

Recognizing Social Cues in Crisis Situations

Di Wang^{*†}, Yuan Zhuang^{*†}, Marina Kogan[†], Ellen Riloff[‡]

[†]Kahlert School of Computing, University of Utah

[‡]Department of Computer Science, University of Arizona

{orpheus, yyzhuang, kogan}@cs.utah.edu, riloff@cs.arizona.edu

Abstract

During crisis situations, observations of other people’s behaviors often play an essential role in a person’s decision-making. For example, a person might evacuate before a hurricane only if everyone else in the neighborhood does so. Conversely, a person might stay if no one else is leaving. Such observations are called *social cues*. Social cues are important for understanding people’s response to crises, so recognizing them can help inform the decisions of government officials and emergency responders. In this paper, we propose the first NLP task to categorize social cues in social media posts during crisis situations. We introduce a manually annotated dataset of 6,000 tweets, labeled with respect to eight social cue categories. We also present experimental results of several classification models, which show that some types of social cues can be recognized reasonably well, but overall this task is challenging for NLP systems. We further present error analyses to identify specific types of mistakes and promising directions for future research on this task.

Keywords: Corpus, Social Media Processing, Other

1. Introduction

During natural disasters and catastrophic events, people frequently use social media platforms to share information about what is happening. Consequently, researchers in the field of *crisis informatics* develop methods to enable near real-time information gathering from social media during crisis situations, which is essential for government officials and affected individuals to respond as quickly and effectively as possible. Historically, crisis informatics research has prioritized collecting information that helps individuals, organizations, and authorities make informed decisions and respond appropriately to emergency situations (Vieweg, 2012; Verma et al., 2011; Zade et al., 2018). Thus far, most research has focused on identifying information that supports the operational needs of government officials and first responders.

In comparison, little research has focused on automatically identifying information that affects people’s *perception* of a situation and their individual decision-making. According to theoretical risk analysis research (Lindell and Perry, 2012), an individual’s response to a natural disaster or emergency situation is often based on cues from social contexts and the environment, as well as warning messages transmitted through communication channels.

In this paper, we aim to identify *social cues*, which are observations of other people’s behavior that can affect a person’s perception of an event and their behavior. For example, when someone learns that their neighbors have evacuated in preparation for

a natural disaster, they are more likely to consider evacuation for themselves. Similarly, if one hears that other people are stocking up on food and water, they are more likely to buy supplies themselves.

With the emergence of social media, one’s observation of others also includes the observation of other people’s social media activities. This suggests that even when people do not directly observe other people’s behavior, their observation of other people’s social media activity can serve as social cues for their own risk assessment and decision-making. Social cues are also informative for officials to better understand the public’s risk perception and probable behavioral response.

The goal of our research is to introduce a new natural language processing task to automatically identify statements on social media that can serve as a *social cue* during a crisis situation. This task is interdisciplinary in nature, as our definition of a social cue is based on prior work in social science and crisis informatics. Specifically, we define eight social cue categories that capture different types of behaviors and reactions that can influence other people’s perceptions and decision-making: *Apathy*, *Change of Plans*, *Emotional Reaction*, *Fact Sharing*, *Official Directives*, *Physical Actions*, *Suggestions*, and *Other*. We create an annotated dataset of 6,000 tweets that were posted before an expected hurricane and manually labeled them with respect to these eight social cue categories. We then trained several types of classification models to automatically assign these social cue labels to tweets and evaluated their performance. We find that some social cue categories can be identified reasonably well, while others are more challenging and warrant

* These authors contribute equally.

further research.

Overall, we make the following contributions:

- We propose a new task of identifying social cues in social media posts. We define eight types of social cue categories and introduce a new dataset of 6,000 tweets that are manually annotated for social cue categories. The dataset is made publicly available at <https://github.com/yyzhuang1991/Crisis-Social-Cues>.
- We create classification models to automatically label social media posts with respect to the eight social cue categories.
- We present experimental results to evaluate classification performance on this task, and analyze the results to provide insights on promising directions for future work on this challenging problem.

2. Background

2.1. Crisis Informatics

Crisis informatics is an interdisciplinary field at the intersection of social science and computing that explores the use of Information and Communication Technologies (ICTs) during emergencies (Hagar and Haythornthwaite, 2005; Palen et al., 2009). Specifically, crisis informatics research highlights the importance of the information generated and shared by the affected populations (Hughes et al., 2008; Liu et al., 2008; Palen et al., 2009; Sutton et al., 2008). Studies often focus on information supporting situational awareness (Zade et al., 2018; Vieweg, 2012; Verma et al., 2011), which refers to the ability of responders, decision-makers and the public to understand what is happening in a crisis situation and how it is evolving. However, such emphasis on information supporting situational awareness has generally prioritized the information that could support the operational needs of government officials or first responders. Yet, the information that could benefit the affected individuals in their risk perception has not been studied quantitatively.

2.2. Social Cues

According to theoretical risk analysis research (Lindell and Perry, 2012), an individual's risk perception during a disaster is mainly influenced by information from official warnings, environment cues (e.g. wind gusts, flooding, street closures) and social cues. Among them, social cues are the observations of others' behavior that could inform people's decision-making about how to stay safe or take protective

actions in crisis (Lindell and Perry, 2012). For example, noticing that neighbors are evacuating is an important social cue that may affect whether people decide to evacuate themselves. Social cues have long been recognized as an important aspect of people's risk assessments (Lindell and Perry, 2012; Demuth et al., 2018; Mileti and Sorensen, 1990; Golding and Krinsky, 1992; Renn, 2008) that lead to appropriate decisions. As social cues can potentially amplify or attenuate risk assessments, they also reflect the prominence of environmental risks (Kasperson et al., 1988, 2003).

Compared with other information types deemed important in the crisis informatics literature, such as information aiding in situational awareness, social cues are more focused on the social environment. Unlike environmental cues, social cues do not always contain actionable information that is most often associated with situational awareness. Yet, they contain information about what others are doing and therefore, directly or indirectly, about how other people are perceiving the situation. Disaster sociology (Dynes, 1969) and crisis informatics have long recognized that actions and perceptions of others, especially those socially proximate, have a substantial effect on people's perception of risk and deployment of protective actions (Kinatader et al., 2018; Metaxa-Kakavouli et al., 2018; Fu et al., 2021). Thus, identifying social cues will further benefit individuals in assessing their social environment and support them in protective decision-making during emergencies.

In the context of crisis communication on social media, previous work has exemplified three key types of social cues through Twitter narratives: cues from peers, cues from businesses, and cues from government (Demuth et al., 2018). However, these three types of social cues did not incorporate the online context. For example, when people observe other people's activities on social media, their risk perception would also be affected. Such information was not well captured by existing annotations for information types (Zahra et al., 2020; Mazloom et al., 2018; Imran et al., 2016; McCreadie et al., 2019; Alam et al., 2021) or classification tasks. Thus, in this paper we extend the definition of social cues to incorporate the online conversations around the environmental hazards and propose a classification task for categorization of social cues.

2.3. Related NLP Tasks

While distinct, our task is related to several NLP tasks. For example, identifying the *Emotional Reaction* category requires understanding the speaker's opinions (Pang et al., 2002; Hu and Liu, 2004; Pak and Paroubek, 2010; Liu et al., 2015) and emotions (Alm et al., 2005; Strapparava and Mihalcea, 2008; Mohammad and Turney, 2010) towards

crises. Speech acts recognition (Cohen et al., 2004; Jeong et al., 2009) is also relevant to our task, as requests and directives are prevalent in tweets in the *Official Directives* and *Suggestions* categories. In addition, social cues in tweets are usually represented by mentions of positive or negative situations during crisis (e.g., “the school is canceled,” “gas price soars”), which have been previously studied by researchers (Deng and Wiebe, 2014, 2015; Ding and Riloff, 2016, 2018; Zhuang et al., 2020; Zhuang and Riloff, 2023).

3. Social Cue Identification Task

We propose a new task of identifying social cues in social media posts related to crisis situations. Social cues can affect how individuals assess their own risk and adapt their behavior. For example, social cues may affect people’s decisions about whether to evacuate, stock up on food, go to work, and board up windows in their home, etc. Social cues also help government officials predict how people are likely to behave, which is important for effective emergency management. For example, social cues can help government organizations and first responders channel their resources most effectively as a disaster unfolds, or to adapt official messaging to promote different behaviors (e.g., more forcefully emphasize the need to evacuate), or to provide emotional support to boost the morale of affected populations.

In this section, we define our classification task in terms of eight social cue categories. We also describe the creation of a new dataset of 6,000 tweets related to Hurricane Harvey and discuss our manual annotation effort to label these tweets with respect to the social cue categories.

3.1. Data Collection

To create a dataset for this task, we collected English tweets that were posted prior to or during Hurricane Harvey, which was a major Hurricane in the 2017 Atlantic hurricane season in the U.S. We collected the data with Twitter’s historical PowerTrack API using hurricane-related keywords and location names in the affected area. We focused on the period between August 24 and August 26 to include people’s preparation period before Harvey’s landfall on August 25 as well as people’s response shortly after landfall. We then focused on the tweets containing geo-location data, to select tweets that came from the affected area, as determined by a bounding box of the affected region. This process resulted in 34,113 local relevant tweets, of which we randomly sampled 6,000 for manual annotation.

3.2. Social Cue Category Definitions

Our work was motivated by prior work in crisis informatics that studied risk assessment and risk communication (Demuth et al., 2018), where social cues were defined as observations of other people’s behavior and other information from the social environment. We used results from that research as the starting point for our social cue definitions. One key distinction, however, is that earlier work defined social cues in terms of direct observations of other people’s behavior. Our research extends the definition of social cues to include descriptions of other people’s behavior and feelings as discussed on social media.

The first two authors then manually coded 100 random samples based on whether the message represents information from the social environment such as cues from people (e.g., family, friends, neighbors), business (e.g., affected stores, business closures, shops where people go to stock up on goods) and government (e.g., facilities affected by government orders, emergency response organizations, etc.). After a series of joint discussion sessions between all authors, the definition of social cues was updated to focus on information that can serve as a potential signal that could affect other people’s perception of the disaster situation (e.g., hurricane) and their decision-making. Then through iterative coding of 500 random samples and a series of joint discussion sessions, the coding scheme for social cues was further defined, refined and calibrated over the course of several months. Finally, the coding rules were adjusted to capture social cues as they occur in online contexts. The random samples used to develop the coding scheme were not used for later experiments. Our final annotation guideline could be found at <https://github.com/yyzhuang1991/Crisis-Social-Cues>.

In total we identified eight different types of social cues, which are summarized below. In addition, Table 1 includes two tweets for each category to show examples of how social cues are expressed in natural language.

Physical Actions: activities that are a direct response to a disaster situation. For example, evacuating one’s home or calling 911.

Emotional Reaction: emotions expressed toward the disaster event or an event directly resulting from the disaster. For example, expressions of anxiety about an impending storm.

Apathy: statements which suggest that one is unconcerned about or not taking a disaster seriously. For example, dismissing need to evacuate.

Change of Plans: statements indicating that someone has changed their plans because of an impending or current disaster. For example, cancelling a planned trip.

SOCIAL CUE CATEGORY	EXAMPLE TWEET
Physical Actions Actions as a response to the disaster. This includes both protective actions and coping behavior by humans.	T1: First time I actually had to buy supplies for the oncoming hurricane/tropical storm. 2 stores were actually out of water... T2: Hunker down for Harvey, y'all. We're stocking up on water, food & cerveza!
Emotional Reaction Emotional expressions or reactions about the disaster. The target of the emotion needs to be related to the disaster. If a tweet has emotion but also belongs to another category, prioritize the other category.	T3: Harvey is so getting on my nerves and messing with my education 😞😞😞 T4: can't sleep because of this hurricane shit. I just want my wife 😞
Apathy Information showing that people do not prioritize protective actions, do not care about the event or explicitly mention they will not do anything as a response to the disaster.	T5: I honestly have no worries about this hurricane knowing Houston it's gonna make a sharp right before it gets to us just watch. T6: While people continue to get all worked up over this #Harvey I'm going to go enjoy a beer. <url >
Change of Plans People/organizations change their plans as a response to the disaster.	T7: We will be closed this weekend. Hopefully we will reopen on Monday with normal business hours. Stay safe. #hurricaneharvey #houheights <url > T8: Hurricane Harvey got SAT dang near shut down!!! ✈️ #hurricaneharvey
Official Directives Information that reveals decisions or actions from official sources such as evacuation orders, absence of evacuation orders, etc. This does not include factual reporting.	T9: Aransas Pass under mandatory evacuation, #Harvey expected to hit as a category 3. <url > T10: FEMA has granted Governor Abbott's Request for Presidential Disaster Declaration.@PO-TUS <url>#HurricaneHarvey #Harvey
Suggestions Recommendations of actions as a response to the disaster. This does not include factual reporting or commercial promotions.	T11: Charge all phones and check all flashlights now people... San Antonio should be in final steps to prepare for this event #harvey T12: Galveston get your gas and supplies now. Lots of rain for us <url>
Fact Sharing Factual information that conveys how other people are preparing or responding to the disaster (without actions).	T13: Water is GONE at Kroger on Westheimer and Elmside #HurricaneHarvey @KHOUweather @KHOU <url> T14: wanna know what a hurricane in Houston is like? our gas stations are out of gas. OUT.
Other Other social cues such as status updates, or explicit risk perceptions, etc.	T15: Probably going to get swept away in the hurricane. If so, luv you all T16: Hurricane is supposed to be bad stay safe ! And don't forget your pets ❤️ They're fam too

Table 1: Social Cue Categories and Example Tweets

Official Directives: messages from official sources giving orders, guidance, or actionable information regarding a disaster situation. Also included in this category: social media posts by individuals that are sharing messages from official sources. For example, announcing evacuation orders.

Suggestions: statements by individuals (not official sources) that contain actionable advice. For example, advising others to finish tying down their boats because the winds seem to be picking up.

Fact Sharing: This category generally captures

factual information that describes what other people are doing in preparation for or in response to a disaster. For example, someone might report that lines at the gas station are extremely long.

Other: There are some messages that explicitly reveal how people are behaving or perceiving a disaster situation, but do not fall into the previous categories. The *Other* category includes miscellaneous additional types of social cues. For example, someone might update their status on social media to indicate that their home was not affected by

flooding. As we will show in the next section, a relatively small number of tweets fall into the *Other* category, so the majority of social cues are covered by the seven more specific categories.

Besides the eight social cue categories, we also define another category, **NotSC**, for tweets that do not contain any social cues. This resulted in nine distinct categories for annotation.

Overall, we found that only a small percentage of the instances exhibited multiple social cues. Furthermore, most of these cases involved the *Emotion* category along with another type of social cue. Emotion (or more generally, sentiment) recognition is a separate task that has been well-studied in NLP, so we decided to prioritize the other types of social cues in our work. For this reason, our annotation guidelines specify that a tweet that exhibits a non-Emotion social cue should be labeled for that social cue category, even if it also expresses an emotion. With this decision, we felt that most tweets did warrant just a single social cue label.

However, there are still occasional tweets that exhibit multiple non-Emotion social cues. For these cases, we asked the annotators to select what they judged to be the primary social cue based on the annotation guidelines.

3.3. Gold Standard Annotations

The first two authors independently labeled the same set of 1,000 tweets from our dataset. We measured their inter-annotator agreement (IAA) using Cohen’s Kappa (McHugh, 2012). The resulting IAA score was $\kappa = 0.81$, indicating strong agreement between the two annotators. The annotators then adjudicated their disagreements, and we used these 1,000 labeled tweets as our *test set*. We then asked each annotator to annotate 2,500 tweets to collect an additional 5,000 labeled tweets. Finally, we randomly partitioned these 5,000 labeled tweets into a *training set* of 4,000 tweets (80%) and a *development set* of 1,000 tweets (20%).

Our final annotated dataset contains 3,992 (66.5%) *NotSC* tweets and 2,008 (33.5%) social cue tweets. Table 2 shows the distribution of social cue categories among the social cue tweets. The most common type of social cue is *Emotional Reaction* and many tweets contain social cues pertaining to *Physical Actions* and expressions of *Apathy* toward the hurricane. We make our dataset publicly available at <https://github.com/yyzhuang1991/Crisis-Social-Cues>.

4. Classification Models

Our task is to categorize a tweet into one of the 9 categories (8 social cue categories plus the *NotSC* category). We developed two different types of

Category	Count	Percentage
<i>Emotional Reaction</i>	815	41.9%
<i>Physical Actions</i>	342	17.6%
<i>Apathy</i>	309	15.9%
<i>Change of Plans</i>	192	9.9%
<i>Suggestions</i>	120	6.2%
<i>Fact Sharing</i>	87	4.5%
<i>Official Directives</i>	80	4.1%
<i>Other</i>	63	3.1%

Table 2: Distribution of Social Cue Categories

classification models for this task. First, we created classification models by fine-tuning large language models on our gold standard training data. We explored the use of three different language models for this task.

Second, we investigated the idea of creating pipelined models that consist of a binary classifier (is the tweet a social cue or not) followed by a multi-class classifier that further categorizes the tweets that were predicted to be social cues. These classifiers were also fine-tuned on our gold training data, and we created classifiers using three different language models.

4.1. Data Preprocessing

We preprocess a tweet before feeding it into a classification model. To preprocess a tweet, we first use the preprocessing pipeline provided by (Nguyen et al., 2020) to tokenize the tweet and replace usernames and urls with special tokens.¹ We found that emojis could sometimes contribute to social cues. For example, an emoji of a sad face could indicate emotions, and an emoji of food could indicate preparation actions. To utilize emojis, we convert each emoji into the corresponding textual description using the python library *emoji*, and insert the special tokens $\langle \text{EMOJI} \rangle$ and $\langle / \text{EMOJI} \rangle$ around the tokens to indicate that the tokens describe an emoji.² As hashtags could also indicate social cues (e.g., *#preparation* might imply preparation actions), we add special tokens $\langle \text{HASHTAG} \rangle$ and $\langle / \text{HASHTAG} \rangle$ around each hashtag to indicate that the token is a hashtag. We also replace the word “Harvey” with the special token $\langle \text{HURRICANE} \rangle$ so the learned model could generalize to data of other hurricanes in the future. In addition, we add special tokens $\langle \text{OTHER-HURRICANE} \rangle$ $\langle / \text{OTHER-HURRICANE} \rangle$ around each mention of other hurricanes such as Katrina and Irma to indicate that

¹The codes could be found at <https://github.com/VinAIRResearch/BERTweet>

²The *emoji* library could be found at <https://pypi.org/project/emoji/>

other hurricanes are mentioned.³ Finally, we add the special tokens [CLS] and [SEP] to the beginning and the end of the tweet.

4.2. Fine-tuned Language Models

Our first set of classification models were created by fine-tuning three language models with our gold standard training data:

1. BERT-base (Devlin et al., 2019)
2. RoBERTa-base (Liu et al., 2019)
3. BERTweet-base (Nguyen et al., 2020), which is a RoBERTa-based language model pre-trained specifically over English tweets.

During the fine-tuning, we first encode the input sentence with the language model. Then we pass the embedding of CLS token through a linear classifier layer to perform classification.

4.3. Pipeline Classification Models

In addition, we experimented with pipeline systems. Specifically, a pipeline system contains two steps. In the first step, we fine-tune a language model to identify whether a tweet contains a social cue, which is a binary classification task. In the second step, another language model is fine-tuned to label tweets identified to contain social cues in the first step with one of the eight social cue categories.

By subdividing the task into two sub-tasks, a pipeline system might better capture the difference between social cues and *NotSC* in the first step and improve the overall performance. In our experiments, we explored pipeline systems with the above language models: BERT, RoBERTa and BERTweet. We consistently used the same language model in the two steps in each pipeline system.

5. Evaluation

In this section, we evaluate the performance of the classification models on social cue categorization, and we also present a variety of analyses to better understand the strengths and weaknesses of the models. For our experiments, we report the micro- and macro-averaged Precision, Recall and F1 scores over the test set, averaged across 3 different runs.

In our experiments, we set the batch size as 32 and the sequence length as 128 for all classification models. For the learning rate, we explored 1e-5, 2e-5 and 3e-5. For the number of training epochs, we

explored 10, 15 and 20. We chose the values based on the model performance over the development set.

5.1. Experimental Results

Method	Macro			Micro		
	Pre	Rec	F1	Pre	Rec	F1
Pipelined Model						
<i>BERT</i>	56.8	47.7	50.4	73.4	73.4	73.4
<i>RoBERTa</i>	52.1	47.9	49.34	72.2	72.2	72.2
<i>BERTweet</i>	56.8	57.5	56.5	76.4	76.4	76.4
Single Classifier						
<i>BERT</i>	57.8	53.0	54.5	74.4	74.4	74.4
<i>RoBERTa</i>	54.3	52.8	53.0	72.9	72.9	72.9
<i>BERTweet</i>	63.4	58.3	59.3	79.1	79.1	79.1

Table 3: Experimental Results

We present the experimental results in Table 3. The rows 1-3 show the performance of pipeline models with BERT, RoBERTa and BERTweet. The best system among them is the pipeline with BERTweet, which achieves a 56.5% macro-F1 and a 76.4% micro-F1 scores. The main improvement is due to the substantial increase in the macro-Recall score (by at least 9.6 absolute points).

The rows 4-6 show the performance of single classifiers, where a single language model is fine-tuned to perform classification with the nine categories. Overall, the single fine-tuned BERTweet achieves the best performance among all systems, reaching a 59.3% macro-F1 and a 79.1% micro-F1 scores. Compared to other single classifiers, it improves both precision and recall scores substantially. Compared to the second best model (the pipeline with BERTweet), the main improvement is in the precision score. Nevertheless, its performance is still far from perfect. This suggests that our task is challenging and there is still much room for improvement.

Table 3 also suggests that fine-tuning a single language model is always better than the corresponding pipeline model for our task. With BERT and RoBERTa, the pipeline models produce slightly lower macro-Precision (by 1-2 points) but substantially lower macro-Recall (by 5.3 and 4.9 points correspondingly) than the corresponding single classifiers. This is probably because these two pipeline systems under-label much less social cues in the first step than the single classifiers, which results in much fewer identified social cues overall. Conversely, the pipeline with BERTweet produces a slightly lower macro-Recall score (by 0.8 point) but a much worse macro-Precision score (by 6.6 points) than the single fine-tuned BERTweet. This indicates that the pipeline model with BERTweet over-labels social cues in the first step. Overall, the experimental results show that breaking up our task into two sub-tasks does not help the model

³To match names of other hurricanes, we collected the names of hurricane between 1950 and 2017 from https://www.aoml.noaa.gov/hrd/hurdat/All_U.S._Hurricanes.html

better capture the difference between social cues and *NotSC*.

To better understand the model performance, we also show the breakdown scores of our best model, the single fine-tuned BERTweet, for each category in Table 4. The first two rows show the model performance of the binary classification between social cues and *NotSC*⁴. The rest of Table 4 shows the breakdown scores for the eight social cue categories. Among all social cue categories, the classifier achieves the best F1 score of 83.5% for *Change of Plans*. However, the F1 scores for all other categories are all below 70%, indicating that the classifier still struggles with most of the social cue categories. For *Change of Plans*, *Official Directives*, and *Emotional Reaction*, the recall scores are higher than the corresponding precision scores. For *Suggestions*, *Physical Actions*, *Apathy*, *Fact Sharing*, and *Other*, the recall scores are lower than the corresponding precision scores. Interestingly, we observe that the performance of a social cue category is not always influenced by its size in the dataset and the classifier achieves better performance for some social cue categories with less training data. For example, the performance on *Change of Plans* and *Official Directives* is much better than the performance on *Physical Actions* and *Apathy*, even if the training data in the former two categories is much less according to Table 2.

Category	Pre	Rec	F1
BINARY			
<i>NotSC</i>	87.8	89.2	88.5
<i>Social Cue</i>	79.4	77.0	78.1
BREAKDOWN			
<i>Change of Plans</i>	81.0	86.3	83.5
<i>Official Directives</i>	64.4	75.0	69.1
<i>Emotional Reaction</i>	61.9	69.2	65.3
<i>Suggestions</i>	62.7	57.9	60.1
<i>Physical Actions</i>	57.9	51.6	54.6
<i>Apathy</i>	56.3	51.1	53.6
<i>Fact Sharing</i>	64.5	37.0	46.9
<i>Other</i>	34.4	7.7	12.3

Table 4: Breakdown of Model Performance.

5.2. Analysis

To understand our model’s behavior, we manually analyzed the predictions of the BERTweet-based classifier. First, Table 5 shows examples of tweets that were correctly labeled. As shown in the previous section, the classifier performs best on the *Change of Plans* and *Official Directives* categories. We observed that these types of social cues are

Emotional Reaction: <i>Taylor flooded so bad 2 years ago from a regular storm. This hurricane kinda got me worried bout my folks at home when it hits.</i>
Physical Actions: <i>Car is filled up with gas, have plenty of water, a bottle of wine and food in the pantry! #HurricaneHarvey #StormWatch #flood</i>
Apathy <i>Hurricane Harvey coming, but I’m still gonna go to the gym.</i>
Change of Plans: <i>My first day of classes for graduate School were canceled due to hurricane Harvey.</i>
Suggestions: <i>Be safe and seek shelter from #Harvey. @KHOU has steady updates.</i>
Fact Sharing: <i>Either the gas stations around here have no gas, or you’ll wait in line for 2 hours just to get a drop of gas. #hurricaneharvey</i>
Official Directives: <i>Evacuation orders issued in advance of Hurricane Harvey via @ABC13Houston</i>

Table 5: Correctly Assigned Social Cues

often expressed in similar ways. For example, *Change of Plans* tweets are likely to use words such as “cancel”, “close” or “delay”. Many *Official Directives* tweets share common syntactic constructions (e.g., “X-order is issued” or “Person-X urges Action-Y”). As a result, the model does not require a large amount of training data to learn these categories.

In the following sections we analyze the classifier’s mistakes in terms of two types of errors: 1) over-labeling social cues; 2) under-labeling social cues. We omit the category of *Other* due to its small size in the dataset.

Over-labeling Social Cues First we examined cases where the classifier predicted a social cue category but was wrong. The majority of these cases were tweets that did not contain any social cue at all (i.e., the gold label is *NotSC*).

One common case is labeling too many tweets as an *Emotional Reaction* social cue. The classifier tends to assign this label when a tweet contains any emotional expression, even if the emotion is not directed toward Hurricane Harvey. For example, the following tweet was mislabeled as an *Emotional Reaction* social cue:

Not happy with her new home away from home. No cats to chase and the cows here are anti

⁴We calculated the binary classification scores based on the predictions with the 9 categories.

social #HurricaneHarvey

This tweet mentions the hurricane in a hashtag, but the emotion “*not happy*” is directed towards *her new home* rather than *Hurricane Harvey*.

We also found that over-labeling frequently occurred with the *Apathy* category. The *Apathy* social cue is warranted when people mention that they are prioritizing eating, drinking or entertainment over preparations for a looming disaster (e.g., “... *nothing keeps me from going to my local bar #Harvey2017 #HurricaneHarvey*”). But the classifier seems to have overgeneralized and often predicts an *Apathy* social cue when food, drink, or entertainment is mentioned in any context. For example, the tweet below was mislabeled as *Apathy* likely because it mentions the phrase “Super Bowl”:

@USER is in yet another Super Bowl for Weather fans! Great job Cantore! @USER is not leaving my ATX side tonight.

There are also cases where the classifier confuses different social cue categories, but this is much less common. Two categories that sometimes get confused are *Apathy* category and *Physical Action*, probably because activities mentioned in *Apathy* tweets can sometimes look like disaster preparation actions. One example is the tweet below, which is a true *Apathy* social cue:

This so true my daddy just bought 3 24 packs for out lil Hurricane party

This tweet was mislabeled as a *Physical Action* social cue probably because it mentions “buying” something, and “buying” is often mentioned in *Physical Action* social cues when people talk about stockpiling food, water, or batteries before a storm.

Under-labeling Social Cues Next, we examined cases where a social cue was not recognized. Table 6 shows the two most frequently assigned *incorrect* labels for each social cue category, along with the percentage of mislabeled tweets that were assigned each label. For instance, the first row of Table 6 indicates that among those tweets in the *Emotional Reaction* category that were mislabeled, 61.5% of them were classified as *NotSC* and 21.3% as *Apathy*. The key take-away from this table is that, for all social cue categories except *Change of Plans*, failing to recognize a social cue (i.e., predicting *NotSC*) is the classifier’s biggest problem, and confusing social cue categories is a smaller issue.



For *Physical Action* social cues, one reason for low recall is that preparation actions can be very diverse. For instance in Table 7, Example 1 is a *Physical Action* social cue because people were catching flights out of the city. But the dataset contains other examples with many different activities,

Category	Top 2 Incorrect Labels
<i>Emotional</i>	NotSC (61.5%), Apathy (21.3%)
<i>Physical Actions</i>	NotSC (60.0%), Apathy (22.2%)
<i>Apathy</i>	NotSC (47.7%), Emotional (34.1%)
<i>Change of Plans</i>	Emotional (42.9%), NotSC (35.7%)
<i>Suggestions</i>	NotSC (83.3%), Emotional (16.7%)
<i>Fact Sharing</i>	NotSC (61.8%), Physical Action (26.5%)
<i>Official Directives</i>	NotSC (66.7%), Suggestions (33.3%)

Table 6: Confusion Table. For each social cue category, we show the top 2 labels that it is incorrectly assigned and the percentage of mislabeled tweets with that label.

1. *#Harvey: Loved ones say goodbye as they catch last flight out of CC @CCIntAirport (Physical Actions)*
2. *In Texas News!!! Harvey is a Cat 3 now?? Since when?? (Emotional Reaction)*
3. *Someone throw an EDM hurricane party and book me. (Apathy)*
4. *Someone pls pick up these pups i swear whenever im back in the valley ill take them off ur hands but please dont let these poor souls die :((Suggestions)*
5. *Driving through Houston from Medical Center to downtown and this city is already a #ghosttown before #HurricaneHarvey arrives #Harvey2017 (Fact Sharing)*
6. *Harvey playing with my money , could 've worked today (Change of Plans)*
7. *State officials increase #readiness levels with #hurricane and storm surge issued for Gulf Coast. #Harvey (Official Directives)*

Table 7: Social Cues that are not Recognized

including wearing rain boots, boarding up windows, driving to another city, and ordering a car service before the hurricane comes. It is challenging for the classifier to recognize the wide range of relevant activities given the limited amount of training data. Furthermore, activities can be described in many different ways. One example is the tweet “Got Water ?  Batteries ?  Don’t forget to drop in and pick up some hurricane reading too !”, where preparations are described as checklist.

For *Emotional Reaction* social cues, we found that the classifier often struggles with tweets that implicitly express an emotion toward the hurricane. For example, shock can be expressed with question marks and exclamation points, as in Example 2 in Table 7. We also found that negative emotions are often expressed with sarcasm (e.g., *It would make perfect sense for me to book my wedding venue on the Shoreline of CorpusChristi before a cat 4 hurricane #HurricaneHarvey*). However, detecting sarcasm is still a challenging task.

Similarly, *Apathy* social cues can be difficult to

recognize because they often express an apathetic attitude implicitly. For instance in Table 7, Example 3 expresses the tweeter’s desire for a hurricane party, which is usually a gathering of people who do not take the hurricane seriously. Another issue is that apathetic attitudes can be suggested by actions which may not seem apathetic. For example, a person is likely apathetic toward the hurricane if they get a haircut in a barber shop as a hurricane approaches, because they are prioritizing the haircut over preparing for the hurricane. Understanding apathy can require complicated reasoning, which is still challenging for current NLP models.

6. Conclusion and Future Work

In this work, we present the first study aimed at recognizing social cues on social media during crises. Identifying social cues would support decision-making of government officials and affected individuals. As part of the contribution to the research community, we release a benchmark dataset of 6,000 tweets manually annotated with fine-grained social cue categories, which could be found at <https://github.com/yyzhuang1991/Crisis-Social-Cues>. The benchmark dataset would encourage future research and benefit government’s decision-making in practice.

We evaluate several benchmark classification models and provide in-depth analysis to identify main challenges in our task. Our analysis suggests several potential directions for future work that may improve the task performance. First, automatically harvesting preparation actions with weakly supervised methods in social media could potentially overcome the model’s insufficient knowledge of preparation actions. Another potential direction is to incorporate existing emotion-recognition models to better detect implicitly expressed emotions and apathetic attitudes, which are prevalent in our model’s errors. In addition, it could also be beneficial to leverage data sampling (e.g., upsampling for smaller and downsampling for large categories) and data augmentation (e.g., back-translation) techniques, which could mitigate the data-hungry problem and the imbalanced data distribution.

Our work contributes to understanding and modeling of socio-behavioral responses in safety-critical crisis situations, contributing to a growing thread of research that relies on NLP to get abreast of computational social science problems of societal import. In addition to research contributions, we believe such approaches hold a promise of delivering practical solutions (Cercas Curry et al., 2023; Plaza-del arco et al., 2023). In future crisis events, models trained to identify social cues may be useful for official disaster response agencies such as Joint In-

formation Centers (JIC) or Emergency Operations Center (EOC) to gather information about disaster preparation and perception by the public and coordinate people and resource accordingly. They can also provide regular users with an automated tool to filter out information about the how other people are responding to the disaster and thus elevate their risk perception and facilitate decision-making.

7. Ethics Statement

In conducting this research, we adhered to ethical principles and practices. University IRB has deemed the social media trace data exempt. In the shared dataset, the identifiers are not released to ensure user privacy.

8. Bibliographical References

- Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. [Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing](#).
- Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. [Emotions from text: Machine learning for text-based emotion prediction](#). In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Amanda Cercas Curry, Giuseppe Attanasio, Debora Nozza, and Dirk Hovy. 2023. [MilaNLP at SemEval-2023 task 10: Ensembling domain-adapted and regularized pretrained models for robust sexism detection](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2067–2074, Toronto, Canada. Association for Computational Linguistics.
- William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. [Learning to classify email into “speech acts”](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316, Barcelona, Spain. Association for Computational Linguistics.
- Julie L. Demuth, Rebecca E. Morss, Leysia Palen, Kenneth M. Anderson, Jennings Anderson, Marina Kogan, Kevin Stowe, Melissa Bica, Heather Lazrus, Olga Wilhelmi, and Jen Henderson. 2018. [“sometimes da #beachlife ain’t always da wave”: Understanding people’s evolving hurricane risk communication, risk assessments, and](#)

- responses using twitter narratives. *Weather, Climate, and Society*, 10(3):537 – 560.
- Lingjia Deng and Janyce Wiebe. 2014. [Sentiment propagation via implicature constraints](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 377–385, Gothenburg, Sweden. Association for Computational Linguistics.
- Lingjia Deng and Janyce Wiebe. 2015. [Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 179–189, Lisbon, Portugal. Association for Computational Linguistics.
- 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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT/NAACL 2019)*.
- Haibo Ding and Ellen Riloff. 2016. [Acquiring knowledge of affective events from blogs using label propagation](#). In *AAAI Conference on Artificial Intelligence*.
- Haibo Ding and Ellen Riloff. 2018. [Weakly supervised induction of affective events by optimizing semantic consistency](#). In *AAAI Conference on Artificial Intelligence*.
- Russell Rowe Dynes. 1969. *Organized Behavior in Disaster: Analysis and Conceptualization*. Disaster Research Center. Monograph series. Disaster Research Center, Ohio State University.
- Meiqing Fu, Rui Liu, and Yu Zhang. 2021. [Do people follow neighbors? an immersive virtual reality experimental study of social influence on individual risky decisions during evacuations](#). *Automation in Construction*, 126:103644.
- D. Golding and S. Krinsky. 1992. *Social Theories of Risk Edited by Sheldon Krinsky and Dominic Golding*.
- Chris Hagar and Caroline Haythornthwaite. 2005. [Crisis, farming & community](#). *The Journal of Community Informatics*, 1(3).
- Minqing Hu and Bing Liu. 2004. [Mining and summarizing customer reviews](#). *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.
- Amanda Lee Hughes, Leysia Palen, Jeannette N. Sutton, Sophia B. Liu, and Sarah Vieweg. 2008. ["site-seeing" in disaster: An examination of on-line social convergence](#).
- Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. [Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages](#).
- Minwoo Jeong, Chin-Yew Lin, and Gary Geunbae Lee. 2009. [Semi-supervised speech act recognition in emails and forums](#). In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1250–1259, Singapore. Association for Computational Linguistics.
- J. Kasperson, Roger Kasperson, Nick Pidgeon, and Paul Slovic. 2003. The social amplification of risk: Assessing fifteen years of research and theory. *The social amplification of risk*, 1.
- Roger E. Kasperson, Ortwin Renn, Paul Slovic, Halina S. Brown, Jacque Emel, Robert Goble, Jeanne X. Kasperson, and Samuel Ratick. 1988. [The social amplification of risk: A conceptual framework](#). *Risk Analysis*, 8(2):177–187.
- Max Kinatader, Brittany Comunale, and William H. Warren. 2018. [Exit choice in an emergency evacuation scenario is influenced by exit familiarity and neighbor behavior](#). *Safety Science*, 106:170–175.
- Michael Lindell and Ronald Perry. 2012. The protective action decision model: theoretical modifications and additional evidence. in: Risk analysis, vol 32(4). *Risk Anal : Off Publ Soc Risk Anal*, 32:616–632.
- Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. [Fine-grained opinion mining with recurrent neural networks and word embeddings](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443, Lisbon, Portugal. Association for Computational Linguistics.
- Sophia Liu, Leysia Palen, Jeannette Sutton, Amanda Hughes, and Sarah Vieweg. 2008. In search of the bigger picture: The emergent role of on-line photo sharing in times of disaster.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pre-training approach](#). *ArXiv*, abs/1907.11692.
- Reza Mazloom, Hongmin Li, Doina Caragea, Muhammad Imran, and Cornelia Caragea. 2018. Classification of twitter disaster data using a hybrid feature-instance adaptation approach. In *ISCRAM*.

- Richard McCreadie, Cody L. Buntain, and Ian Soboroff. 2019. Trec incident streams: Finding actionable information on social media. In *ISCRAM*.
- Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.
- Danaë Metaxa-Kakavouli, Paige Maas, and Daniel P. Aldrich. 2018. How social ties influence hurricane evacuation behavior. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).
- Dennis Mileti and John Sorensen. 1990. Communication of emergency public warnings: A social science perspective and state-of-the-art assessment.
- Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.
- 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*, pages 9–14, Online. Association for Computational Linguistics.
- Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Leysia Palen, Sarah Vieweg, Sophia Liu, and Amanda Hughes. 2009. Crisis in a networked world. *Social Science Computer Review - SOC SCI COMPUT REV*, 27:467–480.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.
- Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.
- Ortwin Renn. 2008. *Risk Governance: Coping With Uncertainty in a Complex World*.
- Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *ACM Symposium on Applied Computing*.
- Jeannette Sutton, Leysia Palen, and Irina Shklovski. 2008. Backchannels on the front lines: Emergent uses of social media in the 2007 southern california wildfires. *Proceedings of the 5th International ISCRAM Conference*.
- Sudha Verma, Sarah Vieweg, William J. Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth Mark Anderson. 2011. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In *ICWSM*.
- Sarah Vieweg. 2012. Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications.
- Himanshu Zade, Kushal Shah, Vaibhavi Rangarajan, Priyanka Kshirsagar, Muhammad Imran, and Kate Starbird. 2018. From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW):195:1–195:18.
- Kiran Zahra, Muhammad Imran, and Frank O. Ostermann. 2020. Automatic identification of eyewitness messages on twitter during disasters. *Information Processing & Management*, 57(1):102107.
- Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2020. Affective event classification with discourse-enhanced self-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5608–5617, Online. Association for Computational Linguistics.
- Yuan Zhuang and Ellen Riloff. 2023. Eliciting affective events from language models by multiple view co-prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3189–3201, Toronto, Canada. Association for Computational Linguistics.

9. Language Resource References

N/A