Towards an Interpretable Approach to Classify and Summarize Crisis Events from Microblogs

Thi Huyen Nguyen
L3S Research Center
Hanover, Germany
nguyen@l3s.de

Koustav Rudra*
Indian Institute of Technology
(Indian School of Mines)
Dhanbad, India
koustav@iitism.ac.in

ABSTRACT

Microblogging platforms like Twitter have been heavily leveraged to report and exchange information about natural disasters. The real-time data on these sites is highly helpful in gaining situational awareness and planning aid efforts. However, disaster-related messages are immersed in a high volume of irrelevant information. The situational data of disaster events also vary greatly in terms of information types ranging from general situational awareness (caution, infrastructure damage, casualties) to individual needs or not related to the crisis. It thus requires efficient methods to handle data overload and prioritize various types of information. This paper proposes an interpretable classification-summarization framework that first classifies tweets into different disaster-related categories and then summarizes those tweets. Unlike existing work, our classification model can provide explanations or rationales for its decisions. In the summarization phase, we employ an Integer Linear Programming (ILP) based optimization technique along with the help of rationales to generate summaries of event categories. Extensive evaluation on large-scale disaster events shows (a). our model can classify tweets into disaster-related categories with an 85% Macro F1 score and high interpretability (b). the summarizer achieves (5-25%) improvement in terms of ROUGE-1 F-score over most state-of-the-art approaches.

CCS CONCEPTS

• **Information systems → Clustering and classification**; **Summarization**.

KEYWORDS

Classification, Summarization, Interpretability, Crisis Events

ACM Reference Format:
Thi Huyen Nguyen and Koustav Rudra. 2022. Towards an Interpretable Approach to Classify and Summarize Crisis Events from Microblogs. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3485447.3512259

*Research was primarily conducted while affiliated to L3S Research Center.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3512259

1 INTRODUCTION

Crisis events work as a trigger for a large volume of real-time information over social media such as Twitter. Local people and authorities post a lot of updates from the ground. Some previous studies [3, 43, 44] have shown the vital role of the Twitter resource in enhancing emergency situational awareness and planning aids. However, in disaster situations, crisis-related messages are immersed in massive sentimental and irrelevant tweets. During disaster events, humanitarian organizations may want to obtain information in multiple categories, such as infrastructure and utilities damage, caution and advice, injured and dead people, etc. Besides, Twitter users also want to quickly get brief information about the events without being overwhelmed with massive data. To fulfill the needs of these organizations and effectively cope with large-scale disasters, it is necessary to develop automated methods to classify tweets into different humanitarian categories and then summarize those tweets in real-time.

All existing crisis-specific classification and summarization approaches primarily focus on performance measures, but they did not pay any attention to their decision-making processes. However, such critical systems need to be interpretable in nature [32–34] so that decision-makers can use them for the purpose. Besides, in many applications, users prefer simple models with high interpretability. It, therefore, brings to forefront the trade-off between accuracy and interpretability of a model. Despite advances in Natural Language Processing [6] and interpretable Deep Learning models [7, 32, 34, 48], interpreting classification of short, noisy tweets has not been explored. In this work, we aim for a classification model in crisis domain to be interpretable by design. We observe that there are short snippets in tweets, so-called explanations/rationales[1] [7], which provide sufficient evidence to support classification outputs. For example, "*03 Dec 2012 – At least 475 people are killed after Typhoon Bopha, makes landfall in the Philippines*", the phrase "*At least 475 people are killed*" captures essential and sufficient information to classify the tweet to a category about injuries and death. Furthermore, we show that the use of rationales helps improve summarization results of crisis events.

This paper presents an interpretable classification and summarization framework to classify and summarize tweets during disaster events. In classification phase, we develop a crisis-related microblog classifier based on the idea proposed by Zhang et al. [48]. First, we extract rationales based on a BERT-based multi-task learning approach [4]. Then, the extracted rationales are used to predict class labels of tweets. Our model is interpretable by design, which

[1]These two terms are used interchangeably throughout the paper.

is transparent to users about the interpretability of predicted rationales. In the summarization phase, the categorized tweets and rationales are used as the input of an Integer Linear Programming (ILP) framework to summarize tweets. Our summarizer optimizes multiple criteria with flexible constraints, which aim to satisfy different needs of end-users. Experiments on two long-ranging natural disaster events show that our multi-task learning approach achieves high classification performance along with high-quality rationales for the model decisions. Besides, the proposed summarization method surpasses various state-of-the-art baselines in terms of ROUGE-1 F-score and informativeness with human judgment. To the best of our knowledge, this is the first study on interpretable classification-summarization approach on crisis-related microblogs.

The major contributions in this work are listed below:

- We provide the first human annotations of "rationales" on two crisis datasets. The datasets will be shared with research community.
- Using the annotated data, we develop a classification and summarization framework[2], which is interpretable in classification decisions and makes use of extracted rationale information to generate summaries of disaster events in near real-time.
- Our classification experiments indicate that our classification method is interpretable by design with about 83% Token-F1 score on rationale extraction task and high classification results on crisis datasets.
- Extensive summarization experiments show the superior performance of our summarization model compared to various baselines. The generated summaries have 5-25% higher ROUGE-1 F-score than baseline methods and are more informative in terms of human evaluation.

2 RELATED WORK

This section briefly reviews prior works on tweet classification and summarization closely related to ours.

Tweet classification during disaster events: Classification of disaster events has attracted great attention of the research community [1, 17, 18, 21, 24, 27, 36, 43, 44]. Approaches range from traditional supervised classification methods such as Support Vector Machine (SVM), Naïve Bayes (NB) to recent deep learning and embedding-based models. Vermal et at. [44] employed bag-of-words classification models to automatically detect messages that may contribute to situational awareness. Later, Rudra et al. [36] introduced low-level lexical and syntactic features to classify tweets. Nguyen et. at [27] proposed a neural network-based approach for the same. These studies mainly apply binary classification methods to identify situational tweets. Imran et al. [10] proposed AIDR to classify situational tweets into multiple classes such as 'infrastructure', 'injured people', 'missing people', etc. Generally, the prior works only focus on improving classification performance without providing explanations of models' decisions. In recent times, interpretability of the models has become a requirement, and researchers proposed various approaches [7, 13, 39, 45, 48, 50] to address this requirement. Inspired by these works, our paper aims to both classify tweets at fine-grained levels of information types and provide *human-understandable explanation* of classification decisions in disaster domain.

Tweet Summarization: In times of disaster events, it is essential to summarize tweets timely so that government authorities can grasp the situation promptly for rapid responses and assistance. Several approaches for real-time summarization of tweet streams have been proposed [29–31, 42, 51]. Olariu [30] introduced a tri-gram graph model to generate abstractive summaries of Twitter events incrementally. Nguyen et al. [29] proposed a diversified ranking algorithm on a graph to represent tweets, detect sub-events and then produce extractive summarization of evolving events from tweet streams. Besides, a few works have attempted to generate summaries of disaster events [15, 16, 28, 35, 36, 38, 40]. Kedzie et al. [16] presented an extractive summarization system that predicts sentence salience and then uses a clustering algorithm to select updates for disaster events. However, the paper focuses on well-written news articles of disasters instead of short, noisy Twitter texts. Rudra et al. [36] proposed a real-time extractive summarization technique for disaster events, yet only focused on general summaries rather than class-level summaries. Later, the authors employed the AIDR platform [10] to classify tweets into different humanitarian classes and then introduced an extractive summarization method for class-level summarization of disasters [37, 38].

A few works [14, 22, 47, 49] have shown the great potential of recent pre-trained models in summarization tasks. However, these studies consider the specific traits of news articles to design summarization models. Besides, some recent proposed BERT-based summarization models have constraints on the length of input texts (i.e., number of sentences in input documents) and computation time, so it is not effective and robust for disaster situations with millions of input tweets. Side by side, tweets have different characteristics and evolve over time. Our approach shows superior performance with these studies on the noisy, short text datasets of Twitter.

3 DATASET

Humanitarian Class	THAGUPIT	NEQUAKE
Caution and advice	467	NA
Infrastructure damage	421	425
Injured or dead people	NA	451
Affected people and evacuations	495	508
Rescue, donation efforts	409	636
Other useful information	434	433
Emotional support and irrelevant	500	500

Table 1: Labeled data of two disaster events. NA indicates that the class is absent or merged with another class.

We consider tweets posted in three days of the following two publicly available crisis datasets from CrisisNLP [11].

i. Typhoon Hagupit (THAGUPIT): an intense tropical cyclone, known as Typhoon Hagupit in Philipines. The dataset includes 0.21M tweets posted between December 06 and 08, 2014.

ii. Nepal Earthquake (NEQUAKE): a devastating earthquake in Nepal. This dataset consists of 1.19M tweets posted between April 25 and 27, 2015.

Around 2000 tweets from each dataset are labeled by crowd workers into different humanitarian categories [11], such as "injured or

[2]Our code will be made publicly available at https://github.com/HPanTroG/Bert2Bert.

Class	Event	Tweet text
Caution and advice	THAGUPIT	@USER: Super Typhoon Hagupit strengthens with 178 mph max winds as storm tracks toward Philippines.
Infrastructure damage	NEQUAKE	Nepal Earthquake: RT @USER: Kathmandu airport closed following 7.8 #NepalEarthquake.
Injured or dead people	NEQUAKE	RT @USER: Nearly 1,805 dead in Nepalś killer quake, India mounts massive rescue operation
Affected people and evacuations	NEQUAKE	RT @USER: We are a local tampa family and my son is #missing due to the #NepalEarthquake [url]
Rescue, donation efforts	THAGUPIT	#WorldVision is prepared to respond to 55,000 people with emergency essentials. #RubyPH [url]
Other useful information	THAGUPIT	NOW ON ANC: Pagasa update on Typhoon #RubyPH via ANC Alerts
Emotional support or irrelevant	THAGUPIT	R-evenge of the\nU-nfinished\nB-usiness of\nY-olanda\n\nHAHAHAHAHAHA xD stay safe mo guys

Table 2: Examples of tweets from various humanitarian classes, the highlighted snippets are rationales.

dead people", "infrastructure and utilities damage", "caution and advice", etc. These categories are defined and used by United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA). Nevertheless, we have observed many tweets that are wrongly annotated in those datasets. For example, the tweet *"*In real time * #NepalEarthquake India : So sad .. Bangladesh : That wasn't No ball.."* is marked as "infrastructure and utilities damage" in Nepal Earthquake, or *"RT @MyJaps: Stay safe everyone. ŏFŏFŏF #RubyPh"* is labeled as "caution and advice" in Typhoon Hagupit dataset. Besides, such annotations do not contain any rationale labels. Our rationales are short snippets that convey important information for the classification decision. A tweet can contain multiple non-consecutive snippets as rationales. All in all, we perform another round of annotation to revise labels, make them more accurate and annotate the rationale data.

Unlike some previous works that only consider classes with a sufficient number of tweets [38, 40], we take into consideration tweets of all classes. However, we merge some small classes that report similar information and create a new label for the tweets as "affected people and evacuations" so as to capture all important information. In THAGUPIT, three classes "missing, trapped and found people", "displaced people and evacuations" and "injured and dead people" are merged (there are not so many reports of injuries or death in flood events). Similarly, in NEQUAKE, two classes, "missing, trapped and found people" and "displaced people and evacuations" are merged (reports about injuries and death are prevalent in such events and should be kept as a separate class). The final set of classes is listed in Table 1. We illustrate examples of tweets in the pre-defined classes, along with rationales in Table 2.

4 THE PROPOSED METHOD

This section presents our proposed method for interpretable classification and summarization of disaster events.

4.1 Overview

We consider our classification-summarization approach in the following context. Given a large stream of tweets in chronological order during disaster events, we aim to classify incoming tweets into humanitarian classes with human-understandable explanations and generate summaries of class-level tweets. Figure 1 presents the overview of our framework. Tweets are pre-processed and fed into a BERT-based multi-task learning model that jointly trains two tasks: tweet classification and rationale/explanation extraction of the classifier. Next, the extracted rationales are employed to again classify tweets into humanitarian classes. The second classification step ensures that the model relies on extracted rationales to make predictions. Finally, the set of labeled tweets along with rationales

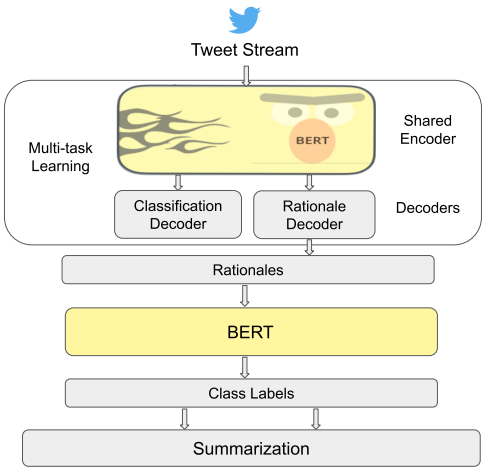

Figure 1: An overview of our interpretable classification and summarization framework.

are utilized as inputs of our summarization model. In this paper, we use an Integer Linear Programming (ILP) algorithm to extract salient, non-redundant tweets as summaries. Due to the large-scale disaster events, we allow users to generate snapshot summaries of a specific time interval and a defined length limit.

4.2 BERT based Multi-task Classification Pipeline: BERT2BERT

4.2.1 Data preparation. Our initially labeled data is imbalanced, the majority of tweets belong to the "emotional support or irrelevant" class, while some other classes have only a few tweets. To efficiently supervise our BERT-based interpretable classifier, we decide to gather more data for small classes and annotate rationale information. Firstly, we randomly sample and manually label new data of each event so as to obtain roughly 400 labeled tweets in each class. Next, rationales are annotated. Besides, we sub-sample irrelevant tweets to make our data more balanced. The final classes and number of labeled tweets used for our training process are shown in Table 1.

4.2.2 Rationale Identification and Classification. Our pipeline model is a BERT-based supervised encoder-decoder network with two learning stages. In the first stage, we extract rationales based on a multi-task learning structure that jointly classifies tweets into humanitarian classes and identifies rationales in the tweets using a BERT encoder and two decoders. The second stage ignores classification labels in the first stage and applies another BERT encoder to generate the classification prediction based on the extracted rationales alone. We formalize the classification as follows:

Input: Given a set of tweets T, each $t \in T$ is represented as $t =< t_1, t_2, .., t_n >$, where t_i is a BERT-based tokenized token in t.
Stage1 Output (Tweet class + Rationale tokens):

- **Output Task 1 (Classification decoder)**: Label $l \in L$ of any given tweet $t \in T$, where L: set of humanitarian classes in Table 1.
- **Output Task 2 (Rationale decoder)**: Token label $r =< r_0, r_1, .., r_n >$, where $r_i \in \{0, 1\}$ to specify whether a token t_i is a part of rationale information ($r_i = 1$).

Stage2 Output (Classification decoder): Final label $l \in L$ of any given tweet $t \in T$, where L is the set of humanitarian classes.
BERT Encoder. We employ BERTweet model [26] to encode input data. Our input tweets are first tokenized and split into a sequence of tokens of the form [CLS] t_1 t_2 .. t_n, where [CLS] is a special token added to mark the beginning of a tweet. We also keep the correspondence between a word and its tokens to later retrieve original words. Rationale labels are assigned to each tokenized token. BERTweet trains a masked language model to generate encoding vectors. Input tokens are padded to a maximum length of 128 - maximum sequence length of BERTweet [27], in each mini-batch. The final hidden state corresponding to the first token [CLS] is used as the aggregate representation of a tweet. BERT Encoder generates embeddings of size 768 dimensions for input tokens. An example of a tokenized tweet in BERT Encoder and our pipeline model is illustrated in Figure 2.
Classification Decoder. Our classification decoder generates a class label for each input tweet. The model is trained by appending a fully connected layer with Softmax on top of the final hidden vector in the encoder, corresponding to the first input token [CLS]. We compute a standard cross-entropy loss between the predicted probability p and the true labels y.

$$\mathcal{L}oss_{cd} = -\sum_{l=1}^{|L|} y_l log(p_l) \tag{1}$$

where, |L| is number of class labels, $y_l \in \{0, 1\}$ - binary indicator if the current tweet t belongs to class label $l \in L$. p_l is the predicted probability that tweet t is of class label l.
Rationale Decoder. The rationale extraction task is formalized as a binary classification task over input tokens. Given a sequence of tokens in an input tweet, the rationale decoder assigns a binary label to each token, which indicates whether the token is a part of the rationales. In this step, we append a Gated Recurrent Unit (GRU) layer followed by an output layer with Sigmoid function to the last hidden token embedding layer of the shared encoder. The GRU layer helps to capture the dependency between input tokens, yet has fewer parameters than a long short-term memory (LSTM). The presence of rationales can be sparse in some classes, i.e., around 20%-30% words (excluding mentions, URLs) in tweets of "caution and advice" contribute rationale information. To address the class imbalance, we use a weighted binary cross-entropy loss function [5], in which weights are proportional to token probabilities in the input tweets. The loss value of the rationale decoder is as follow:

$$\mathcal{L}oss_{rd} = -\sum_{i=1}^{|N|} \frac{|N_{y_i}|}{|N|}(y_i log(p_i) + (1 - y_i) log(1 - p_i)) \tag{2}$$

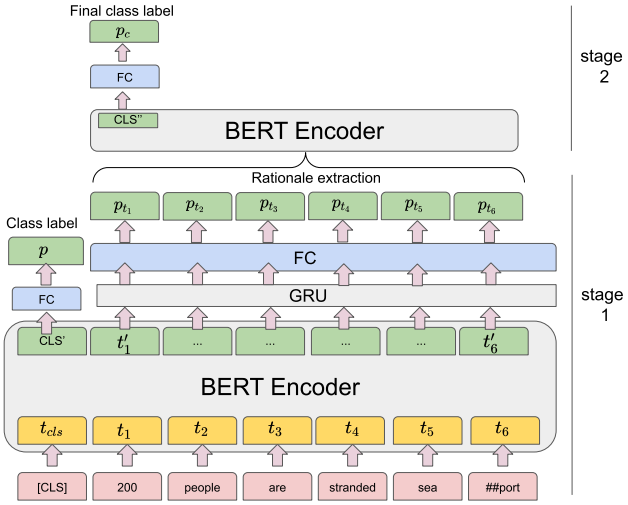

Figure 2: Our BERT2BERT model with example of an input tweet. FC indicates a fully connected layer.

where y_i and p_i are the true label and prediction value of i-th token respectively, $y_i \in \{0, 1\}$, $|N|$ is the length of the tweet, $|N_{y_i}|$ is the number of tokens with label y_i.
Stage1 Prediction. In the first stage, our BERT-based multi-task classifier jointly optimizes losses in the above two decoders. Formally, the overall loss function is defined as follow:

$$\mathcal{L}oss = \mathcal{L}oss_{cd} + \alpha \mathcal{L}oss_{rd} \tag{3}$$

where, α is the weight value to regulate losses of the two tasks.

The output of the rationale decoder is at token level. We merge split sub-tokens to retrieve the original words and word-level labels through max-pooling.
Stage2 Prediction. In this stage, we only consider rationale tokens of the tweets, mark other ones with a special character '*' and feed them to the second BERT classifier. The classification decoder of stage 2 generates the final class labels of tweets.

4.3 Tweet Summarization

In this section, we propose a method to summarize tweets of different humanitarian classes. First, we apply our trained classification model to generate labels and rationales on data of our three event dates. We observe that the extracted rationales cover the essential content of tweets. Side by side, numerals also play a key role. Thus, our summarization method aims to optimize the coverage of the rationales and numerals.

Given a stream of tweets along with tweet labels and rationale snippets in a humanitarian class, we build a model to generate summaries of any user-specified time window. We employ an Integer Linear Programming (ILP) framework for our summarization task. Considering a time window of T tweets, a summary of a desired length M words is generated by optimizing the following ILP objective function:

$$max(\sum_{j=1}^{T} t_j + \sum_{i=1}^{U} S(i).u_i) \tag{4}$$

where: $t_j \in \{0, 1\}$ indicates whether a tweet j is chosen. U is the number of unique rationale words and numerals in T tweets, $u_i \in \{0, 1\}$ specifies whether a rationale word or numeral i is chosen. $S(i)$ indicates the importance of a word i computed using logarithm of document frequency.

The objective function is optimized with following constraints:

- The summary length should contain at most M words, where M is specified by users.

$$\sum_{j=1}^{T} t_j \cdot Length(j) \leq M \tag{5}$$

- If the objective function selects a rationale word or numeral i in the summary, i.e., if $u_i = 1$, then it should select at least one *tweet* containing that word i.

$$\sum_{j \in Z_i} t_j \geq u_i, i = [1 \cdots U] \tag{6}$$

  where Z_i is the set of tweets containing the word i.
- All rationale words/numerals in a *tweet j* must be included in the summary if *tweet j* is selected for the summary.

$$\sum_{i \in R_j} u_i \geq |R_j| \times t_j, j = [1 \cdots T] \tag{7}$$

  where R_j is the set of rationale words/numerals in tweet j.

The above constraints consider both number of *tweets* (through the t_j variables) and number of important rationale words or numerals (through the u_i variables). Hence, our ILP-based summarizer takes care of multiple requirements, i.e., informativeness, diversity, redundancy, etc. We ensure that the most important informative words get selected in summary, and the optimization function does not get any benefit by selecting the same word multiple times. Overall, this process selects a set of tweets that form an informative and diverse summary. We validate our results in Section 6.

We employ the GUROBI Optimizer [8] to solve the ILP. After that, the set of *tweets j* such that $t_j = 1$, represent the summary at the current time window. We define our proposed **RA**tionale word-based **T**weet **SUM**marization approach as **RATSUM**.

5 CLASSIFICATION RESULTS: BERT2BERT

5.1 Baseline models

There exist no previous work on the interpretable classification of crisis-related tweets that is similar to our study. Hence, we compare the performance of our disaster classification model with the following previous baselines:

(1) **SVM**: A strong and supervised baseline [3, 11, 27] for the classification of crisis events. AIDR [10] also adopted a similar strategy.
(2) **RoCNN** [27]: A robust classification of crisis-related data on social networks using Convolutional Neural Network (CNN) with pre-trained word embeddings.
(3) **BERT-CLS** [26]: BERTweet model with a sequence classification head on top [9].
(4) **BERT-GRU**: BERTweet model combined with a GRU + Attention layer and a final output layer with Softmax. We apply additive attention formulation proposed by Bahdanau et al. [2] and extract top-k tokens with the highest attention weights as rationales. The value k is set to the average rationale length of human groundtruth for each category in each dataset. Tokens are then merged into original words to obtain final rationales through max-pooling.
(5) **BERT-MTL**: Our model with only first stage prediction.

5.2 Evaluation Metrics

We use Macro F1 score to evaluate prediction results of the classification models. Besides, we report how well our generated rationales agree with those marked by humans (rationale groundtruth) using Token-F1 metric. Basically, token precision measures the fraction of relevant rationale tokens (words) among the generated tokens, while token recall is the fraction of correctly retrieved rationale tokens among the groundtruth tokens. The Token-F1 reports the trade-off between token precision and token recall.

5.3 Experimental settings

We evaluate our model and baseline methods using a 5-fold cross-validation setting. We follow pre-processing or other setting steps in original papers for SVM and RoCNN. For BERT-based models, we pre-process tweets by removing mentions, URLs and then convert tweets to lower case. At each cross-validation run, we sample training, validation, and test sets with ratios 70%, 15%, 15%, respectively. The validation set is used for early-stop settings and hyper-parameters tuning of our model and all the baselines. BERT-based models are trained with the same setting of 10 epochs, AdamW optimizer [23] with an initial learning rate of 2e-5, and batch size of 16. The bidirectional GRU layer has a hidden size of 128. We specify a grid of candidate values in the range [1e-2, 4e-1] for our hyper-parameter α and compute average F1-scores of classification and rationale extraction tasks with respect to each candidate on validation sets. We select the hyper-parameter that results in the highest mean F1-score (average of Macro-F1 and Token-F1) over five runs on validation sets for test evaluation and new data prediction. The best hyperparameters α on both THagupit and NEquake are 0.07.

5.4 Classification Results

We report average scores on test sets over 5-fold cross-validation in Table 3. It is not surprising that BERT-based models return superior performance than the traditional machine learning approaches, such as SVM and RoCNN. BERT-GRU achieves high Macro F1, yet low Token-F1 scores on both the datasets. It is consistent with conclusions of previous studies [12, 41] that attentions do not provide a faithful explanation for classification decisions. BERT-MTL and BERT2BERT have the same Token-F1 score since they share the same encoder-decoder structure. Among all the methods, BERT-MTL has the highest classification Macro F1. However, one cannot surely say whether the model relies on rationales for its prediction. BERT2BERT gets high classification performance, and it is transparent to users that the model is interpretable by design, extracted rationales alone are sufficient for correct classification prediction. Our model also performs well (F1 $\geq$ 0.80) for each individual class.

Model	THAGUPIT		NEQUAKE	
	Macro F1	Token-F1	Macro F1	Token-F1
SVM	0.802	-	0.799	-
RoCNN	0.814	-	0.834	-
BERT-CLS	0.852	-	0.865	-
BERT-GRU	0.850	0.508	0.875	0.642
BERT-MTL	0.857	0.820	0.880	0.856
BERT2BERT	0.847	0.820	0.869	0.856

Table 3: Average F1 score over 5 fold cross-validation, '-' indicates that rationales are not extracted by a given method.

5.5 Faithfulness of Rationales

In this section, we evaluate the faithfulness of our rationales in terms of *comprehensiveness* and *sufficiency* [7]. We run the second stage of BERT2BERT with two different input settings and compute the two metrics as follows:

1. Comprehensiveness: Earlier, we train the classifier with the input tweet t_i. In this part, we train the classifier again with 5-fold cross-validation using $t_i \backslash r_i$, that is, the original input with r_i (rationales) replaced by a special character *. Finally, we evaluate the performance of both the input settings on the test set. For example, "*at least 13 dead after avalanches at mount everest*" and "* * * * *after avalanches at mount everest*" present the original and modified data. Next, we measure comprehensiveness as Macro F1(t_i) - Macro F1($t_i \backslash r_i$). High comprehensiveness indicates that rationales highly influence the model performance.

2. Sufficiency: In this case, we train the classifier using only rationales r_i (other tokens are replaced by *). Finally, we apply the model trained on the original text and the current one on test data and measure sufficiency as follows: Macro F1(t_i) - Macro F1(r_i). A low sufficiency score means our rationales are adequate for the model to make predictions.

In Table 4, the comprehensiveness score shows that our predicted rationales are important for classification. Specifically, the prediction performance drops significantly on both datasets when we mask rationales in the input text. Besides, the sufficiency scores are 0.005% and -0.004% on THAGUPIT and NEQUAKE, respectively. This ensures that extracted rationales are adequate for the model to make predictions. Compared to human rationales, higher comprehensiveness and the higher sufficiency of predicted rationales reflects that our extracted rationales are covering more tokens, yet some are false positive. The average ratio of extracted rationale words in input tweets is higher than that of human rationale words by 11%. The token-precision on THAGUPIT and NEQUAKE are 77% and 83%, respectively. Meanwhile, the token-recall are 95% and 94% correspondingly on THAGUPIT and NEQUAKE. Thus, there is still the scope for token-precision improvement.

5.6 Agreement between first and second stage prediction

Our BERT2BERT returns two different classification outputs - one in stage 1 and the other in stage 2. We measure the agreement/similarity between the two predicted label sets in terms of accuracy. The average agreement/accuracy scores between the the two predicted label sets are 90.7% and 92.2% on THAGUPIT and NEQUAKE respectively. The disagreement cases are mainly from tweets with mixture of information, i.e, "*RT @USER: In Sindhupalchok alone, death reaches 1,300. 90% homes destroyed, desperate wait for help. [url] #NepalE…*". The high agreement shows that our rationale extraction in stage 1 is effective for the final classification.

Dataset	Comprehensiveness↑		Sufficiency↓	
	Human Rationales	Predicted Rationales	Human Rationales	Predicted Rationales
THAGUPIT	0.218	0.294	-0.066	0.005
NEQUAKE	0.283	0.406	-0.097	-0.004

Table 4: Faithfulness of rationales.

6 SUMMARIZATION RESULTS: RATSUM

In this section, we evaluate our generated summaries in both quantitative and qualitative ways.

6.1 Groundtruth summaries

We employ five volunteers to prepare class-level summaries for each day of the events. In the summarization step, we ignore two classes that are not important from a situational point of view, such as "other useful information" and "emotional support and irrelevant". In total, we need to create 4 (class) x 3 (day) = 12 class-level summaries for each event. Volunteers were first asked to prepare summaries of 200 words (excluding #, @, URLs) independently. Next, we iteratively choose tweets selected by most volunteers until we reach a length limit of 200 words to form the groundtruth.

6.2 Baseline models

We consider both disaster-specific and recent deep learning-based neural summarization methods as baselines.

1. TSUM4ACT [28]: A Pagerank-based extractive summarization method for Twitter disaster events. It uses LDA to detect sub-topics before summarizing tweets.

2. APSAL [16]: An affinity clustering-based extractive summarization method for summarization of disaster-related news articles.

3. COWTS [36]: An unsupervised, extractive summarization model of crisis events on Twitter.

4. MOO [40]: An extractive summarization method for Twitter disaster events by jointly optimizing several objective functions.

5. BERTSUM: The recent supervised summarization model for news articles. It formulates the summarization problem as a classification task to identify sentences in the final summary.

6. PACSUM [49]: The strong unsupervised summarization method for news articles. It builds a sentence similarity graph using fine-tuned BERT embeddings and selects sentences with the highest centrality scores in the summary.

6. BERT-GRU: Our summarization model using the extracted rationales of the BERT-GRU classifier.

The first four strategies are disaster specific approaches, PACSUM and BERTSUM are neural BERT embedding-based approaches. For all the models, we generate summaries of length M = 200 words.

6.3 Evaluation metrics

We measure the summarization performance in both quantitative and qualitative ways.

Groundtruth based evaluation: We use a popular ROUGE toolkit for evaluation [19]. Following baselines and previous works on Twitter summarization [16, 29, 36, 42, 51], we choose ROUGE-1

Model	ROUGE-1 F-score (THAGUPIT)											
	Caution and advice			Affected people, evacuations			Infrastruture damage			Rescue, donation efforts		
	06/12/2014	07/12/2014	08/12/2014	06/12/2014	07/12/2014	08/12/2014	06/12/2014	07/12/2014	08/12/2014	06/12/2014	07/12/2014	08/12/2014
RATSUM	**0.574**	**0.647**	0.516	**0.642**	**0.615**	**0.641**	**0.516**	**0.483**	**0.609**	0.528	**0.657**	0.535
TSUM4ACT	0.327	0.419	0.461	0.314	0.356	0.253	0.328	0.303	0.363	0.485	0.401	0.376
APSAL	0.333	0.370	0.423	0.434	0.369	0.383	0.397	0.439	0.421	0.447	0.412	0.317
COWTS	0.544	0.621	**0.561**	0.639	0.574	0.624	0.487	0.469	0.526	0.465	0.594	0.552
MOO	0.330	0.297	0.343	0.386	0.340	0.290	0.337	0.274	0.292	0.394	0.262	0.324
BERTSUM	0.352	0.364	0.431	0.397	0.397	0.368	0.395	0.345	0.398	0.415	0.383	0.327
PACSUM	0.417	0.378	0.467	0.392	0.333	0.408	0.424	0.396	0.389	0.512	0.538	0.545
BERT-GRU	0.465	0.408	0.515	0.454	0.417	0.335	0.442	0.366	0.440	**0.567**	0.511	**0.602**
Model	ROUGE-1 F-score (NEQUAKE)											
	Injured or dead people			Affected people, evacuations			Infrastruture damage			Rescue, donation efforts		
	25/04/2015	26/04/2015	27/04/2015	25/04/2015	26/04/2015	27/04/2015	25/04/2015	26/04/2015	27/04/2015	25/04/2015	26/04/2015	27/04/2015
RATSUM	0.521	**0.564**	0.404	0.529	**0.526**	**0.556**	**0.581**	**0.580**	**0.472**	**0.644**	**0.651**	**0.576**
TSUM4ACT	0.336	0.295	0.294	0.446	0.359	0.346	0.422	0.347	0.231	0.390	0.383	0.314
APSAL	0.372	0.336	0.376	0.329	0.307	0.291	0.448	0.323	0.246	0.382	0.363	0.312
COWTS	**0.539**	0.476	0.359	**0.548**	0.439	0.390	0.538	0.409	0.386	0.456	0.459	0.549
MOO	0.372	0.303	0.339	0.278	0.355	0.238	0.333	0.273	0.297	0.300	0.228	0.300
BERTSUM	0.377	0.393	0.379	0.350	0.326	0.421	0.415	0.391	0.380	0.418	0.309	0.305
PACSUM	0.409	0.345	0.327	0.515	0.389	0.446	0.402	0.492	0.460	0.473	0.440	0.300
BERT-GRU	0.501	0.536	**0.422**	0.451	0.506	0.554	0.441	0.556	0.373	0.553	0.608	0.522

Table 5: ROUGE-1 F-score of summarization models. The best scores are in bold, the second bests are in brown color.

F-score for evaluating summaries. ROUGE-1 score has shown to be the most consistent with human assessments [20].

Human evaluation: We asked five volunteers to evaluate summaries generated by our model and all the baselines by answering two questions. **Q1**. For each summarization method, we generate 12 summary instances per dataset (hence, 24 instances in total). We give volunteers summaries returned by different methods and ask: Which summary is more informative about the event. This measures the coverage of information in summaries. A summary that contains more informative sentences is considered to have higher information coverage. **Q2**. We give two versions of RATSUM summaries (i). with highlighted rationale words, (ii). without highlighting, and ask volunteers which version they prefer. This evaluates whether the highlighted text reflects important content and helps end-users comprehend the situation better.

Datasets	Model	Q1	Q2
THAGUPIT	RATSUM	47%	100%
	COWTS	19%	NA
	APSAL	6%	NA
	TSUM4ACT	11%	NA
	PACSUM	17%	NA
NEQUAKE	RATSUM	83%	100%
	COWTS	8%	NA
	APSAL	3%	NA
	TSUM4ACT	3%	NA
	PACSUM	2%	NA

Table 6: The fraction of responses that a method is preferred by users. NA indicates that the question is not asked for a given method.

6.4 Summarization Results

6.4.1 Groundtruth-based evaluation. Table 5 shows the ROUGE-1 scores for 24 summary instances returned by our model and all the baselines. Though ROUGE-1 metric includes precision, recall, and F-score, we observe quantitatively similar patterns in all these scores. Hence, we report only F-score in the table. In most cases, RATSUM performs better than all the baseline approaches. On average, our summarization model outperforms COWTS, BERT-GRU, PACSUM by 5%, 8%, 14% respectively. The remaining baselines such as APSAL, TSUM4ACT, MOO and BERTSUM fall behind RATSUM with a large margin of more than 18% in term of average ROUGE-1 F-score. We also perform Wilcoxon signed-rank test [46] between RATSUM and other baselines. The performance of RATSUM turns out to be significantly better than the baselines with 95% confidence interval ($p - value < 0.05$). Side by side, this trend also holds for ROUGE-2 and ROUGE-L.

6.4.2 Human Evaluation. As MOO and BERTSUM shows low performance compared to other models, and BERT-GRU applies the same method as ours, we do not give the results of these models to volunteers to reduce workload. For each dataset, we get 60 responses to a given question (5 volunteers x 12 summary instances). Table 6 illustrates the fraction of responses. In THAGUPIT dataset, 47% of respondents find our generated summaries more informative. The second and third informative models are COWTS and PACSUM. It is generally consistent with the above groundtruth-based evaluation results. In NEQUAKE dataset, 83% of respondents prefer our model in terms of informativeness. It is significantly higher than the evaluation on THAGUPIT dataset. We observe that the NEQUAKE dataset is much bigger, each category covers more sub-events. The human evaluation and our observation indicate that RATSUM tends to work well on large datasets with many subtopics. Table 6 also illustrates the high preference of highlighted text. 100% of volunteers think the highlighting is useful and more user-friendly. We illustrates an example of 100-word summaries generated by RATSUM and COWTS in Table 7. RATSUM is shown in the format with highlighted rationales.

6.5 Discussion on Performance

In this section, we discuss possible reasons why our model is superior to the baseline methods. The disaster-specific summarization

Reports indicate **80% homes near #Nepal #Earthquake epicenter collapsed**. CARE's responding. **Some of Nepal's world heritage sites are damaged or destroyed** in earthquake. **India Flights to Kathmandu put on hold**: **Domestic airlines today put on hold their services** t... #business #kerala. The 7.9 earthquake dat hit nepal has **dstroyed buildings, cellphone netwrks r down nd power is out** #MSGHe… Initial pictures after #Nepalquake show **major damage to buildings and structures**. **Nepal earthquake devastation could cost billions**: Here's how to help. **#Tibet severely affected** by #NepalEarthquake; **houses collapsed, communications cut off**. **Nepal declares state of emergency** after killer quake.	Reports indicate 80% homes near #Nepal #Earthquake epicenter collapsed. CARE's responding Terrible news from Nepal. Donations here. Pic of devastated Palace area taken 10 days ago. Witnesses: Some buildings collapse in Nepal capital after 7.7 quake: By Gopal Sharma and Ross Adkin KATHMANDU (Reuters) - Nepal urged... Devastating visuals of destruction in Nepal....thoughts,prayers and all protective energies for this tragic loss of life..... Katmandu's poorly constructed buildings worsen quake outcome. Nepal earthquake devastation could cost billions: Here's how to help. Nepal Earthquake: Extensive Destruction, Rising Death Toll. Still can't believe what I witnessed in #NepalQuake today. History crumbling, a nation in despair.

Table 7: An example of 100-word summaries (excluding #, @, URLs) generated from tweets in "infrastructure damage" class (NEQUAKE 26/04) by RATSUM and COWTS.

baselines generally perform worse than RATSUM due to various reasons. TSUM4ACT [28] clusters tweets to sub-topics and selects the most informative ones in each cluster using a Pagerank-based method. The model assumes that all clusters are equally important and select the same number of tweets in each cluster. This assumption might not be valid in disaster scenarios, in which some sub-topics might cover more critical information than others. APSAL [16] selects tweets based on specific features of sentences in news articles such as sentence position or language models representing the language of disasters. These features are usually missing in noisy, short texts of Twitter datasets. BERT-GRU falls short behind our model due to the low quality and instability of extracted rationales, as we discussed in Section 5.4. It obtains the best performance for a few summaries, and the remaining cases are significantly worse than RATSUM. Finally, COWTS [36] considers nouns, numerals, and main verbs as important words and tries to cover these words in summaries. However, in some cases, other words (i.e., adjectives) also play an essential role in disaster-related tweets. RATSUM works better because it does not only look at words separately but considers informative phrases of rationales in the context of tweets. COWTS behaves quite competitive with RATSUM model on small or less diverse datasets.

Our embedding-based summarization baselines show high computational complexity and low performance in the summarization of large-scale short texts. MOO [40] generally prefers long sentences with high TF-IDF scores. The extracted tweets by MOO are also redundant due to the drawback of Word Move Distance (WMD) based dissimilarity strategy. Besides, the computation of WMD scores is expensive. Next, the supervised model BERTSUM [49] falls short in our experiment due to the difference in specific traits of well-written news articles and tweets. BERTSUM and some supervised neural summarization models [25] grow parameters with the length of the input documents. Therefore, it fits well for news articles, but not large tweet sets. We adapt the model by breaking down our tweet datasets into sub-documents. However, BERTSUM faces another challenge of highly imbalanced data, with only a few tweets are in the groundtruth summary. Another embedding-based summarizer, PACSUM generates less diverse summaries than RATSUM. PACSUM is specifically designed for news articles, it learns similarity between input texts by fine-tuning BERT on news articles datasets. The model builds a directed graph for sentence selection under the assumption that relative positions of sentences influence the centrality, i.e., preceding sentences are more central. However, the assumption is not true for a set of equally important tweets on Twitter. Besides, it is also computationally expensive to extract BERT-based similarity scores for all pairs of tweets when building the PACSUM graph.

Class	#Tweets	Macro F1	Token-F1
Infrastructure damage	164	83.95	86.09
Injured or dead people	166		
Affected people and evacuations	157		
Rescue, donation efforts	161		
Other useful information	168		
Emotional support or irrelevant	185		

Table 8: Performance of BERT2BERT on Mexico dataset.

6.6 Discussion on generalization.

Our model requires intensive initial labor work for rationale annotation. However, it can generalize well on new data. To observe the ability of our approach for generalization, we download 1000 labeled tweets of the recent Mexico earthquake event [1] and evaluate both classification performance and rationale extraction. We first manually check labels and then annotate rationale snippets. Then, we train BERT2BERT model on 100% NEQUAKE dataset with 10 epochs and the predict class labels and extract rationales for evaluation. The performance on new data are shown in Table 8. Although we do not use any in-domain data of Mexico dataset for training, our model achieves good performance on both tweet classification and rationale extraction tasks.

7 CONCLUSION

This paper presents an interpretable classification and summarization framework for disaster events on Twitter. We leverage an interpretable by design approach to develop BERT2BERT classifier for crisis-related microblogs. Our evaluation shows the efficacy of BERT2BERT over baseline methods. We also show that the extracted rationales are beneficial for the summarization of tweets. Our RATSUM summarizer turns out to be good for both informativeness and human understanding. The model is robust, simple, yet able to generate informative summaries in near real-time. For future work, we would like to use NLP tools in pre-processing tweets to handle the current misclassification cases and further improve our rationale prediction task. We are also interested in exploring embedding-based methods to design a novel, robust and effective tweet summarizer. The pre-trained embedding techniques have shown great potential in the summarization of news articles, but are very expensive and perform poorly on large-scale evolving tweet streams. Therefore, the study of neural summarization methods on Twitter datasets is a promising direction. Besides, we are planning to deploy an interpretable framework to assist rescue agencies, NGOs.

ACKNOWLEDGMENTS

This work was partially funded by the DFG Grant NI-1760/1-1, and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 832921.

REFERENCES

[1] Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Graph Based Semi-supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM)*.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[3] Mark A. Cameron, Robert Power, Bella F Robinson, and Jie Yin. 2012. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12 Companion)*.

[4] Rich Caruana. 1997. Multitask Learning. *Rich Caruana* (1997).

[5] N. Chawla, K. Bowyer, L. Hall, and P. Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* (2002), 321–357.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[7] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 4443–4458.

[8] gurobi. 2015. Gurobi – The overall fastest and best supported solver available. http://www.gurobi.com/

[9] Huggingface. 2021. Hugging Face – The AI community building the future. https://huggingface.co/

[10] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. 2014. AIDR: artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14 Companion)*.

[11] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the 10th Language Resources and Evaluation Conference (LREC)*.

[12] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (ACL)*.

[13] S. Jain, S. Wiegreffe, Y. Pinter, and B. C. Wallace. 2020. Learning to Faithfully Rationalize by Construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 4459–4473.

[14] Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao, and Shi Wang. 2020. Neural Extractive Summarization with Hierarchical Attentive Heterogeneous Graph Network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[15] Chris Kedzie, Fernando Diaz, and Kathleen R. McKeown. 2016. Real-Time Web Scale Event Summarization Using Sequential Decision Making. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI*, Subbarao Kambhampati (Ed.). 3754–3760.

[16] Chris Kedzie, Kathleen McKeown, and Fernando Diaz. 2015. Predicting Salient Updates for Disaster Summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (IJCNLP)*.

[17] Prashant Khare, Grégoire Burel, Diana Maynard, and Harith Alani. 2018. Cross-Lingual Classification of Crisis Data. In *Proceedings of the International International Semantic Web Conference (ISWC)*.

[18] Hongmin Li, Doina Caragea, and Cornelia Caragea. 2021. Combining Self-training with Deep Learning for Disaster Tweet Classification. In *Proceedings of the 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

[19] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceddings of Workshop on Text Summarization Branches Out (with ACL)*.

[20] Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language (NAACL)*.

[21] Junhua Liu, Trisha Singhal, Lucienne T.M. Blessing, Kristin L. Wood, and Kwan Hui Lim. 2021. CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT)*.

[22] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[23] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceddings of the International Conference on Learning Representations (ICLR)*.

[24] Reza Mazloom, Hongmin Li, Doina Caragea, Cornelia Caragea, and Muhammad Imran. 2019. A hybrid domain adaptation approach for identifying crisis-relevant tweets. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 2 (2019), 1–19.

[25] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*.

[26] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

[27] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks. In *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*.

[28] Minh-Tien Nguyen, Asanobu Kitamoto, and Tri-Thanh Nguyen. 2015. TSum4act: A Framework for Retrieving and Summarizing Actionable Tweets During a Disaster for Reaction. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*.

[29] Thi Huyen Nguyen, Tuan-Anh Hoang, and Wolfgang Nejdl. 2019. Efficient Summarizing of Evolving Events from Twitter Streams. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*.

[30] Andrei Olariu. 2014. Efficient Online Summarization of Microblogging Streams. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

[31] Miles Osborne, Sean Moran, Richard McCreadie, Alexander Von Lunen, Martin Sykora, Elizabeth Cano, Neil Ireson, Craig Macdonald, Iadh Ounis, Yulan He, Tom Jackson, Fabio Ciravegna, and Ann O'Brien. 2014. Real-Time Detection, Tracking, and Monitoring of Automatically Discovered Events in Social Media. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[33] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*.

[34] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In *Nature Machine Intelligence*.

[35] Koustav Rudra, Niloy Ganguly, Pawan Goyal, and Saptarshi GhoshACM Transactions on the Web. 2018. Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters. *ACM Transactions on the Web* (2018).

[36] Koustav Rudra, Subham Ghosh, and Niloy Ganguly. 2015. Extracting Situational Information from Microblogs during Disaster Events: a Classification-Summarization Approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM)*.

[37] Koustav Rudra, Pawan Goyal, Niloy Ganguly, Muhammad Imran, and Prasenjit Mitra. 2019. Summarizing Situational Tweets in Crisis Scenarios: An Extractive-Abstractive Approach. In *IEEE Transactions on Computational Social Systems*.

[38] Koustav Rudra, Pawan Goyal, Niloy Ganguly, Prasenjit Mitra, and Muhammad Imran. 2018. Identifying Sub-events and Summarizing Disaster-Related Information from Microblogs. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*.

[39] Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. EXPLAGRAPHS: An Explanation Graph Generation Task for Structured Commonsense Reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (ACL)*. 7716–7740.

[40] Naveen Saini, Sriparna Saha, and Pushpak Bhattacharyya. 2019. Multiobjective-Based Approach for Microblog Summarization. *IEEE Transactions on Computational Social Systems* (2019).

[41] Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the //Association for Computational Linguistics (ACL)*.

[42] Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. 2013. Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval)*.

[43] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is Out There: Looking for Help from Tweets during a Large Scale Disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

[44] Sudha Verma, Sarah Vieweg, William J. Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram1, and Kenneth M. Anderson. 2011. Natural

Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*.

[45] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[46] Wikipedia. 2021. Wilcoxon signed-rank test. https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

[47] Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve1. 2020. TED: A Pretrained Unsupervised Summarization Model with Theme Modeling and Denoising. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*.

[48] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and Predict, and then Predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM)*.

[49] Hao Zheng and Mirella Lapata. 2019. Sentence Centrality Revisited for Unsupervised Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[50] R. Zhong, S. Shao, and K. McKeown. 2019. Fine-grained sentiment analysis with faithful attention. In *arXiv preprint arXiv:1908.06870*.

[51] Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. 2012. Towards Real-Time Summarization of Scheduled Events from Twitter Streams. In *Proceedings of the 12th ACM conference on Hypertext and Hypermedia (HT)*.