Mining Risk Behaviors from Social Media for Pandemic Crisis Preparedness and Response

Yasas Senarath[1], Steve Peterson[2], Hemant Purohit[3], Amanda L. Hughes[4], and Keri K. Stephens[5]

[1] George Mason University, VA, USA ywijesu@gmu.edu
[2] Community Emergency Response Team, Montgomery County, MD, USA stevepeterson2@gmail.com
[3] George Mason University, VA, USA hpurohit@gmu.edu
[4] Brigham Young University, UT, USA Amanda_Hughes@byu.edu
[5] University of Texas at Austin, TX, USA, keristephens@austin.utexas.edu

Abstract. The significance of social media in crisis management is well-known, as evident from the vast literature generated in the last decade. Prior research has studied public behavior for various contexts ranging from evacuation planning to supporting response operations and providing emotional support for recovery. However, the majority of these research studies have focused on natural hazard events and there is less exploration for pandemic crises. In this paper, we present a novel social media mining approach for detecting risk behaviors at scale to support crisis preparedness and response efforts of city services, using the COVID-19 pandemic as a case study. In collaboration with eight Community Emergency Response Teams of the Washington D.C. Metro region, we first defined a behavior schema of *risk-preventing* and *risk-taking* behavior types for social media content that have the potential to inform the response operations for pandemic crises. We then developed a classification approach to automatically infer the risk behavior from social media posts using machine learning models and Twitter data collected during the period of March to May 2020. Our experimentation demonstrates a feasible automated solution to rapidly filter social media content with high efficiency (AUC up to 88%), which provides technological capabilities to assist future pandemic crisis management efforts of city services.

Keywords: Crisis Informatics · Social Media Mining · Risk Behavior.

1 Introduction

The adoption of social media in daily life is ubiquitous. This trend offers opportunities to understand and characterize public behavior through systematic analysis of social media content during the times of crises [2,9]. In that context, social media mining [18] techniques have been explored to timely model and study user behavior for applications in various governmental and non-governmental organizations [15,10]. For instance, social media content has potential value for crisis

Table 1. Examples of social media posts for risk behavior detection with relevance to crisis management operations of city services. *(posts rephrased for anonymity)*

	Message	**Risk Class**
T1	People are still not obeying the dc stay at home order.	*Risk-taking*
T2	Why were u in store without a mask? Don't u care about the safety of ur fellow man	*Risk-preventing*
T3	Me, a dependent, watching people get their stimulus checks <link >	*Irrelevant*

response operations [12,13] and community relief [5], ranging from sharing behaviors like requesting or offering help to providing situation updates on-ground and revealing unsubstantiated rumors [6,14].

From the perspective of city emergency services, however, there are key challenges in effectively leveraging social media content for pandemic crises. First, given the majority of the prior disaster research on social media focused on natural hazards, the relevant public behaviors of interest to pandemic crisis response are under-explored. Second, while user behavior on social media has been investigated for public health in general contexts (e.g., smoking cessation), there is an opportunity to study behaviors that are especially relevant for city emergency services.

For pandemic crisis management in cities, understanding the implications of public mobility on the resource allocation of emergency services is crucial (e.g., post T2 in Table 1). The differences in public behaviors around decision-making involving mobility-induced risk is an attribute of the *risk attitude* [1] of a user that we focus on in this study. Studies in public health have found that greater risk-averse attitude is associated with a reduction of human mobility [3,17] that can eventually constrain the resources of city emergency services. The COVID-19 pandemic introduced a range of health and safety risks that provide a unique opportunity to study such behaviors of the public and their implications for the emergency services' resources of the cities.

In this paper, we explore the feasibility of a social media mining approach to study risk behaviors of the public with implications for the crisis preparedness and response of city emergency services. Using data collected from Twitter, we partnered with eight Community Emergency Response Teams (CERTs) in the Washington D.C. Metro region to define and label a large dataset and develop a novel classification approach to detect relevant risk behaviors from social media posts (see examples in Table 1). The proposed modeling scheme leveraged a combination of lexical and distributional semantics-based features. When compared against various baselines, the proposed approach showed an effective classification capability that provides a feasible solution for city emergency services.

In the remainder of the paper, we first describe the background, followed by an explanation of our methodology with the risk behavior schema and classification framework. Lastly, we discuss the experimental results.

2 Background and Related Work

2.1 Social Media Mining

The abundance of user-generated content on social media platforms like Twitter and Facebook has led to opportunities for big data analytics using the rich technologies and tools of data mining. According to Zafarani et al. [18], social media mining is "the process of representing, analyzing, and extracting actionable patterns from social media data." These patterns can be studied for content, users, social network, user interactions, as well as user behavior such as posting content, joining groups, or following/linking another user or pages/brands. Applications [8] of social media mining have been explored in several domains from advertising to public health to activism to the times of crisis as described next.

2.2 Crisis Informatics and Social Media

Crisis informatics is the study of data that is created around crisis events and how that data is shared and used. Much of the research in this area focuses on social media and its role in generating rich, openly shared user-generated content that can be useful to the public, emergency response agencies, and other relief organizations [9,14]. Of most relevance to this paper are studies that discuss how social media data can contribute to situational awareness during a crisis event [15,16]. Social media provides a platform for local citizens affected by an event to share on-the-ground information that can help emergency responders better understand the circumstances of the event and how to respond. In this study, we draw on the knowledge and experience of local CERT volunteers to extract this valuable information from the large volume of available social media posts shared during the COVID-19 pandemic.

2.3 Risk Management in Crises

Risk can be defined in many ways, so here we use Stern and Fineberg's [4] broad definition of risk as the "things, forces, or circumstances that pose danger to people or what they value." Protective actions are a type of desired behavioral outcome, and in prior research on pandemic communication, scholars have identified a host of actions related to hygiene, contact with others, hand-washing, getting vaccinated, following advice, and seeking healthcare. [7] Risk-preventing actions can be considered a broader category that not only considers protective actions, but also actions people take to prevent the spread of a disease like COVID-19. These theoretical foundations of risk behaviors inform the design of the behavior schema used in this study as detailed in the next section.

3 Method

This section describes the components of the social media mining method explored in this study. First the risk behavior schema, as described in 3.1, is

defined, followed by a description of the data collection and annotation process for relevant labels from the defined schema. We then describe the classification models to infer the behavior labels from social media posts automatically.

3.1 Risk Behavior Schema

We rely upon emergency management experts in defining what constitutes relevant public behaviors of risk attitudes. A certified emergency manager from the Washington D.C. Metro region identified the requirements of city emergency services for information filtering from social media. These requirements informed the two major types of behavioral information that had the most potential to increase the situational awareness (shared perception of the elements in the environment) of emergency managers. In particular, we defined the relevant behavior schema of {*risk-preventing*, *risk-taking*, *irrelevant*} behaviors for social media content after considering the practitioner-provided information needs by our CERT collaborators.

Risk-preventing (referred to as *Prevention* here on) behavior refers to the situation where a user indicates an intent to support crisis preparedness policies, such as promoting mask-wearing practices and maintaining social distancing for the COVID-19 pandemic. *Risk-taking* (referred as *Risk* here on) behavior refers to the contrary situation where a user acts to undermine the rules and policies of crisis preparedness set by the city emergency services, such as criticizing mask-wearing practices and encouraging crowding with dissent for social distancing.

3.2 Data Collection and Labeling

We collected data from Twitter using its Streaming API to study risk behaviors in the COVID-19 pandemic crisis during the period of March to May 2020. The Twitter Streaming API provides the ability to filter data by restricting the post's (tweet) location of origin, or the keywords it contains. We used geo-fencing based location filtering to retrieve tweets that originated from the Washington D.C. Metro region, i.e. U.S. National Capital Region. We were able to collect approximately 2.1 Million tweets through the streaming API. We then used a keywords-based filter criterion to collect potentially-related tweets for COVID-19 by using a seed set of 1521 keywords (to be shared in the data release) that was curated with the help of CERT volunteers.

Labeling Process: The annotation interface presented CERT volunteers with one tweet at a time and sought their judgement for whether the tweet content a.) was relevant to the COVID-19 response, and b.) contained mention of Prevention and/or Risk behaviors. 14,000 unique tweets were randomly sampled from the dataset for labeling with CERT volunteers. In total, 39 CERT volunteers annotated 13,584 tweets that were relevant to the Washington D.C. Metro region.

Inter-annotator Agreement: Each tweet was annotated by a minimum of two annotators. In addition to the volunteer annotations for each tweet, a certified

Table 2. Average pairwise Inter-annotator agreement measures on labeled data.

Task	Cohen's kappa	Percent Agreement
Relevancy Labeling	64 %	89 %
Prevention Labeling	53 %	90 %
Risk Labeling	53 %	90 %

emergency manager provided the final decision on tweets where two annotators disagreed on relevance. The average Cohen's kappa and percent agreement for relevancy, Prevention, and Risk behavior classes were calculated separately and indicated in Table 2. Based on the *fair* annotation agreement, this analysis indicates that the task of identifying risk behavior from social posts is a complex cognitive task for human annotators despite training. This might be due to the limited contextual information found in the typically brief text of tweets.

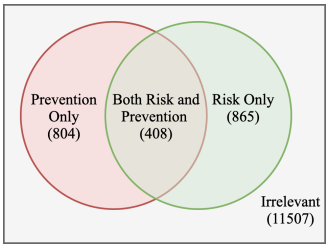

Fig. 1. Label distribution and intersection between *Prevention* and *Risk* behaviors.

Dataset: From the resulting annotated dataset of tweets, all annotations were combined and majority-voting (i.e. more than or equal to two votes for each categorical label) was employed to identify the final label for a given tweet. Figure 1 shows the distribution of the number of tweets pertaining to each class. The annotated instances of risk behaviors (*Risk* and *Prevention*) were considered mutually exclusive to the *Irrelevant* class. However, instances with the *Prevention* and *Risk* class labels were not mutually exclusive (multi-label setting), meaning such behaviors could be observed together in the same tweet. Moreover, due to the nature of real-world data we observe a large number of irrelevant tweets.

3.3 Classification Framework

Figure 2 provides a high-level overview of the classification framework proposed in this study. The framework is divided into two sections: relevancy classification and behavior classification. Relevancy classification precedes behavior classification as indicated in Figure 2, which helps in contextualizing the risk behavior expressed in the tweet. This hierarchical approach enables us to systematically

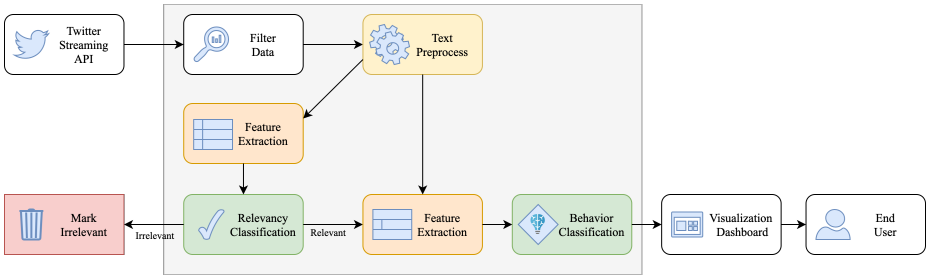

Fig. 2. High-level architecture of the proposed classification framework.

scope the problem according to the practitioner's information needs. In the following subsections, we describe each classifier, detailing the task, features used, and the model specifications.

Relevancy Classification: The task of relevancy classification is to determine whether a provided tweet text is relevant in determining risk behavior (i.e. Risk or Prevention). Formally, this can be specified as a supervised learning task to determine binary target label *Relevant* or *Irrelevant*.
Features: We explore a diverse set of features in this study given the task complexity to determine risk behavior from text:

- *Lexical Features*
  - $tf-idf$: Bag-of-words feature generated by representing *term frequency–inverse document frequency* of words in the vocabulary containing all unique words in the dataset. It is a feature representation strategy used commonly in information retrieval that takes into account the importance of words in a document (tweet).
- *Distributional Semantics-based Features*
  - E_{mean}: The mean of pre-trained word embeddings of the word tokens in an input tweet computed as a feature vector; this feature can provide a quantified representation of the generic sense of words learned from large external text corpus and potentially represent the contextual meaning of risk behaviors expressed in a text. We used 200-dimensional `GloVe` embeddings [11].

In addition to the above features, we also have two variants of lexical features with and without *stopwords*.

Modeling: Since this is the first work using the novel dataset, we evaluate the performance of multiple classification strategies used in the literature. Specifically, we use the following classifiers:

- **LR**: Logistic Regression (Maximum Entropy) classifier.
- **SVM**: Support Vector Machine (SVM) classifier with linear kernel.

We employ the implementations of these algorithms from the *scikit-learn* python library and use the default parameters for training the models.

Behavior Classification: The task of behavior classification is to determine whether a provided (relevant) tweet text contains risk-preventive behavior or/and risk-taking behavior. Formally, this can be specified as a multi-label supervised learning task of determining target label of *Risk* or/and *Prevention* for a given tweet.
Features: The same set of features are identified for training and evaluating the behavior classification model. However, it is essential to note that the behavior feature extraction process is independent of the feature extraction in Section 3.3 to maintain the flexibility of using different sets of features for two classification tasks as indicated in Figure 2.
Modeling: We explore multiple classification models for behavior classification. While the models are similar to the ones discussed under Section 3.3, the key difference here is the use of the one-vs-the-rest (OvR) paradigm for classifiers that do not support multi-label classification.

4 Experimental Setup and Evaluation

We conduct an extensive analysis of various modeling schemes for both classification tasks described above separately by using different algorithms with varied combinations of lexical and semantic features:

- [***M1***] - $tf - idf$: This scheme includes only the $tf - idf$ representation of the pre-processed tweet as features.
- [***M2***] - E_{mean}: This scheme includes only the distributional semantics-based E_{mean} feature.
- [***M3***] - $tf - idf + E_{mean}$: This scheme concatenates the features of *M1* and *M2*.

Evaluation: We use a 10-fold cross validation setting to evaluate the models and use the standard performance measures of precision, recall, F1 score, and AUC from machine learning literature. For fair comparisons, we use stratification in dividing the folds to maintain the same percentage of samples of each target label similar to the overall dataset.

5 Results and Discussion

5.1 Model Performance

Relevancy Classification: Table 3 shows the performance of different model schemes trained using two classifiers. One clear observation is that scheme *M1* outperforms schemes *M2* and *M3* significantly. $tf - idf$ features might already capture the patterns of corpus-specific knowledge for the irrelevant content in contrast to the relevant content for risk behaviors in the feature space effectively, and thus, the patterns from the domain-agnostic distributional semantics features may not contribute much, as also evident from the performance of model *M2*.

Table 3. 10-fold cross-validation results for relevancy classification model schemes. Average cross-validation scores are presented with standard deviation for F1 and AUC.

Classifier	Model Scheme	Precision	Recall	F1	AUC
LR	M1	0.64	0.84	**0.73±0.01**	**0.88±0.01**
	M2	0.42	0.80	0.55±0.01	0.80±0.01
	M3	0.61	0.84	0.71±0.02	0.87±0.01
SVM	M1	0.69	0.78	**0.73±0.01**	**0.86±0.01**
	M2	0.42	0.80	0.55±0.01	0.80±0.01
	M3	0.69	0.77	0.73±0.02	0.85±0.02

Behavior Classification: Table 4 shows the performance of different model schemes for behavior classification task. The performance is higher when the distributional semantics features are concatenated with $tf - idf$ features. The observed increase in performance could be attributed to the semantics captured by word embeddings for contextual representation of behavior, which might not be easily captured through only lexical features extracted from the given corpus alone. In this way, the hybridization of the two types of features is likely to improve the model generalizability for risk behavior detection.

Overall, logistic regression performance was better for both relevancy and behavior classification tasks. Lastly, our analysis of the performance of models using lexical features with and without stopwords (AUC 0.76 vs. 0.74 respectively) shows an interesting pattern that keeping stopwords is a better approach for behavior classification task.

Table 4. 10-fold cross-validation results for behavior classification model schemes. Average cross-validation scores are presented with standard deviation for micro-averaged F1 and AUC scores.

Classifier	Model Scheme	Micro Average				Risk				Prevention			
		Precision	Recall	F1	AUC	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC
LR	M1	0.82	0.77	0.79±0.05	0.76±0.05	0.83	0.80	0.82	0.77	0.80	0.74	0.77	0.74
	M2	0.80	0.74	0.77±0.03	0.74±0.04	0.82	0.76	0.79	0.74	0.79	0.72	0.75	0.73
	M3	0.83	0.79	**0.81±0.02**	**0.77±0.03**	0.83	0.81	0.82	0.77	0.82	0.76	0.79	0.77
SVM	M1	0.79	0.76	0.77±0.02	0.73±0.03	0.81	0.78	0.80	0.75	0.76	0.73	0.75	0.71
	M2	0.79	0.73	0.76±0.04	0.73±0.05	0.81	0.76	0.78	0.73	0.78	0.70	0.74	0.72
	M3	0.80	0.78	**0.79±0.03**	**0.74±0.03**	0.82	0.81	0.81	0.77	0.77	0.74	0.75	0.72

5.2 Effect of Task Complexity

Table 5 shows some examples where the best model outperforms the baseline model (M1); although we also observed cases where both of those models predicted incorrectly. For example, the second tweet in the table shows an instance related to prevention, which has not been correctly interpreted by the baseline model. While for the first tweet, it is partially interpreted by any of these two models, possibly due to the fact that it contains risk-related keywords but overall

Table 5. Examples where the best proposed model scheme predicts correctly/incorrectly in contrast to the baseline M1.

Post	True Label	Prediction	
		Baseline (M1)	Best Model
Seeing people come out of a store with gloves on and using those same gloves to drive around and use their phone makes me question so much	Prevention,Risk	Risk	Prevention
Spent the afternoon at the National Arboretum yesterday, keeping a healthy distance from other patrons. COVID or no COVID, the blooms are still just as beautiful.	Prevention	None	Prevention

preventive sense. This shows the need for more contextual representation of the input text.

6 Conclusion

This paper presented a novel social media mining approach to extract risk behaviors of the public with implications for the crisis preparedness and response of city emergency services. In collaboration with CERTs of the Washington D.C. Metro region, we defined a novel risk behavior schema for social media content and created a labeled dataset for the research community. Using this dataset, we developed and evaluated a classification framework to extract such relevant risk behaviors from social media posts against various baseline models. The experimental results show that our approach can provide an effective classification capability to rapidly filter social media streams for risk behaviors to assist crisis response operations of city emergency services.

A limitation of our dataset is that there is only a few relevant examples for training state-of-the-art models such as deep neural networks. Moreover, given the complexity of the multi-label task for detecting risk behaviors, we experimented with only a few modeling schemes, which can be further fine-tuned to improve the performance. We plan to extend our labeled dataset and evaluate the effects of different hyper-parameters on the performance of each classification task in the future. As shown in the error analysis and the annotation agreement results, risk behavior detection is a complex task. Thus, we also plan to explore both deep learning models and additional features such as psycholinguistics categories to improve context representation for better recognition of patterns during model training. Additionally, we plan to explore common behavior patterns in the large unlabeled data with the help of unsupervised learning.

Acknowledgement: This work was partially supported by grants from the National Science Foundation # 2029719, 2029692, & 2029698.

References

1. Blais, A.R., Weber, E.U.: A domain-specific risk-taking (dospert) scale for adult populations. Judgment and Decision making **1**(1) (2006)

2. Castillo, C.: Big crisis data: social media in disasters and time-critical situations. Cambridge University Press (2016)
3. Chan, H.F., Skali, A., Savage, D.A., Stadelmann, D., Torgler, B.: Risk attitudes and human mobility during the covid-19 pandemic. Nature Scientific Reports **10**(1), 1–13 (2020)
4. Council, N.R., et al.: Understanding risk: Informing decisions in a democratic society. National Academies Press (1996)
5. Glasgow, K., Vitak, J., Tausczik, Y., Fink, C.: With your help... we begin to heal: Social media expressions of gratitude in the aftermath of disaster. In: SBP-BRiMS. pp. 226–236. Springer (2016)
6. Li, J., Stephens, K.K., Zhu, Y., Murthy, D.: Using social media to call for help in hurricane harvey: Bonding emotion, culture, and community relationships. International Journal of Disaster Risk Reduction **38**, 101212 (2019)
7. Liu, B.F., Austin, L., Lee, Y.I., Jin, Y., Kim, S.: Telling the tale: the role of narratives in helping people respond to crises. Journal of Applied Communication Research **48**(3), 328–349 (2020)
8. Liu, H., Morstatter, F., Tang, J., Zafarani, R.: The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. International Journal of Data Science and Analytics **1**(3), 137–143 (2016)
9. Palen, L., Anderson, J., Bica, M., Castillo, C., Crowley, J., Díaz, P., Finn, M., Grace, R., Hughes, A., Imran, M., et al.: Crisis informatics: Human-centered research on tech & crises (2020)
10. Paul, M.J., Sarker, A., Brownstein, J.S., Nikfarjam, A., Scotch, M., Smith, K.L., Gonzalez, G.: Social media mining for public health monitoring and surveillance. In: Biocomputing. pp. 468–479. World Scientific (2016)
11. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014)
12. Peterson, S., Stephens, K., Purohit, H., Hughes, A.: When official systems overload: A framework for finding social media calls for help during evacuations. In: ISCRAM. pp. 867–875 (2018)
13. Purohit, H., Peterson, S.: Social media mining for disaster management and community resilience. In: Big Data in Emergency Management: Exploitation Techniques for Social and Mobile Data, pp. 93–107. Springer (2020)
14. Reuter, C., Hughes, A.L., Kaufhold, M.A.: Social media in crisis management: An evaluation and analysis of crisis informatics research. International Journal of Human–Computer Interaction **34**(4), 280–294 (2018)
15. U.S. Homeland Security: Using social media for enhanced situational awareness and decision support (2014), https://www.dhs.gov/publication/using-social-media-enhanced-situational-awareness-decision-support
16. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: CHI. pp. 1079–1088 (2010)
17. Xu, P., Cheng, J.: Individual differences in social distancing and mask-wearing in the pandemic of covid-19: The role of need for cognition, self-control and risk attitude. Personality and Individual Differences **175**, 110706 (2021)
18. Zafarani, R., Abbasi, M.A., Liu, H.: Social media mining: an introduction. Cambridge University Press (2014)