Narratives and Needs: Analyzing Experiences of Cyclone Amphan Using Twitter Discourse

Ancil Crayton[1] , **João Fonseca**[3] , **Kanav Mehra**[5] , **Michelle Ng**[2] , **Jared Ross**[1] , **Marcelo Sandoval-Castañeda**[4] and **Rachel von Gnechten**[2]
[1]Booz Allen Hamilton, [2]International Water Management Institute, [3]NOVA Information Management School (NOVA IMS), [4]New York University Abu Dhabi, [5]Independent Researcher
ancil.crayton@ucdconnect.ie, jpfonseca@novaims.unl.pt, {jaredrossj, kanav.mehra6}@gmail.com, {m.ng, r.vongnechten}@cgiar.org, marcelo.sc@nyu.edu

Abstract

People often turn to social media to comment upon and share information about major global events. Accordingly, social media is receiving increasing attention as a rich data source for understanding people's social, political and economic experiences of extreme weather events. In this paper, we contribute two novel methodologies that leverage Twitter discourse to characterize narratives and identify unmet needs in response to Cyclone Amphan, which affected 18 million people in May 2020.

1 Introduction

With wind speeds gusting up to 200 kilometres per hour, Cyclone Amphan was the first super cyclone to form in the Bay of Bengal since 1999 [Hansen, 2020]. It made landfall in West Bengal, India on May 20, 2020 before tracing a destructive path northward to Bangladesh [Beer, 2020]. Along the way, Cyclone Amphan damaged nearly 3 million houses, 18,000 square kilometres of agricultural lands and 449,000 electric poles, leaving 18 million affected people in its wake [of Red Cross and Societies, 2020; of West Bengal, 2020]. It was the costliest cyclone in the history of the North Indian Ocean with a reported 13.2 billion USD in damages in the state of West Bengal alone [of West Bengal, 2020].

Extreme weather events are expected to increase in magnitude and frequency due to the impacts of climate change; Cyclone Amphan is one such example. The Bay of Bengal's unprecedentedly high sea surface temperature, which is linked to anthropogenic climate change, likely contributed to Cyclone Amphan's speed and energy [Mukherji, 2020]. Unfortunately, warming ocean temperatures will intensify more cyclones and hurricanes in the future — both in the Bay of Bengal and beyond. Thus, it is imperative to develop and refine approaches for responding to extreme weather events that draw upon all available tools.

In the case of Cyclone Amphan, response efforts were complicated by COVID-19. For instance, in addition to the typical heavy rains and obstructed roads, responders had to cope with restricted mobility due to India's nationwide lockdown, limitations on shelter capacities due to social distancing measures and the need to obtain, use and distribute personal protective equipment [of West Bengal, 2020; of Red Cross and Societies, 2020]. On-the-ground response efforts by governments, disaster relief organizations and civil society are crucial and life-saving after extreme weather events. Could online data serve as an additional tool to supplement on-the-ground efforts, particularly when they are hindered? Although online data cannot paint a complete or representative picture of offline realities, it could help fill knowledge gaps when there are challenges reaching affected people, such as those caused by COVID-19.

We took Cyclone Amphan as our use case in exploring the potential for Twitter content to target relief efforts in response to extreme weather events. We first aimed to characterize how collective knowledge about Cyclone Amphan was produced on Twitter. Considering Twitter is a decentralized microblogging platform, anyone can add their commentary to an issue, influence its narrative and build new layers of interpretation for on-the-ground realities.

In this paper, we harness Twitter data to inform and assist on-the-ground relief efforts. Specifically, we propose two methodologies to (1) identify who and what is shaping the narratives around Cyclone Amphan and (2) identify unmet needs of people affected by Cyclone Amphan.

2 Natural Disasters and Social Media

Social media platforms serve as massive repositories of real-time situational and actionable data during man-made or natural emergencies, such as extreme weather events. For instance, people can use social media to organize volunteer or donation campaigns in support of on-the-ground relief efforts or directly contact relevant organizations via their official social media accounts [Imran *et al.*, 2015]. Consequently, social media posts regarding extreme weather events vary broadly, from people sharing personal experiences and opinions, to emergency response agencies posting updates, warnings and information about relief efforts.

Social media content can be used to characterize people's experiences of extreme weather events and trace how narratives took shape through collective knowledge production. Several existing studies focus on developing efficient and scalable methods for extracting important, actionable information from social media content using a range of techniques based on natural language processing (NLP), text mining and network analysis [Imran *et al.*, 2013b; Imran *et al.*, 2013a;

Imran *et al.*, 2014]. A large volume of work in this domain focuses on developing classifiers to categorize tweets as informative or uninformative using learning-based approaches [Alam *et al.*, 2018; Hernandez-Suarez *et al.*, 2019; Zhang and Vucetic, 2016] or matching-based approaches [To *et al.*, 2017; Mehra and Chandra, 2017]. Due to the unavailability of large labeled datasets, unsupervised or semi-supervised learning methods [Alam *et al.*, 2018; Zhang and Vucetic, 2016; Arachie *et al.*, 2020] are preferred over supervised learning based approaches, which tend to not generalize well across different crisis events. To the best of our knowledge, we provide the first usage of zero-shot text classification to assign tweets to multiple classes relevant to the impacts of extreme weather events. This method not only utilizes the abundantly available unlabeled Twitter data (tweets, or microblogs) and circumvents the need for supervised learning, but also facilitates the application of a pre-trained language model that makes the overall method generalizable across different crisis events.

Another active direction of research is the identification of important sub-events within a large-scale extreme weather event for efficient management of relief efforts [Rudra *et al.*, 2018]. Recent work proposes an unsupervised learning-based framework using semantic embeddings of noun-verb pairs from tweets to detect sub-events [Arachie *et al.*, 2020]. Both extractive and abstractive methods of text summarization have been studied to effectively summarize the huge volumes of microblogs posted during emergency events [Rudra *et al.*, 2015; Rudra *et al.*, 2016; Mehra and Chandra, 2017; Dutta *et al.*, 2018]. Researchers have also explored methods to classify microblogs posted during natural disasters by sentiment and information utility [Ragini *et al.*, 2018; Zhang and Vucetic, 2016]. However, our paper aims to develop a more holistic analysis of social media content for disaster management. We combine methods of sentiment analysis and classify tweets into multiple relevant classes based on the information they convey. Besides adding more context to the tweets as well as the classes they are placed in, our methodology explores the emotional spectrum associated with the classes.

While the content analysis of microblogs during emergency events has been extensively discussed, the analysis of social network structure along with patterns of user behaviour on Twitter during extreme weather events is still relatively unexplored [Pourebrahim *et al.*, 2019]. In this paper, we intend to address this gap by studying the embedded social network structure and comparing user behavior across different sets of users to clearly distinguish between individuals who shaped the dominant narrative and those who were marginalized.

3 Approach

3.1 Dataset

We extracted around 470,000 tweets using the Twitter API. We targeted tweets from May 1st, 2020 to June 15th, 2020, which cover the build-up through the aftermath of Cyclone Amphan. The main languages targeted in our query were English, Odia, Hindi and Bengali, but some tweets in other languages were extracted as well. We also filtered out terms related to other catastrophes happening in the area during the same time period, such as Cyclone Nisarga, which hit the Indian subcontinent at the beginning of June.

3.2 Preprocessing data

Given the multi-language approach of our query, the first step in the preprocessing pipeline is to translate non-English tweets into English using the Google Translate API. The Twitter API identifies and tags the language of each Tweet. We take advantage of this attribute such that only non-English tweets are translated, thus reducing computational costs.

The next step is to remove URLs and reserved words from the content of the tweets. This includes hashtags, emojis and words like 'RT' or 'FAV'. Then we remove all remaining punctuation and change any uppercase letters to lowercase. As a final step, we remove all stop words found in the text of the tweets, following NLTK's stop words list for English [Loper and Bird, 2002], and lemmatize the remaining words using NLTK's WordNetLemmatizer.

3.3 Feature Extraction

The next goal is to extract information from the Twitter data. This process is divided into 4 independent tasks:

Sentiment analysis

The extraction of data information regarding the writer's sentiment is a common, well-studied NLP task. In this work, the sentiment of each tweet is not known beforehand. Therefore, we leverage the Valence Aware Dictionary and Sentiment Reasoner (VADER) model [Hutto and Gilbert, 2014] to capture the sentiment of each tweet. VADER determines the sentiment of a document through a rule-based approach using a sentiment lexicon (i.e., a list of lexical features labelled according to their semantic orientation) to determine how positive/negative a specific tweet is [Hutto and Gilbert, 2014]. It is advantageous in the context of this paper as it was developed for social media text.

Point-of-view extraction

We use point-of-view extraction to classify tweets as first person, second person or third person, with the goal of determining whether the user experienced Cyclone Amphan personally. It is done by iterating over all tokens in the tweet's text and identifying any word matching a list of pronouns mapped to first, second or third person speech. Examples of first person speech could be tokens such as 'I', 'my' and 'our'; for second person, this may include 'you' or 'your'; and, finally, for third person, it may include 'them', 'they' or 'it'.

Our classification strategy is to assign a tweet to be first, second or third person based on the order of precedence, respectively. Therefore, if the tweet contains any first person pronouns, it is designated to be first person point of view. If the tweet contains second person pronouns, but does not contain any first person pronouns, it is assigned to be second person. Finally, if the tweet contains third person pronouns, but does not contain second or first person pronouns, it is classified as third person.

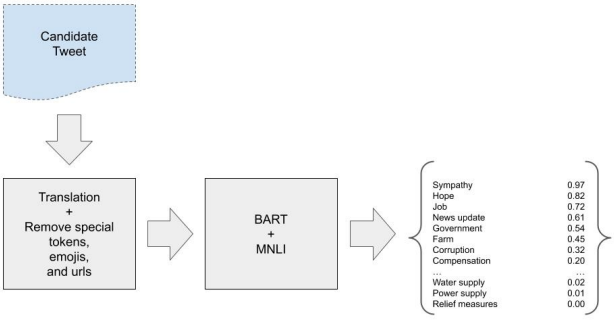

Figure 1: Zero-shot text classification pipeline.

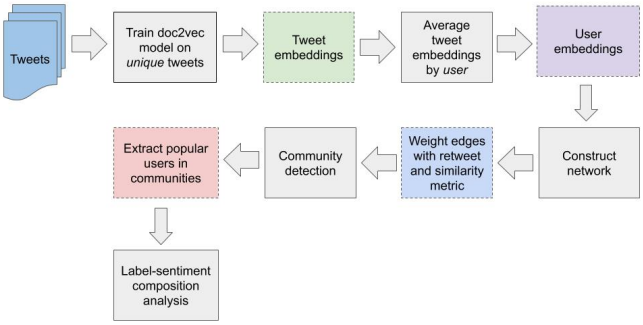

Figure 2: Pipeline for identifying the narrative and who's shaping it.

Zero-shot text classification

In order to extract critical information and derive actionable insights from large volumes of social media content generated during extreme weather events, it is imperative to effectively categorise the microblogs into distinct classes.

Due to the unavailability of large annotated training sets, we present a novel application of zero-shot text classification for a multi-label classification of tweets that is not only scalable but also generalizable across different crisis events. Researchers have proposed an approach of using a pre-trained Natural Language Inference (NLI) sequence-pair classifier as a zero-shot text classifier [Yin *et al.*, 2019]. The model considers a sequence input (tweet) as the premise and each candidate topic label as a hypothesis. For our purpose, we use the zero-shot classification pipeline implementation available in the Transformers package that uses a large BART model [Lewis *et al.*, 2019] pre-trained on the MNLI dataset [Williams *et al.*, 2018]. This pipeline is shown in Figure 1.

After experimenting with different combinations, we settle on a comprehensive set of 24 specific labels that cover a wide variety of information: {'sympathy', 'criticism', 'hope', 'job', 'relief measures', 'compensation', 'evacuation', 'ecosystem', 'government', 'corruption', 'news updates', 'volunteers', 'donation', 'cellular network', 'housing', 'farm', 'utilities', 'water supply', 'power supply', 'food supply', 'medical assistance', 'coronavirus', 'petition', 'poverty'}.

The entire set of unique tweets is fed to the zero-shot classifier after basic preprocessing steps outlined in section 3.2. However, to preserve more text and conform the data closer to the training data of the BART model, we omit the case folding, stopword removal and lemmatization steps for this task. The classifier yields a confidence score ranging between 0 and 1 for every tweet-label pair. To ensure minimum overlap and maintain exclusivity, each tweet is assigned to a topic label if the confidence score associated with the pair is above a certain threshold, say α. We experiment with a set of values for α and observe the best results with $\alpha = 0.7$. We discuss validation/hyperparamter tuning for α in Appendix B.

User and Tweet embeddings

Embeddings are vector representations of either words, documents (tweets) or a set of documents (the user). They allow the conversion of non-numerical data (text) into a n-dimensional space, where the relationships among words, tweets and/or users is preserved. There are many methods directed toward vector representation of words. Amongst the most popular methods is one-hot encoded representations, distributed representations, Singular Value Decomposition, continuous bag of words and skip-gram model.

Tweet vectorization is done using a skip-gram model known as the Doc2Vec algorithm [Le and Mikolov, 2014]. This choice was motivated by not only its popularity and computational efficiency, but also its capacity to maintain a logical spatial structure among tweets, both regarding the tweets' corpus and their underlying topic and sentiment. The Doc2Vec model is trained using unique tweets and replies in order to avoid the bias toward highly retweeted tweets that would come from keeping duplicate text. This results in a model trained on approximately 113,000 documents over 50 epochs. Tweets containing rare words in the dataset's corpus (i.e., words appearing twice or less) are rejected for training. The output are 200-dimensional embeddings for each tweet in our dataset.

The user embeddings are based on the tweet embeddings, with a process that involves averaging the tweets/retweets belonging to a given user [Hallac *et al.*, 2019]. This allows the analysis of the type of discourse and opinion shared among users in a 200-dimensional space.

3.4 Analysis

In this section we address our research questions by combining the preprocessed data and extracted features [1].

Identifying narratives and influential users

The pipeline for identifying narratives and influential users in the dataset is shown in Figure 2. We address this question through the usage of user vectors as described in subsection 3.3 as a means of positioning users in a two dimensional space. The projection of the 200-dimensional embeddings was done using t-SNE [Van Der Maaten and Hinton, 2008], resulting into 2-dimensional coordinates used to position each user (i.e., nodes) in the network graph. The network's edges are assigned based on the number of retweets and/or replies among users, which are then weighted by dividing this number with the users' euclidean distance (using the original 200-dimensional embeddings). These users can now be grouped into different communities using two different methods: 1)

[1] All functions, algorithms, experiments, and results reported are provided at the Github repository of the project: https://github.com/ancilcrayton/solve-iwmi.

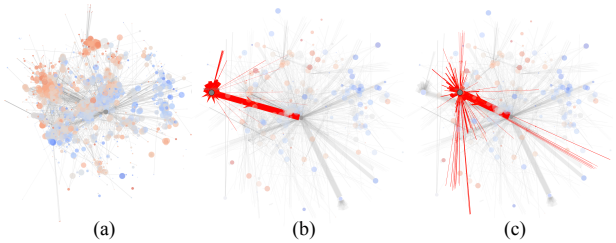

Figure 3: Twitter's user network, based on tweets related to Cyclone Amphan. Legend: (a) Sampled user network, node size varies according to the number of tweets related to Amphan, (b) with influential user Priyanka Gandhi Vadra highlighted, (c) influential user Narendra Modi highlighted. In both (b) and (c) node size varies according to the number of followers. All figures are sampled to 2000 edges.

discourse-based, where the clustering is done on the embedding features and 2) community-based, done through network clustering methods. For both clustering methods, the most popular users within each cluster are identified based on centrality measures and the number of followers the user has. Lastly, we analyze these users' discourse based on labels and average sentiment associated to these users.

Figure 3 depicts different visualizations of a sample of 2000 edges of the user network. In all the different visualizations, the color reresents the user's mean sentiment, where red and blue nodes are users with positive and negative mean sentiment scores, respectively. Figure 3a shows the network with node sizes proportional to the number of tweets related to cyclone Amphan. The analysis of the resulting network revealed the consistency of the user embeddings with their mean sentiment score. Users with a positive mean sentiment score tend to be on the top-left region of the graph, whereas users with a negative mean sentiment score tend to be on the opposite region. Specifically, the analysis of the network based on the number of followers distinguished key players in the definition of the narratives within the network, such as political users and news agencies. Specifically, Priyanka Gandhi Vadra (shown in Figure 3b), shows a similar type of discourse to the one of Narendra Modi (shown in Figure 3c), but with an opposite sentiment score (Priyanka Gandhi has a mean sentiment score -0.59, whereas Narendra Modi has 0.43). The discourse types followed by the two users are reflected based on their spatial proximity in the network. Priyanka Gandhi Vadra is the general secretary of the All India Congress Committee, an opposition party to the one of Narendra Modi, Bharatiya Janata Party.

Identifying negative experiences and unmet needs

The pipeline for identifying negative experiences and unmet needs is outlined in Figure 5. The first step is to identify the topics discussed in each tweet by assigning labels via zero-shot text classification. In order to only analyze tweets from users who were personally affected by Cyclone Amphan, the data is then filtered to include only first person tweets using the point-of-view analysis. At this point, we determine to which topic each affected individual is most sensitive. The sentiment analysis results enables understanding of whether

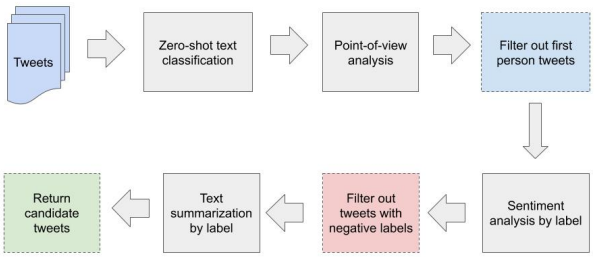

Figure 4: Pipeline for identifying negative experiences and unmet needs.

the individual's view is either positive or negative. [2]

We narrow down our focus to the labels that have a negative median sentiment with the assumption that negative experiences are more likely to suggest unmet needs. We then report representative tweets using extractive summarization techniques to identify the dominant themes within each label.

Our summarization method expands upon recent work that proposes an unsupervised graph-based summarization algorithm specifically for microblogs using a tweet-similarity graph over the tweet vectors generated from Doc2Vec embeddings [Dutta *et al.*, 2018]. We present a slight modification by choosing representative tweets returned from the resulting connected components based on a (maximum) score, which takes into account centrality and tweet significance as follows:

$$Score = C + S - Si, \tag{1}$$

where C represents the degree centrality, Si is the sentiment score of the tweet, and S is the node size that is calculated as $S = log(tweet\_length)$ where $tweet\_length$ is the number of tokens in the tweet. We subtract the sentiment score to prioritize tweets with negative sentiment values.

We create K-length summaries for labels with negative median sentiment scores, where K represents the number of representative tweets resulting from the text summarization method. The exact number to be considered can be adjusted by the researcher, policymaker or relief organization interested in learning about the experiences.

A sample of our initial results are reported in Table 1. These four tweets are selected through the text summarization algorithm on first-person tweets over labels that have negative median sentiment scores - 'housing', 'farm', and 'criticism'. Our approach allows us to identify cases of home damage, including flooding and major destruction specifically in the Siddha Galaxia Oceania and Khejuri II blocks of Kolkata, India, respectively. We discuss our initial plan to validate this methodology in Appendix B.

4 Conclusion

In this paper, we contribute two methodologies for analyzing Twitter content to characterize experiences of Cyclone Amphan, including the identification of unmet needs and the

[2]Refer to Appendix A for more information on the distribution of sentiment scores per label.

Full Text	Label	Sentiment	Location
@siddhagroup Plz don't fool people. We r residents of Siddha Galaxia Oceania block. We r suffering from poor quality windows, bedrooms of residents flooded during Amphan cyclone. Lifts are not working since Amphan cyclone. No update from Siddha when the lifts will be repaired. Shame on u.	Housing	-0.8589	Kolkata, India
This year, we will undergo huge crisis of food n income bcoz of Rabi crop loss n delay of kharif crop bcoz of Cyclone & Covid 19. We request @CMO_Odisha @Food_Odisha @stscdev to support us in this distress n provide crop insurance, food grains n crop storage @MoSarkar_Odisha	Farm	-0.7964	Odisha, India
@TimesNow After amphan the situation in Kolkata: No power for five days in the Tollygunge area. TMC goons beating up women for protesting. The councellor of ward 96 making fake promises. Please telecast this in your channel. We want a response.	Criticism	-0.6124	Kolkata, India
@PMOIndia @narendramodi Khejuri Block II in East Medinipur District,West Bengal is completely deatroyed caused to Amphan Cyclone Yesterday. Almost 250 Homes has been destroyed completely. I request to all Administrator to look into this area so that Khejuri Block II gets Proper help this time atleast. https://t.co/rcDUnf552W	Housing	-0.6326	Kolkata, India

Table 1: Sample of tweets from the negative experiences and unmet needs analysis. These results are from text summarization algorithm on first-person tweets over labels that have negative median sentiment scores - 'housing', 'farm', and 'criticism'. We set $K = 50$.

collective production of narratives. We recognize that social media does not provide a complete or representative picture of extreme weather events, especially in low-resource environments where people may not have access to technology. For instance, in 2019, Internet penetration in West Bengal, India was at 29%; of rural Internet users in India, 72% were male [Internet and of India, 2019]. While these methodologies should not be used alone, they can supplement existing on-the-ground efforts, particularly when in-person needs assessments are hindered, e.g. due to COVID-19. We anticipate these methodologies to be helpful for policy-makers, disaster relief organizations, researchers and members of civil society who wish to leverage an additional tool to better understand the impacts of extreme weather events and better focus their efforts. This will become increasingly important as climate change amplifies the magnitude and frequency of extreme weather events.

Acknowledgements

This project was completed as part of the Data Science for Social Good (DSSG) Solve for Good program. We would like to thank Andrew Bell, Rayid Ghani, and Jessica Toth–members of the DSSG Solve for Good team–for their assistance during this project. We would also like to thank Simon Langan at the International Water Management Institute for his support of and feedback on our work. Finally, we thank the International Water Management Institute for graciously funding data collection from the Twitter API.

Disclaimer

The views expressed in this paper are those of the authors and not of their affiliations.

References

[Alam *et al.*, 2018] Firoj Alam, Shafiq Joty, and Muhammad Imran. Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. In *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 2018.

[Arachie *et al.*, 2020] Chidubem Arachie, Manas Gaur, Sam Anzaroot, William Groves, Ke Zhang, and Alejandro Jaimes. Unsupervised Detection of Sub-Events in Large Scale Disasters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):354–361, 2020.

[Beer, 2020] Tommy Beer. 500,000 families may be homeless due to devastation from cyclone amphan. https://www.forbes.com/sites/tommybeer/2020/05/22/500000-families-may-be-homeless-due-to-devastation-from-cyclone-amphan/, 2020. Section: Business.

[Dutta *et al.*, 2018] Soumi Dutta, Vibhash Chandra, Kanav Mehra, Asit Kumar Das, Tanmoy Chakraborty, and Saptarshi Ghosh. Ensemble Algorithms for Microblog Summarization. *IEEE Intelligent Systems*, 2018.

[Hallac *et al.*, 2019] Ibrahim R. Hallac, Semiha Makinist, Betul Ay, and Galip Aydin. User2Vec: Social Media User Representation Based on Distributed Document Embeddings. In *2019 International Conference on Artificial Intelligence and Data Processing Symposium, IDAP 2019*. Institute of Electrical and Electronics Engineers Inc., sep 2019.

[Hansen, 2020] Kathryn Hansen. Amphan batters india, bangladesh. https://earthobservatory.nasa.gov/images/146749/amphan-batters-india-bangladesh, May 2020.

[Hernandez-Suarez *et al.*, 2019] Aldo Hernandez-Suarez, Gabriel Sanchez-Perez, Karina Toscano-Medina, Hector Perez-Meana, Jose Portillo-Portillo, Victor Sanchez, and Luis Javier García Villalba. Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. *Sensors (Switzerland)*, 2019.

[Hutto and Gilbert, 2014] C. J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 2014.

[Imran *et al.*, 2013a] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Extracting information nuggets from disaster- Related messages in social media. In *ISCRAM 2013 Conference Proceedings - 10th International Conference on Information Systems for Crisis Response and Management*, 2013.

[Imran *et al.*, 2013b] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. Practical extraction of disaster-relevant information from social media. In *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[Imran *et al.*, 2014] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. AIDR: Artificial intelligence for disaster response. In *WWW 2014 Companion - Proceedings of the 23rd International Conference on World Wide Web*, 2014.

[Imran *et al.*, 2015] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in Mass Emergency: A survey. *ACM Computing Surveys*, 2015.

[Internet and of India, 2019] Internet and Mobile Association of India. India internet 2019, 2019.

[Le and Mikolov, 2014] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *31st International Conference on Machine Learning, ICML 2014*, 2014.

[Lewis *et al.*, 2019] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[Loper and Bird, 2002] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

[Mehra and Chandra, 2017] Kanav Mehra and Vibhash Chandra. Summarizing microblogs for emergency relief and preparedness. In *CEUR Workshop Proceedings*, 2017.

[Mukherji, 2020] Aditi Mukherji. Cyclone amphan and covid-19: the recipe for a cascading disaster. http://www.iwmi.cgiar.org/2020/06/cyclone-amphan-and-covid-19-the-recipe-for-a-cascading-disaster/, June 2020.

[of Red Cross and Societies, 2020] International Federation of Red Cross and Red Crescent Societies. Operation update report india: Cyclone amphan. https://reliefweb.int/report/india/india-cyclone-amphan-operation-update-report-dref-n-mdrin025, July 2020.

[of West Bengal, 2020] State Inter Agency Group of West Bengal. Joint rapid need assessment report on cyclone amphan, June 2020.

[Pourebrahim *et al.*, 2019] Nastaran Pourebrahim, Selima Sultana, John Edwards, Amanda Gochanour, and Somya Mohanty. Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. *International Journal of Disaster Risk Reduction*, 37(May):101176, 2019.

[Ragini *et al.*, 2018] J. Rexiline Ragini, P. M.Rubesh Anand, and Vidhyacharan Bhaskar. Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, 2018.

[Rudra *et al.*, 2015] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. Extracting situational information from microblogs during disaster events: A classification-summarization approach. In *International Conference on Information and Knowledge Management, Proceedings*, 2015.

[Rudra *et al.*, 2016] Koustav Rudra, Siddhartha Banerjee, Niloy Ganguly, Pawan Goyal, Muhammad Imran, and Prasenjit Mitra. Summarizing situational tweets in crisis scenario. In *HT 2016 - Proceedings of the 27th ACM Conference on Hypertext and Social Media*, 2016.

[Rudra *et al.*, 2018] Koustav Rudra, Pawan Goyal, Niloy Ganguly, Prasenjit Mitra, and Muhammad Imran. Identifying sub-events and summarizing disaster-related information from microblogs. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, pages 265–274, 2018.

[To *et al.*, 2017] Hien To, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi. On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based. *Proceedings - 2017 IEEE 3rd International Conference on Multimedia Big Data, BigMM 2017*, pages 330–337, 2017.

[Van Der Maaten and Hinton, 2008] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

[Williams *et al.*, 2018] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018.

[Yin *et al.*, 2019] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach, 2019.

[Zhang and Vucetic, 2016] Shanshan Zhang and Slobodan Vucetic. Semi-supervised Discovery of Informative Tweets During the Emerging Disasters. 2016.

A Unmet Needs

A.1 Sentiment Scores by Label

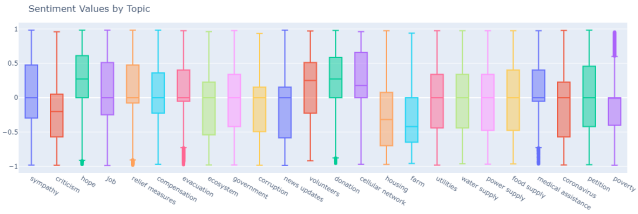

Figure 5: Box Plot for Sentiment Scores by Label.

Tweets are assigned labels via zero-shot text classification, as explained in Section 3.3. We group tweets by labels and analyze the distribution of sentiment scores for each label. Accordingly, Figure 5 reveals that our method successfully captured the general sentiment associated with each label. Labels that usually depict an optimistic outlook exhibit a trend of positive sentiment scores as seen in - ”hope”, ”relief measures”, ”compensation”, ”sympathy”, ”volunteers”, and ”donation”. Similarly, the analysis also highlights labels with negative median sentiment scores - ”criticism”, ”farm”, and ”housing”. We focus on these labels to identify unmet needs and negative experiences. It is interesting to observe that labels - ”ecosystem”, ”corruption”, ”news updates”, and ”poverty” exhibit strong negative trends, highlighting the damaging impact of the cyclone.

B Validation of Results

As this is currently a work in progress, we plan to further validate our results. Our pipeline for identifying negative experiences and unmet needs contains many components that

could use independent validation, however we identify two key components that we would like to prioritize in our validation procedure: zero-shot text classification and returning candidate tweets.

B.1 Validating Zero-Shot Text Classification Model

The output of our pipeline to identify unmet needs and negative experiences depends heavily on the labeling of the tweets using a zero shot text classification approach [Yin *et al.*, 2019]. As the model returns a confidence score for each label between 0 and 1, we must select a threshold α above which we assign the tweet to the relevant label. But how do you validate the results of essentially 'predicting' a label when there is no ground truth label? Our planned solution to this is to leverage a crowdsourcing platform, such as AWS Mechanical Turk.

We will leverage a crowdsourcing platform to compute human confidence scores that can be used as ground truth values compared to the model's confidence scores. By computing a measure of loss using a suitable loss function, we can find the value of α that minimizes this loss and select that as our threshold value. This will allow us to select more tweets that are closer to human intuition. A sketch of the steps to this process are as follows:

1. Sample a subset of k tweets from the corpus of n tweets
2. Allow m users to select which labels in the label set L characterize each tweet (binary - 1 if label characterizes tweet, else 0)
3. Compute confidence scores for each label by computing the proportion of 1's
4. Get confidence score for each label $l \in L$ for all k tweets using zero-shot text classification model
5. Calculate loss using a relevant loss function (such as least squares) for each label
6. Select the threshold value α that minimizes the average loss over all L labels

We hope to release the results of this procedure once the validation process is completed.

B.2 Evaluating Output of Negative Experiences and Unmet Needs Pipeline

In our evaluation process, it is useful to characterize how useful the tweets are in mitigating disaster relief. Unfortunately, this is a complex task, which requires concretely defining what is meant by 'useful'. On top of this, it is important to define an audience for whom the results would be useful for. Although this is a work in progress, we have preliminary developed a procedure for evaluating these results.

Audience

In our evaluation procedure, we hope the results would be useful for disaster relief and/or humanitarian institutions, such as The Red Cross or United Nations Development Programme (UNDP).

Defining 'Useful'

We believe that the output of the pipeline would be considered useful if the candidate tweets returned are *actionable*. We consider a candidate tweet to be *actionable* if it meets the following two criteria:

(i) Conveys damage or unmet need following relevant catastrophe

(ii) Contains location for action

Criterion (i) is important for the institution to identify the help needed and what resources are available to meet this need. Criterion (ii) is needed for determining where to allocate the resources, determining local resources available, and developing a strategy for on-the-ground relief efforts.

Evaluation Procedure

In this case, it is difficult to develop a statistical metric that measures whether a candidate tweet is useful based on the criteria defined in the previous subsection. Therefore, we believe crowdsourcing would be useful in this case as well. However, contrary to the zero shot text classification, we would limit the 'crowd' to being those working at research-for-development, humanitarian, and disaster relief institutions, such as the International Water Management Institute (IWMI), The Red Cross, and UNDP.

The evaluation process could follow the steps below:

1. Return k tweets for each label l in the label set L with a negative median sentiment for first person tweets
2. Allow m evaluators to select if the tweet conveys damage or an unmet need and whether the location is available or can be inferred by the text (1 if yes, else 0)
3. Compute confidence score for each criterion in each label by computing the proportion of 1's, yielding $P_i \in [0, 1]$ for $i \in \{unmet, location\}$.
4. Calculate the harmonic mean of P_{unmet} and $P_{location}$, yielding $P_{harmonic} = 2 * \frac{(P_{unmet} * P_{location})}{P_{unmet} + P_{location}}$.
5. Use $P_{harmonic}$ as the metric to establish a benchmark and compare future model iterations

We also believe the scale of this evaluation procedure (controlled by the m) could be much smaller than in the previous case as utilizing more domain experts would likely give a lower variance in resulting scores. The choice of k is arbitrary but will depend on the institution's capacity.