An Evaluation of Twitter Datasets from Non-Pandemic Crises Applied to Regional COVID-19 Contexts

Shivam Sharma
New Jersey Institute of Technology,Newark
ss4354@njit.edu

Cody Buntain*
New Jersey Institute of Technology, Newark
cbuntain@njit.edu

ABSTRACT

In 2020, we have witnessed an unprecedented crisis event, the COVID-19 pandemic. Various questions arise regarding the nature of this crisis data and the impacts it would have on the existing tools. In this paper, we aim to study whether we can include pandemic-type crisis events with general non-pandemic events and hypothesize that including labeled crisis data from a variety of non-pandemic events will improve classification performance over models trained solely on pandemic events. To test our hypothesis we study the model performance for different models by performing a cross validation test on pandemic only held-out sets for two different types of training sets, one containing only pandemic data and the other a combination of pandemic and non-pandemic crisis data, and comparing the results of the two. Our results approve our hypothesis and give evidence of some crucial information propagation upon inclusion of non-pandemic crisis data to pandemic data.

Keywords

Covid19, Twitter, Trecis, cross-validation, machine learning, transfer learning.

INTRODUCTION

In recent years, social media has seen an immense increase in the number of users and content. It has transformed from being just a platform to update your social life to being a medium to communicate crucial information. Social media has emerged as a much faster source of mass media communication, especially during the time of a crisis like wildfire or earthquake, as compared to other mediums like newspapers, articles, etc. During such times, platforms such as Twitter have especially evolved into a crucial source of information as well as being a medium to request for aid. This evolution of social platforms into sources of crisis-related data has encouraged the Crisis Informatics community to include this data in development and improvement of tools for the aid of victims of such crises.

In recent times, we have witnessed a rare crisis, namely, the COVID-19 pandemic. Pandemic and other health related crises are very different from the usual crises like earthquakes and wildfires. These are much rarer and affect a larger portion of society. Pandemics are also among the crises which occur for a long period of time.

Due to the need for quarantining in such times, there is an expected increase in the content posted on the social media, ranging from important information shared by health agencies to social awareness posts. This immense data for a crisis with such a unique nature raises multiple questions about its utility for crisis informatics. Our approach towards this data is based on evaluating the importance of the information extracted from non-pandemic crisis data in identifying and classifying different types of COVID-19 content available on social media.

Thus, we can define the main question we aim to address through this work as:

Research Question : *Whether data from non-pandemic crises is valuable for developing models for pandemic-type events.*

We approach this question by conducting experiments on the labeled Twitter data provided by TREC-IS. We hypothesize that some core information remains consistent throughout all the crises. If this hypothesis holds,

*corresponding author

including non-pandemic crisis data with pandemic data should improve our model's ability to identify and classify different types of pandemic data. To test our hypothesis, we perform a cross validation experiment on the labeled crisis data provided by TREC-IS and compare the results over a held-out set of pandemic-only data for two different types of training sets, one containing only COVID-19 tweet data and the other consisting of a collection of data from multiple crises, including COVID-19. Our results suggest that including non-pandemic data with pandemic-only training data improves the model's ability to classify COVID-19 tweets.

RELATED WORKS

With increased use of social media, and increased access to the internet, communicating necessary high priority information through such means has become a common practice. Research done by Vieweg (2012) suggests that Twitter has been used excessively to communicate such high priority information in the times of crises and such posts contain necessary information which can help improve situational awareness about these crises.

Work done by Munro and Manning (2012) and Kanhabua and Nejdl (2013) suggests that the information extracted from Twitter varies across crises, and is also poorly correlated to other information sources like SMS. However, research done by Tierney (1993), Buntain and Lim (2018) and Olteanu et al. (2015) show evidence of some level of consistency in the tweets across crisis events of differing types. Multiple studies (Buntain and Lim 2018; Olteanu et al. 2015; Sutton et al. 2015; Verma et al. 2011) have been conducted to analyse the importance of extracting information across differing crises and using this information to improve the classification of data into different categories. Verma et al. (2011) uses a combination of hand-annotated and automatically-extracted linguistic features to automatically detect messages that may contribute to situational awareness across a set of three different crises, namely, floods, wildfire and earthquake. Their work shows evidence that simulating information propagation through linguistic features from prior events improves the model's ability to classify data for a new event.

Encouraged by these evidences of importance of information propagation across different types of crises, we decided on selecting a dataset which would contain labeled social media data across multiple crisis events including COVID-19 pandemic. Social Media for Emergency Relief and Preparedness (SMERP) (Ghosh et al. 2017) and TREC Incident Streams (TREC-IS) (McCreadie et al. 2020) are workshops providing standardized dataset containing labeled tweets for multiple crisis events. Standardized test collections in TREC-IS in particular allow researchers to develop and evaluate systems across a variety of crises, assuming a core set of 25 types of information are consistently present regardless of the crisis type. In 2020, however, and in response to the pandemic, TREC-IS introduced a separate task specifically aimed at COVID-19 tweets; since the workshop lacked training data on any similar public-health crisis, organizers separated COVID-19 contexts to support a focus on the pandemic, separate from other crises. This separation demonstrates the uncertainty in the community about whether and to what degree does our understanding of crises apply to COVID-19, when its geographic and temporal dimensions vary so widely from many other crises. At the same time, the work cited above generally demonstrates some consistency across information shared during crises, regardless of their underlying type. This paper aims to unify these views by asking whether the COVID-19 contexts need to be separated from other non-pandemic type event data and introducing the following hypothesis:

Hypothesis: *Including labeled crisis data from a variety of non-pandemic events will improve classification performance in COVID-19 events compared to models trained solely on pandemic-type events, where we have fewer data but more specific and relevant information types.*

To test this hypothesis, we describe a cross-validation experiment for comparing the results of different models against a held-out set consisting only of COVID-19 data. Multiple studies give evidence of improved classification ability of Deep Neural Networks (DNNs) over classical Machine Learning models. Research done by Hettiarachchi and Ranasinghe (2020), ALRashdi and O'Keefe (2019), Kumar et al. (2020) and Nguyen et al. (2016) form a consensus regarding improved performance of these DNNs in crisis response tasks.

Encouraged by these evidences we decided to test our hypothesis based on the results shown by these DNNs. Transformer models like BERT (Devlin et al. 2019) are a special type of DNNs which have shown state-of-the-art performances in major Natural Language Generation (NLG), Natural Language Understanding (NLU) and Natural Language Processing (NLP) tasks. Since this work is focused on using only the tweet text to test the above stated hypothesis, we plan on using these transformer models for our cross-validation experiments.

METHODOLOGY

To answer the research question described above and to test our hypothesis we devise a cross-validation experiment, performed on a multi-label information type label classification task, to compare the results given by a set of models

Table 1. Event Type Data Distribution

Event Type	Labeled Tweets
Wildfire	7,303
Earthquake	128,70
Flood	8,342
Typhoon	14,530
Shooting	10,701
Bombing	3,239
Accident	210
COVID-19	14,835

trained on two different sets of train data, one consisting of only the COVID-19 tweets and the other containing COVID-19 as well non-pandemic event tweets. These models were then in turn evaluated against a held-out set of COVID-19 only tweets.

This experiment requires labeled tweet data for both, COVID as well as non-COVID events, machine learning models for classifying these tweets and metrics to evaluate the performance of these models. We discuss each of these in turn below.

Cross Validation Experiment Description

To test our hypothesis, we created a cross-validation test. If results on a held-out set of COVID-19 tweets given by a model trained on just the COVID-19 tweets are better than the results on the same held-out set given by a model trained on the complete crisis data, which includes COVID as well as non-COVID type event data, then we can conclude that pandemic-type events should be separated from the rest of the crises. On the other hand, if this does not happen, that is, if the results by the model trained on just the COVID-19 tweets are worse than the results by the model trained on complete crisis data, then we can say that the inclusion of general crises like wildfires, floods, blasts etc. does not offset the performance of the model. That is, we can conclude that there is some form of information propagation which occurs upon inclusion of non-pandemic type crisis data which is crucial for better classification of pandemic type tweets

We performed a 10-fold cross-validation with 90-10 data split on the information type label prediction task for TREC-IS. We split the COVID-19 tweets into training and validation sets using 10-fold cross-validation. These validation sets containing only the COVID-19 tweets are termed as “held-out set”, and the performance of every model would be evaluated against this “held-out set”. For getting the results for COVID-19 only model, we trained the model on the COVID-19 training set for 2 epochs and predicted the information labels for the held-out set. For getting the results for complete crisis data, we trained the models on the complete training data excluding the tweets already present in the held-out set, for 2 epochs, and validated the performance against the held-out set. To prevent same tweets from repeating in the training and the validation set, only the training set contained the augmented tweets.

To evaluate the results we used the average accuracy, average precision, average recall and the average F1 score for the 10 fold cross validation. We used the “macro” averaging method to calculate the precision, recall and F1 scores, which does not take into account any imbalance in the labeled dataset. This would help ensure that the pipeline with the best results would be able to take into account the heavy imbalance in the real world also.

Dataset

To test our hypothesis, we use the dataset made available through TREC-IS, up through 2020-A. As seen in table 1, the track included COVID-19 data for three United States regions, namely, Washington DC, Washington State and New York. In total, this dataset includes labels for 14835 COVID-19 tweets and 57195 non-COVID tweets.

Models

This section outlines the Neural Language Model Pipeline we used to test our hypothesis. Figure 1 shows the completed pipeline. Given a tweet text, the pipeline first processes and cleans the text. For training the model, the pipeline augments data in the training dataset and trains a transformer model. For a test tweet, the transformer models takes as input the processed tweet text and returns a one hot-encoded list of corresponding labels. As this task was a multi-label classification task, there can be more than one labels predicted for a single input tweet. As seen from Figure 1, the pipeline can be split into two sections after the initial data processing:

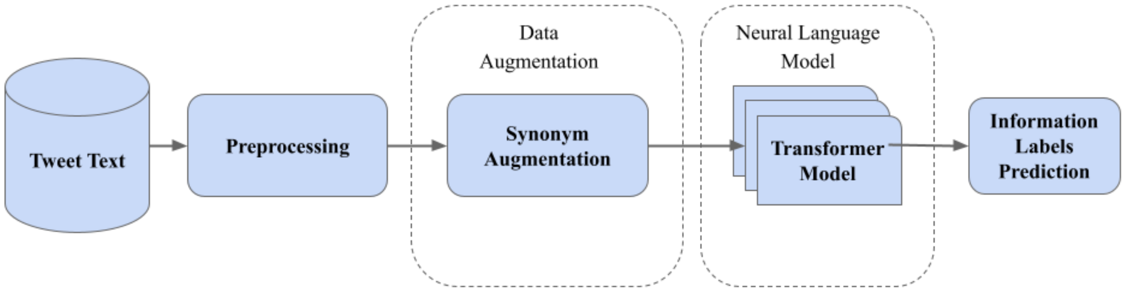

Figure 1. Neural Language Model Pipeline

1. Data Augmentation
2. Neural Language Model

Processing of the tweet text mainly includes data cleaning and refining. The text is first cleaned of all punctuation and special characters like #, @, etc. We replace all the URLs, mentions and hashtags from the text with "URLHERE", "MENTIONHERE" and "HASHTAGHERE" tags, respectively.

Data Augmentation

The information type labels provided by TREC-IS can be majorly divided into two sections, Critical Importance type and Low Importance type. The information type labels classified as being of Critical Importance contain the tweets which have a higher priority for an emergency responder than the tweets belonging to information type labels classified as being of Low Importance. Information type labels such as Search and Rescue, Volunteers, etc can be termed as being of higher importance than labels like Multimedia Share or Sentiments.

Our initial analysis of data showed evidence of heavy imbalance between these two types of labels. We found a majority of tweets belonging to Low Importance information type categories. To counter this imbalance, we considered using the oversampling strategy SMOTE (Chawla et al. 2002) to increase the number of Critical Importance Information Type categories. After a few tests of comparing different scores like precision, recall and F1 score for different oversampling strategies like SMOTE through imbalance-learn library, we decided on "generating" new tweets of Critical Importance Information categories.

The text augmentation was done on the idea of replacing specific words in a tweet with their synonyms, so as to "generate" new tweets for training data which do not deviate from their original meaning but still have some semantic differences from the original tweet. Since the tweets are related to disasters and calamities, we decided to replace the verbs with their synonyms as those would be the words which would have higher weight in the sentence. After some tests, we decided on replacing one single verb with top four synonyms, as more than that and the synonym would sometimes change the meaning of the text. This approach gave us more than four thousand new tweet texts to be used in training with Critical Importance Information type labels.

While this pipeline introduces complexity into the research design, we have submitted this pipeline to the TREC-IS 2020-A edition (SHARMA and BUNTAIN n.d.), where it outperformed the non-augmented approach and was more competitive with other participants. Repurposing this existing system has accelerated this research, however, and the hypothesis, if true, should hold regardless.

Natural Language Models

After processing and augmenting the data, we pass the text to a NLM for training and prediction. To remove the potential biases introduced by model selection, we run this experiment across a set of 9 different types of pre-trained transformer models. These 9 models, namely, ALBERT (Lan et al. 2019), BERT (Devlin et al. 2019), CamemBERT (Martin et al. 2020), DsitilBERT (Sanh et al. 2019), ELECTRA (Clark et al. 2020), Longformer (Beltagy et al. 2020), RoBERTa (Liu et al. 2019), XLM (Lample and Conneau 2019) and XLNet (Yang et al. 2019), were shortlisted from the vast list of transformers due to the fact that they are the basic general purpose models for NLP tasks. Table 2 outlines these models with their corresponding pre-trained weights used. We used the simpletransformers (Rajapakse 2019) library for model implementation.

Table 2. Model and Pre-Trained Weights Used for Cross-Validation Test

Model Name	Model Pre-Trained Weights
ALBERT	albert-base-v2
BERT	bert-base-uncased
CAMEMBERT	camembert-base
DISTILBERT	distilbert-base-uncased
ELECTRA	google/electra-base-discriminator
LONGFORMER	allenai/longformer-base-4096
ROBERTA	roberta-base
XLM	xlm-mlm-en-2048
XLNET	xlnet-base-cased

Table 3. 10-Fold Cross-Validation Results for COVID-19-Only Versus Full Dataset. Summarized in the median row, one can see models based on the full dataset significantly outperform those trained only on COVID-19 data.

	Average Accuracy		Average Precision		Average Recall		Average F1-Score	
Model Name	**COVID**	**FULL**	**COVID**	**FULL**	**COVID**	**FULL**	**COVID**	**FULL**
ALBERT	0.670	0.601	0.114	0.234	0.057	0.100	0.061	0.126
BERT	0.631	0.501	0.208	0.253	0.072	0.109	0.089	0.139
CAMEMBERT	0.644	0.399	0.029	0.131	0.037	0.067	0.033	0.081
DISTILBERT	0.633	0.497	**0.235**	0.244	0.086	0.105	0.108	0.136
ELECTRA	0.677	0.473	0.052	0.201	0.042	0.075	0.038	0.097
LONGFORMER	0.627	0.622	0.127	0.298	0.071	**0.178**	0.081	**0.208**
ROBERTA	0.640	**0.623**	0.231	**0.309**	**0.127**	0.178	**0.149**	0.207
XLM	**0.693**	0.191	0.028	0.096	0.040	0.054	0.033	0.064
XLNET	0.635	0.617	0.215	0.296	0.118	0.168	0.138	0.200
Median	**0.640**	0.501	0.127	**0.244**	0.071	**0.105**	0.081	**0.136**

RESULTS

Table 3 shows the results for the cross-validation experiment discussed above. The scores mentioned are for the different models trained on the two types of datasets. The "COVID" dataset contains labeled tweets for COVID-19 only. The "FULL" dataset contains labeled tweets for COVID as well as non-COVID events. These results are evaluated against a held-out set containing COVID-19 tweets only.

As observed from the table, even though the average accuracy for the models trained on only COVID-19 data are higher than the models trained on the complete dataset, the average precision, recall, and consequently the average F1 scores are significantly higher for the models trained on the complete crises dataset.

It is evident from the table that the cross-validation results are in consensus with our proposed hypothesis, thus demonstrating that there is some underlying information which is consistent throughout all the crises, due to which including the general non-pandemic type events to COVID-19 significantly improved the models ability to classify COVID-19 tweets.

DISCUSSION AND FUTURE WORK

The significant improvement in the scores for the cross validation using non-pandemic type events as well as COVID-19 tweets over the model solely trained on COVID-19 tweets is evidence that there is significant value in including information from general crises into models for COVID-19 or other pandemic type events. Our work gives evidence that including general non-pandemic event data to pandemic event data simulates crucial information propagation which is essential for better classification of pandemic events data.

One interesting aspect of our results is the opposite results shown by the accuracy metric. Contrary to our hypothesis, and to the other matrices, the average accuracy for models trained on COVID-only data is higher than the accuracy for the models trained on the complete (non-COVID plus COVID) dataset. Due to heavy imbalance in the dataset, we can expect to find more predictions for some classes than others, which would, thus, affect the accuracy metric. This was the reason for including metrics like precision, recall and F1 score to compare our results. However, this leads us to an interesting question that whether including the general crises data improves the performance for specific information types. We aim to analyze which information types are positively affected by inclusion of general crisis data and which ones are negatively affected by it.

Our aim for this article was to validate this claim. We aim to build on this for our future work. We aim to analyze whether the information from a natural disaster like wildfires or floods should be given the same importance as the information from man-made disasters like bombings or shootings when predicting on a natural disaster.

We compared the cross-validation performance showed by models trained on two different dataset, against a held-out set of COVID-only dataset. Both of the training datasets, one containing COVID-only data and the other containing COVID as well as general crisis data, had pandemic data in them. We plan on conducting an experiment with dataset containing only non-pandemic type crisis data. This would further help us to understand the effects of inclusion of pandemic-type data.

The nature of pandemic-type events raises a question on the priority scoring of the tweets. In general crises like wildfires and bombings, we have a clear concept of which tweet can be classified as High Priority tweets, like the ones asking for aid or volunteers. But since pandemics are such long lasting events, and affect the people at a much slower rate, it would be hard to say which type of tweets will be of High Importance. By High Importance we mean the tweets or posts which should be seen by some emergency responders. It is hard to define a set of tweets type which needs to be seen by some emergency responders during a pandemic-type event. We aim to study the information types which can be classified as High Priority during such an event.

Our cross-validation test gives strong evidence that inclusion of general non-pandemic type events improves classification for pandemic type event data. But this raises a question on whether the reverse is true also, that is, does inclusion of pandemic type event data improves our model's ability to classify general non-pandemic type event data. Though, based on the evidence shown by our cross-validation test, we can expect our hypothesis to hold true for reverse condition, but we aim to analyze this by conducting a similar cross-validation experiment as described above.

CONCLUSION

In this work, we aimed at answering the question on whether we can group a pandemic-type event, like the COVID-19 pandemic, with other general crises like wildfires and earthquakes for information retrieval tasks for social media posts. This question arises due to the unique nature of such health related crises, like the time period, the number of people affected over time, the area affected by it and the kind of content people post online related to this crisis. We hypothesized that including labeled data from a variety of non-pandemic events would improve our model's ability to classify pandemic-only data over models trained solely on pandemic-type events.

We created a cross-validation experiment to test our hypothesis for the TREC-IS dataset and the results showed evidence that our hypothesis holds true. We observed that there was a significant improvement in model performance, evaluated by generating average accuracy, precision, recall and F1 scores, after we included the non-pandemic events (general crises data) to the pandemic type events (COVID-19 data).

In conclusion, our work gives the evidence of some underlying core information which remains consistent throughout all crises. We aim to continue our research on further questions raised by our findings.

LIMITATIONS

One limitation to our work is the removal of extra information from tweet text during pre-processing. For initial analysis, we removed the information contained by mentions (@) and hashtags (#) and replaced them with custom tags like "MENTIONHERE" and "HASHTAGHERE". Our reason for this is that these words (mentions and hashtags) will not in themselves provide any extra information towards the information type classification task. However, we plan on integrating this additional source of information and observe the effects to our performance.

REFERENCES

ALRashdi, R. and O'Keefe, S. (2019). "Deep learning and word embeddings for tweet classification for crisis response". In: *arXiv preprint arXiv:1903.11024*.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). "Longformer: The long-document transformer". In: *arXiv preprint arXiv:2004.05150*.

Buntain, C. L. and Lim, J. K. R. (Nov. 2018). "#pray4victims: Consistencies in Response to Disaster on Twitter". In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). "Electra: Pre-training text encoders as discriminators rather than generators". In: *arXiv preprint arXiv:2003.10555*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (June 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, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Ghosh, S., Ghosh, K., Ganguly, D., Chakraborty, T., Jones, G., and Moens, M.-F. (Aug. 2017). "ECIR 2017 Workshop on Exploitation of Social Media for Emergency Relief and Preparedness (SMERP 2017)". In: *ACM SIGIR Forum* 51, pp. 36–41.

Hettiarachchi, H. and Ranasinghe, T. (2020). "Infominer at wnut-2020 task 2: Transformer-based covid-19 informative tweet extraction". In: *arXiv preprint arXiv:2010.05327*.

Kanhabua, N. and Nejdl, W. (Jan. 2013). "Understanding the diversity of tweets in the time of outbreaks". In: *WWW 2013 Companion - Proceedings of the 22nd International Conference on World Wide Web*, pp. 1335–1342.

Kumar, A., Singh, J. P., Dwivedi, Y. K., and Rana, N. P. (2020). "A deep multi-modal neural network for informative Twitter content classification during emergencies". In: *Annals of Operations Research*, pp. 1–32.

Lample, G. and Conneau, A. (2019). "Cross-lingual language model pretraining". In: *arXiv preprint arXiv:1901.07291*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). "Albert: A lite bert for self-supervised learning of language representations". In: *arXiv preprint arXiv:1909.11942*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., Clergerie, É. de la, Seddah, D., and Sagot, B. (July 2020). "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7203–7219.

McCreadie, R., Buntain, C., and Soboroff, I. (2020). *Incident Streams 2019: Actionable Insights and How to Find Them*.

Munro, R. and Manning, C. D. (2012). "Short Message Communications: Users, Topics, and in-Language Processing". In: *Proceedings of the 2nd ACM Symposium on Computing for Development*. ACM DEV '12. Atlanta, Georgia: Association for Computing Machinery.

Nguyen, D. T., Joty, S., Imran, M., Sajjad, H., and Mitra, P. (2016). "Applications of online deep learning for crisis response using social media information". In: *arXiv preprint arXiv:1610.01030*.

Olteanu, A., Vieweg, S., and Castillo, C. (2015). "What to Expect When the Unexpected Happens: Social Media Communications Across Crises". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '15. Vancouver, BC, Canada: Association for Computing Machinery, pp. 994–1009.

Rajapakse, T. C. (2019). *Simple Transformers*. https://github.com/ThilinaRajapakse/simpletransformers.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108*.

SHARMA, S. and BUNTAIN, C. (n.d.). "Improving Classification of Crisis-Related Social Media Content via Text Augmentation and Image Analysis". In: ().

Sutton, J., Gibson, C., Phillips, N., Spiro, E., League, C., Johnson, B., Fitzhugh, S., and Butts, C. (Nov. 2015). "A Cross-Hazard Analysis of Terse Message Retransmission on Twitter". In: *Proceedings of the National Academy of Sciences* 112.

Tierney, K. J. (1993). "Disaster preparedness and response: Research findings and guidance from the social science literature". In:

Verma, S., Vieweg, S., Corvey, W., Palen, L., Martin, J., Palmer, M., Schram, A., and Anderson, K. (Jan. 2011). "Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency." In:

Vieweg, S. E. (2012). "Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications".

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). "Xlnet: Generalized autoregressive pretraining for language understanding". In: *arXiv preprint arXiv:1906.08237*.