Learning Faithful Attention for Interpretable Classification of Crisis-Related Microblogs under Constrained Human Budget

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

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

ABSTRACT

The recent widespread use of social media platforms has created convenient ways to obtain and spread up-to-date information during crisis events such as disasters. Time-critical analysis of crisis data can help human organizations gain actionable information and plan for aid responses. Many existing studies have proposed methods to identify informative messages and categorize them into different humanitarian classes. Advanced neural network architectures tend to achieve state-of-the-art performance, but the model decisions are opaque. While attention heatmaps show insights into the model's prediction, some studies found that standard attention does not provide meaningful explanations. Alternatively, recent works proposed interpretable approaches for the classification of crisis events that rely on human rationales to train and extract short snippets as explanations. However, the rationale annotations are not always available, especially in real-time situations for new tasks and events. In this paper, we propose a two-stage approach to learn the rationales under minimal human supervision and derive faithful machine attention. Extensive experiments over four crisis events show that our model is able to obtain better or comparable classification performance (~86% Macro-F1) to baselines and faithful attention heatmaps using only 40-50% human-level supervision. Further, we employ a zero-shot learning setup to detect actionable tweets along with actionable word snippets as rationales.

CCS CONCEPTS

• **Information systems** → **Clustering and classification**; • **Computing methodologies** → **Semi-supervised learning settings**; **Transfer learning**.

KEYWORDS

Interpretability, Classification, Crisis Events, Twitter

ACM Reference Format:
Thi Huyen Nguyen and Koustav Rudra. 2023. Learning Faithful Attention for Interpretable Classification of Crisis-Related Microblogs under Constrained Human Budget. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583861

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583861

1 INTRODUCTION

Disaster events create disruption among communities in the vicinity of affected zones. Natural disasters are usually long-ranging in nature and create an impact over large geographic areas. With the advent of social media and the easy acceptability of smartphones, a large number of multimodal information gets posted. This information scatters across multiple dimensions such as news, updates, actionable information, etc. [5]. Further, situational updates also come from different humanitarian classes (i.e., infrastructure damage, missing people, injured persons, volunteer info, etc.) [1, 12, 13]. The availability of information makes the decision-making tasks of NGOs and governments easier. However, a huge volume of information also becomes a bottleneck; hence, a streamlined mode of updating across different categories is desired by the agencies. Lots of existing approaches have already tried to assist NGOs, but without proper justification/explanation, such systems can hardly be used in real-life critical scenarios.

In recent times, interpretability has become an essential component in model building. Tasks that have an impact on society and human lives are crucial in nature, and explainability is an important component of model building in such cases. There exists a series of works for the classification and summarization of crisis events, detection of actionable content, and different information types [5, 23, 41], etc. However, none of them focus on the explainability aspect of the model. Recently, Nguyen et al [26–28] proposed interpretable-by-design approaches to classify crisis tweets into different humanitarian categories. Their methods identify the class information and tokens responsible for determining that class. They asked humans to provide explanation tokens along with class-level annotations. For example, the tweet "*RT @USER:* ***Three people from #Taiwan died*** *in #MexicoEarthquake, Chinese embassy in Mexico confirms https://t.co/2Ig19YnCbs*" is labeled as 'injuries or death' and words in bold are annotated as rationales. While this approach shows a promising direction toward interpretable crisis systems, human-level annotation also adds a bottleneck toward the scalability of the method and its application toward new unseen events. On the other hand, some sets of approaches tried to use attention weights as a mode of explanation [3, 10]. However, recent studies pointed out the flaws in considering attention weights as a proxy for explanation [14, 42]. The debate is still ongoing [4].

This brings two open challenges into the framework — (a). How could we learn faithful attention weights that could represent explanations with high confidence, and (ii). How to develop interpretable models under the given human budget, i.e., limited annotated data.

In this paper, we try to address the above-mentioned challenges and design a two-stage framework that exploits the power of semi-supervised learning. Next, we incorporate the distance between

Move People
Authorities urge residents to seek higher ground immediately as rains lash the north after Cyclone Kenneth #mozambique #naturaldisaster https://t.co/nSFnEQDdsA https://t.co/3B84ljp8oY

Emerging Threat
READ: Families in #Mozambique face new floods and fight for their lives. Nicholas Finney comments on #CycloneKenneth via @MailOnline: 'we have grave fears for the thousands of families currently taking shelter. https://t.co/V68H9RJixX

Figure 1: Example of actionable tweets, actionable labels are in bold, rationales are in blue.

attention weights and predicted probabilities of rationales into our customized loss function to make the attention weights faithful. Evaluation on four different disaster events shows that this would help to alleviate almost 50% human annotation budget and learn faithful attention weights. Further, we extend the idea of zero-shot learning to directly transfer the knowledge acquired in the humanitarian classification model to the actionable tweet detection problem. Results suggest that if the source and target tasks are related and from a similar application area (e.g., crisis), such zero-shot learning setup can be directly applied to the target task. This direct transfer helps to identify the actionable classes and related rationale tokens from the tweets. Examples of actionable tweets, class labels, and rationale snippets are illustrated in Figure 1. We apply our proposed approach to the following disaster events. — (i). Nepal Earthquake, (ii). Typhoon Hagupit, (iii). Mexico Earthquake, and (iv). Cyclone PAM. Overall, our contributions are as follows:

- We propose a faithful attention-based model for the classification of tweets posted during crisis events (FAC-BERT)[1].
- We evaluate the faithfulness of FAC-BERT attention under limited human rationales as supervision. Experimental results suggest that 40-50% human annotated rationales are good enough to get a performance similar to a fully supervised model (100% annotated rationales).
- Our customized loss helps FAC-BERT learn faithful attention heatmaps. We obtain an average 20% improvement across our four datasets in learning faithful attention weights through FAC-BERT customized loss over the generic cross-entropy-based loss function (details Section 5.4).
- Our FAC-BERT can identify rationales in actionable tweets, and such rationales have a contribution of 24.6% in actionable tweet detection (i.e., removing those rationales results in a performance drop of 24.6%).

2 RELATED WORK

This section briefly overviews works on the classification of crisis-related microblogs and model interpretability.

2.1 Classification of crisis events

Classification of crisis events has been a topic of increasing interest and attracted growing research attention. Various approaches have been proposed to classify tweets during crisis events. Verma et al. [40] applied standard machine learning models such as Naive Bayes and Maximum Entropy to identify tweets that contribute situational awareness during crisis events. The authors utilized both hand-annotated and automatically extracted features for classification. Similarly, Rudra et al. [31] used a Support Vector Machine (SVM) but considered low-lexical and syntactic features. Besides, multiple studies also employed traditional methods for the classification of crisis events [11, 12].

A substantial number of previous works focused on deep learning models with pre-trained embeddings for crisis-related data classification [16, 20, 22, 25]. Nguyen et al. [25] employed a Convolutional Neural Network (CNN) with word embeddings pre-trained on Google news or crisis datasets. Manna and Nakai [22] compared the performance of Neural Network models with pre-trained word embeddings and traditional machine learning approaches in the classification of crisis-related tweets. Recently, Transformer-based models [39] have been proposed and archived superior performance compared to previous approaches. Liu et al. [20] introduced a robust Transformer for crisis classification. The authors ran experiments on two classification tasks, namely crisis recognition, crisis detection, and showed that the model achieves better performance than conventional word embedding-based methods. Besides, Nguyen and Rudra [26, 27] developed multi-task transformer-based approaches for tweet classification. Moreover, TRECIS track [36] designed challenges across multiple years to classify tweets into 25 information types and predict tweet priorities. The high-performance runs also tend to be transformer-based models [5].

2.2 Model Interpretability

Nowadays, interpretation of black-box models becomes an essential requirement for any kind of task that has an impact on society or human lives [18]. Broadly, there are three ways to achieve the interpretability of black-box models — (i). pre-modeling, (ii). in-modeling, and (iii). post-modeling approaches [15]. *Pre-modeling* approaches mostly deal with data understanding, visualization, dimensionality reduction, etc. *Post-modeling* approaches deal with understanding the trained black-box models. Most of the approaches fall into this category. Such post-modeling methods could be grouped based on model usage and scope of explanations [15, 19]. As per model usage, explanation models are either model agnostic, i.e., they don't have access to model parameters, or model intrinsic, where model parameters are accessible. Many interpretable strategies were proposed under the model usage category, such as feature, gradient-based importance, etc. (details are covered in [15, 19]). As per the scope of explanation, approaches are categorized into local and global types. In local, the objective is to explain a single instance, whereas, in global mode, the objective is to explore the overall decision process of a model. However, post-modeling approaches are unreliable and could be easily fooled [34]. In-modeling approaches try to make the model inherently interpretable. Popular approaches are like decision trees [29], rule-based models [9, 17], etc. In recent times, researchers also proposed explainable deep learning models that learn the task along with explanations [8, 43].

Attention based explanations: Badhanu et al [2] introduced the attention concept in machine translation tasks that helps to identify the importance of individual tokens. Following this idea, many researchers proposed attention as a way of modeling explanations [3, 10]. However, recent studies [14, 30, 42] showed that attention is not an explanation. Tutek et al [7, 38] analyzed reasons behind the failure of attention weights as a transparency tool. On a similar note, Chrysostomou and Aletras [7] tried to improve the

[1]Our code will be available at https://github.com/HPanTroG/FAC-BERT

faithfulness of attention-based explanations with task-specific information for text classification. Unlike previous works, we learn faithful rationales under limited human supervision that cover the human comprehension/readability part into account. Hence, we consider the aspect of comprehension part, i.e., consecutiveness. To the best of our knowledge, this is the first study to learn faithful attention with help from human rationales.

Interpretable approaches for crisis-related tweet classification: Many studies tried to classify crisis-related tweets into different humanitarian categories [31, 32, 40]. However, such works did not focus on the interpretability aspect that hinders the utility of such models. Recently, Nguyen et al [26, 27] proposed explainable approaches to classify tweets into humanitarian classes and summarize the information. However, the major limitation of these models is that they assume 100% human annotated rationales, i.e., rationales for each training instance. This bottlenecks in the direct application of models to new events and tasks. In this paper, we try to overcome the limitations and propose a two-stage approach that could learn within a given human budget constraint.

In contrast to prior works, we assume rationales are present only for k% of the training data instead of the entire training dataset. Thus, our objective is to learn under the given rationale budget. Further, we learn a mapping between the tokens' attention weights and probabilities to be rationales. This helps in learning faithful attention weights. Finally, we leverage the similarities between related tasks in crisis domain and show the application of the humanitarian classification model over actionable class detection.

3 METHODOLOGY

This section describes the detailed architecture of our faithful attention-based classification model.

3.1 Problem Formulation

Given a small set of tweets $T = \{twt_1, twt_2, .., twt_m\}$, along with labels $L = \{l_1, l_2, .., l_m\}$, $l_i \in C$, where C is the set of humanitarian classes (i.e., infrastructure damage, affected people, rescue, etc.). We assume that we also have access to human rationales for a small set of tweets $S = \{twt_i\} \subset T$. Rationales are short snippets from original texts that are marked as having supported the class label. Here, we consider each tweet twt as a list of words $twt = \{w_1, w_2, .., w_k\}$. If the tweet twt is provided with human rationales, we then have labels $y = \{y_1, y_2, .., y_k\}$ assigned for every word, $y_i \in \{0, 1\}$ specifies whether a word is a part of rationales ($y_i = 1$). Our aim is to take the limited human rationales as little supervision to design a Faithful Attention-based Classification model (FAC-BERT) of tweets during crisis events.

3.2 Overview

As discussed above, our goal is to develop an interpretable classification approach with little supervision of human rationales. Many previous works [14, 33, 42] have argued that attention is not explanation. Hence, we also aim at finding a way to make the attention become a faithful explanation. Our model learns a mapping from the annotated rationales to machine attention. We achieve this by proposing a hierarchical learning structure that predicts the probabilities of each word being rationales. Then we align these probabilities with attention weights to inform tweet classification.

As a first step, we apply BERTweet [24] to tokenize input tweets and generate token embeddings. These embeddings are fine-tuned on a token classification task that predicts whether a token is a part of rationales with probabilities. These values are used to guide machine attention. Then, we apply a weighted sum of token vectors to obtain tweet vectors for tweet-level classification. The weights are learned to reflect the importance of each token to the output decision. Our model is able to obtain high classification performance and faithful attention. We refer the model as Faithful Attention-based BERTweet Classification (FAC-BERT). The detailed training process of our FAC-BERT classifier is described below.

3.3 Model architecture

Figure 2 illustrates the architecture of our FAC-BERT model. It consists of two training phases with a shared BERTweet encoder. Note that our approach is different from muti-task learning setups in some previous work [27, 28]. Our two phases are not trained simultaneously. The second phase takes the information and last checkpoint from the first phase and continues to train its own task.

BERTweet encoder [24]. We use BERTweet as a shared encoder for our learning phases. BERTweet is a language model pre-trained on a large-scale dataset of English tweets. First, each tweet is tokenized into tokens of the form $[CLS]tok_1tok_2..tok_n$, where $[CLS]$ is a special symbol added in front of every input instance. An unknown word from BERTweet vocabulary can be split into several tokens. Input sequences are padded to the same length, which is the maximum length of tweets in each learning batch. Then, we feed the tokenized data to the BERTweet encoder and obtain token embeddings of size 768 dimensions x^{ij}_{tok} for each token tok_i in tweet twt_j. The token representations are fine-tuned in the first learning phase and then aggregated to form tweet representation for classification in the second training phase.

Token-level training (Phase 1). This step takes token embeddings as inputs and trains a binary classifier to predict which tokens are part of rationales. We append a GRU (Gated Recurrent Unit) followed by a fully connected layer with a Sigmoid function on top of BERTweet token embeddings. Initially, rationale labels are assigned at the word level. To train our model, we map labels to token level, where each tokenized token has the same label as its original word. Later, at the evaluation step, we retrieve word-level labels by applying max pooling on token labels. We employ the binary cross-entropy loss function for token-level classification.

$$Loss_{tok} = BCELoss(y_i, p_i) \quad (1)$$

where y_i is the true label, p_i is the predicted probability of token tok_i to be rationale. Recall that we aim at learning with little supervision of human rationales. Hence, the above loss function is only averaged over k% of tweets in training set with human rationales. After the training completes, we obtain fine-tuned token embeddings and the probabilities of tokens to be rationales for all tweets in the training set.

Tweet-level training (Phase 2). This step predicts the class label of input tweets. Token vectors are summed up to obtain tweet representations. In our FAC-BERT, the sum of token vectors is the attention-based weighted sum. Specifically, we apply an attention

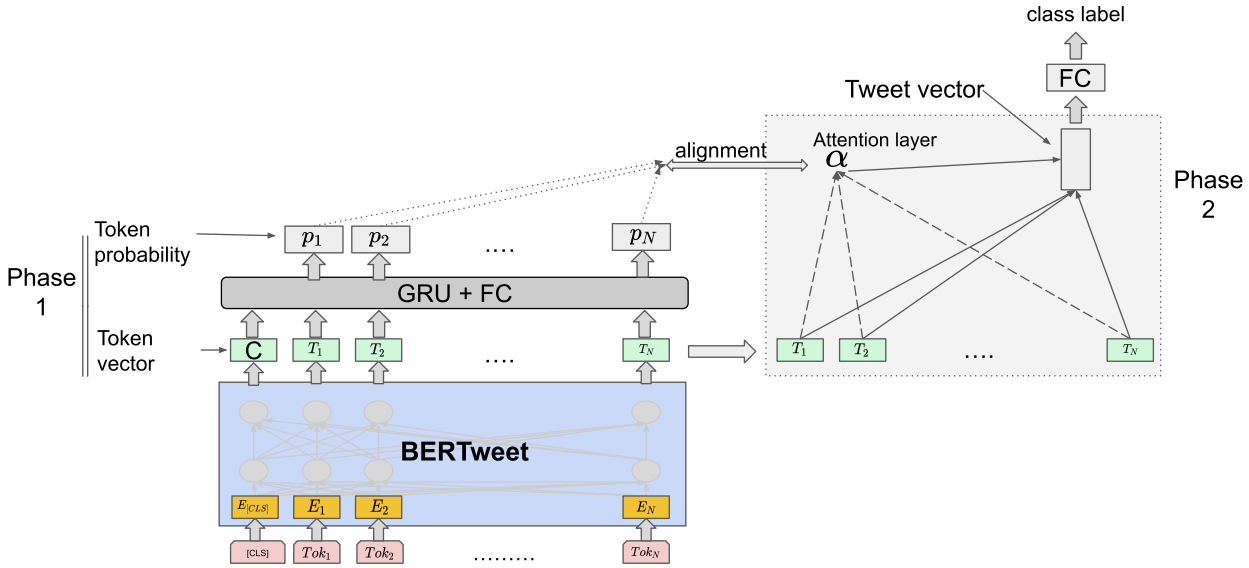

Figure 2: FAC-BERT - Our faithful attention-based classification model

layer [2] on top of fine-tuned BERTweet token embeddings. The attention weights α_{ij} are computed by a softmax function as follow:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{|N|} exp(e_{ik})} \tag{2}$$

Where e_{ij} is the output score of a feedforward neural network model [2], which captures the alignment between input at position j and output i, |N| is the length of the considering tweet. The representation of each tweet twt_i is then the weighted sum over token embeddings:

$$x_{twt}^{i} = \sum_{j=1}^{|N|} \alpha_{ij} x_{tok}^{j} \tag{3}$$

The tweet embeddings are fed into a fully connected softmax layer to predict class labels. Besides, we want attention weights to mimic human rationales so that the attention accurately reflects the true reasoning behind a prediction (i.e., tokens with high attentions highly influence the model decision). Hence, we minimize the distance between attention weights α_j in a tweet twt and probabilities p_j of tokens to be a rationale that is learned in the first phase:

$$d(\alpha_j, p_j) = max(0, 1 - cosine(\alpha_j, p_j)) \tag{4}$$

The above distance is interpolated with the weighted cross-entropy classification loss to form the final loss function of the tweet-level training step:

$$Loss_c = -\sum_{l=1}^{|L|} w_j * y_{jl} log(p_{jl}) + \lambda \sum_{i=1}^{|N|} d(\alpha_i, p_i) \tag{5}$$

where |L| is the number of unique class labels, y_{jl} and p_{jl} are the true label and predicted value of tweet twt_j having class label l, w_j is the inverse weighted probability of label occurrence in the dataset. In case of a balanced dataset, w_j is set to 1 for all classes. |N| is the token length of the tweet.

Humanitarian class	#Tweets			
	NeQuake	MexQuake	THagupit	CPam
Rescue & donation efforts	636	381	411	398
Infrastructure damage	425	390	421	396
Injured & dead people	451	395	-	-
Caution & advice	-	-	469	404
Affected people & evacuations	508	399	502	396
Other useful information	433	399	431	364
Emotional or irrelevant	497	438	493	411

Table 1: Labeled datasets. - if the class is absent.

When training the class label prediction task, we fix parameters of top layers (GRU+FC) at the token-level training phase.

4 EXPERIMENTAL SETUP

4.1 Datasets

We consider the following four natural disaster events from recent studies [26, 27] for our evaluation.

Nepal Earthquake (NeQuake): An intensive earthquake occurred on 25 April 2015.

Mexico Earthquake (MexQuake): A powerful earthquake happened in Mexico on September 19, 2017.

Typhoon Hagupit (THagupit): The intense tropical cyclone that impacted Philippines in December 2014.

Cyclone PAM (CPam): A tropical cyclone in the South Pacific Ocean occurred in March 2015.

Each dataset contains about 2000 tweets with humanitarian classes and rationales. The details of datasets and humanitarian classes are shown in Table 1.

4.2 Baseline methods

We compare our model with the following classification models, which include both typical classification approaches and recent interpretable crisis-related classification models.

Model	In-domain								Cross-domain (Train‖Test)							
	NeQuake		MexQuake		THAGUPIT		CPAM		MexQuake‖NeQuake		NeQuake‖MexQuake		CPAM‖THAGUPIT		THAGUPIT‖CPAM	
	Macro-F1	Token-F1	Macro-F1	Token-F1	Macro-F1	Token-F1	Macro-F1	Token F1	Macro-F1	Token-F1	Macro-F1	Token-F1	Macro-F1	Token-F1	Macro-F1	Token-F1
SVM	0.799	-	0.738	-	0.802	-	0.768	-	0.661	-	0.679	-	0.523	-	0.524	-
Robust-CNN	0.833	-	0.787	-	0.817	-	0.843	-	0.730	-	0.683	-	0.671	-	0.602	-
LCL	0.865	-	0.850	-	0.856	-	0.864	-	0.849	-	0.835	-	0.819	-	0.800	-
BERTweet	0.864	-	0.851	-	0.852	-	0.888	-	0.851	-	0.837	-	0.822	-	0.853	-
BERT2BERT(-2stg)	0.874	0.857	0.855	0.826	0.857	0.820	0.891	0.868	0.847	0.839	0.841	**0.862**	0.808	0.831	0.851	**0.873**
BERT2BERT	0.862		0.836		0.847		0.861		0.842		0.829		0.815		0.818	
RACLC(-2stg)	**0.890**	**0.868**	**0.869**	**0.874**	**0.865**	**0.847**	**0.896**	**0.893**	**0.855**	**0.851**	**0.849**	**0.862**	**0.829**	**0.833**	**0.858**	0.867
RACLC	0.869		0.842		0.845		0.871		0.850		0.832		0.819		0.813	
FAC-BERT-50%-$p1$	0.876	0.848	0.851	0.845	0.853	0.826	0.871	0.873	0.834	0.794	0.826	0.844	0.794	0.806	0.832	0.854
FAC-BERT-50%-$p2$		0.850		0.844		0.822		0.869		0.791		0.846		0.805		0.853

Table 2: Performance evaluation. - if a model does not extract rationales

- SVM: An effective classification baseline for classification of crisis events [6, 13].
- Robust-CNN [25]: A Convolutional Neural Network based approach with pre-trained word embeddings for classification of crisis events.
- BERTweet [24]: BERTweet with a linear classification on top of the first [CLS] token embedding.
- LCL [35]: A classification model that relies on label-aware contrastive loss.
- BERT2BERT [27]: An interpretable by design approach for classification of crisis events. The model employs a multi-task learning strategy to train and predict class labels and rationales simultaneously. BERT2BERT(-2stg) is the variant BERT2BERT, which is not interpretable by design.
- RACLC [26]: A recent contrastive learning-based approach for classification of crisis events. It applies a contrastive multi-task learning approach to boost the performance of class label and rationale prediction tasks. RACLC(-2stg) is a variant of RACLC, which is not interpretable by design.

4.3 Evaluation Metrics

4.3.1 Groundtruth based evaluation. We evaluate how good our predicted class labels and rationales are compared to human annotations. For classification performance, we measure Macro-F1 score. Similarly, we measure the agreement between extracted rationales and human rationales using Token-F1 metric. First, Token-precision is computed to show the fraction of relevant rationale words among all predicted rationales. Next, Token-recall measures the fraction of correctly extracted rationale words among the total number of human rationale words. Then, we combine the two scores by taking their harmonic mean Token-F1.

4.3.2 Model Faithfulness. One might argue that a model can have a high agreement with human rationales (plausibility), but does not reflect the true internal reasoning. We measure to what extent the extracted rationales influence the model decision by using the following metrics.

Comprehensiveness [8]. This metric measures how much the classification performance drops when extracted rationales are removed/masked from the original inputs. Given X, R and $X \backslash R$ are original examples, predicted rationales (non-rationales are marked by '*'), and predicted non-rationales (rationales are marked by '*'), respectively. We compute comprehensiveness score as follows.

$$\textit{Comprehensiveness} = \textit{Macro-F1(X) - Macro-F1(X} \setminus \textit{R)}$$

The higher comprehensiveness shows the high influence of the predicted rationales on the classification performance.

Sufficiency [8]. This metric evaluates performance differences when using only rationales and the original input texts.

$$\textit{Sufficiency} = \textit{Macro-F1(X) - Macro-F1(R)}$$

The lower sufficiency is better since it shows that only predicted rationales are sufficient for a model to make predictions.

4.4 Model Details and Hyperparameters

We evaluate our model and all the baselines using a 5-fold cross-validation setup. At each run, we apply a stratified sampling method to obtain train/valid/test sets with ratios 70%/15%/15% respectively. All the baseline models are run with configurations from original papers. To train our FAC-BERT, we pre-process data by converting tweets to lowercase and removing mentions, URLs. Our method is trained for 10 epochs, and the batch size is 16. The GRU layer has a hidden size of 128. We optimize the model using AdamW optimizer [21] with a learning rate of 2e-5. Besides, we specify a list of candidates for the hyper-parameter λ and select the one that obtains consistently good performance (average Macro-F1 and Token-F1) with a 5-fold setting on validation sets. After fine-tuning, we set $\lambda = 0.5$ for all the datasets since it generally performs the best in the majority of cases across different validation runs. Another hyperparameter is k, i.e., the percentage of human-annotated rationales required to successfully train the model. We set $k = 50\%$ to compare performance with other baseline models and also observe FAC-BERT performance with varying k.

5 CLASSIFICATION RESULTS

This section presents the performance of our proposed approach. We consider both in-domain and cross-domain evaluation. In in-domain classification, training and testing data come from the same event. For cross-domain evaluation, we train the model on one dataset and apply it to another dataset of the same event type. For example, the model trained on NeQuake dataset is used to predict class labels and human rationales on MexQuake dataset. Recall that FAC-BERT consists of two-phase learning. In the first phase ($p1$), it predicts binary labels for tokens, i.e., whether a token is a rationale or not. The second phase ($p2$) learns faithful attention weights, it does not predict any binary classification of tokens (rationale/not rationale). To evaluate Token-F1 of the second phase, we extract the same number of rationale tokens as in the first phase prediction. Note that the objective of this paper is not to improve class labels

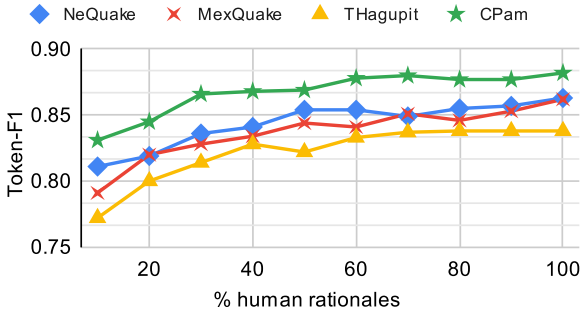

Figure 3: In-domain evaluation with various percentages of human rationales

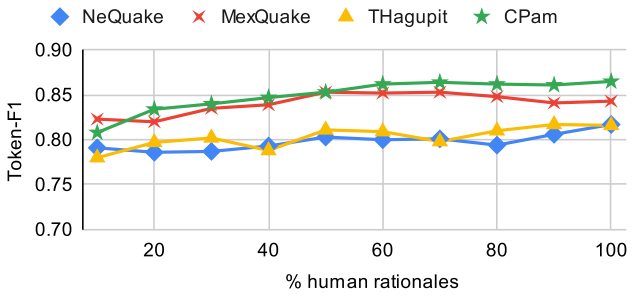

Figure 4: Cross-domain evaluation with various percentages of human rationales

or human rationale prediction tasks. Rather, we want to answer the following two questions:

(1) How well our FAC-BERT performs under limited human supervision?
(2) How to learn faithful machine attention?

In the following sections, FAC-BERT-k%-$p1$ and FAC-BERT-k%-$p2$ are used to indicate the performance of FAC-BERT with rationales extracted from the *first phase* and *second phase* respectively. The value k specifies the percentage of human rationales used during the training process.

5.1 In-domain evaluation

We evaluate the performance of classification models on the test set of the same event on which the models are trained on. Table 2 compares the performance between different classification models. We show the prediction results of FAC-BERT using $k = 50\%$ human annotated rationales. FAC-BERT achieves competitively equal Macro-F1 with other baselines such as BERT2BERT or RACLC. Compared to the best Token-F1 returned by RACLC [26] with 100% human rationale supervision, FAC-BERT that uses 50% human rationales only get a slight drop (i.e., $< 3\%$). It is also interesting that adding the regularization part in the loss function of phase 2 helps to align the rationale tokens learned in phase 2 with phase 1. That's why we get almost similar Token-F1 scores between the two phases.

5.2 Cross-domain evaluation

This section evaluates the classification performance when the prediction is made on a similar event dataset that was not used for training. We compare the cross-domain performance between FAC-BERT using 50% rationale supervision with baseline methods in Table 2. The Macro-F1 is equal to or slightly worse than some other baselines. Both the two phases of FAC-BERT return similar Token-F1 values. It is observed that FAC-BERT-50%-$p2$ got 6%, 1.6%, 2.8% and 2% drops than the best Token-F1 (RACLC) on NeQuake, MexQuake, THagupit and CPam respectively.

5.3 How does performance of FAC-BERT vary with budgeted human rationales (k)

Table 2 shows the performance of FAC-BERT with $k = 50\%$ human annotated rationales. In this section, we would like to explore the variation in model performance under different human budgets.

Variation in in-domain scenario: Figure 3 shows the Token-F1 scores of our second phase prediction with varying percentages of human rationales. Interestingly, when we use only 10% human rationale labels, we obtain pretty good performance (i.e., 83.1% Token-F1 on CPam dataset). The Token-F1 increases significantly when we vary the percentage of human rationales from 10% to 50%. Then, the Token-F1 gets improved slowly when more human rationales are added. The result shows that by using 10% or around 200 instances with human rationales, our FAC-BERT can obtain more than 75% Token-F1 for all the datasets. Besides, FAC-BERT obtains 80% Token-F1 for all the datasets when 20% or around 400 instances with human rationale supervision are used for training. Unlike previous studies [26, 27] only focus on performance improvement of both Macro-F1 and Token-F1, this evaluation gives a guideline on how much rationale data is needed to train a good interpretable classification model on crisis domain.

Varying the human rationales doesn't harm the performance of the humanitarian class label prediction task. FAC-BERT obtains similar Macro-F1 score as shown in Table 2 with different k% human rationales. Side by side, the Token-F1 performance also reaches a quite stable point with 50% human rationale labels. The gain is not significant beyond this point, and results only slightly improve when more human rationales are added.

Variation in cross-domain scenario: Similar to the in-domain scenario, we feed FAC-BERT with an increasing percentage of human rationales for supervision and observe the performance change in cross-domain. Figure 4 illustrates Token-F1 values extracted from the second phase of FAC-BERT. Using 10% human rationales archives more than 75% Token-F1. The rationale prediction results improve when more human rationales are added, but not as significant as in case of in-domain evaluation. Starting from 50%, adding more human rationales slightly boost the Token-F1.

5.4 Influence of the alignment between human rationales and machine attention

So far, we observe the variation in performance under different annotated rationale budgets. In this section, our objective is to learn the role of alignment regularizer in the loss function (Eqn. 5). We observe how our loss function with the alignment between rationale prediction and attention weight helps to improve the

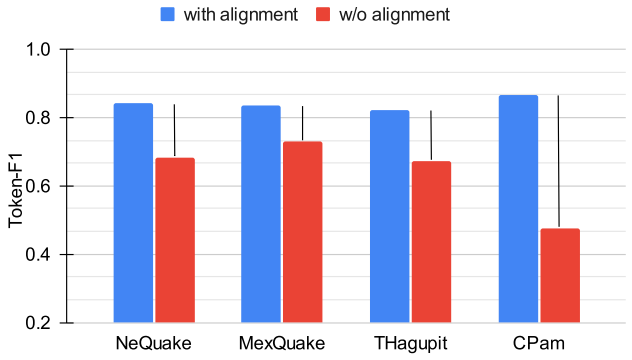

Figure 5: In-domain average Token-F1 with/without weights alignment in the loss function. Vertical black lines indicate drops in the performance/Token-F1.

faithfulness of machine attention. First, we take predicted rationales from the second phase learning (attention weight-based rationales), namely FAC-BERT-p2, with % human rationales vary in range $k \in [10, 20, .., 100]$. This is done for both cases, which are with and without distance alignment between attention weight and rationale probability in FAC-BERT loss function. Here, we report the average result over different k values. Figure 5 illustrates the impact in in-domain evaluation. We also obtain similar patterns in cross-domain evaluation. When there is no attention alignment in the loss function ($\lambda = 0$ in Section 3), the average Token-F1 score returned by FAC-BERT-p2 decreases significantly. More specifically, it drops by 10.9%, 14.5%, 16.2% and 39.3% on MexQuake, THagupit, NeQuake and CPam, respectively, compared to ones using attention alignment. Besides, the prediction of FAC-BERT-p2 without attention alignment varies greatly across datasets. It predicts rationales poorly on CPam dataset with an average of 47.4% Token-F1. We observe that without alignment, the standard attention might give high attention weights to unimportant words that are not supportive evidence for output labels. As an example, for the tweet "*50k children at risk in #vanuatu* ***after*** *devastation of* ***#cyclonepam*** ***. please help*** *respond …*", FAC-BERT-p2 *without attention alignment* correctly predicts the tweet as "affected people & evacuation"; however, it assigns the highest weights to words in bold. By using our regularized loss, the model reassigns the highest weights to the following bold words "***50k children at risk*** *in #vanuatu after devastation of #cyclonepam please help respond …*".

5.5 Model Faithfulness

We evaluate whether the extracted rationales can be seen as explanations for the output decision of FAC-BERT. There is a high overlap between the rationales obtained in two phases ($p1$ and $p2$). For brevity, we only show the comprehensive and sufficiency of predicted rationales from the second phase ($p2$), which is based on attention weights ($\lambda = 0.5, k = 50\%$). The scores are computed when we learn rationales from 50% rationale supervision. Generally, FAC-BERT obtains high comprehensiveness scores for all datasets. This illustrates the huge drop in Macro-F1 when the predicted rationales are removed from the original inputs. In other words, the predicted rationales are important for FAC-BERT to make decisions. Besides, the low sufficiency of FAC-BERT indicates that the predicted rationales alone are sufficient for FAC-BERT to classify tweets. Compared to the best classification model RACLC, which uses 100% human rationales, FAC-BERT obtains better comprehensiveness on two earthquake datasets. However, FAC-BERT has higher sufficiency than RACLC, but the difference is only less than 2%. By using 50% human rationale supervision, FAC-BERT attention is competitively faithful compared to RACLC using 100% rationale supervision.

Dataset	Comprehensiveness ↑		Sufficiency↓	
	RACLC	FAC-BERT	RACLC	FAC-BERT
NeQuake	0.352	0.378	-0.005	0.017
MexQuake	0.259	0.365	0.009	0.025
THagupit	0.349	0.265	0.007	0.018
CPam	0.403	0.352	0.017	0.002

Table 3: Comprehensiveness and Sufficiency

6 APPLICATION OF FAC-BERT IN DETECTION OF ACTIONABLE TWEETS

In this section, our objective is to explore the power of transfer learning over the related tasks of the same application area. In Section 5, we observed that FAC-BERT gives promising results under a given amount of annotated rationale data. This section tries to answer the question, "what would happen if we deploy the humanitarian classification model over actionable tweet detection?".

6.1 Data Collection

The recent Text Retrieval Conference (TREC) Incident Streams track [5] has released datasets for the classification of crisis-related tweets into fine-grained information types. Besides, the track identifies six actionable information types: Requests for Goods/Services, Requests for Search and Rescue, Calls to Action for Moving People, Reports of Emerging Threats, Reports of Significant Event Changes and Reports of Services becoming available. An 'actionable' tweet contains crucial information and immediate alert that might be useful for individuals and stakeholders to pay more attention. Those actionable information types/classes are the most difficult ones for classification models to predict due to the scarcity of labeled data.

We consider all tweets of earthquake or typhoon events from TRECIS 2021 training data [5]. Actionable tweets that belong to no more than one actionable class are selected. This set is quite small in number, which consists of 10% of the collected data. There are six actionable classes in the dataset. Apart from that, we randomly sample 100 tweets that do not contain any actionable labels for each event type and filter out the other tweets from our dataset. The details of the collected dataset are shown in Table 4. The last column shows the size of each class in our final actionable dataset. Generally, the dataset is quite imbalanced, classes such as 'EmergingThreats' or 'ServiceAvailable' have more tweets. Meanwhile, only a few tweets report information about 'MovePeople' or 'GoodsService'. This imbalance poses a challenge for classification models.

Information Type/Class	Earthquake	Typhoon	Total
ServiceAvailable	747	397	1144
SearchAndRescue	168	4	172
MovePeople	6	66	72
EmergingThreats	545	1632	2177
NewSubEvent	126	561	687
GoodsServices	56	45	101
Others	100	100	200

Table 4: A dataset of actionable information types

6.2 Actionable tweet classification using FAC-BERT

In this section, we study the application of our proposed model FAC-BERT in two aspects:

(1) How well FAC-BERT is able to extract actionable snippets from actionable tweets?
(2) How well our proposed FAC-BERT performs on a new dataset with a new problem setup?

The major bottleneck that hinders the direct application of FAC-BERT over actionable tweet classification is the nonavailability of human-annotated rationales. TREC-IS does not have rationale snippets of tweets. Hence, to answer the first question, we apply the idea of transfer learning, i.e., directly apply FAC-BERT-$p1$ on the actionable tweets to gather the rationales. We train the model using the 100% rationale dataset provided for the humanitarian class identification problem. We trained two different models for two different events (earthquake and typhoon), i.e., NEQUAKE and MEXQUAKE datasets are used to train (FAC-BERT-100%-$p1$) and extract the rationale snippets from the actionable tweets of an earthquake event. Similarly, typhoon datasets (THAGUPIT and CPAM) are used to train and extract rationale snippets from the actionable tweets of typhoon category. The two phases of FAC-BERT obtain similar rationale prediction performance; hence, we just simply predict rationales on the actionable dataset from the first FAC-BERT learning phase (FAC-BERT-$p1$).

Now, we have the actionable class labels of tweets, and rationale snippets for each of the tweets gathered using FAC-BERT-$p1$ trained on humanitarian class-related rationales. Next, we directly follow the model architecture (described in Section 3) to detect the class of actionable tweets and learn faithful attention-based rationale snippets. As we obtained the rationales through transfer learning, we used 100% of the rationale data in phase 1 and tried to align attention weights in phase 2. FAC-BERT will assign to each tweet an actionable class label.

6.3 Results and Evaluations

We evaluate the performance of FAC-BERT on the classification of actionable tweets under the same configuration as in Section 4.4. Our FAC-BERT obtains 0.599 Macro-F1 in classification of actionable tweets. This result is significantly better than the leaderboard performance of 0.2784 Macro-F1. Although it is not a fair comparison since the test set is different. However, the results suggest that the task itself is quite difficult, and transfer learning-based applications such as FAC-BERT would help in getting good performance and learning faithful attention-based rationales.

	X	X\R	R
Macro-F1	0.599	0.353	0.529

Table 5: Macro-F1 FAC-BERT - classification of actionable tweets with different input settings.

Faithfulness of Actionable Rationales: As we don't have any human annotation for rationales, we used transfer learning to gather the rationale snippets in actionable tweets. As mentioned above, we trained the models based on humanitarian class-based rationales and retrieved the rationales for actionable tweets. Here, we evaluate "how well our FAC-BERT is able to extract actionable snippets from the actionable tweets". For that, we simply consider rationales extracted by the FAC-BERT-$p1$ trained on previous earthquake or typhoon and feed the second learning phase with three different input settings when classifying actionable tweets, which are original texts (X), input texts with rationales marked by '*' (X\ R) and input texts with non-rationales marked by '*'(R). This is similar to comprehensiveness or sufficiency evaluation.

Table 5 shows that when we mask out rationales, Macro-F1 score significantly decreases (i.e., from 59.9% to 35.3%). That means the zero-shot predicted rationales cover important content of the original tweets that FAC-BERT relies on to make predictions. Besides, when we replace non-rationales with a wild character '*', FAC-BERT performs slightly worse than the setting with original input texts. Using only rationales is sufficient to obtain decent performance.

7 CONCLUSIONS

This paper introduces a faithful attention-based classification model. We learn to derive high-quality and faithful attention heatmaps from little human rationale supervision. We conduct experiments on different datasets of short texts from microblogs during crisis events. Experimental results show that our attention heatmaps are highly aligned with human rationales. Besides, the learned attention weights can be considered as faithful explanations, which effectively reflect the reasons for the model's decision. We also vary the size of human rationales supervision to observe the effectiveness of our model in both in-domain and cross-domain classification. Further, we also show the application of our proposed model in a new setup, i.e., the detection of actionable tweets and actionable snippets. As the next step, we will evaluate the faithfulness of our attention-based explanations as a gray-scale measure of attention weights using decision flips. Besides, we aim to investigate and improve the faithfulness of attention-based explanation with a zero-shot learning setup (i.e., without human rationale supervision). We believe this kind of zero-shot learning setup helps in contributing data and new problems to TREC-IS [36] and crisisFACTS [37] tracks.

ACKNOWLEDGMENTS

This work is 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. 101021866. Further, this work is supported in part by the Science and Engineering Research Board, Department of Science and Technology, Government of India, under Project SRG/2022/001548. Koustav Rudra is a recipient of the DST-INSPIRE Faculty Fellowship [DST/INSPIRE/04/2021/003055] in the year 2021 under Engineering Sciences.

REFERENCES

[1] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*. 465–473.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[3] Francesco Barbieri, Luis Espinosa Anke, Jose Camacho-Collados, Steven Schockaert, and Horacio Saggion. 2018. Interpretable emoji prediction via label-wise attention LSTMs. In *Proceedings of the 2018 conference on empirical methods in natural language processing*. 4766–4771.

[4] Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is Attention Explanation? An Introduction to the Debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3889–3900.

[5] Cody Buntain, Richard McCreadie, and Ian Soboroff. 2022. Incident Streams 2021 Off the Deep End: Deeper Annotations and Evaluations in Twitter. In *19th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2022, Tarbes, France, May 22-25, 2022*. 584–604.

[6] Mark A. Cameron, Robert Power, Bella Robinson, and Jie Yin. 2012. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume)*. 695–698.

[7] George Chrysostomou and Nikolaos Aletras. 2021. Improving the faithfulness of attention-based explanations with task-specific information for text classification. *arXiv preprint arXiv:2105.02657* (2021).

[8] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron 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 2020, Online, July 5-10, 2020*. 4443–4458.

[9] Maria Jose Gacto, Rafael Alcalá, and Francisco Herrera. 2011. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences* 181, 20 (2011), 4340–4360.

[10] Reza Ghaeini, Xiaoli Z Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. *arXiv preprint arXiv:1808.03894* (2018).

[11] Tuan-Anh Hoang, Thi-Huyen Nguyen, and Wolfgang Nejdl. 2019. Efficient Tracking of Breaking News in Twitter. In *Proceedings of the 11th ACM Conference on Web Science, WebSci 2019, Boston, MA, USA, June 30 - July 03, 2019*. 135–136.

[12] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. 2014. AIDR: artificial intelligence for disaster response. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*. 159–162.

[13] 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 Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*.

[14] 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: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. 3543–3556.

[15] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–38.

[16] Jens Kersten, Anna M. Kruspe, Matti Wiegmann, and Friederike Klan. 2019. Robust filtering of crisis-related tweets. In *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, València, Spain, May 19-22, 2019*.

[17] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[18] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 131–138.

[19] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. 2022. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (2022), 1–59.

[20] Junhua Liu, Trisha Singhal, Lucïënne T. M. Blessing, Kristin L. Wood, and Kwan Hui Lim. 2021. CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding. In *HT '21: 32nd ACM Conference on Hypertext and Social Media, Virtual Event, Ireland, 30 August 2021 - 2 September 2021*. 133–141.

[21] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

[22] Sukanya Manna and Haruto Nakai. 2019. Effectiveness of Word Embeddings on Classifiers: A Case Study with Tweets. In *13th IEEE International Conference on Semantic Computing, ICSC 2019, Newport Beach, CA, USA, January 30 - February 1, 2019*. 158–161.

[23] Richard McCreadie, Cody Buntain, and Ian Soboroff. 2020. Incident Streams 2019: Actionable Insights and How to Find Them. In *17th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2020, May 2020*. 744–760.

[24] 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, EMNLP 2020 - Demos, Online, November 16-20, 2020*. 9–14.

[25] Dat Tien Nguyen, Kamla Al-Mannai, Shafiq R. 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 Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*. 632–635.

[26] Thi Huyen Nguyen and Koustav Rudra. 2022. Rationale Aware Contrastive Learning Based Approach to Classify and Summarize Crisis-Related Microblogs. In *CIKM 2022: The 31st ACM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, October 17 - 22, 2022*.

[27] Thi Huyen Nguyen and Koustav Rudra. 2022. Towards an Interpretable Approach to Classify and Summarize Crisis Events from Microblogs. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*. 3641–3650.

[28] Thi Huyen Nguyen, Miroslav Shaltev, and Koustav Rudra. 2022. CrisICSum: Interpretable Classification and Summarization Platform for Crisis Events from Microblogs. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*. 4941–4945.

[29] Petra Perner. 2011. How to interpret decision trees?. In *Industrial Conference on Data Mining*. Springer, 40–55.

[30] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913* (2019).

[31] Koustav Rudra, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 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*. 583–592.

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

[33] Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 2931–2951.

[34] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[35] Varsha Suresh and Desmond C Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. *arXiv preprint arXiv:2109.05427* (2021).

[36] TREC. 2020. TREC Incident Streams: Enabling emergency services with social media data. https://www.dcs.gla.ac.uk/~richardm/TREC_IS/

[37] TREC. 2022. Crisis FACTS. https://crisisfacts.github.io/

[38] Martin Tutek and Jan Šnajder. 2020. Staying true to your word:(How) can attention become explanation? *arXiv preprint arXiv:2005.09379* (2020).

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). http://arxiv.org/abs/1706.03762

[40] Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James Martin, Martha Palmer, Aaron Schram, and Kenneth Anderson. 2011. Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. 385–392.

[41] Congcong Wang, Paul Nulty, and David Lillis. 2021. Transformer-based Multi-task Learning for Disaster Tweet Categorisation. In *18th International Conference on Information Systems for Crisis Response and Management, ISCRAM 2021, Blacksburg, VA, USA, May 2021*. 705–718.

[42] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. 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 2019, Hong Kong, China, November 3-7, 2019*. 11–20.

[43] 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*. 418–426.