

Social Media and Crowdsourcing Assessment of Weather Impacts

Project Report

GCXE19M010



Environment and
Climate Change Canada

Environnement et
Changement climatique Canada

Social Media and Crowdsourcing Assessment of Weather Impacts

About	4
Executive summary	5
1. Introduction	7
2. Literature Review	9
2.1 Social Media To Report Extreme Weather	9
2.2 Computational Techniques to Leverage Social Media for Crisis Managers	10
2.3 Gaps and Caveats	19
3. Grant activities	22
3.1 Scoping the state-of-the-art	22
3.2 Developing the infrastructure	26
3.3 Building a data pipeline	28
3.4 Conducting a social network analysis	32
3.5 Automatically classifying social media during disasters	37
3.6 Adapting social media classification to Canadian snowstorms	45
3.7 Dashboard	54
4. Project Conclusions and Outcomes	57

About

About SMaCAWI

To fulfill the grant and contract, an interdisciplinary team, called Social Media Crowdsourcing Assessment of Weather Impacts (SMaCAWI) was assembled. It was led by Professor Renee Sieber at McGill University. Team members include Andrei Mircea, Rosie Zhao, Mikael Brunila, Sam Lumley, Lucia Berger, Sarah Greenidge, Hannah Ker, Stefan Morgan and Drew Bush.

About the authors

This report was produced by Renee Sieber, Andrei Mircea, Rosie Zhao, Mikael Brunila and Sam Lumley.

Acknowledgements

This research was funded by a grant and contract from Environment and Climate Change Canada (ECCC), Social Media Crowdsourcing Assessment of Weather Impacts. We extend our thanks and appreciation to the many ECCC colleagues who provided invaluable input and feedback throughout the project.

Executive summary

Purpose of research

In this grant and contract, we investigated how crisis managers can effectively use social media for real-time situational awareness. We had four objectives.

1. A literature review on how crisis managers and public agencies are using social media contributed by the public related to extreme weather
2. A methodology for assessing where do most people in Canada and the United States get their news on extreme weather
3. A methodology for the use of artificial intelligence (AI) modelling that allows us to measure public reaction to weather events. This includes two repositories of training data and two NLP models
4. Outreach to share relevant information with diverse communities and training so students can be familiar with ECCC concerns.

Method and scope

In this project, we built upon the literature of social media responses to extreme weather and recent developments in artificial intelligence. We developed supervised classification models that automatically filter and organize Tweets into salient categories to leverage noisy social media data. We augmented Tweet categorization with a social network analysis (SNA) to assess key influencers on public responses to extreme weather. We developed unsupervised ML techniques to identify categories for snowstorms. We then applied these models in various case studies of extreme weather events, including the March 2019 Nebraska Flood and the January 2020 Newfoundland-Labrador Snowstorm. We developed automatic and human methods to evaluate our models and compare their performance to previous methods. Lastly, we created a prototype dashboard allowing crisis managers to visualize geolocated and categorized tweets in real-time during extreme weather events.

Findings and outputs

The ML models developed and applied throughout this project improved upon methods previously reported in the literature. These models allowed us to assess the impact of extreme weather events by automatically categorizing tweets that, for example, report “Infrastructure and utilities damage”. In a case study of the 2019 Nebraska flood,¹ we combined these models with SNA to characterize how affected people interact with and respond to various sources, including

¹ Romascanu, A., Ker, H., Sieber, R., Greenidge, S., Lumley, S., Bush, D., Morgan, S., Zhao, R., & Brunila, M. (2020). Using deep learning and social network analysis to understand and manage extreme flooding. *Journal of Contingencies and Crisis Management* <https://doi.org/10.1111/1468-5973.12311>

organizations, elected officials and celebrities. In a case study on a 2020 Newfoundland-Labrador Snowstorm,² we extended these models to automatically discover novel topics that are specific to snow-related events and relevant to crisis managers, helping fill a significant gap in the crisis management literature. Lastly, our prototype dashboard³ enabled us to explore how our findings could be applied to real-world applications in crisis management. The code for these three publications and supporting infrastructure has been made publicly available as open-source software.

This G&C was successful in terms of outreach and training. Over the course of the project from March 2018 to March 2021, SMaCAWI trained nine research assistants: graduate students, undergraduate students, and a postdoctoral fellow. Our findings were presented at conferences, including CMOS 2020. We gave two presentations to ECCC staff. We showcased our work at the World Meteorological Organization's (WMO) HiWeather Impact Workshop and at an applied computer science conference. One student gave a talk in a computer science seminar series; another used our data repository. Our methods are currently being tested in a computer science course on AI in Climate Change. Team members developed connections to MILA, the Quebec AI Institute, which is one of the top AI research institutes in Canada. Finally, McGill University showcased the ECCC as part of its own public relations.

Outlook

We showed that using AI to automatically classify social media content can be useful for assessing the impact of extreme weather events and the reactions of affected people. In particular, this approach can effectively structure large amounts of content to help crisis managers extract real-time information on how people are responding to an extreme weather event. We also identified machine learning (ML) expertise and computational infrastructure that was required to build and maintain as a key bottleneck. Lastly, our work demonstrated the need for human involvement at every stage of this process to limit algorithmic bias and ensure the use of artificial intelligence is aligned with the problem being addressed.

² Brunila, M., Zhao, R., Mircea, A., Lumley, S., & Sieber, R. (2021). Bridging the gap between supervised classification and unsupervised topic modelling for social-media assisted crisis management. Adapt-NLP Workshop. European Chapter of the Association for Computational Linguistics (EACL) April 21, 2021. <https://www.aclweb.org/anthology/2021.adaptnlp-1.5/>

³ Mircea, A. (2020). Real-time Classification, Geolocation and Interactive Visualization of COVID-19 Information Shared on Social Media to Better Understand Global Developments. <https://doi.org/10.18653/v1/2020.nlpCOVID19-2.37>

1. Introduction

As climate change increases the frequency of extreme weather events and the vulnerability of affected people, effective crisis management is becoming increasingly important for mitigating the negative effects of these crises. In the crisis management literature, social media has been identified as a useful source of information for crisis managers to gauge reactions from and communicate with the public, increase situational awareness, and enable data-driven decision-making.

Led by Renee Sieber and funded by Environment and Climate Change Canada, the Social Media Crowdsourcing Assessment of Weather Impacts (SMaCAWI) team was created to determine how social media platforms can be used to understand public reactions to extreme weather events.

This report documents the objectives, activities and findings from the SMaCAWI G&C. In the Literature Review section, we review the state of the art in computational approaches to leverage social media for management of extreme weather events. This includes key terms and concepts in natural language processing (NLP) and social media network analysis, which form the primary methods we used for extracting information from social media. In the Grant Activities section, we describe the objectives, methods and outcome of each activity. Finally, we conclude with key takeaways and outlooks for future research.

We investigated how crisis managers can effectively use social media for real-time situational awareness. We addressed this through four objectives:

1. A literature review on how crisis managers and public agencies are using social media contributed by the public related to extreme weather
2. A methodology for assessing where people in Canada and the United States get their news on extreme weather
3. A methodology for use of artificial intelligence (AI) modelling that allows us to measure public reaction to weather events. This includes two repositories of training data and two NLP models
4. Outreach to share relevant information with diverse communities and training so students can be familiar with ECCC concerns.

These objectives were operationalized through seven grant activities:

1. Scoping the state-of-the-art
 - Assess state-of-the-art research literature
 - Identify options for historic/live access to Tweets

2. Developing the infrastructure
 - Build a server/workstation for deep learning and data streaming/processing
3. Building a data pipeline
 - Develop API for live and historic harvesting (scraping) of Twitter during extreme weather events
 - Build data storage solution for harvested tweets
 - Explore different data scraping methodologies (e.g., keywords, hashtags, geotags, accounts)
 - Test live scraping case study on extreme weather event
4. Automatically classifying social media during disasters
 - Explore different ML techniques for automatic tweet labeling
 - Perform case study to characterize extreme weather events
5. Conducting social network analysis
 - Identify official accounts on Twitter and potential to measure reactions
 - Characterize flow of information
6. Adapting social media classification to Canadian snowstorms
 - Perform human evaluation of automatic tweet labelling to identify limitations
 - Improve usefulness of automatic tweet labelling to Canadian crisis managers
7. Building a dashboard to the models
 - Build an interface to allow crisis managers to harvest, classify and visualize Tweets

2. Literature Review

This section describes the current literature as well as key terms and concepts covered in this project.

2.1 Social Media To Report Extreme Weather

Social media plays an increasingly important role in how people, governments, organizations, and institutions interact and respond during crisis events.^{4,5} For example, the Australian Government Crisis Coordination Centre used information from Twitter to provide citizens with enhanced situational awareness during emergencies.⁶ In a 2012 earthquake in Indonesia, an early tweet from Indonesia's Meteorological, Climatological, and Geophysical Agency effectively mobilized citizens and government agencies to diffuse important information about the disaster early on in its development.⁷

The ubiquity of mobile phones and the accessibility of social media platforms like Twitter allow citizens to document events for a public audience with an immediacy that would not be possible using traditional media outlets.⁸ Traditional media figures, such as journalists and politicians, also have increased their use of social media during crises to share information.⁹ In addition to communication, social media has served to bind people together and increase trust between social institutions and citizens to collectively address crises.¹⁰ Social media, like Twitter data, holds such importance to managing crises that the related fields of *crisis informatics* and *crisis analytics* emerged to understand the various ways in which that data can create near real-time situational awareness within the public.^{11,12,13}

⁴ The original version of this section appears in our paper, Romascanu, et al. (2020).

⁵ Reuter, C., Stieglitz, S., & Imran, M. (2019). Social media in conflicts and crises. *Behaviour & Information Technology*, 0(0), 1–11. <https://doi.org/10.1080/0144929X.2019.1629025>

⁶ Cameron, M. A., Power, R., Robinson, B., & Yin, J. (2012). Emergency Situation Awareness from Twitter for Crisis Management. *Proceedings of the 21st International Conference on World Wide Web*, 695–698. <https://doi.org/10.1145/2187980.2188183>

⁷ Chatfield, A., & Brajawidagda, U. (2012). Twitter tsunami early warning network: A social network analysis of Twitter information flows. *ACIS 2012 : Location, Location, Location : Proceedings of the 23rd Australasian Conference on Information Systems 2012*, 12.

⁸ Murthy, D. (2011). Twitter: Microphone for the masses? - Dhiraj Murthy, 2011. *Media, Culture & Society*, 33(5), 779–789.

⁹ Vis, F. (2013). Twitter as a Reporting Tool for Breaking News. *Digital Journalism*, 1(1), 27–47. <https://doi.org/10.1080/21670811.2012.741316>

¹⁰ abs/1610.01561

Solnit, R. (2009). *A Paradise Built in Hell: The Extraordinary Communities that Arise in Disasters* Viking. New York.

¹¹ Cameron et al., (2012).

¹² Helsloot, I., & Groenendaal, J. (2013). Twitter: An Underutilized Potential during Sudden Crises? *Journal of Contingencies and Crisis Management*, 21(3), 178–183. <https://doi.org/10.1111/1468-5973.12023>

¹³ Qadir, J., Ali, A., ur Rasool, R., Zwitter, A., Sathiaselalan, A., & Crowcroft, J. (2016). Crisis analytics: Big data-driven crisis response. *Journal of International Humanitarian Action*, 1(1), 12. <https://doi.org/10.1186/s41018-016-0013-9>

Near real-time social media use during a crisis generates a wealth of valuable information. However, the large volume of data and irrelevant content have made it difficult for crisis managers to use.¹⁴ Social media data is largely unstructured and challenging to represent in traditional databases. Analyses that leverage social media can be slow to scale during high traffic making it hard to generate useful findings within the timescales of crisis events.^{15,16} Whereas such analyses can inform decision-making in the long term; analyses may struggle to keep pace with the dynamic and evolving needs that an effective crisis response demands.¹⁷

Computational approaches have the potential to address the practical challenges of processing and analyzing social media data in fast evolving crises. ML, for example, allows the categorization of text or the attribution of sentiment. These techniques can enable crisis managers to more easily filter through large volumes of data.¹⁸ Social network analysis (SNA) also can provide insights into the complex networks through which information diffuses on social media platforms. This method allows crisis managers to identify key information sources in their social networks.¹⁹ These computational approaches are increasingly available (e.g., in software libraries) and seemingly fast and easy-to-use. However, the metrics used to assess effectiveness can be misunderstood by, the systems can be challenging to set up and interpret. This drives our work.

2.2 Computational Techniques to Leverage Social Media for Crisis Managers

Let us first consider what is reported on the use of ML for crisis management. Existing research has investigated the unique challenges that crises pose when attempting to apply ML models across multiple events.²⁰ ML is designed to be generalizable, so that once trained on one or more standardized instances, an ML model can be transferred to similar incidents. Previous research efforts have developed ML models to analyze unstructured text content of messages generated on

¹⁴ Hiltz, S. R., Kushma, J., & Plotnick, L. (2014). Use of Social Media by U.S. Public Sector Emergency Managers: Barriers and Wish Lists. ISCRAM. <https://doi.org/10.13140/2.1.3122.4005>

¹⁵ Qadir et al., (2016).

¹⁶ Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39, 156–168. <https://doi.org/10.1016/j.ijinfomgt.2017.12.002>

¹⁷ Hiltz et al., (2014).

¹⁸ Nguyen, D. T., Joty, S., Imran, M., Sajjad, H., & Mitra, P. (2016). Applications of Online Deep Learning for Crisis Response Using Social Media Information. *ArXiv:1610.01030 [Cs]*. <http://arxiv.org/abs/1610.01030>

¹⁹ Gupta, A., Joshi, A., & Kumaraguru, P. (2012). Identifying and characterizing user communities on Twitter during crisis events. *Proceedings of the 2012 Workshop on Data-Driven User Behavioral Modelling and Mining from Social Media*, 23. <https://doi.org/10.1145/2390131.2390142>

²⁰ Li, H., Caragea, D., Caragea, C., & Herndon, N. (2018). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1), 16–27. <https://doi.org/10.1111/1468-5973.12194>

social media during crisis events.^{21,22} For example, ML has been used to classify text content according to the stages of a crisis²³ and the sentiment of the public reacting to a crisis.²⁴ Researchers also have built ML models to map content based on location markers in the text.^{25,26} Reynard and Shirgaokar used ML to assist with geolocation and sentiment classification of tweets to help guide resource allocation during a natural disaster.²⁷ Nguyen et al. used a type of deep learning—deep learning is that type of AI that involves neural networks—called a convolution neural network (CNN) to attempt to automatically detect the level of crisis damage from images.²⁸

ML has been framed as an effective approach for filtering out noisy and irrelevant information to accommodate the speed of real-time social media data during a crisis.^{29,30,31,32} It is easy to overlook the amount of effort and computational expertise required to effectively apply ML, resources that might not be easily or quickly accessible to crisis managers. As part of their work on CNN for text, Nguyen et al. described the challenges of interpreting short messages using ML because small messages like tweets contain less data, are often informal and unstructured, using slang or misspellings.³³ Much of the work in the use of computational methods has focused on NLP models to automatically classify tweets into finer-grained and categories that can be more salient to crisis managers and affected people in rapidly evolving situations.^{34,35} Derczynski et al.

²¹ Buscaldi, D., & Hernandez-Farias, I. (2015). Sentiment analysis on microblogs for natural disasters management: A study on the 2014 genoa floodings. *Proceedings of the 24th International Conference on World Wide Web*, 1185–1188.

²² Reynard, D., & Shirgaokar, M. (2019). Harnessing the power of machine learning: Can Twitter data be useful in guiding resource allocation decisions during a natural disaster? *Transportation Research Part D: Transport and Environment*. <https://doi.org/10.1016/j.trd.2019.03.002>

²³ Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., Schram, A., & Anderson, K. M. (2011). Natural language processing to the rescue? Extracting “situational awareness” tweets during mass emergency. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 385–391. Association for the Advancement of Artificial Intelligence.

²⁴ Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. *Sentiment Analysis and Ontology Engineering*, 313–340.

²⁵ Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, 80, 56–71.

²⁶ Ghahremanlou, L., Sherchan, W., & Thom, J. A. (2015). Geotagging twitter messages in crisis management. *The Computer Journal*, 58(9), 1937–1954.

²⁷ Reynard & Shirgaokar (2019).

²⁸ Nguyen, D., Alam, F., Ofli, F., & Imran, M. (2017, April 9). Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises. *Proceedings of the 14th ISCRAM Conference – Albi, France*.

²⁹ Imran et al., (2013)

³⁰ Emmanouil, D., & Nikolaos, D. (2015). Big data analytics in prevention, preparedness, response and recovery in crisis and disaster management. *The 18th International Conference on Circuits, Systems, Communications and Computers (CSCC 2015)*, Recent Advances in Computer Engineering Series, 32, 476–482.

³¹ Nguyen et al., (2016).

³² Rao, R., Plotnick, L., & Hiltz, R. (2017, January 4). Supporting the Use of Social Media by Emergency Managers: Software Tools to Overcome Information Overload. *Proceedings of the 50th Hawaii International Conference on System Sciences*. <https://doi.org/10.24251/HICSS.2017.036>

³³ Nguyen et al. (2016).

³⁴ Ragini, J. R., & Anand, P. R. (2016). An empirical analysis and classification of crisis related tweets. *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 1–4.

³⁵ Schulz et al., (2014)

noted the need to meticulously hand-label 70,000 tweets prior to developing the ML model, into “informativeness” and “actionable” classes, utilizing and paying crowdworkers (70,000, as the authors note, is relatively small for training ML).³⁶ Training such classification models typically requires large-scale annotated corpora of crisis-related tweets such as that made available by Imran et al., which covers a variety of countries and natural disasters including flooding, tropical storms, earthquakes, and forest fires.³⁷ Indeed, effective training of ML models has been described as “a black art that requires years of experience to acquire”,³⁸ due to their complexity and long training times, as well as the as-of-yet still limited understanding of the different interactions between elements of a crisis. ML is still far from off-the-shelf for real-time or near real-time crisis management.

It would be useful to describe three main types of NLP, all three of which were used in our research.

2.2.1 Supervised text classification

NLP relies heavily on ML classification techniques, which can be broadly divided into two types of algorithms, supervised and unsupervised. The difference between these algorithms can be illustrated by considering how they would handle one of the oldest and best known statistical datasets, which is the Iris dataset collected by the statistician Ronald Fisher in 1936.³⁹ The Iris data consists of 50 samples from each of three different flower species of iris along with measurements of the length and width of the sepals and petals.

A supervised learning algorithm would be shown a sample of the dataset, for example five of each of the iris types. This is the training set, and the unsampled data is the test set. The algorithm would then learn to estimate how likely it is that an iris belongs to each of the three classes, given the iris features, which are the petal and sepal lengths and widths. Probabilistically speaking, the algorithm learns to estimate the iris label y conditioned on the different features x . How well the algorithm performs is then evaluated by showing it the iris features x of the test set and letting it guess what the label y of each iris is based on what is shown to them without learning and what it learned from the training set. An extremely good model would achieve over 90 percent accuracy on the test set.⁴⁰

Supervised learning can be extended to the domain of text. Instead of the size of sepals and petals, the learning can come from the words contained in a document. A famous dataset for this purpose is 50,000 reviews from the movie review website IMDB, half of which are labelled

³⁶ Derczynski, L., Meesters, K., Bontcheva, K., & Maynard, D. (2018). Helping Crisis Responders Find the Informative Needle in the Tweet Haystack. ArXiv:1801.09633 [Cs]. <http://arxiv.org/abs/1801.09633>

³⁷ Imran et al. (2016)

³⁸ Smith, (2018), p. 1

³⁹ Fisher, R. A. (1936). The Use Of Multiple Measurements In Taxonomic Problems. *Annals of Eugenics* (renamed *Annals of Human Genetics*). 7, 2: 179-188

⁴⁰ Collingwood, L. & J. Wilkerson, J. (2012). Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods. *Journal of Information Technology & Politics* 9:3, 298-318. DOI: 10.1080/19331681.2012.669191

positive and the other half negative.⁴¹ By teaching an algorithm which words tend to appear in the two different types of reviews, it can learn to classify reviews into the two categories.

To know which categories a given text belongs to, the model must be trained initially and that training data must be labelled. This labelling is most often done by human annotators and is a labour intensive process. Researchers decide on the categories (labels), the training data and then ask individuals to decide which content (think individual tweets or social media posts) would be best described by a single category. The supervised learning then “learns” by associating the words and word combinations that were important to that category. In the domain of crisis management, the most widely used training dataset is the CrisisNLP dataset developed and maintained by the Qatar Computing Research Institute.⁴² The dataset contains tweets from several natural disasters across the world with each tweet labelled into one out of nine categories. By looking at words as well as emojis, links and other textual features in the tweets, supervised learning teaches itself to estimate to which of the nine categories any given tweet belongs.

2.2.2 Unsupervised topic modelling

The other main type of ML algorithm is unsupervised learning. Unsupervised learning is a bottom-up approach in which the model learns directly from the test dataset. No predefined categories nor labelled datasets are needed for the algorithm. There is no overhead in terms of training the dataset, which is primarily a human annotation activity. However, one criticism of supervised learning is that it is not easily generalizable for other events.⁴³ To return to the iris dataset described in the previous section on supervised classification. If we did not know which three types of irises there were, we could use an unsupervised classifier to divide the dataset into clusters of similar irises. In this sense, unsupervised models are often used to find patterns and classes in datasets that are novel or lack annotation. Unsupervised models are therefore faster to implement but they can lack clear actionable intelligence because the clusters are not obvious from the beginning and are entirely dependent on the input dataset.

In NLP, unsupervised ML can broadly be divided into clustering and topic modelling.⁴⁴

Clustering models rely on multidimensional representations of words or documents to find some number of classes in the data. Topic models, on the other hand, are given a set of documents and then learn not only to separate them into groups based on the words used in them, but also which “latent” keywords best characterize each group or “topic”.

⁴¹ Manek, A.S., Shenoy, P.D., Mohan, M.C. *et al.* (2017). Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. *World Wide Web* 20, 135–154. <https://doi.org/10.1007/s11280-015-0381-x>

⁴² Imran, M, et al. 2015. Processing Social Media Messages in Mass Emergency: A Survey. *ACM Computing Surveys* 47(4), 1-38.

⁴³ Yu, M., Huang, Q., Qin, H., Scheele, C., Yang, C. (2019). Deep learning for real-time social media text classification for situation awareness—using Hurricanes Sandy, Harvey, and Irma as case studies. *International Journal of Digital Earth* 12(11),1230-1247, DOI: 10.1080/17538947.2019.1574316.

⁴⁴ Li, X., Lei, L. (2021). A bibliometric analysis of topic modelling studies (2000–2017). *Journal of Information Science* 47, 2, 161–175. <https://doi.org/10.1177/0165551519877049>

As an example of the power of unsupervised learning, researchers used topic modelling to examine a corpus of approximately 15 million tweets from various parts of the world referencing climate change and global warming.⁴⁵ They used the automated processing from this test dataset to identify political polarization in the word clusters or topics used to describe extreme weather events. The authors associated terms such as “climate change” with terms suggesting anthropogenic causes as opposed to “global warming”, which was associated with natural cycles. Like supervised learning, unsupervised learning can infer sentiment through its clustering.

There have been proposed methodologies where unsupervised learning or classification could assist supervised classification in crisis management. Imran and Castillo suggested using the popular unsupervised approach called Latent Dirichlet Allocation (LDA) method to investigate “Miscellaneous” categories.⁴⁶ LDA is considered latent because it reveals unseen patterns (topics) potentially deep in the dataset.⁴⁷ In this way, unsupervised could be used to augment existing sets of categories.

2.2.3 Deep Learning Methods

Deep learning, or the use of neural networks, represent an important class of ML models that have largely overtaken the field of NLP as a result of their increased performance on a wide variety of benchmark tasks.⁴⁸ Neural networks are composed of probabilistic mesh of ‘neurons’ that exist in layers; neurons interact with each other in and across layers. The neurons are tuned and refined as the data passes back and forth among them to meet the expectations of the data or the rules given to them.⁴⁹ Classification algorithms, which comprise supervised and unsupervised learning, are usually trained on massive amounts of pre-designated data. This improvement in performance metrics such as text classification accuracy comes at the cost of reduced interpretability, often referred to as the black box problem, whereby neural networks can offer no explanation for their predictions. This opacity becomes particularly problematic in understanding model failures in performance or classification.⁵⁰ The field has largely relied on the evaluation of models on benchmark tasks to quantify progress, which is further discussed below. The use of neural networks, often referred to as deep learning, has also stood out from previous ML approaches based on the observation that improvements in performance are largely attributable

⁴⁵ Al-Rawi, A., Kane, O., Bizimana, A-J. 2021. Topic modelling of public Twitter discourses, part bot, part active human user, on climate change and global warming. *Journal of Environmental Media* 2(1): 31–53
https://doi.org/10.1386/jem_00039_1

⁴⁶ Imran, M., Castillo, C. (2015). Towards a Data-driven Approach to Identify Crisis-Related Topics in Social Media Streams. Proceedings of the International World Wide Web Conference Committee, May 18–22, 2015, Florence, Italy. <http://dx.doi.org/10.1145/2740908.2741729>. ACM.

⁴⁷ Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research* 3, 993-1022.

⁴⁸ Manning, C. D. 2015. “Last Words: Computational Linguistics and Deep Learning.” *Computational Linguistics* 41(4): 701–7.

⁴⁹ Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. Cambridge, US: MIT Press.

⁵⁰ Castelvechi, Davide. 2016. “Can We Open the Black Box of AI?” *Nature News* 538(7623): 20.

to increases in neural network and training dataset sizes, instead of algorithmic advances.⁵¹ This has led to the emergence of a dominant paradigm in NLP: the use of pretrained language models, discussed below.

There are promising advances in the use of deep learning for leveraging crisis-related social media. We mentioned Nguyen et al. on their use of convolution to detect damage from images.⁵² There are efforts to blend computer vision techniques in image detection with unsupervised NLP for crisis event detection.⁵³ Considerable effort is placed on improving semantic understanding, whether augmenting traditional ML with CNN⁵⁴ or Long Short-Term Memory (LSTM).⁵⁵ Concerned about generalizing across the specifics of different hurricanes, geographers Yu et al. use CNN to provide some transfer learning functionality.⁵⁶ Imran's team 'weighed' in on graphs—graphs form the basis of SNA—to lend some of the advantages of supervised classification to any social media content that does not have training data to structure the results.⁵⁷ These papers demonstrate considerable interest in deep learning approaches for event detection, situational awareness and crisis management. But there remains a large noise-to-signal problem and an interpretability challenge as these ensemble approaches become ever more complex.

2.2.4 Pretrained language models

Machine learning models tend to perform better as they are given more data. The more exposure they have had to language, the better they are able to perform at a given task, such as classification, because they 'understand' the basic rules of language, for instance, how words tend to co-occur. Major breakthroughs have been achieved in general purpose models that are trained on vast amounts of text data.⁵⁸ These pretrained "language models" can then be "finetuned" for custom purposes, by training them on a specific dataset. For example, our project used a language model that was trained on millions of documents and billions of words to have a

⁵¹ Sutton, Rich. (2019). The Bitter Lesson. <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>. March 13, 2019.

⁵² Nguyen et al. (2017).

⁵³ Abavisani, M., Wu, L., Hu, S., Tetreault, J., Jaimes, A. (2020). Multimodal Categorization of Crisis Events in Social Media. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14679-14689

⁵⁴ Caragea, C., Silvescu, A., Tapia, A.H. (2016). Identifying informative messages in disaster events using convolutional neural networks. International Conference on Information Systems for Crisis Response and Management, pp. 137-147.

⁵⁵ Sit, M. A., Koylu, C., Demir, I. (2019). Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane Irma. *International Journal of Digital Earth* 12(11), 1205-1229. <https://doi.org/10.1080/17538947.2018.1563219>

⁵⁶ Yu et al. (2019).

⁵⁷ Alam, F., Joty, S., & Imran, M. (2018). Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. Proceedings of the International AAAI Conference on Web and Social Media, 12(1).

⁵⁸ Hirschberg, J. and C. D. Manning. (2015). Advances in natural language processing. *Science* 349 (6245), 261-266. DOI: 10.1126/science.aaa8685

general understanding of the English language.⁵⁹ We then finetuned the model on the CrisisNLP dataset, so that it became particularly sensitized to the patterns in short-texts like tweets with the particular domain of crisis related events. By using language models like this, researchers can leverage a much broader sense of language that is hard to achieve on relatively limited datasets.

Pretrained language models are similar to word embedding models such as word2vec or GloVe, in that they model the distributional semantics of text (i.e., the meaning of a word can be determined from the context in which it appears). Word embedding models achieve this by learning/creating static vector representations of words, such that they are more similar for co-occurring words. In contrast, pretrained language models use neural networks to generate dynamic vector representations that learn a word's surrounding context, such that they are more similar to the static vector of the corresponding word. In practice, these dynamic contextual representations have been shown to capture a wider variety of sentence-level linguistic phenomena.⁶⁰

Whereas language models are very powerful and often extremely useful; they also have potential drawbacks.⁶¹ Training a custom model is too costly for most people and using a pre-trained model can introduce unpredictable biases, since we do not know exactly what data a given model was originally trained on. Additionally, language models are many times less interpretable than more simple models. If we want to understand why a NLP model classified a text a certain way, language models are not always the best choice.

2.2.5 Evaluating Machine Learning Techniques

Because the field is so dynamic, the use of metrics and benchmark datasets has played a central role in evaluating progress in the field of ML. A novel method's improvement on metrics, such as classification accuracy, is typically taken as empirical evidence that the new method is better than existing alternatives and thus a contribution to the field.⁶² As deep learning methods approach human-like performance on benchmark tasks, the scope of evaluation has expanded. For example, the use of human evaluators to assess and compare model outputs is becoming more widespread, despite its increased cost, as performance metrics on benchmark datasets paint a very limited picture and may not align with real-world applications. In NLP, the need for

⁵⁹ Devlin, J., Chang, M-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1 (Long and Short Papers). DOI: 10.18653/v1/N19-1423

⁶⁰ Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. "A Primer in BERTology: What We Know About How BERT Works." *Transactions of the Association for Computational Linguistics* 8: 842–66.

⁶¹ Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y. (2016). Exploring the Limits of Language Modeling. <https://arxiv:1602.02410>.

⁶² Yu, et al. (2019).

human evaluation is further cemented due to known issues with biases in pretrained language models which may not be otherwise captured.⁶³

Kumar et al. performed a comparative analysis of conventional ML and deep learning to find the best tweet classifiers in crisis reporting in social media.⁶⁴ They tested numerous ML methods that should be familiar to several readers, including Support Vector Machine, Random Forest, Logistic Regression, K-Nearest Neighbors, Naive Bayes and Decision Trees. They tested deep learning classifiers, including CNN and Long-Short-Term-Memory (LSTM). The authors deployed two kinds of word embeddings, GloVe, a common NLP algorithm, and Crisis. (Embeddings are typical in NLP as they lump or cluster words from the dataset/corpus that occur near each other in “vector space”. The assumption is that these clusters represent concepts.) They used a training dataset to compare the systems.⁶⁵ They found that Gradient Boosting, a type of decision tree algorithm, performed the best. The authors also noted that “Crisis embedding performed best in case of earthquake and GloVe embedding performed best in case of wildfire.”⁶⁶ This shows how, even in ostensibly automated methods, evaluation of even the sub components, like word embeddings, is important.

2.2.6 SNA

NLP approaches can be applied to analyse the text of social media posts; however, they may not capture the dynamics of how text content spreads among social media users. Nguyen et al. provide a framework to contextualize crisis applications; however, they do not account for how the structures of social media affect the spread of the data.⁶⁷ In addition, by using a publicly available dataset to both train and test their model, they also did not account for the amount of work that is required to collect a dataset (although they attempt to ameliorate this for images).⁶⁸ The dataset also leaves out potentially important Twitter information by omitting retweets, as well as leaving out the network that is built upon repeated retweets. Because they found a single step process suggested in other applications, was onerous to do in real-time, Rudra et al. created a two-step architecture for categorizing Twitter responses to earthquakes in real-time (the first step creates a micro summary).⁶⁹

⁶³ Bender, E. M, Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency pp. 610–623. <https://doi.org/10.1145/3442188.3445922>

⁶⁴ Kumar, A., Singh, J., & Saumya, S. (2019, November 14). A Comparative Analysis of Machine Learning Techniques for Disaster-Related Tweet Classification.

⁶⁵ Alam, F., Ofli, F., Imran, M. (2018). Crisismmd: Multimodal twitter datasets from natural disasters. Proceedings of the 12th International AAAIConference on Web and Social Media, June 2018.

⁶⁶ Kumar, et al. (2019), p. 227.

⁶⁷ Nguyen et al. (2016).

⁶⁸ Nguyen et al. (2017).

⁶⁹ Rudra, K., Banerjee, S., Ganguly, N., Goyal, P., Imran, M., & Mitra, P. (2016). Summarizing Situational and Topical Information During Crises. ArXiv:1610.01561 [Cs]. <http://arxiv.org/>

SNA has frequently been used to investigate the interactions between social media users during crises.^{70,71,72,73,74} For example, a network of retweets and replies between Twitter accounts during the 2017 Storm Cindy in the US revealed accounts that functioned as dominant information sources.⁷⁵ Silver and Andrey found that weather information originated from weather professionals and enthusiasts, who acted as key stewards; whereas, the public “engaged in the dialogue predominately by retweeting and by sharing personal observations of the storm.”⁷⁶ Distinct communities of users were found on Twitter during Hurricane Irene, the England riots, and the 2011 Virginia earthquake.⁷⁷ Such analyses reveal how information is exchanged among numerous connected actors (i.e., user accounts) as a crisis unfolds.⁷⁸ Crisis data from Twitter is particularly well-suited to such analyses as user interactions (retweets, mentions, and replies) can be readily accessed by researchers.

We argue that SNA could be used to augment ML approaches to analyzing textual social media data (e.g., a Tweet or status update). In SNA, a collection of tweets can be converted into a traditional network or graph structure, which consists of nodes and edges. Figure 1 illustrates the network, where a node can be any entity (a person, an organization) and the edge describes the connection among the entities. For us, the node is a user account. The edges connecting these nodes are informational exchanges between accounts, as represented by mentions, retweets, and replies.

Node centrality is a commonly studied feature of social networks and consists of a number of properties that capture the prominence of a node within its network. In this case, the relative centrality of a Twitter account (node) within the network is considered to be indicative of that account’s influence over the flow of information within the network. Multiple measures of node centrality have been used to identify influential Twitter accounts for the purpose of communicating information about crises.⁷⁹

⁷⁰ Chatfield & Brajawidagda (2012).

⁷¹ Gupta, et al., (2012).

⁷² Hagen, L., Keller, T., Neely, S., DePaula, N., & Robert-Cooperman, C. (2018). Crisis Communications in the Age of Social Media: A Network Analysis of Zika-Related Tweets. *Social Science Computer Review*, 36(5), 523–541. <https://doi.org/10.1177/0894439317721985>

⁷³ Silver, A. & Andrey, J.. (2019). Public attention to extreme weather as reflected by social media activity. *Journal of Contingencies and Crisis Management* 27, 4: 346-358.

⁷⁴ Kim, J., Bae, J. J., Hastak, M. (2018). Emergency information diffusion on online social media during storm Cindy in U.S. *International Journal of Information Management*, 40, 153-165, <https://doi.org/10.1016/j.ijinfomgt.2018.02.003>.

⁷⁵ Kim et al. (2018).

⁷⁶ Silver & Andrey (2019), p. 349.

⁷⁷ Gupta et al., (2012).

⁷⁸ Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research*, 18(4), 323–342. [https://doi.org/10.1016/S0740-8188\(96\)90003-1](https://doi.org/10.1016/S0740-8188(96)90003-1)

⁷⁹ Hagen et al., (2018).

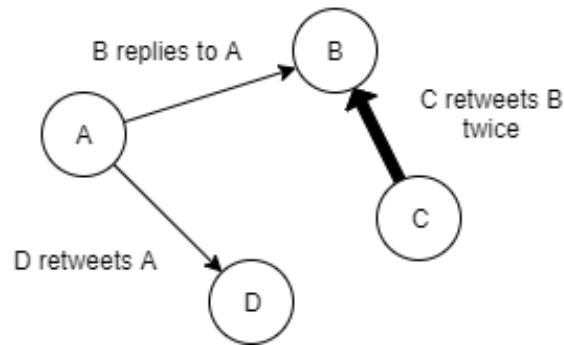


Figure 1. An example network. Note that edges may also be weighted according to the strength of interactions between user accounts.

2.3 Gaps and Caveats

Recent research has explored ways to ensure that the insights gained from social media can lead to informed decision-making during crisis management.^{80,81} Simply performing a computational analysis is insufficient. The concept of ‘actionability’ offers a useful lens for evaluating the ways that social media data can be useful to the work of crisis managers.^{82,83} Variables such as time, responder role, and trustworthiness of information may impact the extent to which that information can be effectively used by crisis managers.⁸⁴ Within the field of crisis informatics, Zade et al. suggest a shift from obtaining information that contributes to improved situation awareness and towards information that is directly actionable for crisis managers.⁸⁵ These discussions highlight the need for a deeper consideration of how the results of data analyses such as NLP and SNA can be applied in practice to the work of crisis management.

Even as the three models of NLP—unsupervised, supervised, and language models—and SNA are increasingly common in the crisis informatics literature, there is a greater need for thorough discussions of technical uncertainty, and trustworthiness and timeliness of results. As researchers continue to explore and develop new analytic techniques to aid with crisis management, there is a need for continued discussions on the practicality and real-world applicability of these approaches.

⁸⁰ Rao et al., (2017).

⁸¹ Zade, H., Shah, K., Rangarajan, V., Kshirsagar, P., Imran, M., & Starbird, K. (2018). From Situational Awareness to Actionability: Towards Improving the Utility of Social Media Data for Crisis Response. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), 195:1-195:18. <https://doi.org/10.1145/3274464>

⁸² Ferrario, M. A. A., Simm, W., Whittle, J., Rayson, P., Terzi, M., & Binner, J. (2012, May 20). Understanding Actionable Knowledge in Social Media: BBC Question Time and Twitter, a Case Study. *Sixth International AAAI Conference on Weblogs and Social Media*. Sixth International AAAI Conference on Weblogs and Social Media. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4642>

⁸³ Zade et al., (2018)

⁸⁴ Zade et al., (2018).

⁸⁵ Zade et al. (2018).

We should not leave this literature review without noting the limitations of relying on social media as a source of extreme weather impacts for crisis management. This constitutes a type of noise that is not always amenable to computational remedies.⁸⁶ We need to consider the context in which the public is contributing as that will affect the subsequent ability to computationally analyze it and utility for crisis managers:

In a crisis, someone may be reporting what they see in a ‘citizen journalism’ style, while also alerting friends and relatives to their wellbeing, while also recirculating both verified and unverified reports of others: how are we to categorize or interpret the ‘value’ of these messages?⁸⁷

Crisis managers’ use of Twitter assumes that a broad swath of the public has installed the Twitter app and has the luxury of time in which to report. It assumes all age groups have smart phones and sufficient data plans. There is a sampling bias as well as responses to disasters reported on social media can vary considerably by socio-economic characteristics such as age, race, income, level of education and ethnicity.^{88,89} Emergency services can be directed to middle and high income areas instead of areas where the need is greater, for instance in poorer neighbourhoods where the infrastructure is in less adequate shape. The use of social media, when it needs to be communicated via Internet or cell phone, assumes that the telecommunications infrastructure has not collapsed due to the event, for example from weather effects (e.g., iced phone lines) or as a consequence of increased cell phone usage. It also assumes the electrical grid is active, which is necessary to power phones (after batteries are exhausted) and computers.⁹⁰ Crawford and Finn call this a ‘signal problem’: where a dataset like Twitter supposedly serves as an accurate representation of the affected public, but there is missing data, “with little or no signal coming from particular communities.”⁹¹ Social media also has been shown to contain considerable media bias, which can allow for the spreading of rumours and fake information, which is why it is important to infer the source of weather information.^{92,93} Finally, the platform itself can impose limits, which then impacts how well someone can describe the degree to which they are impacted by a crisis in the limited number of characters or words (e.g., the 280 character limit of tweets).

⁸⁶ Crawford, K., Finn, M. (2015). The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal* 80, 491–502. <https://doi.org/10.1007/s10708-014-9597-z>

⁸⁷ Crawford, Finn. (2015), p. 496.

⁸⁸ Yuan, F., Li, M., Liu, R., Zhai, W., Qi, B. (2021). Social media for enhanced understanding of disaster resilience during Hurricane Florence. *International Journal of Information Management* 57, Article 102289. <https://doi.org/10.1016/j.ijinfomgt.2020.102289>

⁸⁹ Longley, P. A., Adnan, M., & Lansley, G. (2015). The Geotemporal Demographics of Twitter Usage. *Environment and Planning A: Economy and Space*, 47(2), 465–484. <https://doi.org/10.1068/a130122p>

⁹⁰ <https://www.cnet.com/news/hurricane-sandy-disrupts-wireless-and-internet-services/>

⁹¹ Crawford, Finn. (2015), p. 497.

⁹² Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A. (2013). Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. Proceedings of the 22nd International Conference on World Wide Web, pp. 729–736. <https://doi.org/10.1145/2487788.2488033>

⁹³ Fowler, B. M. (2017). Stealing thunder and filling the silence: Twitter as a primary channel of police crisis communication. *Public Relations Review* 43, 4, 718–728. <https://doi.org/10.1016/j.pubrev.2017.04.007>

3. Grant activities

In this section we describe the objectives, methods, results and outcomes for each activity completed as part of the G&C. These included:

1. A scoping phase
2. Infrastructure development
3. Building a data pipeline to collect tweets
4. Automatically classifying tweets during disasters, with a U.S. example to ensure a sufficient volume of analyzable content.
5. Social network analysis to identify sources of weather information
6. Adapting tweet classification to Canadian snowstorms
7. Building a dashboard to the models

3.1 Scoping the state-of-the-art

We carried out a scoping phase to assess the state-of-the-art in the field and determine methods for crowdsourcing weather information.

3.1.1 Scoping via a review of the literature

We reviewed the literature to assess leading research and methodologies. The review covered domain research (i.e., extreme weather, social media, crisis management) and computational methods (i.e., NLP, text mining, network analysis) to ensure we employ the most effective means for achieving our goals. We created an annotated bibliography of more than 50 peer reviewed journal articles as well as publications in the grey literature.⁹⁴ A particular focus was on patterns of information diffusion and dissemination in social networks, specifically focusing on Twitter. The annotated bibliography was created as a first research product that served as the basis for the literature reviews in the project's publications and summarized in Section 2 of this report.

We had initially determined that we should begin with unsupervised classification. However, as a result of our literature review, we found that supervised learning was considered state-of-the-art in detecting and then reacting to extreme weather events. This was due to the issue of actionable situational awareness around extreme weather events. Supervised classification offers more rapid response because the model was pretrained and it offered more actionable intelligence because the social media responses would be grouped into predefined categories. This greatly reduced the noise to signal problem, as unsupervised learning classifications were entirely dependent on the

⁹⁴ <https://smacawi.github.io/docs/annotated-bibliography.pdf>

corpus and shifted by type of weather (e.g., snow, flood). Encouraged by its flexibility in handling new types of events and more recent technical developments,⁹⁵ we returned to unsupervised learning, as well as to language models, to handle events not captured in traditional supervised learning of crises. These are not mutually exclusive; towards the end of the project we integrated these techniques to improve on performance.

3.1.2 Social media and dataset options

We created and applied a survey of instances of extreme weather events in the U.S. and Canada on Twitter as the starting point for the SNA so we could map the full range of actors on the datasets we collected. We identified and used several open source datasets including several used by CrisisNLP. We surveyed various social media options, but found that many others (e.g., Facebook and Instagram) did not have readily accessible application programming interfaces (API) to their data. We decided to use Twitter data. Twitter appeared to be better established as a source of information in ML research and was likely to provide a greater volume of relevant, time-sensitive information during crises.

Because of our focus on supervised learning we needed to find a sufficiently large training dataset. This led us to CrisisNLP⁹⁶, the most widely used dataset for managing situational awareness and responding to extreme weather events.

3.1.3 Scoping for SNA

In early conversations with ECCC personnel, the desire was to identify the public's reactions—their sentiment—to official ECCC weather announcements. But we found that ECCC did not always know from where people were getting their weather information. This was not for want of trying: ECCC had apparently surveyed members of the public and found the public had little idea of the source to which it was reacting. We conducted an initial assessment of official North American weather sources. We broadened our scope beyond Canada because (a) the majority of Canadians live near the U.S. border and are likely to get weather information from both Canadian and U.S. sources and (b) many private sector weather firms cover both U.S. and Canadian events. We conducted three surveys during 2019:

1. **Survey of Weather News Sources on Twitter:** We completed a survey of sources of weather in the U.S. and Canada on Twitter and compiled a 21-item list. Included in this was both official government accounts and those run by commercial news networks such as The Weather Channel. For each account in the list, we recorded: (1) number of followers, (2) geographic coverage, (3) language of communication, and 4) notes on the style of content in tweets. We also compiled lists of twitter accounts posting on weather,

⁹⁵ Arachie, C., Gaur, M., Anzaroot, S., Groves, W., Zhang, K., Jaimes, A. (2020). Unsupervised Detection of Sub-Events in Large Scale Disasters. The Thirty-Fourth AAAI Conference on Artificial Intelligence. 34 (01), 354-361

⁹⁶ <https://crisisnlp.qcri.org/>

including 142 Twitter accounts from the National Weather Service, 66 popular non-commercial/non-governmental accounts, and seven commercial accounts. The account lists and database of news weather accounts can be accessed on our public GitHub repository.⁹⁷

2. **Survey of ECCC's Twitter Accounts:** To assess the impact of ECCC's extreme weather alerts on social media, we first scraped all the tweets from ECAAlert accounts,⁹⁸ along with user engagement metrics such as number of likes, Retweets, and replies. User engagement with these accounts was very limited, with most accounts having little to no followers. Similarly, the provincial ECCCWeather channels⁹⁹ were found to have very few followers, except for the BC (~16k, likely due to the wildfires at the time) and QC (~3k) accounts. Interestingly, every other account set their twitter accounts as being "protected," rendering it more difficult to view and share tweets and gain followers.
3. **Background Research on National Weather Alerting Systems:** We completed background research on national weather alerting systems for Canada and the U.S. In particular, we examined the Integrated Public Alert and Warning System (IPAWS) infrastructure in the U.S. and the National Public Alerting System (NPAS) in Canada. A particular focus included understanding how these alert systems functioned, what international data standards existed for publishing alerts to them (with a focus on the CAP-XML format), and what best practices exist for publishing official weather alerts to Twitter. The reports for Canadian and U.S. accounts can be found at our public GitHub repository.¹⁰⁰

3.1.4 Scoping methods for analysing tweets

To test our ability to scrape/collect, clean, and process datasets that will be used in each of our analyses, we utilized several open source datasets including several from CrisisNLP¹⁰¹ as well as Twitter weather-related sentiment.¹⁰² We found these limited in their utility to ECCC, so we built a tool on top of Twitter's API to capture live Tweets during extreme weather events that used keywords and geofencing to capture as many relevant Tweets for a given event. To better define keywords for a given type of weather, we built a tool to cluster related terms from the ECCC glossary,¹⁰³ using word embeddings and t-Distributed Stochastic Neighbor Embedding (t-SNE) to simplify the visualization. We determined that NLP plus SNA would be methods best suited to extracting useful information from Twitter data. We successfully tested these tools during a

⁹⁷ <https://github.com/smacawi/smacawi.github.io/tree/main/data>

⁹⁸

<https://www.canada.ca/en/environment-climate-change/services/weather-general-tools-resources/subscribe-to-twitter-alerts.html>

⁹⁹ <https://twitter.com/search?f=users&vertical=default&q=eccc%20weather&src=typd>

¹⁰⁰ <https://github.com/smacawi/smacawi.github.io/tree/main/docs>

¹⁰¹ <https://crisisnlp.qcri.org/>

¹⁰² <https://www.kaggle.com/c/crowdfower-weather-twitter/data>

¹⁰³ http://climate.weather.gc.ca/glossary_e.html

significant series of snowstorms from February 9-12 2019 in the U.S. (see Figure 2).¹⁰⁴ A more robust version of this approach was subsequently used to capture tweets related to the floods in the midwest U.S. in March 2019 and then a major snowstorm in Newfoundland-Labrador in January 2020.

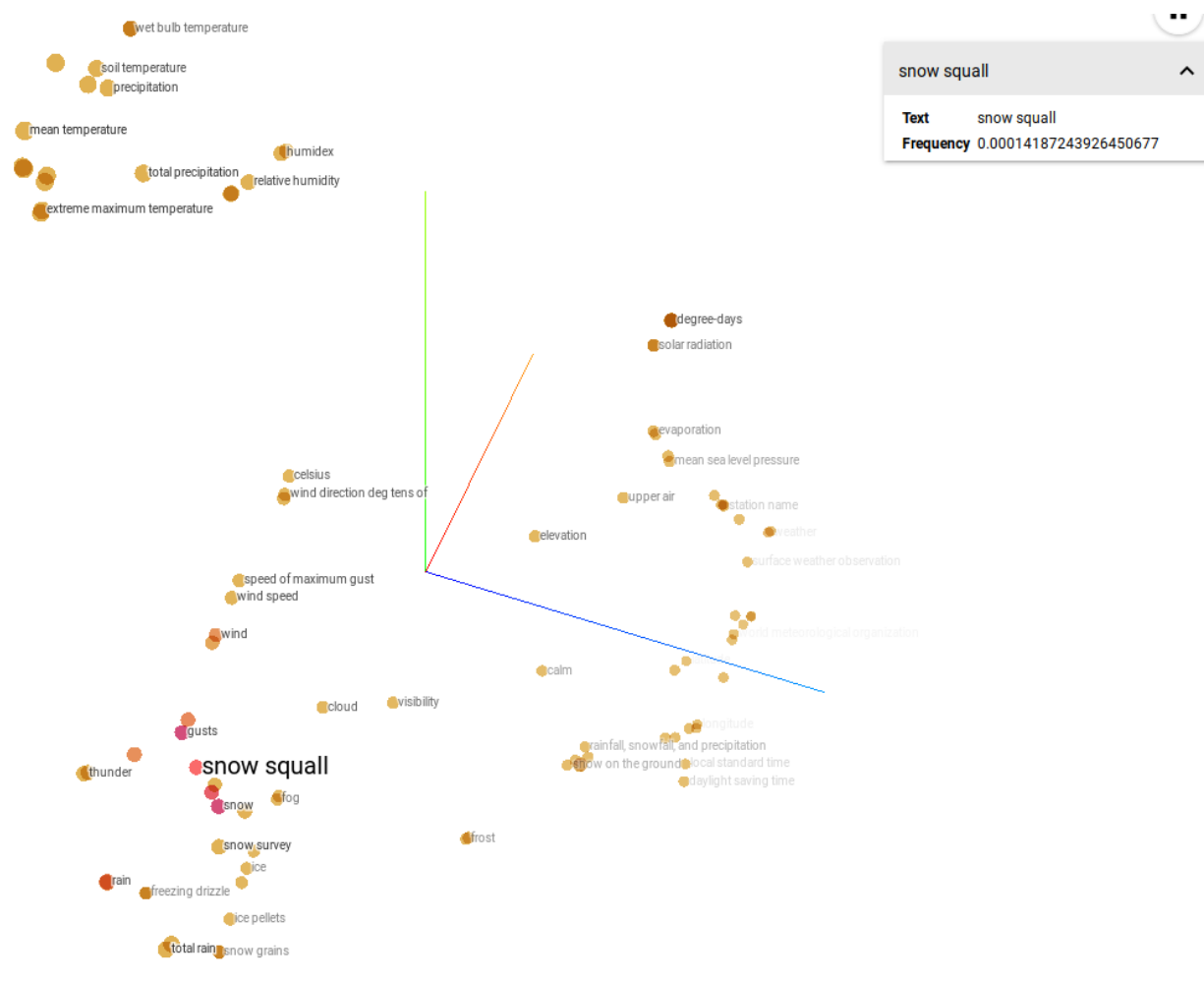


Figure 2. Clustering (using the glossary of terms created by ECCC) of live Tweets captured during a series of significant snowstorms in mid-February 2019 using the tool we built on top of Twitter's API.

3.1.5 Conclusions and directions

We identified several gaps in existing research and practice in the use of social media data for environmental management. These included

- A lack of robustness/objectiveness of off-the-shelf tools

¹⁰⁴ <https://smacawi.github.io/data/snowmageddon.csv>

- Supervised classification is the state-of-the-art in the crisis management community
- The availability of a large training dataset is a motivating factor
- There is an emphasis in the literature on actionable content and improving on the signal to noise problem

This motivated us to focus on NLP as a primary method to address our research goals. It also motivated us to extend the existing proposal to leverage different kinds of NLP and to develop a dashboard to make the models more readily accessible.

3.2 Developing the infrastructure

Developing the appropriate computational infrastructure to conduct this project was a crucial first step because working with social media data and ML has high computational requirements. The important criteria of this infrastructure include (1) data storage capabilities, (2) computational power, and (3) secure remote access.

A comparison of potential infrastructure solutions revealed that we could most cost-effectively meet our criteria by building a server with consumer-grade components. In contrast, cloud-based solutions would have had higher costs throughout the lifetime of the project, without even taking into account the risk of runaway costs (e.g., for system administration of cloud-based servers). Similarly, commercial-grade components would have had limited benefits, as consumer-grade components are optimized for our deep learning-based workflow, at a fraction of the cost.

3.2.1 Data storage

The volume of social media data, current and projected, is beyond most desktop personal computer capacities and requires modern means of data storage. Storage is required both for principal data sources (e.g., Twitter), but also for analysis results. Storage solutions with fast read-write speeds are also essential when working with large amounts of data to prevent bottlenecks during analysis. With this in mind, we built our server using a high-speed 1Tb solid state drive with a motherboard allowing storage expansion as needed.

3.2.2 Processing units and memory

Deep learning and ML often can be optimized with graphical processing units (GPUs) to efficiently and scalably perform the highly-parallel tensor computations involved in neural networks. Consumer-grade GPUs have become the workhorse of artificial intelligence research due to their cost-efficiency and modern software libraries being optimized for them. We chose an Nvidia GTX 2080Ti, the top-of-the-line consumer-grade GPU at the time. In particular, we selected this option because of its large VRAM of 12Gb, which is necessary to run the large deep learning models adapted to NLP.

The Central Processing Unit (CPU) is an important component for all computational tasks in our workflow. We opted for the Intel Core i7-8700 due to its balance of cost, multiprocessing capabilities, and high single thread speed. Although the 6 cores allow us to efficiently parallelize text preprocessing and other computational tasks, the single thread performance allows us to better handle large amounts of live data in real-time.

We built our system with 32Gb of Memory (RAM) to enable holding large amounts of data in memory so we can more efficiently operate on the data. While error correcting code (ECC) RAM is typical in commercial applications, our workflow did not require it. So we could greatly reduce cost by opting yet again for consumer-grade RAM. Once again, the motherboard allowed RAM expansion as needed, but this amount of memory was found to be sufficient.

3.2.3 Secure remote access

We built our server with the Ubuntu 18.04 LTS operating system to streamline remote access through secure shell protocol (SSH). Remote access in turn facilitated monitoring of multi-day experiments, transferring data and experiment results, and collaborating on code. To ensure secure SSH access, we only allowed local Internet Protocol (IP) connections so that the user has to use McGill University's Virtual Private Network (VPN) to connect remotely.

3.2.4 Outcomes

As a result of this activity, sufficient computational infrastructure was set up and made available, which successfully enabled the exploration of data and compute-intensive ML workflows on consumer-grade hardware. This is particularly important in making our work more accessible for future research and application by ECCC and crisis managers.

Despite the relative accessibility of consumer-grade hardware used in this activity, many managers may not have access to a GPU (e.g., a crisis manager attempting to run software on their laptop or phone). As a result, the reliance on GPUs for training and deploying our models is a potential limitation that requires further research to address whether it is possible to develop models with equivalent performance that can scale without GPUs.

3.2.5 Outputs

The Infrastructure Development activity indirectly contributed to the outputs of every other grant activity because they rely on the computational infrastructure. We argue that the specifications described in this section can act as a cost-effective blueprint for future work requiring appropriate computational infrastructure to apply the methods developed in this grant.

3.3 Building a data pipeline

To study how people react on social media during extreme weather events, we had to develop a robust pipeline to harvest relevant content from the social media platform of choice, Twitter. This presented various challenges, including:

- Develop API for live and historic harvesting (scraping) of Twitter during extreme weather events
- Build data storage solution for harvested tweets
- Explore different data scraping methodologies (e.g., keywords, hashtags, geotags, accounts)
- Test live scraping case study on extreme weather event

To make the adoption of our tool easier, we opted for the Standard tier of the Twitter API,¹⁰⁵ which is freely available and easy to obtain access.¹⁰⁶ The primary limitations of this tier are the volume of tweets that can be harvested, the available querying parameters, and the period for which historic scraping is possible. We found that we could work well within these limitations and could not justify the added cost of the paid tiers, which can total thousands of dollars per month.

We then built a custom API on top of the Twitter API to better suit our purposes, specifically to unify live/historic scraping, and automatically extract/construct SNA information. This exists in an easy-to-use python library.¹⁰⁷ We designed the library to make it easy to handle authentication, query-specification and data storage in a SQLite database with json files. The library also includes a custom pipeline for metadata extraction to convert tweets to a unified data format that facilitates the ML approach and the SNA.

We explored different scraping methods that would create a corpus (dataset) of potentially relevant content. This is how the classification in supervised and unsupervised learning typically works: one first does a filtering pass to create the corpus and then one classifies the content within the corpus. The corpus needs to be relevant to the issue (i.e., the extreme weather event), the geographic location (i.e., capturing content from individuals and infrastructure directly impacted by the event) and the temporal period (e.g., during the actual extreme weather event, or the one-week aftermath of the event). Identifying the geographical location of a social media poster is particularly challenging. Less than two percent of tweets are deliberately geocoded, in which the user identifies their home location or the location of the tweet.¹⁰⁸ (That location is

¹⁰⁵ <https://developer.twitter.com/en/docs/twitter-api/v1>

¹⁰⁶ <https://developer.twitter.com/en/apply-for-access>

¹⁰⁷ <https://github.com/smacawi>

¹⁰⁸ Schlosser, S., Toninelli, D., Cameletti, M. (2021). Comparing Methods to Collect and Geolocate Tweets in Great Britain. *Journal of Open Innovation: Technology, Market, Complexity* 7, 44. <https://doi.org/10.3390/joitmc7010044>

stored in the metadata of the tweet.) Consequently practitioners miss the majority of content that is relevant to the impacted area (e.g., individuals flooded out of an area) if they rely on geocoding. Crisis managers are likely to be only interested in individuals directly impacted; whereas, people from around the world may discuss, report, or offer help far outside the affected areas.

Table 1. Description of tweet attributes (metadata) extracted and stored as part of data pipeline

Field	Description
Tweet information	
id	Unique Tweet ID that allows retrieval of the original Tweet object through the Twitter API
text	Text content of the Tweet, used in e.g., text classification or topic modelling
hashtags	Hashtags included in Tweet, useful for filtering content
created_at	Time (UTC) of Tweet publication, useful for time series analysis
coordinates	User-provided geotagging of Tweet, useful for geolocation but rarely provided
place	User-provided coarse-grained location ¹⁰⁹ of Tweet, useful for geolocation
User information	
user_id	Unique user ID that allows retrieval of various data through Twitter API
user	Unique user account name
user_name	Displayed user name
user_followers_count	Number of followers, useful for identifying influential accounts or filtering bots
user_friends_count	Number of friends, useful for identifying influential accounts or filtering bots
user_favourites_count	Number of favourites, useful for filtering bots
user_statuses_count	Number of statuses, useful for filtering bots
user_listed_count	Number of lists user belongs to, useful for filtering bots
user_location	User-provided location of account, useful for geolocation but often inaccurate
user_verified	User account verification, useful for identifying influential accounts or filtering bots
user_created_at	User account creation date, useful for filtering bots

¹⁰⁹ <https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/geo#place>

Network analysis metadata

user_mentions_id	Ids of users mentioned in Tweet
user_mentions	Account names of users mentioned in Tweet
reply_to_tweet_id	Tweet id of Tweet being replied to
reply_to_user	User of Tweet being replied to
reply_to_user_id	User id of Tweet being replied to
reply_to_tweet_text	Text content of Tweet being replied to
rt_text	Text content of Tweet being retweeted
rt_id	Tweet Id of Tweet being retweeted
rt_user	User account name of Tweet being retweeted
rt_user_id	User id of Tweet being retweeted
qt_text	Text content of Tweet being quoted
qt_user_id	User id of Tweet being quoted
qt_user	User account name of Tweet being quoted
qt_id	User id of Tweet being quoted
source_id	Tweet id of parent Tweet for network analysis
source_user	User account name of parent Tweet for network analysis
source_user_id	User id of parent Tweet for network analysis
source_text	Text content of parent Tweet for network analysis
edge_type	Relation between Tweet and parent Tweet (mention, reply, retweet, or quote)

Other methods can be used to obtain location (e.g., specifying IP addresses) but they also can distort the actual location (e.g., position the individual at the location of the service provider). Plus it can be quite expensive to obtain the Twitter tier of API that provides a sufficient geographic resolution. Often researchers will resort to the content of the tweet. Geocoders like the Geonames API can be used to identify location from words in the body of the media post.¹¹⁰ We have found this approach to be quite client-heavy. We have chosen to use hashtags as a lightweight way to capture areas or events.

The initial test of our method, which confirmed the use of hashtags and demonstrated the robustness of our pipeline, is described in Section 3.1.4.

3.3.1 Results and Analysis

To assess our data pipeline, we performed two case studies, one on a major flood in the U.S. State of Nebraska, in March 2019, and one on a large blizzard in northeast Canada in January

¹¹⁰ <http://www.geonames.org/export/geonames-search.html>

2020. We found that the initial filtering by geolocation was inefficient due to the large number of irrelevant geotagged tweets and of relevant but non-geotagged tweets. Searching for weather-related keywords also caused a large noise-to-signal ratio. For example, using “blizzard” as a keyword led to over half of harvested tweets being irrelevant, due to a series of highly publicized layoffs at a company called Blizzard. We found the optimal strategy was to use event-specific hashtags that emerge as an extreme weather event unfolds. While this may require constant monitoring and adjusting in the live setting, it is straightforward to apply in the historic setting when scraping tweets from the previous days or hours.

The rest of the steps are described in detail in subsequent activity sections.

3.3.2 Outcomes

The resulting data pipeline is illustrated in Figure 3. It contains the workflow for collecting tweets, their preprocessing, the selection of a training dataset (for supervised) and then applying testing data, the selection of an untrained corpus (for unsupervised) and subsequent classification, and the process for SNA.

We developed a python library¹¹¹ for harvesting tweets in both live and historic settings. Through a case study, we identified an adequate querying strategy to minimize the noise-to-signal ratio in harvested tweets by leveraging highly event-specific hashtags and short-term historic harvesting.

Our data pipeline strategy met our needs for automating data harvesting and analysis. It should be noted no use of classification or SNA is automatic. There is considerable need for human intervention, even at the beginning in terms of relevant hashtags and choice of modelling approach. To address this, we suggest future work exploring the automatic identification of relevant search terms as an extreme weather event unfolds. As modelling approaches improve, more performance evaluation can be conducted.

The python library was used throughout the remainder of the project to harvest, clean, and store tweets from extreme weather events. It is also open-source and designed to be easy-to-use by others for future projects.

¹¹¹ <https://github.com/smacawi/twitter-scraper>

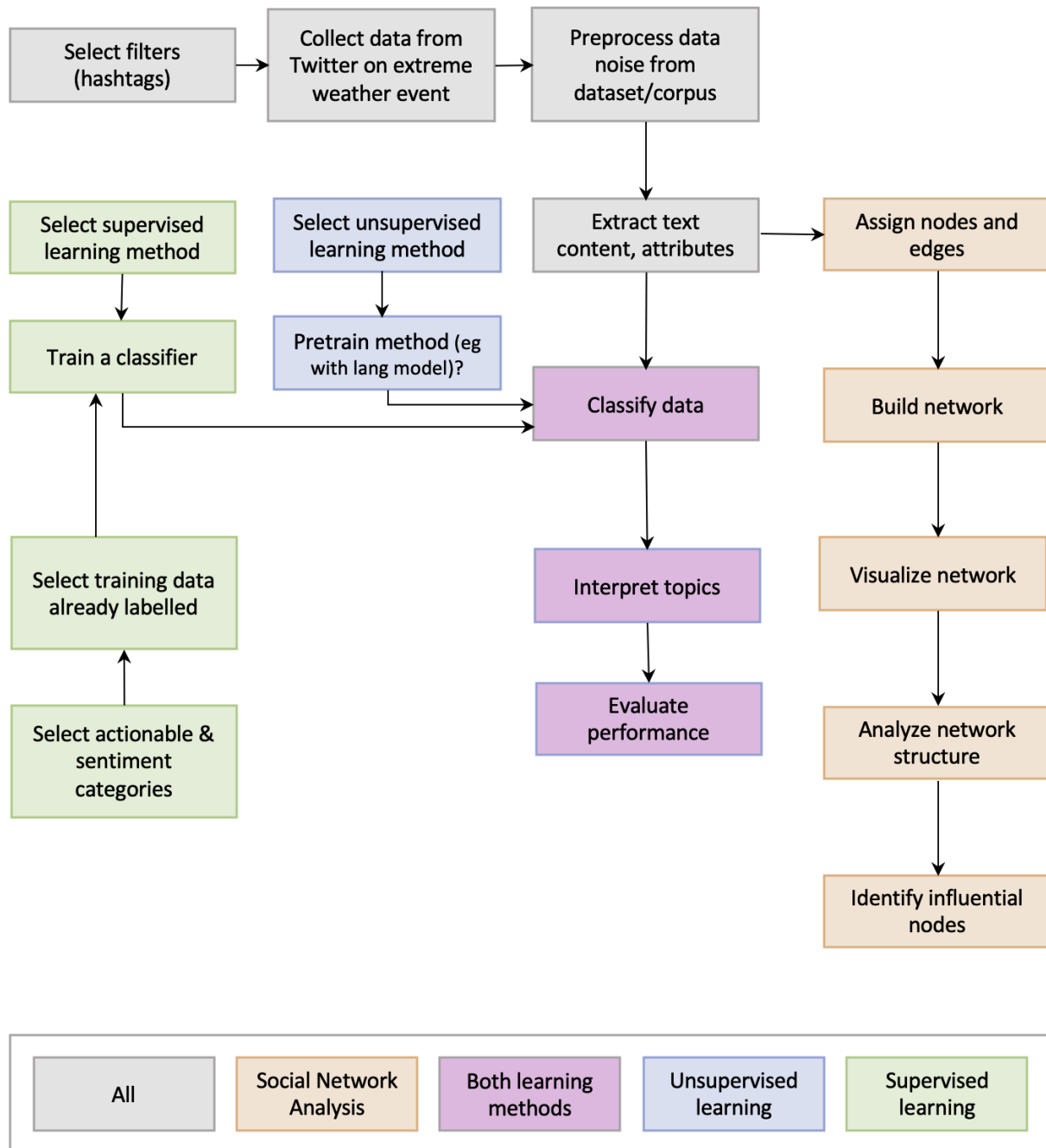


Figure 3. A diagram that illustrates our data pipeline.

3.4 Conducting a social network analysis

As described above, we found that official accounts on Twitter often had limited engagement with affected people during extreme weather events. This led to our second research objective: where do people in Canada and the U.S. get their news on extreme weather? Are those sources official weather-related government agencies? Whether or not the sources are official, is the information flow centralized, originating from a small number of influential accounts? Or is the information decentralized, disseminating in a grassroots manner from the affected people themselves? In this grant activity, we turned to SNA as a tool to characterize the flow of information during extreme weather events, and answer this question in our initial study on the March 2019 Nebraska flood.

3.4.1 Method

From March 22 to March 26, 2019 we used our data pipeline to collect a dataset of 11,982 tweets relating to the damage containment stage of a major Nebraska flood. We used the following hashtags to filter relevant tweets: #nebraskaflood, #flood2019, #nebraskastrong, #missouririver, #nebraskaflood2019, #prayfornebraska. Of these tweets, 1,055 were regular tweets, 9,658 were retweets, 1,239 were mentions, and 30 were replies.

We conducted a SNA to identify the Twitter accounts within our dataset that functioned as key sources of information on Twitter during our period of data collection. We converted our Twitter dataset into a network structure of nodes and edges to model flows of information between Twitter users during the Nebraska flood. One user account corresponded to one node in the network. The edges connecting these nodes were informational exchanges between accounts, as represented by tweets that were mentions, retweets and replies.

To identify influential Twitter accounts in this network we used several measures of node centrality: degree centrality, betweenness centrality and PageRank score. The top ten nodes with the highest values for each centrality measure were reported. The nodes that appeared in at least two out of the three top ten rankings were considered to be among the most central nodes in the network. The network was then broken down into communities based on Blondel et al.'s¹¹² algorithm, which identifies groups of densely-connected nodes in a network that interact with each other more frequently than with nodes from other communities.

¹¹² Blondel et al., 2008

Table 2. Most central nodes in the network. Usernames were anonymized to respect user privacy. Anonymized names were classified as either (N)ews, (P)ublic, (I)nstitution or (C)ommercial. Usernames of verified accounts and official government accounts were preserved.

Degree	Betweenness centrality	PageRank
GovRicketts	GovRicketts	C01
C01	NETAGBohac	P03
P01	N03	P04
N01	I01	P05
NEStatePatrol	NEMAtweets	P06
RBrex34	NENationalGuard	I01
NEMAtweets	NEStatePatrol	P07
cucoachmac	P02	N04
N02	able I02	P08
Barbi_Twins	N01	P09

3.4.2 Results

We constructed a social network of retweets, replies and mentions that had a total of 7,583 nodes and 8,832 edges. Table 2 shows the top ten most central nodes in the network according to the three measures of centrality. The accounts that ranked highly according to our three measures of node centrality (degree, betweenness and PageRank) primarily correspond to official information sources, such as government officials and institutions and news figures. For example, the official Twitter account for Nebraska's Governor Pete Ricketts ranks highly in two of the three centrality measures, as well as accounts for the Nebraska Emergency Management Agency (NEMA) and the Nebraska State Patrol. One influential account is an official account associated with an Omaha reporter.

Finer-grained analysis can also be conducted on these central nodes. For instance, most of the direct connections to Governor Ricketts' account are outbound, indicating that other accounts are interacting with his Tweets, rather than the inverse. Governor Ricketts' account also most frequently lies on the shortest path between any other two accounts in the network, indicating that information shared on his account may reach many communities of Twitter users.

We then analyzed nodes from the largest community. Blondel et al.'s (2008) modularity class algorithm identified 496 distinct communities within the entire network. Out of these 496 communities, the largest community contained a total of 905 nodes, which constituted 11.2

percent of all nodes in the network. Figure 4 is a visualization of this community, where the colour intensity and size of each node correspond to the node's degree.

The most central nodes in this community were identified using the three node centrality measures. Table 3 summarizes the characteristics of the nodes that appear in at least two out of the three top ten centrality rankings.

The characteristics of the highly central nodes in this subset of the network suggest that the information (ie. Tweet content) being shared amongst this group of Twitter accounts is more formalized news about the Nebraska flood. Many of the accounts in Table 3 are official government accounts, rather than personal or commercial accounts. This finding may indicate that, within this cluster of tweets, information about the flood was communicated through official news sources. Some of the accounts identified in Table 3 communicate information at a local (county) level, suggesting that the information contained in the Tweets is potentially more specific to local impacts of the flooding. All accounts with associated locations are within Nebraska, further verifying that the Tweet content is relevant to the Nebraska flooding.

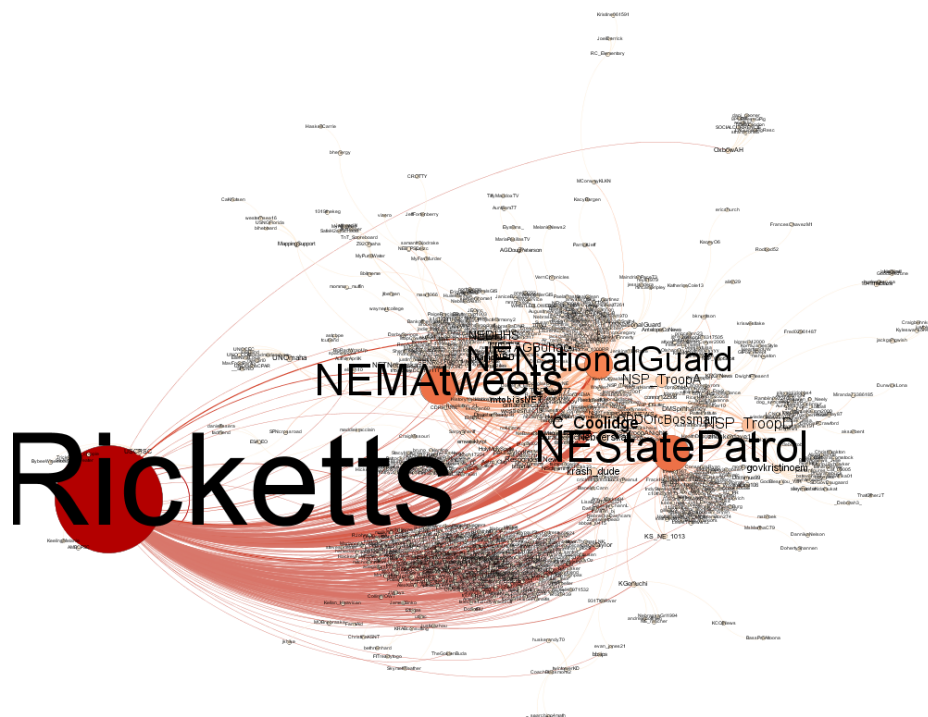


Figure 4. Visualization of the largest communities.

Table 3. Characteristics of most central nodes in largest modularity class.

Twitter account	Number of followers	Type of account	Geographic scope	Associated location
GovRicketts	17,656	Government/ Personal	Regional (state)	Nebraska
NEMAtweets	6,663	Government	Regional (state)	Nebraska
NEStatePatrol	26,100	Government	Regional (state)	N/A
NENationalGuard	5,997	Government	Regional (state)	Lincoln, NE
NSP_TroopD	3,369	Government	Local (counties)	North Platte, NE
NETAGBohac	535	Government/Personal	Regional (state)	N/A
NSP_TroopA	4,170	Government	Local (counties)	Omaha, NE

3.4.3 Outcomes

In our flooding event study, we successfully used SNA on Twitter data to gain insights into how affected people interact on social media and respond to an extreme weather event. Overall, the SNA pointed to the most influential Twitter accounts during the time period in question.

Of importance to ECCC, the predominant influencers were official sources, although not directly weather-related (e.g., a governor instead of a state weather service). The results revealed subtle distinctions between the type of content generated (through unique posts or retweets) from highly influential posters, suggesting different roles that Twitter accounts and official sources played in spreading information during the flood.

Whereas SNA could help us answer where information is flowing; it does not shed light on what information is flowing. For example, a large proportion of tweets during a crisis may be offers of sympathy, which are not as informative as tweets that may describe impacts of an event or share actionable advice. Distinguishing between different such types of content is essential in extracting insightful information from this data. We address this issue in the next grant activity, using NLP to augment SNA by automatically classifying Tweets into actionable and sentiment-related categories.

These findings were published in the *Journal of Contingencies and Crisis Management*.¹¹³ More detailed findings and methodology are also reported in the SNA Report.¹¹⁴

To conclude, this part of the project was an exploration into the social structure of Twitter interactions during an extreme weather event, the identification of major influencers and the ways that information is communicated within a group of actors. This can be used by crisis managers to identify the most influential sources of information during a crisis event and query specific official news sources, such as those from the government. In future research efforts, this analysis could be extended to include a more rigorous and comprehensive characterization of the

¹¹³ Romascanu et al. (2020).

¹¹⁴ <https://smacawi.github.io/docs/social-network-analysis-report.pdf>

notable Twitter accounts, particularly the type of government agency being relied upon for critical weather information and crisis management.

3.5 Automatically classifying social media during disasters

Extracting insightful information from social media during a crisis requires distinguishing between such types of content, which is challenging when it is big data, which is characterized by a high volume, velocity and a potentially high noise-to-signal ratio. This ties in to our third research objective, which is a methodology for use of artificial intelligence (AI) modelling that allows us to measure public reaction to weather events. Here we experiment with supervised learning, which classifies content into pre-arranged categories.

In this activity, we used NLP to automatically classify a large volume of tweets into informative categories that help assess impact and augment SNA to gain novel insights into public response. To achieve this, we leveraged recent advances in NLP and deep learning to reduce the need for costly data cleaning and preprocessing while significantly improving classification accuracy (for the categories in Table 4) compared to previous approaches. To achieve this, we extended the 2019 Nebraska Flood study of the previous grant activity.

Table 4. Categories for crisis-related tweets from CrisisNLP¹¹⁵

Category	Description
Injured or dead people	Reports of casualties and/or injured people due to crisis
Missing, trapped or found people	Reports and/or questions about missing or found people
Displaced people and evacuations	People who have relocated due to the crisis, even for a short time (includes evacuations)
Infrastructure and utilities damage	Reports of damaged buildings, roads, bridges or utilities/services interrupted or restored
Donation needs or offers or volunteering services	Reports of urgent needs or donations of shelter and/or supplies such as food, water, clothing, money, medical supplies or blood; and volunteering services
Caution and advice	Reports of warnings issued or lifted, guidance and tips
Sympathy and emotional support	Prayers, thoughts and emotional support
Other useful information	Other useful information that helps one understand the situation
Not related or irrelevant	Unrelated to the situation or irrelevant

¹¹⁵ Imran et al., (2016), p. 3.

3.5.1 Methods

We used supervised learning in this activity. As mentioned in the literature review, supervised learning is considered state-of-the-art when leveraging social media to address the effects of extreme weather. In supervised learning, the burden of processing is done previously when the model is trained. Model training can be done infrequently or only once and at a much earlier date. Then the input data is classified into the pre-trained categories one item at a time. The system is faster and more responsive to peaks—high volumes and velocities—that happen in these extreme events. Lastly, the categories are far easier to interpret. Depending on the categories chosen, the categories can express action items and sentiment.

As with other supervised learning, we required a training dataset. We trained our models (we used multiple supervised learning models, see below) with a dataset of tweets from crises labeled and published by CrisisNLP.¹¹⁶ We used a subset of the CrisisNLP dataset with approximately 11,038 crowdflower-annotated tweets in English from 19 different crises that took place between 2013 and 2015. Tweets were classified according to the categories identified in Table 4. Our testing data was the same nearly 12,000 tweets used in the previous section on SNA.

Rather than an *a priori* choice of a single classification method, we decided to compare modelling approaches to maximize the supervised learning. As part of our methods, we tested the following architectures: Bidirectional Encoder Representations from Transformers (BERT) with a final dense layer applied to the first hidden state, BERT with an LSTM layer applied to every hidden state, as well as previous state-of-the-art models (LSTM + Crisis, LSTM + GloVe).¹¹⁷

We used a dataset of tweets from crises labelled and published by CrisisNLP to train and evaluate these models. We compared the different models using the performance measures of precision, recall, F1 scores and accuracy on a withheld evaluation set of tweets from the 2014 Pakistan floods. As mentioned in the literature review, we did this in recognition that performance can depend on the word embeddings chosen.¹¹⁸ These analyses were performed on three variations of training data to evaluate transferability to new crises. The first variation trained on data from the same event; the second variation trained on data from a similar event (the 2014 India floods); and the last variation trained on data from all events in the CrisisNLP dataset.

Ultimately, we chose to build our supervised learning atop BERT with the dense layer. BERT is a deep learning language model pre-trained on large amounts of text.¹¹⁹ Our approach is considered a semi-supervised pre-training as it allows BERT to learn generalizable structural information about language so it then can be finetuned on a specific task with less labelled data and better generalize to new datasets (e.g., finetuning for tweet classification on a previous crisis event and

¹¹⁶ Imran et al., (2016).

¹¹⁷ Kumar et al. (2019).

¹¹⁸ Kumar et al. (2019).

¹¹⁹ Devlin et al., (2019).

generalizing to a new event).¹²⁰ Another advantage of BERT over other models is that it uses word-piece tokenization, which can further break down words into “word pieces.” This solves the issue of out-of-vocabulary words commonly encountered with non-standard text in social media and which had previously been solved with less generalized text preprocessing and rule-based approaches. Reports about extreme weather events may contain such regional specific terms.

The final part of the method was to ground truth our choice of model. We sampled a subset of the classified tweets and manually analyzed the model’s predictions.

Table 5. Performance evaluation of various modelling approaches, some of which are NLP only and some of which include language models. These include results for Precision, Recall, F1 score, and validation accuracy. Results for BERT/Dense are bolded for easy reference.

Training Set	Model	Precision	Recall	F1	Accuracy
All events	BERT/Dense	0.77	0.77	0.76	0.77
	BERT/LSTM	0.73	0.72	0.70	0.72
	LSTM + Glove	0.63	0.59	0.59	0.59
	LSTM + Crisis	0.63	0.59	0.57	0.59
Similar event	BERT/Dense	0.60	0.62	0.59	0.62
	BERT/LSTM	0.42	0.48	0.44	0.48
	LSTM + Glove	0.22	0.37	0.27	0.37
	LSTM + Crisis	0.47	0.43	0.39	0.43
Same event	BERT/Dense	0.73	0.74	0.72	0.74
	BERT/LSTM	0.65	0.65	0.61	0.65
	LSTM + Glove	0.57	0.61	0.57	0.61
	LSTM + Crisis	0.61	0.62	0.60	0.62

3.5.2 Results and Analysis

Our modelling approach sorted approximately 12,000 tweets into the nine categories pre-determined by CrisisNLP (Figure 5). Table 6 shows examples of the tweets by category.

¹²⁰ Jawahar et al., (2019)

Table 6. Examples of tweets classified in each category. Usernames were redacted to respect user privacy. Usernames of verified accounts and official government accounts were preserved.

Injured or dead people		
@[REDACTED] @[REDACTED] It's so bad, it's unbelievable. Three people have died, so many livestock, wild animals, countless people have lost everything & are misplaced, & roads & water systems have been devastated.	Worst flooding damage in our state's history – As many as a million calves dead in Nebraska, AT LEAST SOMEBODY IS A HAPPY CAMPER #AOC #GreenNewDeal	Flooding in the midwest has already caused 3 deaths and \$3 billion in damage. These events, once unthinkable, are now commonplace.
#NebraskaFloods #NebraskaStrong #PrayForNebraska	https://t.co/cDTUDPWLH2 #NebraskaFlood #Nebraskaflooding https://t.co/OnM7x1Zwmc	We face a simple choice: decarbonise now, or watch the suffering increase every day. #ClimateCrisis #NebraskaStrong https://t.co/iFeNGS4AEa
Missing, trapped, or found people		
Search turns up no signs of missing Norfolk man #NebraskaFlood2019 https://t.co/sgvhTe7Nvk	A needed smile: It's #NationalPuppyDay, so here is a dog rescued with his family last week by a #Blackhawk helicopter crew. The #NEGuard helped rescue 111 people & 13 pets during the #NebraskaFlood. #NebraskaStrong https://t.co/tdyJkgSv9x	Went to pick up what was left at grandmas yesterday, this poor girl also managed to get stuck in the river behind her house.. Saved her tho! #nebraskastrong #nebraskaflood2019 @ [REDACTED] https://t.co/n5kDdKMfBG
Displaced people and evacuations		
Chaotic evacuations with emergency sirens blaring as the #MissouriRiver rises to the top of the three-story-high levee wall in St. Joseph, Missouri! https://t.co/BCUcxXrF05 via @[REDACTED]	It made the hair on my neck stand up going by #Elwood where cops are at the exits to keep people from getting into town after it's been evacuated. Helicopters are flying around. It's insane. The town is just empty. #Flood2019 #missourifloods https://t.co/xJnJu3n71k	They've sounded the alarm. Mandatory evacuation for Elwood Kansas. Please pray for them. #Flood2019 https://t.co/gDLvJlwBeo
Infrastructure and utilities damage		
Surveyed flood damage in Plattsmouth and many other southeast Nebraska communities along the Missouri River which is so high it's almost impossible to see across to the Iowa/Missouri banks in some places.	There hasn't been much national coverage but, Nebraska is flooded & lives are completely devastated & even lost. #NebraskaStrong	Busy weekend for @DouglasCountyNE road crews as they rebuild washed out roads and shoulders. Please go slow and give them room. #NebraskaSTRONG #NebraskaFlood https://t.co/kPGx1PTgsA
#NebraskaFlood #NebraskaStrong https://t.co/XPsTBM37or		


Donation needs or offers or volunteering services

	@[REDACTED] Thank you to all the volunteers in all the communities that are taking time to give back.	
	A truly amazing Volunteer is: Selfless Generous Helpful Thoughtful Valuable	
Thank you to @[REDACTED] and all the people from across the country providing hay for flood relief!	Patient Kind Giving #volunteers #NebraskaStrong #loveinaction #k9comfortdogs	How to help those impacted by the floods? Drop off supplies and donations at any @[REDACTED] location. Hang in there, everyone. We'll get through this together.
#NebraskaFlood #NebraskaStrong https://t.co/kjGtG3a6hR	#Flooding2019 https://t.co/nYKJB0y0VL	#nebraskastrong https://t.co/Sx1w7vccVV

Caution and advice

We decided to plot all flood warnings since March 14th. Over 35,000 square miles and 5.2 million people covered! #Flood2019 https://t.co/q7axlD5fOi	#Nebraska, the Flood Warning continues for #MissouriRiver areas near #Blair affecting #Harrison and #WashingtonCo. #NebraskaFlood #Flood2019 https://t.co/fJdm0wf4kO	Moderate flooding going on here. Worried about the people in the bottoms. Hopefully it will crest soon. #flood2019
--	--	---

Sympathy and emotional support

This is incredible! \$325,000 and counting!		
So many #NebraskansHelpingNebraskans. #NebraskaStrong #NebraskaStrongDay https://t.co/qhv8g2BMLZ	This is truly devastating. A true natural emergency. Stay strong. My prayers are with you all until we find out how we can help.  #Flooding2019 #Flood19 #Flood2019 https://t.co/v02lpvsDGT	Our thoughts and prayers are with those in Nebraska. As we get the latest updates of videos, pictures, and stories, our hearts break, but we also know Nebraska will get past this. #NebraskaStrong https://t.co/gx5RH4GfPB

Other useful information

Farmers affected by #flooding look ahead to planting #Nebraskastrong	After a flood, what happens to local #realestate markets? #Flood2019	If you or someone you know is being affected by the flood, counseling and information services are available. #NebraskaFlood #NebraskaStrong
--	--	---

Not related or irrelevant

Lunch successfully served! NEFB and @[REDACTED] were in Verdigre today with burgers, hotdogs, water and more! We also dropped off a donation of supplies for the recovery effort. We're honored to help our communities after historic floods. #nebraskastrong #fbproud <https://t.co/uBD2XThpl4>

I really enjoyed my visit to the Black Hills Unity Concert in 2017. People have a sense of humor about themselves that's very Midwestern #lakota #missouririver #greatplains <https://t.co/uXCAkiATSU>

Happy Spring! New calf born on the island-Our cattle are on a high spot next to the river. We can go feed them everyday by boat, hopefully will be able to rescue them soon. 🙏 #Flood2019 #FarmFamily #WeCanMakeIt #PrayForAll <https://t.co/V31GDbrIX9>

Figure 5 shows that, even though the model can rapidly sort social media into classes, a considerable amount remains essentially unclassified: potentially valuable but unable to be sorted. The “Other useful information” category was the largest category of all (40% greater than “Donations”).

The most tweets of the actionable/sentiment categories were those categorized as “Donations,” which occurred 5.75 times as much as the first four categories combined. The volume of tweets in the “Sympathy” category also exceeded the first four combined by 1.2 times. The shift in the stage from the immediacy of the event to its aftermath likely explained the popularity of these categories. Even when we consider the deep learning model's ability to rapidly categorize into nine categories, the existing categories may provide insufficient guidance.

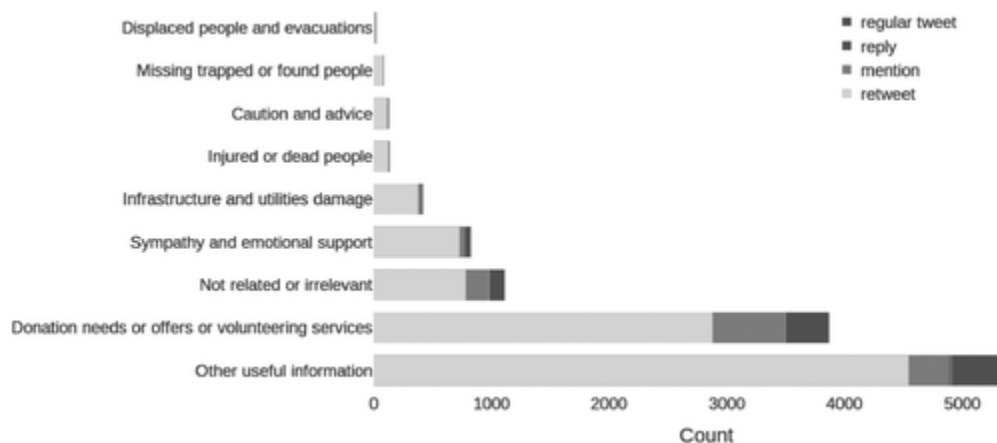


Figure 5. The number of tweets assigned to each category

The vast majority of tweets in our dataset are retweets, supporting the findings of Starbird, et al., who characterize retweet activity during crises as a means for those affected to spread information that they believe to be valuable and trustworthy.¹²¹ When retweets are removed from

¹²¹ Starbird, Palen, Hughes, and Vieweg (2010)

the dataset, the “Donations” category supersedes the “Other useful information” category and outflanks the first four categories by more than 14 times. “Other useful information” comprises almost half (41%) of the unique tweets.

To combine our classification model with the results from the SNA, we selected two accounts that ranked highly in the measures of node centrality and were official information sources: State Governor Pete Ricketts and the Nebraska Emergency Management Agency (NEMA). Figures 6 and 7 show the distribution in categories of the tweets from each account. As shown in Figure 6, tweets from Governor Ricketts primarily related to donations and volunteering services. A large number of his tweets were still classified as “Other useful information.”

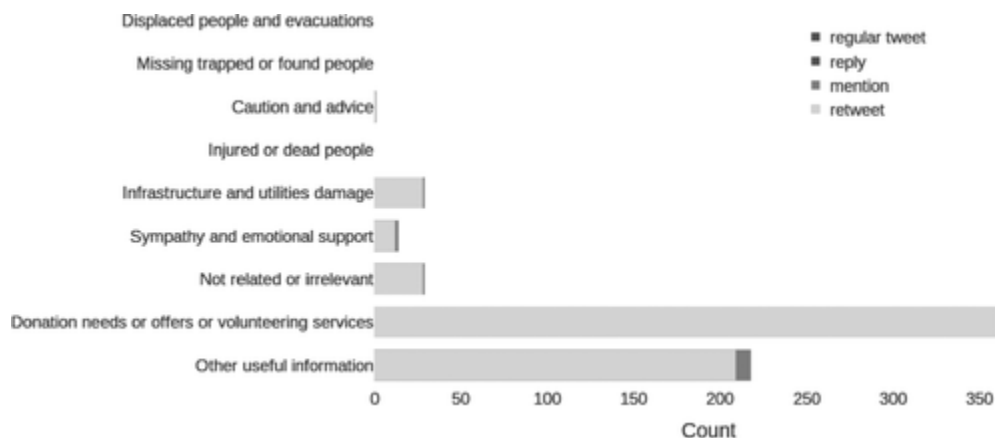


Figure 6. The number of tweets from Governor Ricketts assigned to each category.

Figure 7 shows how many of the tweets from NEMA related to donations and other useful issues. Notably, many of the tweets from NEMA were classified according to the “Sympathy and emotional support” category, which was present to a much lesser extent in the tweets from Governor Ricketts.

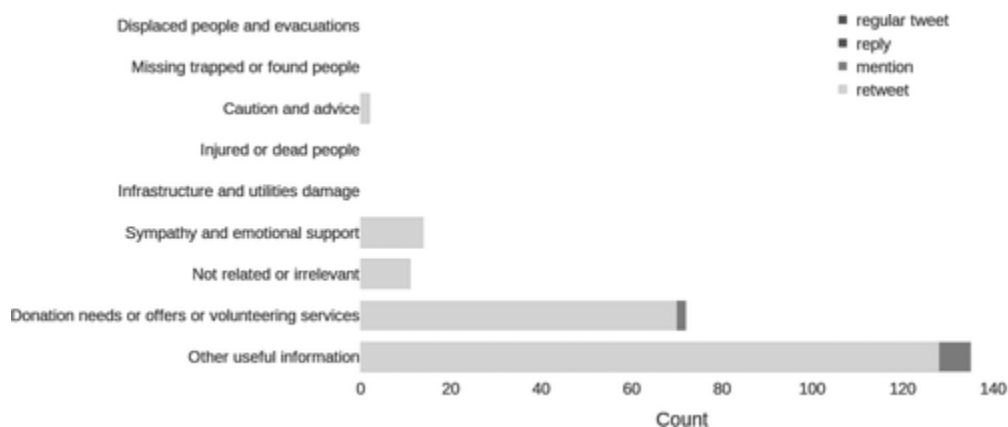


Figure 7. The number of tweets from NEMA assigned to each category, including retweets.

3.5.3 Outcomes

By augmenting NLP with deep learning, our approach extends existing analytical techniques in the field of crisis informatics. Using a labelled dataset from previous crises, we successfully used a deep learning language model to classify tweets from unseen crisis events. This is evinced by our model's significant improvements to automatic tweet classification compared to the previous state-of-the-art models. These additions of AI are innovations to leveraging social media for extreme weather. Furthermore, we successfully combined this model with SNA to gain insights into the flow of information on Twitter during extreme weather events, identifying influential Twitter accounts and the type of information they disseminate. We also chose a training dataset and classification scheme that allowed us to extract actionable items as well as sentiment.

We validated our approach by sampling a representative subset of the classified tweets and manually analyzing our model's predictions. Through this process, which highlights the importance of human involvement,¹²² we found that a supervised approach, even one improved with a language model, was limited by biases in the training data. In particular we found that large amounts of tweets were classified into less-informative 'catch-all' categories such as "Other useful information". Event-specific topics were not captured by the predefined labels in the training data. The work in this activity covered the containment phase of a flooding event in the U.S. Future work should explore extreme weather events in Canada and through the entirety of an event. We addressed both these issues in the next grant activity by developing an approach that combines supervised, unsupervised and language models and then applying it to a Canadian snowstorm.

The findings of this activity were published in the *Journal of Contingencies and Crisis Management*.¹²³ The code used for the publication is publicly available.¹²⁴

The combination of automatic tweet classification and SNA developed in this activity can be used by crisis managers to better characterize the flow of information on Twitter during extreme weather events and assess impact and public response. The tweet classification models that were developed and trained in this activity are also available to download and run for future work or application of these methods.

¹²² Also noted in Nguyen et al. (2016).

¹²³ Romascanu et al. (2020).

¹²⁴ <https://github.com/smacawi/tweet-classifier>.

3.6 Adapting social media classification to Canadian snowstorms

Our case study on the Nebraska flood pointed out the limitations of supervised learning. Although supervised approaches are fast and they work well for classifying tweets from the same event as their training data, they often fail to generalize to novel events. Unsupervised approaches, namely probabilistic topic modelling approaches such as LDA¹²⁵ can overcome this limitation and identify novel categorizations. Unfortunately, these unsupervised methods are typically difficult to apply to tweets due to issues of document length—tweets are short and therefore difficult to create word embeddings—and non-standard language.¹²⁶ The categorizations are entirely dependent on the corpus (consider tweets collected during the event as opposed to in the aftermath of the event). Most importantly, the categories produced by these models can be difficult to interpret by humans, limiting their usefulness.

Our goal in this part of the G&C was to adapt our tested data pipeline and our knowledge of integrating language models with NLP with a Canadian example of extreme weather. How do we retain the benefits of a supervised learning with the flexibility of unsupervised learning's treatment of novel events?

3.6.1 Methods

To address these issues, we proposed a method for the unsupervised clustering of tweets from novel crises using a pre-trained language model finetuned on labeled crisis tweets. Using a method that has become standard practice in NLP, we took the BERT language model that has been trained on billions of documents to gain a general understanding of the English language. We then 'finetuned' it using the labelled data from the above discussed CrisisNLP dataset. By teaching the BERT model to predict labels in the CrisisNLP dataset, we hypothesized that the semantic representations or 'embeddings' of the model would retain some salient information about the language used in tweets during crises, while its general comprehension of English would allow it to identify novel topics from a specifically Canadian event. This hypothesis was proven correct: our model successfully bridged the gap between supervised, trained approaches and unsupervised learning by incorporating knowledge from labeled tweets of past crises while allowing classification of tweets from new crises in novel and relevant topics.

Testing Data: We assessed the transferability of our approach to novel crises by creating a dataset of tweets from Winter Storm Jacob, a severe snowstorm that hit Newfoundland-Labrador, Canada on January 18, 2020. As a result of the high winds and extreme snowfall, 21,000 homes were left without power. A state of emergency was declared in the province as snow drifts as high as 15 feet (4.6 m) trapped people indoors. Following the Newfoundland-Labrador Snowstorm, we collected 21,797 unique tweets from 8,471 users between January 17 and January 22 using the Twitter standard search API with the following filter terms: *#nlwhiteout*,

¹²⁵ Blei, et al. (2003).

¹²⁶ Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. In Proceedings of the first workshop on social media analytics (pp. 80-88).

#nlweather, *#Newfoundland*, *#nlblizzard2020*, *#NLStorm2020*, *#snowmageddon2020*, *#stormageddon2020*, *#Snowpocalypse2020*, *#Snowmageddon*, *#nlstorm*, *#nltraffic*, *#NLwx*, *#NLblizzard*. Based on past experience with the Twitter API, we used hashtags to limit irrelevant tweets (e.g., searching for ‘blizzard’ resulted in half the collected tweets being about the video game company with the same name). We filtered retweets to only capture unique tweets and better work within API rate limits.

Model Development: Our modeling approach can be understood in four steps. First, we trained a model using the BERT language model, which was pretrained on billions of documents.¹²⁷ We further trained our model on CrisisNLP (from the Nebraska flood case study). Second, this ‘finetuned’ model was used to produce ‘embeddings’ of each tweet gathered during Winter Storm Jacob. Embeddings are representations of documents—tweets in this case—that assign each document a point in a high-dimensional vector space. By assigning each document a point, their similarity can be measured using standard clustering algorithms.

Third, given the embeddings for each tweet, we applied a K-Means clustering algorithm¹²⁸ to generate topics. Unsupervised classification can generate numerous semantically distinct clusters or topics, from 10 to 30 in many cases. Most common topics can be “everyone knows; everyone’s retweeting” and thus useless. We chose to limit the optimal number of topics to better integrate with the nine categories recommended by CrisisNLP.

Finally, to extract meaningful keywords from topic clusters generated by our model and ensure their interpretability, we used Term frequency-Inverse document frequency (Tf-Idf) by combining each cluster into one document, returning a score for each word present in the tweets. We then combined activations from the model’s attention layers with Tf-Idf by multiplying the attention values for each token with its Tf-Idf score. Tf-Idf and attention are used to amplify certain categories of words or specific objectives, for example, if there are terms that reflect crises or extreme weather events. They also can be used to de-emphasize high frequency words.

For the sake of comparison, we compared this approach to three other models, including a version of BERT that was not finetuned on CrisisNLP data. These gave us a baseline against which to assess the impact of finetuning.

Model Evaluation: The evaluation of the output given by our FTE model involved both automatic metrics and human feedback. We reported measures comparing the performance of FTE with that of three baseline topic modelling techniques: LDA, Biterm (BTM), and non-finetuned BERT. As mentioned above, LDA is the most popular form of topic modelling, which is the most common form of unsupervised classification. BTM is a topic modelling technique used for short message texts as it is hard to construct word embeddings from

¹²⁷ Devlin et al. (2019).

¹²⁸ Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28(2), 129-137.

‘data-sparse’ content.¹²⁹ We also compared our model on non-finetuned BERT (repeating the steps above without finetuning on the CrisisNLP data), to assess the impact of finetuning. We chose topic coherence metrics, which score individual topics by capturing the semantic similarity between prominent keywords pertaining to that topic.¹³⁰ Coherence is typically used in topic modelling to determine the optimal number of topics. It can be used to inspect how well the topics cohere, in terms of the words that represent each topic being related and forming a coherent set of ‘facts’. Coherence metrics also were chosen based on previous research that found them to be most compatible with human notions of coherence.¹³¹

The human evaluation involved two tasks given to four annotators familiar with concerns of crisis managers. Specifically, we gave two evaluation methods focused on (1) topic keywords and (2) document clustering within topics. To evaluate the quality of topic keywords, annotators were presented with the top 10 keywords for each topic (Table 7) and asked to assign an interpretability score and a usefulness score on a three-point scale. We defined interpretability as good (eight to ten words are related to each other), neutral (four to seven words are related), or bad (at most three words are related). Usefulness in turn considers the ease of assigning a short label to describe a topic based on its keywords and that the label *should be useful for crisis managers*. We score usefulness on a three-point scale: useful, average, or useless.

We then evaluated the clusters. This task assessed—from the perspective of a crisis manager—the interpretability and usefulness of the actual documents clustered within a topic, instead of only analyzing topic keywords as done in previous work. Given an anonymized model, for each topic we sampled 50 sets of four documents within its cluster along with one document—the “intruder”—outside of that topic. For each set of documents, all four annotators were tasked with identifying the intruder from the sample of five documents, as well as assigning a interpretability score and a usefulness score to each sample. Participants could either identify a single tweet as the intruder, or label the intruder as “unsure” to discourage guessing. The interpretability score was graded on a three-point scale: good (3-4 tweets seem to be part of a coherent topic beyond “snowstorm”), neutral, and bad (no tweets seem to be part of a coherent topic beyond “snowstorm”). The cluster usefulness score was similar to the keyword usefulness score, but formulated as a less ambiguous binary assignment of useful or useless for crisis managers wanting to filter information during a crisis.

The annotators performed these two tasks on our FTE model and the BTM and non-finetuned BERT baselines. We chose to exclude the LDA baseline from this evaluation following its low performance with the automated metrics on nine topics (see Figure 8).

¹²⁹ Yan, X., Guo, J., Lan, Y., Cheng, X.. (2013). A biterm topic model for short texts. Proceedings of the 22nd international conference on World Wide Web, pp 1445–1456. <https://doi.org/10.1145/2488388.2488514>

¹³⁰ Fang, A., Macdonald, C., Ounis, I., Habel, P. (2016). Using Word Embedding to Evaluate the Coherence of Topics from Twitter. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Pages 1057–1060 <https://doi.org/10.1145/2911451.2914729>

¹³¹ Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM international Conference on Web search and data mining (pp. 399-408).

3.6.2 Results and analysis

Qualitative insights on tweet content during a Canadian snowstorm: The topic keywords from our FTE model and a baseline BTM model are shown in Table 7. Our model retained some of the original categories from CrisisNLP (Table 4), suggesting that such categories are relevant for the affected citizens of the specific snowstorm at hand. Indeed, the topic suggests the influence of finetuning with clustering of terms like trapped, stranded, ambulance, dead, rescue, permitted, body, and helped (one would need to manually inspect the corpus to determine why hydrant and garbage clustered as well). Topic 3 (outage, campus, widening, advisory, reported, impassable, remaining, thousand, reporting, suspended) suggests Category 4, Infrastructure and utilities damage, although it also implies the ways in which snow and snow removal does not fit the narrative of floods and earthquakes, the predominant extreme events in CrisisNLP. Also note the bottom up emergent quality of unsupervised classification. Even when filtering for nlstorm, one still can get intrusion of other concepts, like the introduction of Ivy Park Adidas shoes in Topic 2.

Interestingly, we found that our approach identifies *novel* unsupervised topics that are distinct from the labels seen during supervised training. For instance, the appearance of meteorological information (Topic 1) and information about power outages and closures (Topic 3) suggests that a significant volume of affected citizens' tweets fell within these topics during the snowstorm, despite not being a "universal" label present in the CrisisNLP categories. Topics 8 and 9 were less clear to annotators, but the former seemed to carry information about how extreme the storm was thought to be and the latter about citizens bundling up indoors with different foods and activities.

Compared to the FTE topic keywords, the topics in BTM and non-finetuned BERT (Table 7) were less semantically meaningful. Biterm is a traditional unsupervised classification and an extension of the LDA model to short texts. Note that the top topics are not particularly useful, which is typical of topic modeling. The first topic collects the obvious; the second clusters popular and political events of the day. Annotators noted that Topic 5 showed information about the need to stockpile provisions (with people, today, storm, need, like, day, grocery, food, open, know). Topic 7 relates to traffic conditions and Topic 8 potentially provides information about a state of emergency and closed businesses. The non-finetuned BERT proved unproductive. Topic 5 (campus, provincial, mayor, remain, operation, pharmacy, region, taxi, advisory) suggests preparations needed as services shutter. If the topic numbers were not restricted then the clusters might be more interpretable but they still might not be actionable.

Table 7. Topic keywords for our FTE model and the BTM baseline used in human evaluation.

					Topics				
Model	1	2	3	4	5	6	7	8	9
FTE	reporting	ivyparkxadidas	outage	assistance	prayer	blowingsnow	trapped	monster	bread
	monster	mood	campus	assist	praying	alert	stranded	meteorologist	song
	snowiest	song	widening	troop	pray	advisory	hydrant	drifting	coffee
	recorded	blackswan	advisory	volunteer	wish	caution	ambulance	perspective	milk
	peak	le	reported	providing	wishing	advised	dead	stormofthecentury	feelin
	temperature	snowdoor	impassable	relief	humanity	stormsurge	garbage	mood	pin
	cloudy	perspective	remaining	aid	brave	wreckhouse	rescue	snowdrift	enjoying
	reported	ode	thousand	request	surviving	surge	permitted	climate	laugh
	equivalent	music	reporting	offering	loved	drifting	body	windy	favorite
	meteorologist	adidasxivypark	suspended	rescue	kindness	avoid	helped	snowdoor	girl
	emergency	eminem	safe	cbcnl	people	like	nltraffic	closed	cm
BTM	state	photo	stay	today	today	storm	road	st	st
	st	click	blizzard	thank	storm	snow	power	tomorrow	winds
	city	learn	newfoundland	people	need	time	st	john	pm
	cityofstjohns	saveng	storm	day	like	day	street	remain	today
	says	michelleobama	canada	home	day	newfoundland	drive	january	km
	declared	sexeducation	weather	work	grocery	going	pearl	update	airport
	john	ken	nlstorm	help	food	blizzard	roads	emergency	yyt
	mayor	starr	warm	storm	open	house	line	state	snowfall
	roads	pin	snowstorm	nltraffic	know	today	mount	today	blizzard
	shareyourweather	pin	thankful	lay	metro	mood	glad	thankful	monster
BERT	ivyparkxadidas	feelin	bank	justintrudeau	campus	hutton	looked	glad	stormsurge
	saturdaythoughts	taxi	yo	save	provincial	cute	apartment	sharing	le
	saturdaymotivation	mayor	mayor	suck	mayor	compound	cloudy	favourite	wreckhouse
	snowdoor	metro	neighbor	radiogregsmith	remain	lovely	law	shareyourweather	newfoundlandlabrador
	bcstorm	en	walked	kettle	operation	spread	snowblowing	grateful	ode
	titanscollections	blowingsnow	wa	anthonygermain	pharmacy	stream	honestly	monster	snowiest
	badboyforlife	ivyparkxadidas	bus	kilbride	region	design	as	feelin	ottawa
	bingo	bus	glad	kaylahounsell	taxi	ivyparkxadidas	eat	neighbor	historic
	blackswan	dannybreennl	pharmacy	campus	advisory	crisis	weird	mom	explore

Automatic evaluation and human evaluation: For the automatic evaluation, we tested coherence across four different numbers of topics: 5, 9, 10 and 15. Figure 8 shows the automated coherence scores for the two models across a range of topic numbers, using the context vector (Cv) coherence metric.¹³² On the automated coherence measures, our FTE model outperformed the LDA and BTM baselines for all number of topics. Whereas Figure 8 shows that the best performing model was the non-finetuned BERT model; human evaluation (reported below) showed that the topics given by this model were completely uninterpretable. This highlights the importance of always pairing automatic evaluation with human evaluation.

¹³² Röder, M., Both, A., & Hinneburg, A. (2015).

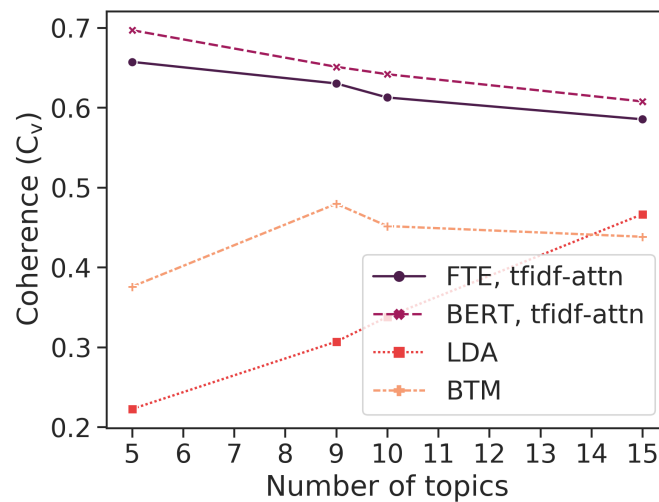


Figure 8. Automated C_v coherence scores for different models, showcasing the improvement of our method over the baselines. Whereas the non-finetuned BERT model achieves better automated coherence than FTE; the topics it produced were completely uninterpretable for humans, highlighting the importance of pairing automatic evaluation with human evaluation.

Recall that the human evaluation aimed to assess the interpretability and usability of the keywords in each cluster, as well as the cluster itself. We only include the results from our FTE model and the BTM model because LDA and non-finetuned BERT returned significantly lower results. Scores on the original three point scale were normalized to fall between 0 and 100 (i.e. 0, 50, 100). Intruder detection was normalized to the same scale with 0 for incorrect detection or unknown intruders and 100 for correct detection. Similarly, Unknown Intruders receive a score of 100 if the annotator declared they could not identify an intruder. Topic Count represents how many topics score above 50; this helped ensure that the small number of high-scoring topics, which would be represented in the average, did not overwhelm a large number of medium-scoring topics. Fleiss's Kappa is a standard statistical measure for inter-rater reliability. The Kappa value assesses how well annotators agreed in scoring the interpretability and usefulness of the keywords and clusters, with values close to 0 indicating weak inter-rater agreement and values close to 100 indicating strong inter-rater agreement. Scores above 20 indicate fair agreement; whereas scores below that are considered poor.

We compared results averaged across annotators for topic keyword evaluations (Table 8) and across samples and annotators for cluster evaluations (Table 9). As you can see from the bolded items in Table 8, the keywords from the FTE model return an overall slight improvement in usefulness and interpretability over the BTM baseline. As seen in Table 9, our FTE model returned a greater number of higher quality topics. The inter-rater agreement also is stronger for FTE when assessing interpretability, although approximately the same for utility to crisis managers. Annotators also were presented with samples that contained intruders, tweets outside the topic cluster. As you can see from the bolded values in the last two rows in Table 9,

annotators were able to identify intruders within the FTE clusters more accurately and with less uncertainty compared to the BTM baseline. Indeed, correct intruder detection was overall higher for FTE and there were more high-performing clusters. Furthermore, the lower score in “Unknown Intruders” for FTE suggests that annotators were more confident in their selection of an intruder and tweets within the same cluster for our FTE model have a stronger semantic relationship.

Table 8. Keyword Evaluation scores averaged across topics, number of topics with average scores greater than 0.5, and inter-rater agreements (Fleiss's Kappa¹³³).

Score	Average Score		Topic Count		Fleiss's Kappa	
	BTM	FTE	BTM	FTE	BTM	FTE
Interpretability	31.94	65.28	1	5	15.01	17.97
Usefulness	27.78	59.72	1	5	12.36	21.55

Table 9. Cluster Evaluation scores averaged across topics, number of topics with average scores greater than 0.5, and inter-rater agreements (Fleiss's Kappa).

Score	Average Score		Topic Count		Fleiss's Kappa	
	BTM	FTE	BTM	FTE	BTM	FTE
Interpretability	50.28	51.53	3	4	11.05	23.45
Usefulness	45.46	46.11	3	5	21.82	21.60
Correct Intruders	35.28	44.17	2	4	25.78	31.50
Unknown Intruders	26.39	8.89	0	0	-	-

¹³³ https://en.wikipedia.org/wiki/Fleiss%27_kappa

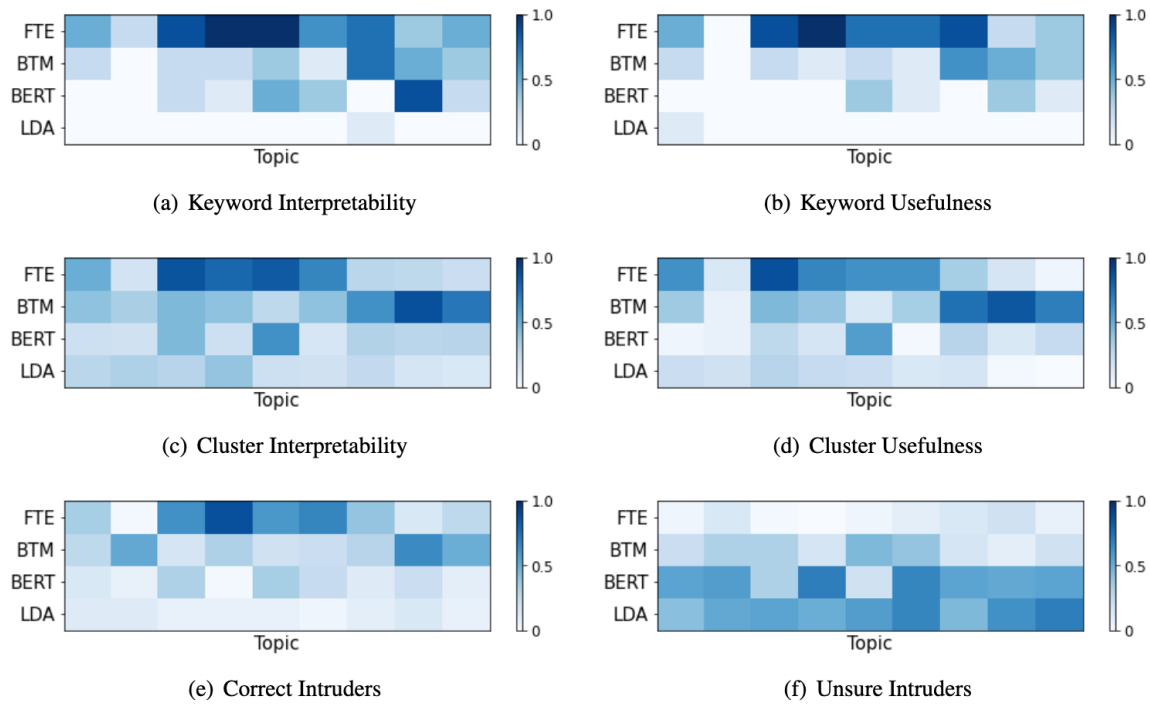


Figure 9. Topic-level results for keyword and cluster evaluations, aligned with topics from Table 7. All scores are rescaled to values between 0 and 1, then averaged across annotators and samples.

Figure 9 also shows that our method outperforms the various baselines on the keyword and cluster evaluation. In particular, the stark improvement over non-finetuned BERT further confirms the importance of the finetuning step in our method. Although the improvements in average scores over BTM reported in Table 9 are marginal, we find that the number of interpretable and useful topic clusters was greater for our approach. Indeed, while the BTM baseline had more semi-interpretable (i.e. only a subset of the sampled tweets seemed related) but non-useful topics, our method had a much clearer distinction between interpretable/useful and non-interpretable/non-useful topics, suggesting that tweets marked as hard to interpret and not useful are consistently irrelevant (Figure 9b). This may be preferable for downstream applications, as it allows users to better filter out irrelevant content.

The annotators also identified intruder tweets in topic samples from FTE more reliably and with less uncertainty, as measured by the number of correct intruders predicted and the number of times an intruder could not be predicted (Figure 9, e-f). Interestingly, BTM topics rated for high interpretability had lower rates of correct intruder detection, suggesting that these topics may seem misleadingly coherent to annotators. Inter-rater agreements as measured by Fleiss' κ further confirm that annotators more often disagreed on intruder prediction and interpretability scoring

for BTM topics. This is undesirable for downstream applications, where poor interpretability of topics can lead to a misinterpretation of data with real negative consequences.

3.6.3 Outcomes

We introduced a new model which leverages both supervised and unsupervised methods to address each approach's shortcomings and make automatic tweet labelling both interpretable and useful for crisis managers. This research continued from the previous exploration with supervised methods applied to CrisisNLP and we explored the adaptation of our models to another type of weather event (snowstorms). Our results also were supported by novel human evaluation tasks of automatic tweet labelling which assessed the interpretability and usability of the model.

We restricted the number of topics to better integrate the unsupervised with the supervised learning. However, topic generation in unsupervised learning is more dynamic. Our work can be extended to explore clusters initialized with a varying number of topics or to explore “subtopic” clusters that form within a single topic. Greater control over the unsupervised learning (specifically, topic modelling) is likely necessary to ensure the retention of certain categories that are common across disasters (e.g., infrastructure damage or volunteering efforts). This could be done through incorporating “term-seeded” categories to guide a topic to have a certain semantic meaning; this has been implemented for other topic modelling approaches like LDA and Biterm.¹³⁴

Our paper presenting the development of our FTE model and its application to snowstorms was presented at the second Adapt-NLP Workshop, organized virtually in April 2021 and in conjunction with the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL).¹³⁵ The FTE model has been implemented for real-time use on a dashboard. We also gave a talk about this portion of the work at the HiWeather Workshop of the WMO.¹³⁶ The dataset of snowstorm tweets and code from both the FTE model and baseline models are available online on the project's github account.¹³⁷

Our contribution was four-fold: we created a dataset of snowstorm-related tweets, developed a model that learned to cluster tweets into novel crisis-relevant topics, assessed the model's ability to extract interpretable keywords and transfer to unseen crisis events and showed that finetuning BERT models on supervised tasks can provide salient information for unsupervised classification

¹³⁴ Li, N., Chow, C. Y., & Zhang, J. D. (2019). Seeded-BTM: enabling biterm topic model with seeds for product aspect mining. 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 2751-2758). IEEE.

¹³⁵ Brunila, et al. (2021).

¹³⁶ <http://www.hiweather.net/>

¹³⁷ <https://github.com/smacawi>

in a related domain. A more thorough technical description of the model architecture and evaluation metrics is included in our paper.¹³⁸

The FTE model was designed to extract useful information from social media and assist crisis managers by using output from pretrained models to drive greater semantic meaning in the cluster output of an unsupervised model. The code is open-sourced and designed to be easy-to-use by others for future projects, which we have currently implemented in a dashboard.

3.7 Dashboard

Crisis managers need to make quick, high-impact decisions based on limited information, but operating ML models can be timely and complicated. After consultation with ECCC, we determined that a web-based user interface would be critical to making the models more accessible and to operate the data pipeline and visualise results. Our objectives for the dashboard were two-fold: (1) allow users to run the data pipeline to filter, harvest, and classify Tweets, and (2) to visualize the Twitter data and model results in real-time.

3.7.1 Methods

We conducted background research and consulted with ECCC to determine requirements for the dashboard. The first main requirement was to allow users to harvest live Tweet data using the Twitter API. It was important to allow users to search and filter incoming Tweets using Twitter API parameters. This required building infrastructure to securely use a user's Twitter API credentials. It also was important to allow access to previously collected data and manage data collection for historical Tweets. To support this, the interface needed to provide secure user authentication and allow multiple user sessions.

Second, the dashboard needed to allow users to run analysis on incoming Tweets. The primary component of this analysis was to run the supervised model classification. A secondary component that was built into a development version of the dashboard was to run the unsupervised models to identify topics for Tweet classification.

Finally, the dashboard needed to visualize the results of the Twitter scraping and classification analysis. This included: (1) geolocating Tweets and visualizing them in an interactive map, (2) visualizing Tweet metadata in a table, including model classifications, and (3) visualizing occurrences of each category of Tweet.

3.7.2 Implementation and codebase

We built a prototype dashboard that classified and geolocated tweets in real-time. The primary components of the system were: text-based geolocation of tweets, classification of tweets and an interactive map to visualize results. Our tweet geolocation and classification allowed us to

¹³⁸ Brunila, et al. (2021).

organize and present useful information on an interactive map that emphasizes the most up-to-date Tweets and makes it easier to navigate the large volume of information in an intuitive way.

The project is structured as a set of standalone APIs and a front end. The standalone APIs are used to interface with the Twitter scraping and ML back end. The front end was implemented using Plotly Dash, a python framework that allows developers to provide point-&-click interfaces to models written in python. Using python libraries made the codebase more consistent with the rest of the project.

The entry point for the application is in our online repository.¹³⁹ Each tab is modularized and also is stored in the github account.¹⁴⁰ The callback for Twitter data is shared (for performance reasons).¹⁴¹

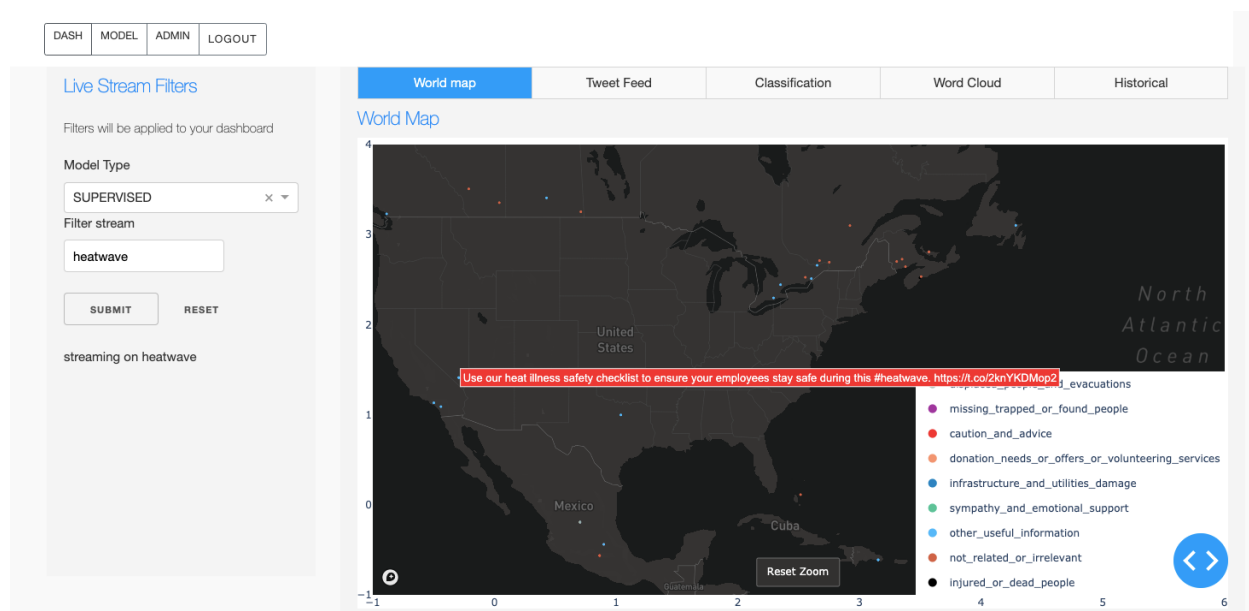


Figure 10. The dashboard with the World map tab selected.

3.7.3 Operation

Once deployed, the dashboard provides a user-friendly interface to the models. After entering user login information, the interface provides multiple tabs for access to the data pipeline and to visualize results. A control panel on the left hand side is used to enter search information to filter incoming tweets and to submit requests through the Twitter API (Figure 10).

¹³⁹ https://github.com/smacawi/dashboard_frontend/blob/master/index.py

¹⁴⁰ https://github.com/smacawi/dashboard_frontend/tree/master/components

¹⁴¹ https://github.com/smacawi/dashboard_frontend/blob/master/callbacks/twitter_feed_callback.py

Once a search has been made the results are populated by Tweets scraped in real time. The tabs provide different views of the data. The first tab, shown in Figure 10, shows Tweets geolocated on an interactive map. The data can be viewed in four other tabs. First, the Tweet Feed tab displays the Tweets in tabular form, including the time the Tweet was recorded, its location and the label it was assigned, as shown in Figure 11. The Classification tab displays the frequency of each category as a coloured bar chart and updates as Tweets are streamed. The Word Cloud tab visualizes word frequency by the size of the words displayed. Finally, the Historical tab allows users to access the historic Twitter API to search for historical Tweets by specifying search terms.

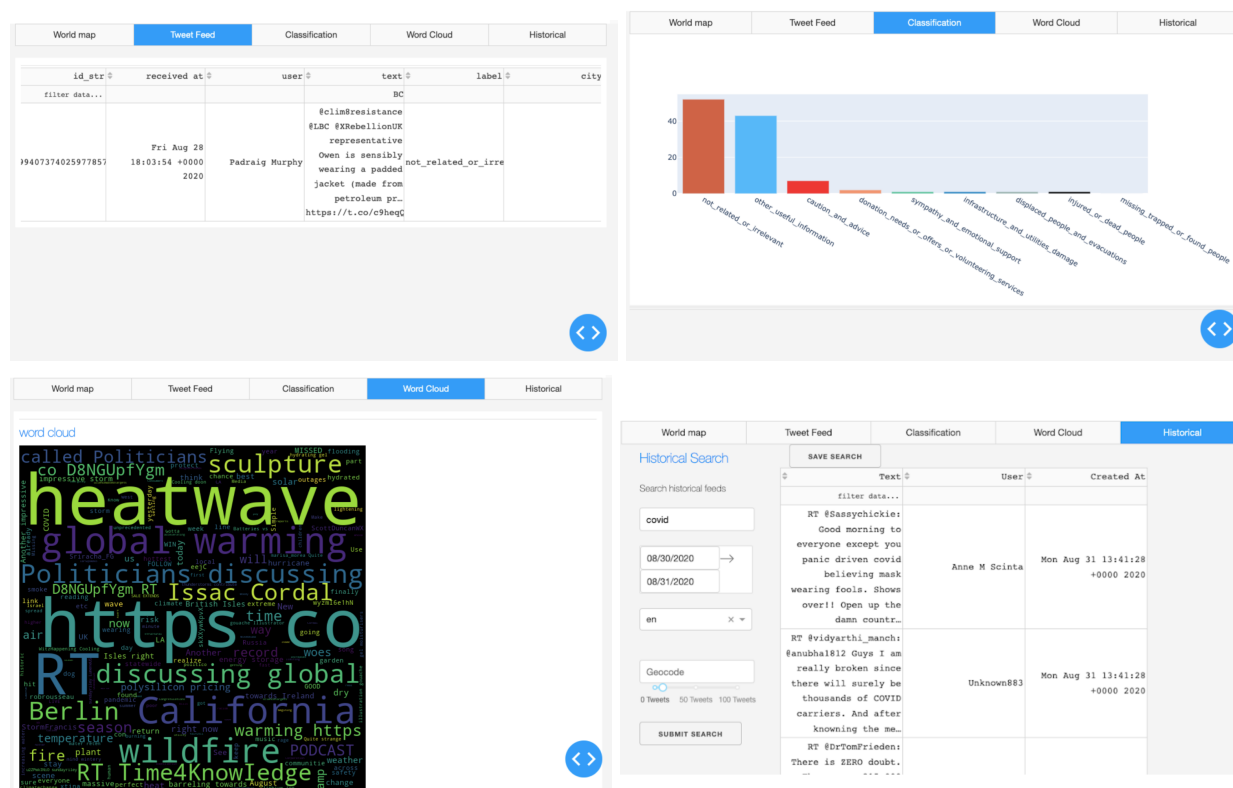


Figure 11. Four dashboard tabs (from top-left clockwise): tabular Twitter data, bar chart of the frequency of each category, word cloud, and historic Twitter search.

3.7.4 Outcomes

We developed a fully functional user interface to our data pipeline to filter, harvest and classify live Tweets. The outputs include an open source repository for both the standalone APIs to access the data pipeline and the dashboard front end.¹⁴² The code is modular and so it can be repurposed for other applications. We also have created a static demo version of the app.¹⁴³

¹⁴² https://github.com/smacawi/dashboard_frontend.

¹⁴³ <https://dash-flask-ec.herokuapp.com> - requires login details username: hello, password: world

Our publication described the creation of an interactive dashboard to visualize results from the NLP/Twitter data pipeline.¹⁴⁴ We gave a presentation to ECCC on our dashboard.¹⁴⁵

While we met the immediate needs for the project, the development of the dashboard prompted discussion of additional features during feedback sessions with Environment Canada, some of which were out of scope of this project. These included: geolocating tweets that do not contain location metadata, integrating tweets as data layers with existing ECCC RADAR systems, and allowing users to correct incorrect categorizations to train the ML model. We have plans to address these issues in the near future and would welcome additional comments from ECCC.

4. Project Conclusions and Outcomes

The SMaCAWI project has shown how social media can be leveraged to better meet the needs of the ECCC and the public during extreme weather events. Individuals often share information about emergency situations, real-time weather conditions, infrastructure damage and cautionary information. It can be challenging for crisis managers and others to make this information actionable, since social media data is typically unstructured, high volume and high velocity. There is a well-acknowledged noise-to-signal problem in this potentially massive amount of content.

To understand public sentiment and obtain actionable information we automated the analysis of public responses from social media data. We assessed the field to determine the state-of-the-art in AI approaches to crisis analytics. We combined ML methods, SNA, deep learning, language models and other NLP techniques to categorize social media data and to identify key social media accounts. We constructed a fully functional data pipeline and applied it in several case studies. We generated a dataset of Canada-specific extreme weather classified Tweets. We developed evaluation methods to assess model performance and showed that they improved upon the state-of-the-art methods for analyzing weather-related social media data. We developed a user interface for our models. As part of this process, we identified the limitations and challenges of applying ML methods in an extreme weather context. We provided a Made-in-Canada solution that also achieved international acclaim. It also was a Made in Quebec solution, conducted at McGill University and leveraging the concentration of AI in Montreal and connections to MILA, of which our team members took part (e.g., graduate school, internships, hackathons).

¹⁴⁴ Mircea, A. (2020). Real-time Classification, Geolocation and Interactive Visualization of COVID-19 Information Shared on Social Media to Better Understand Global Developments. Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. <https://doi.org/10.18653/v1/2020.nlpCOVID19-2.37>

¹⁴⁵ Berger, L. (2020). Social Media and Crowdsourcing Assessment of Weather Impacts: The Dashboard. Demonstration to ECCC staff and interested parties on September 9, 2020. <https://smacawi.github.io/docs/september-pres-dashboard.pdf>

Overall, our project has provided the technical and operational scaffolding for implementing ML models in crisis management situations that are particular to Canada, but could generalize to other contexts as well. Our models and our dashboard could, with some finetuning for robustness, be deployed in real-life situations, provide the opportunity for more efficient and rapid-response crisis management during different types of disruptive weather events. Implementing our models in real-time and on a large scale could provide significant opportunities to better support the public during the distress caused by crises induced through climate change and we hope to see this deployment in the near future.

We have several recommendations for this work. Further exploratory work and practical application of methods can be used to expand analysis of social media to assess public responses to extreme weather. Additional tests can be conducted of different weather conditions in Canada (floods, freezing rain) or of different conditions of the same event (wind, freezing rain, snow). Weather events can be compared in the same area from year to year or within a season. Second, the data pipeline and models developed can be used via the user interface to scrape and classify tweets. More models and more visualizations can be added. Third, the codebase was designed with the intention of being easily modified or extended. Components of the project are documented and highly modular to allow others to add to existing functionality.

Figure 11 summarizes the skills and expertise we recommend are needed to run the models and interact with the code. Once it is deployed, the front end interface to the models requires no experience interacting with code or underlying software. It does, however, require some familiarity to understand the model results and the limitations of the underlying model and pipeline. Experience with command line, python and web technologies would be useful for installing the dashboard and deploying it on a web server. Using the system components requires a greater level of familiarity with the technologies (e.g., python, plotly), and concepts (e.g. ML, NLP). The project must be cloned from the GitHub repository and is installed via the command line. There are several reasons someone may wish to interact with the codebase. Relatively little technical background may be necessary to make minor fixes or adjustments. A deeper understanding of python and the project's architecture may be necessary to modify existing functionality or add new features. Familiarity with ML techniques and software libraries would be useful to retrain the models with new data or to adjust its parameters. The codebase has been documented throughout to encourage its use in other projects and to support further development.

GUI yes	For your supervisor	Have skilled demonstrator who knows features
	For the end user (of the front end)	Have documentation describing features
	For the person doing the filtering*	Understand limits of filters, streams
GUI no	For developer, who may wish to modify functions	Install via command line Add to/modify to existing functionality (eg new viz) Know plot.ly or comparable dashboard engine Know Python, comparable coding language
	For the data scientist, working on the classifications	Retrain model with new datasets/different configurations Use new topic modelling, classification model
	For system administrator, who maintains h/s	Monitor servers, update patches

*probably also needs skills to install

Figure 11. The skills needed for using the codebase or user interface to the models

This project has produced a range of outputs that included an open source codebase, academic publications, presentations and project reports. It also includes highly qualified personnel. Below we list selected outcomes with links to relevant work.

4.1 Codebase

The codebase for this project has been made publicly available as open-source software at <https://github.com/smacawi>. The codebase includes modular repositories to harvest Tweets using the Twitter API, to classify Tweets using supervised, unsupervised and hybrid deep learning models, and to deploy a user interface to run the models and visualize results. The repositories for components of the project are summarized in Table 10. A demo version of the dashboard with dummy data can be found at: <https://dash-flask-ec.herokuapp.com/> (username: hello; password: world).

Table 10. Individual repositories for components of the SMaCAWI codebase

Twitter harvester	https://github.com/smacawi/twitter-scraper
Tweet classifier	https://github.com/smacawi/tweet-classifier
Topic modeling	https://github.com/smacawi/topic-modeler
BERT hybrid models	https://github.com/smacawi/bert-topics
Dashboard APIs	https://github.com/smacawi/dashboard_standalone_apis
Dashboard front end	https://github.com/smacawi/dashboard_frontend

4.2 Publications

Our first publication described the use of ML and SNA to analyze Twitter in crisis management contexts. It was aimed at a crisis management audience. Our second publication described the creation of an interactive dashboard to visualize results from the NLP/Twitter data pipeline. Our third publication is aimed at a ML audience, and describes our work at the intersection of supervised and unsupervised methods for categorizing crisis-related Tweets.

- Romascanu, A., Ker, H., Sieber, R., Greenidge, S., Lumley, S., Bush, D., Morgan, S., Zhao, R., & Brunila, M. (2020). Using deep learning and social network analysis to understand and manage extreme flooding. <https://doi.org/10.1111/1468-5973.12311>
- Mircea, A. (2020). Real-time Classification, Geolocation and Interactive Visualization of COVID-19 Information Shared on Social Media to Better Understand Global Developments. <https://doi.org/10.18653/v1/2020.nlp-covid19-2.37>
- Brunila, M., Zhao, R., Mircea, A., Lumley, S., & Sieber, R. (2021). Bridging the gap between supervised classification and unsupervised topic modelling for social-media assisted crisis management. <https://www.aclweb.org/anthology/2021.adapt-nlp-1.5/>

4.3 Presentations and Other Outreach

We presented this work at several conferences and other venues, including:

- Mircea, A. (2020). CrisisTweetMap, Using Natural Language Processing to categorize and map tweets in real-time during crises. McGill University, McHacks 7 Hackathon, February 2, 2020. url: <https://devpost.com/software/crisistweetmap-txahf2>. Slides: <https://docs.google.com/presentation/d/1F8Bsy6VtaLzS-kj4lmANSnOrUOAiAneO68Qg00o9mUM/edit#slide=id.p>
- Zhao, Rosie. (2020). Deep Dive into Applied ML Research Talk. McGill Artificial Intelligence Society. March 9, 2020. <https://www.facebook.com/events/2842959672438165>
- Sieber, R., 2020, Communicating Winter Storms via Natural Language Processing of Social Media, CMOS Congress 54. June 4, 2020, https://www.iclr.org/wp-content/uploads/2020/07/Renee_Sieber.pdf
- Sieber, R., 2020, Crowdsourcing Winter Storm Sentiment via Natural Language Processing, World Meteorological Organisation HIWeather Workshop Successful citizen science, November 30, 2020. <http://hiweather.net/article/18/132.html>

- Brunila, M., Zhao, R. (2021). Bridging the gap between supervised classification and unsupervised topic modelling for social-media assisted crisis management. Adapt-NLP EACL April 21, 2021

4.4 Reporting

Our work was featured by McGill University public relations:

<https://www.mcgill.ca/newsroom/channels/news/using-artificial-intelligence-manage-extreme-weather-events-327770>

This news release and the connection of our work to ECCC was subsequently reported in several outlets, including

- Canadian Geographic
<https://www.canadiangeographic.ca/article/how-artificial-intelligence-can-help-crisis-managers-respond-during-extreme-weather-events>
- Synced AI and Technology Review
<https://syncedreview.com/2021/01/20/mcgill-researchers-use-ai-to-manage-extreme-weather-events/>
- Carson Daily News
<https://carsondailynews.com/using-artificial-intelligence-to-cope-with-extreme-weather-events-newsroom/>
- Florida News Times
<https://floridanewstimes.com/manage-extreme-weather-using-artificial-intelligence/98547>

4.5 Highly qualified personnel

To fulfill the grant and contract, an interdisciplinary team was assembled. It was led by Professor Renee Sieber at McGill University. Nine students and fellows were cross-trained in crisis management and artificial intelligence related to issues of concern to ECCC. Also of note: this was a diverse team. It was led by a woman. Of the 10 members (including the PI), half were women. The team included an Afro-Canadian and an Asian member.

Postdoctoral Fellow

Drew Bush, Geography, McGill University

PhD Students

Mikael Brunila, Geography, McGill University

Masters Graduate Students

Andrei Mircea, Computer Science, McGill University

Rosie Zhao, Computer Science, McGill University

Sam Lumley, Geography, McGill University

Lucia Berger, MILA

Hannah Ker, Data Science, University College London

Undergraduate Students

Stefan Morgan, Geography, McGill University

Sarah Greenidge, Computer Science, McGill University