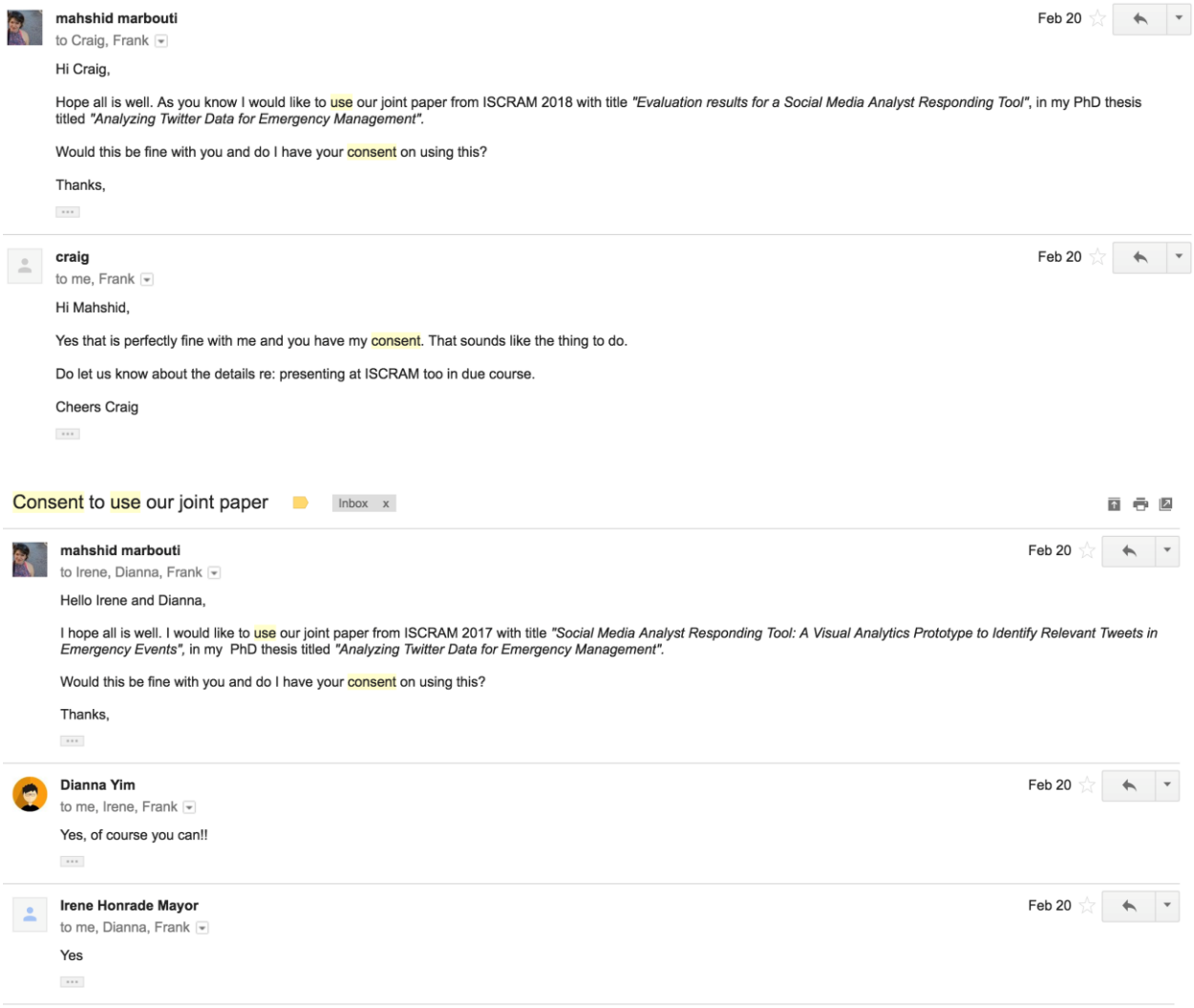
Purpose	Task
	5. Explore through filtered events for the live event (Perform previous tasks for a live event) read and tag tweets and trains the machine learning if applicable. User can iterate through this task until the result is "good enough". Good enough is when user can get a sense of what's going on and the AI tags seems reliable. They can perform tagging each tweet as (Sympathy, Detail, Media) or (SA or Not SA) and expect that AI learn these labels.

While performing this task, please answer the following question:

a. Give an overview of the situation: participants are asked to find relevant messages and events.
b. Find the important topics that people are talking about
c. Find the region most affected by the event
d. Provide a report of the tweets in which people provided detail
e. What are public questions and concerns

Appendix D- Permission of Copy Right

Consent from Co-Authors.



mahshid marbouti Feb 20
to Craig, Frank

Hi Craig,

Hope all is well. As you know I would like to use our joint paper from ISCRAM 2018 with title *"Evaluation results for a Social Media Analyst Responding Tool"*, in my PhD thesis titled *"Analyzing Twitter Data for Emergency Management"*.

Would this be fine with you and do I have your consent on using this?

Thanks,

craig Feb 20
to me, Frank

Hi Mahshid,

Yes that is perfectly fine with me and you have my consent. That sounds like the thing to do.

Do let us know about the details re: presenting at ISCRAM too in due course.

Cheers Craig

Consent to use our joint paper Inbox x

mahshid marbouti Feb 20
to Irene, Dianna, Frank

Hello Irene and Dianna,

I hope all is well. I would like to use our joint paper from ISCRAM 2017 with title *"Social Media Analyst Responding Tool: A Visual Analytics Prototype to Identify Relevant Tweets in Emergency Events"*, in my PhD thesis titled *"Analyzing Twitter Data for Emergency Management"*.

Would this be fine with you and do I have your consent on using this?

Thanks,

Dianna Yim Feb 20
to me, Irene, Frank

Yes, of course you can!!

Irene Honrade Mayor Feb 20
to me, Dianna, Frank

Yes

Appendix E- Consent Form for Requirement Elicitation Study Participants

Name of Researcher, Faculty, Department, Telephone & Email:

Supervisor:

Title of Project:

Requirements Elicitation (For using Information Technology in Emergency Management)

Sponsor:

NSERC CRD

This consent form, a copy of which has been given to you, is only part of the process of informed consent. If you want more details about something mentioned here, or information not included here, you should feel free to ask. Please take the time to read this carefully and to understand any accompanying information.

The University of Calgary Conjoint Faculties Research Ethics Board has approved this research study.

Purpose of the Study

The purpose of this study is a requirements elicitation study for using information technology in Emergency Management. We are interested in learning more about emergency events and would like to gather requirements for the future. If you are interested we invite you to participate in this study.

What Will I Be Asked To Do?

If you agree to participate in this study, you will be asked to participate in the following different research activities:

1. You will be asked to complete a short questionnaire about emergency events and your background experience. This questionnaire should take no more than 5 minutes to complete.
2. You may be asked to participate in either an interview/focus group about emergency events. A researcher will observe, take notes, and videotape or audiotape your responses for approximately 60-90 minutes or until you feel comfortable with your answers.

The whole process was designed to last no longer than approximately 95 minutes, although if you feel that you can discuss longer on certain topics or may have additional insights you think that are important, feel free to talk about it.

Your participation in this research is voluntary. You may refuse to participate altogether or in part. You may withdraw from participation in this study at any time without penalty or loss of benefits

What Type of Personal Information Will Be Collected?

Should you agree to participate we ask to videotape or audiotape you in an interview or focus group, and to audiotape your responses during the interview. Other than these video and audio recordings, no other personal identifying information (such as your name) will be collected. By default, in all written publications and presentations based on this research, you will remain anonymous and your comments from the interviews will be referred to with either a participant number or a pseudonym.

In order to better communicate the results of this research in written publications and presentations, it may be helpful to share video (or still photographs from the video) of you in an interview or focus group. If you grant us permission to share video (or still photographs from the video) of yourself in an interview or focus group, in written publications or presentations of this research, there is a chance that you may be recognized and so we cannot guarantee your anonymity. We will never, however, reveal your name in association with your image.

Please note that, where intended reporting of photographed or videotaped images includes public display, the researchers will have no control over any future use by others who may copy the images and repost them in different formats or contexts, including online.

I grant permission to be audio taped: Yes: ___ No: ___

I grant permission to be video taped: Yes: ___ No: ___

I grant permission to have my company's name used: Yes: ___ No: ___

Are there Risks or Benefits if I Participate?

There is no known harms or risks associated to the participation in this study.

You will NOT be compensated for your time.

What Happens to the Information I Provide?

Participation in this research is completely voluntary and confidential. You are free to discontinue participation at any time during the study. Any information you contribute up to the point at which you choose to discontinue your participation will be retained and used in the study. No one except the researchers will be allowed to see or hear any personally identifiable information unless you have given permission for us to share video or photographs of you in our interview or focus group, in publications or presentations of this research. The audio/video tapes, questionnaires and interview data will be kept on password-protected university computers or in a locked cabinet only accessible by the researchers. The data will be stored for five years, after which it will be permanently erased.

Signatures

Your signature on this form indicates that 1) you understand to your satisfaction the information provided to you about your participation in this research project, and 2) you agree to participate in the research project.

In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. You are free to withdraw from this research project at any time. You should feel free to ask for clarification or new information throughout your participation.

Participant's Name: (please print) _____

Participant's Signature: _____ Date: _____

Researcher's Name: (please print) _____

Researcher's Signature: _____ Date: _____

Questions/Concerns

If you have any further questions or want clarification regarding this research and/or your participation, please contact:

If you have any concerns about the way you've been treated as a participant, please contact an Ethics Resource Officer, Research Services Office, University of Calgary at (403) 210-9863; email cfreb@ucalgary.ca.

A copy of this consent form has been given to you to keep for your records and reference. The investigator has kept a copy of the consent form.

Appendix F- Consent Forms for Evaluation Study Participants

Name of Researcher, Faculty, Department, Telephone & Email:

Supervisors:

Title of Project:

Social Media Analyst Client Usability Study (A visual analytics tool for gaining situation awareness from social media data)

This consent form, a copy of which has been given to you, is only part of the process of informed consent. If you want more details about something mentioned here, or information not included here, you should feel free to ask. Please take the time to read this carefully and to understand any accompanying information.

The University of Calgary Conjoint Faculties Research Ethics Board has approved this research study.

Purpose of the Study:

The purpose of this study is to evaluate the effectiveness of using online training in a social media analyst client tool. The purpose of the tool is to monitor social media data for extracting situation awareness information during an emergency event.

What Will I Be Asked To Do?

If you agree to participate in this study, you will be asked to participate in the following different research activities:

1. You will be asked to complete a pre-study questionnaire. This questionnaire should take less than 5 minutes.
2. You will be asked to perform a practice round of tasks where you will be given a description and demonstration of each task. The practice session should take about 10 minutes.
3. You will then be asked to complete number of tasks. A researcher will observe, take notes, and videotape/audiotape your responses for up to 60 minutes or until you feel satisfied with your answers.
4. After completing the study, you may be asked to complete a post study questionnaire about your experience in the study. This questionnaire will take about 10 minutes.

During the study we encourage you to talk aloud.

The whole process should last approximately 90 minutes. Participation is voluntary and you may withdraw at any time.

What Type of Personal Information Will Be Collected?

If you agree to participate, we ask to videotape you, and to audiotape your comments. Other than these video and audio recordings, no other personal identifying information (such as your full name) will be collected. By default, in all written publications and presentations based on this research, you will remain anonymous and your comments from the interviews will be referred to with either a participant number or a pseudonym.

In order to better communicate the results of this research in written publications and presentations, it may be helpful to share video (or still photographs from the video) of you in an interview. If you grant us permission to share video (or still photographs from the video) of yourself in an interview, in written publications or presentations of this research, there is a chance that you may be recognized and so we cannot guarantee your anonymity. However, steps to hide your identity will be taken such as blurring your face if you choose to remain anonymous. We will never, however, reveal your name in association with your image.

Please note that, where intended reporting of photographed or videotaped images includes public display, the researchers will have no control over any future use by others who may copy the images and repost them in different formats or contexts, including online.

There are several options for you to consider if you decide to take part in this research. You can choose all, some or none of them. Please put a check mark on the corresponding line(s) that grants us your permission to:

I grant permission to save task data:	Yes: ___ No: ___
I grant permission to save questionnaire data:	Yes: ___ No: ___
I grant permission to be audio taped	Yes: ___ No: ___
I grant permission to be video taped:	Yes: ___ No: ___
I grant permission to share the recorded video:	Yes: ___ No: ___
I wish to remain anonymous:	Yes: ___ No: ___

Are there Risks or Benefits if I Participate?

There are no known harms or risks associated to the participation in this study.

What Happens to the Information I Provide?

Participation is completely voluntary, anonymous and confidential. You are free to discontinue participation at any time during the study. No one except the researchers and their supervisors will be allowed to see or hear any of the answers to the questionnaires or task performance. There are no names on the questionnaires. The questionnaires are kept in a locked cabinet only accessible by the researchers and their supervisor. All the collected data will be kept by the investigators indefinitely and it might be used, in anonymized form, with publications and thesis purposes.

Signatures (written consent)

Your signature on this form indicates that you 1) understand to your satisfaction the information provided to you about your participation in this research project, and 2) agree to participate as a research subject.

In no way does this waive your legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. You are free to withdraw from this

research project at any time. You should feel free to ask for clarification or new information throughout your participation.

Participant's Name: (please print) _____

Participant's Signature _____Date: _____

Researcher's Name: (please print) _____

Researcher's Signature: _____Date: _____

Questions/Concerns

If you have any further questions or want clarification regarding this research and/or your participation, please contact:

If you have any concerns about the way you've been treated as a participant, please contact the Research Ethics Analyst, Research Services Office, University of Calgary at (403) 210-9863; email cfreb@ucalgary.ca.

A copy of this consent form has been given to you to keep for your records and reference. The investigator has kept a copy of the consent form.