Information Type Classification with Contrastive Task-Specialized Sentence Encoders

Philipp Seeberger, Tobias Bocklet, Korbinian Riedhammer
Technische Hochschule Nürnberg
`{firstname.lastname}@th-nuernberg.de`

Abstract

User-generated information content has become an important information source in crisis situations. However, classification models suffer from noise and event-related biases which still poses a challenging task and requires sophisticated task-adaptation. To address these challenges, we propose the use of contrastive task-specialized sentence encoders for downstream classification. We apply the task-specialization on the CRISISLEX, HUMAID, and TRECIS information type classification tasks and show performance gains w.r.t. F_1 score. Furthermore, we analyse the cross-corpus and cross-lingual capabilities for two German event relevancy classification datasets.

1 Introduction

User-generated information content on social media has become an important information source in crisis and emergency situations (Reuter et al., 2018). Social media posts immediately provide details about ongoing developments, first-party observations, and other information which would be missed with traditional sources (e.g., official news) (Sakaki et al., 2010). Access to this information content is thereby crucial for situational awareness in order to support official institutions, government organisations, and relief providers (Kruspe et al., 2021).

However, processing this noisy high-volume social media streams is challenging and requires sophisticated methods for automatic reliable detection of information content. To tackle this challenge, recent work has focused on binary, multiclass, and multi-label information type classification approaches (Alam et al., 2018, 2021b; Buntain et al., 2021). For instance, important information categories cover missing and injured people, damaged infrastructure, etc.

Another challenge is the nature of data prevalent in social media and microblogging platforms. For example, a large portion of noisy user-generated texts inherit properties such as a limited number of words, less contextual information, hashtags, and noise (e.g., misspellings, emojis) (Wiegmann et al., 2020; Zahera et al., 2021). Furthermore, event-related biases and entities prevent models from generalizing to unseen disaster events and therefore degrade in performance (Zhang et al., 2021a; Seeberger and Riedhammer, 2022).

These challenges motivates the use of efficient and effective approaches for adapting classifiers to the noisy text domain and the different information type tasks. Recently, contrastive fine-tuning mechanisms attracted research efforts for few-shot and task-specialization settings by adapting language models and sentence encoders (SE) for downstream classification (Vulić et al., 2021; Tunstall et al., 2022; Su et al., 2022). Following this approach, we aim to analyse the contrastive task-specialization in the field of information type classification.

Contributions Our main contributions are as follows: **1)** We introduce the contrastive task-specialization method for information type classification. **2)** We analyse the cross-corpus capabilities of the task-specialized models. **3)** We empirically show the cross-lingual and cross-task transfer capabilities for two German disaster datasets.

2 Method

As discussed in section 1, we follow previous work (Vulić et al., 2021) and aim to fulfill the requirement of effective adaptation by **1)** quickly bootstrapping a general-purpose SE for new domains and tasks via contrastive learning, and **2)** training a classifier on top of the fixed SE.

2.1 Contrastive Task-Specialization

The main idea is to follow the specialization of a general-purpose SE which is pre-trained on a large corpus of sentence pairs. Specializing a universal

180

SE to particular tasks has been proven effective in prior work for multi-class and multi-label scenarios via further fine-tuning by a contrastive loss (Vulić et al., 2021; Zhang et al., 2021b; Vulić et al., 2022). In this way, we can utilize available annotations to achieve task-adaptation to create more accurate encodings for the downstream classification.

Positive and Negative Pairs   For the creation of sentence pairs, we follow prior efficient contrastive and few-shot approaches by implicitly leveraging the information type ids to create positive and negative learning examples (Tunstall et al., 2022). Therefore, we use the sentence pair creation scheme proposed by SETFIT[1] and construct the positive set Pos and negative set Neg by applying n iterations of sentence pair generations. For the multi-label task, we follow (Vulić et al., 2022) by sampling a positive sentence for each label in the label set of sentence s_i.

Contrastive Loss   As contrastive loss, we opt for the Online Contrastive Loss (OCL) (Reimers and Gurevych, 2019; Vulić et al., 2022). This online version of contrastive loss operates with hard in-batch negative pairs and hard in-batch positive pairs and yields the final task-specialized SE. The constrastive learning should attract similar sentences together and push dissimilar sentences apart.

2.2   MLP Classification

A standard approach for classification based on SE's is the Multi Layer Perceptron (MLP) which is stacked on top of a fixed SE. This is much more lightweight than fine-tuning the entire SE but still achieves comparable performance in low-resource settings. We train a MLP classifier composed of a single hidden layer with non-linearity. For the multi-class and multi-label classifier, we use the standard cross-entropy and binary cross-entropy loss, respectively. A threshold θ determines the final classification for the multi-label task by only classifying information types with probability scores $\geq \theta$.

3   Experimental Setup

3.1   Datasets

We experiment with three TWITTER datasets, covering **1)** multi-class and multi-label classification,

2) different information type ontologies, and **3)** numerous diverse event types composed of natural disasters and human-made disasters.

CRISISLEX   The T26 variant of CRISISLEX (Olteanu et al., 2015) includes labeled tweets for 26 crisis events, annotated with seven information types including the category NOT RELATED. This set reflects a wide variety of events about emergencies with approximately 1,000 tweets per individual event. As preprocessing step, we removed tweets with the label NOT APPLICABLE as these contain issues such as "not readable" for the annotator (Olteanu et al., 2015). This task represents a multi-class classification problem.

HUMAID   This collection contains data about 19 events with dataset sizes ranging from approximately 570 to 9500 tweets (Alam et al., 2021a). HUMAID covers eleven categories ranging from NOT HUMANITARIAN to INURED OR DEAD PEOPLE which captures fine-grained information about disasters. We ignore posts with the labels CAN'T JUDGE and MISSING OR FOUND PEOPLE as the latter case is only available for four events. Similar to CRISISLEX, this task is about multi-class classification.

TRECIS   TREC Incident Streams is a multi-label classification task composed of over 70 events with annotations for 25 information types (Buntain et al., 2021). The sample ranges are from 90 to 5900 tweets and highly vary across events in terms of tweets and label distribution. Furthermore, the collection also covers the large-scale public-health event COVID whereby we only focus on general crisis events. For our experiments, we drop COVID events and select the top-30 events with the highest number of posts as events with a few posts only cover a small subset of relevant information types.

Data Splits   For within-corpus classification, we evaluate each method with 5-fold cross-validation (5-fold CV) with disjoint events. Due to the high cost of task-specific annotations, we additionally focus on low-data scenarios for bootstrapping SE's. Therefore, we conduct experiments in the two data configurations **1)** *Low* and **2)** *High*. For the *High*-setup, we use the training and test splits provided by the 5-fold CV method. Then, in the *low*-setup, we randomly sample 10 posts for each information type and event in order to construct the low-

---

[1]Tunstall et al. (2022) introduced SETFIT as an efficient and prompt-free framework for few-shot classification and fine-tunes sentence transformers in a contrastive manner.

resource training sets. Throughout all experiments, the test splits remain the same.

3.2 Models and Hyperparameters

For our evaluation, we use MPNET_{LM}[2] (Song et al., 2020) as language model and MPNET_{SE}[3] as SE variant, transformed by a standard contrastive dual-encoder framework. MPNET_{LM} comprises 12 transformer layers with hidden size $h_T = 768$ and prior work has trained MPNET_{SE} with approximately one billion sentence pairs.

Baseline As baseline, we additionally conduct full end-to-end fine-tuning of the MPNET_{LM} model with a MLP classification head which we denote as MPNET_{LM}+FFT. Following suggested settings (Wang et al., 2021; Alam et al., 2021b; Seeberger and Riedhammer, 2022), we train the baseline models for 15 epochs with the optimizer AdamW, learning rate $2e-5$, weight decay 0.01, batch size 32, and evaluate the best checkpoint selected by a validation set.

Contrastive Task-Specialization In terms of CTS fine-tuning with OCL, we adopt a similar setup. The learning rate of AdamW is set to $2e-5$, weight decay to 0.01, and batch size to 64. For the *High*- and *Low*-setup, we construct sentence pairs with $n = 1$ and $n = 5$, respectively. We fine-tune the models on the sentence pairs for 3 epochs with the warmup ratio of 0.05 and cosine decay.

Classification The classifier consists of a MLP architecture with one hidden layer of size 512 with ReLU as non-linear activation function. We train the classifier for 30 epochs with the opimizer AdamW, learning rate $1e-3$, weight decay 0.01, dropout 0.4, and batch size 32. For multi-label classification, we use the threshold θ of 0.3. We select the best classifier based on a validation split sampled from the training set with the ratio of 0.1.

3.3 Evaluation

For all experiments, we report the micro-averaged and macro-averaged F_1 scores across events. In the *High*-setup, all reported results are averaged across the five folds. For the *Low*-setup, we additionally conduct three runs with random seeds in order to reach more stable results with respect to few-shot sampling.

[2] `microsoft/mpnet-base`
[3] `sentence-transformers/all-mpnet-base-v2`

4 Results and Discussion

The main results are summarized in Table 1, while further cross-corpus and cross-lingual experiments are shown in Table 2 and Table 3. In the following, we discuss the results and findings.

Contrastive Task-Specialization The results in Table 1 reveal that performance gains for the multiclass-classification are achieved via CTS. These performance boosts are across all *Low*- and *High*-setups with respect to the CRISISLEX and HUMAID datasets. With focus on the low-resource setup, we additionally experience significant improvements over the full fine-tuning baseline. In comparison, the gap between MPNET_{SE} and MPNET_{SE}+CTS is consistently higher than the counterpart MPNET_{LM} and MPNET_{SE} which suggests the effectiveness of CTS. However, there are no substantial performance gains or even a decrease for the TRECIS multi-label task. This finding is contrary to the results in the domain of multi-label intent detection (Vulić et al., 2022). We hypothesize the cause are differences in semantic concepts across events and annotations (Seeberger and Riedhammer, 2022). More sophisticated sentence pair sampling techniques, hard-negative mining or the usage of high level information types may tackle these shortcomings.

Cross-Corpus With cross-corpus evaluation we aim to analyze other important aspects of CTS fine-tuning. We hypothesize that similar information type ontologies lead to better classification performances by transfering the fine-tuned knowledge about semantically similar information types. Therefore, we trained the SE's on the source corpus and only trained the MLP classifier with the fixed SE on the target corpus. The results in Table 2 indicate an improvement for the datasets CRISISLEX and HUMAID which share similar information type ontologies. However, the knowledge transfer for TRECIS does not maintain improvements or even leads to worse results. We believe the reason for this observation is two-fold. Firstly, the results of Table 1 suggest that the obtained embedding representations are less semantically discriminative than the pre-trained language model for multi-label classification. This may result into a worse cross-corpus knowledge transfer. Secondly, the information type ontologies of CRISISLEX and HUMAID differ from TRECIS. While CRISISLEX and HUMAID share most of the information types, the

Variant	CRISISLEX		HUMAID		TRECIS	
	Low	**High**	**Low**	**High**	**Low**	**High**
MPNET$_{LM}$+FFT	$\underline{54.5}_{5.0}$ / $\underline{46.7}_{5.7}$	$62.9_{4.9}$ / $\mathbf{55.0}_{3.6}$	$\underline{63.6}_{4.0}$ / $\underline{57.6}_{3.6}$	$\mathbf{73.2}_{2.8}$ / $\underline{66.3}_{3.6}$	$18.1_{1.2}$ / $14.3_{0.5}$	$29.1_{2.1}$ / $22.8_{2.6}$
MPNET$_{LM}$	$49.9^{*}_{4.7}$ / $42.9^{*}_{5.4}$	$56.7^{*}_{5.9}$ / $48.6^{*}_{5.4}$	$56.4^{*}_{4.3}$ / $52.2^{*}_{3.3}$	$64.2^{*}_{5.8}$ / $59.5^{*}_{4.5}$	$\mathbf{32.9}^{*}_{3.0}$ / $\underline{25.2}^{*}_{3.1}$	$\mathbf{30.4}^{*}_{2.4}$ / $\mathbf{23.1}_{3.3}$
MPNET$_{SE}$	$51.7^{*}_{4.0}$ / $44.1^{*}_{4.8}$	$56.7^{*}_{5.6}$ / $49.0^{*}_{5.8}$	$58.4^{*}_{4.4}$ / $53.1^{*}_{2.6}$	$65.6^{*}_{4.1}$ / $60.1^{*}_{3.2}$	$32.1_{2.2}$ / $\underline{24.3}^{*}_{3.7}$	$30.1_{2.3}$ / $\underline{22.9}_{3.6}$
MPNET$_{SE}$+CTS	$\mathbf{56.6}^{*}_{4.9}$ / $\mathbf{49.2}^{*}_{5.5}$	$\mathbf{63.0}_{5.4}$ / $\underline{54.3}_{5.7}$	$\mathbf{66.7}^{*}_{3.1}$ / $\mathbf{60.6}^{*}_{3.0}$	$72.7^{*}_{2.9}$ / $\mathbf{67.4}_{3.7}$	$\underline{32.7}^{*}_{3.1}$ / $\mathbf{25.3}^{*}_{2.8}$	$29.4_{3.0}$ / $22.6_{3.1}$

Table 1: Overall results for event micro-averaged F_1 (x100%) and macro-averaged F_1 (x100%) scores with standard deviations. **Bold** numbers indicate the best performance whereas underlined numbers denote the second best performance in each column. Results with * are significantly different from MPNET$_{LM}$+FFT (p-value < 0.05).

ontology of TRECIS differs with fine-grained information types such as NEW SUB EVENT, EMERGINGTHREATS, and FACTOID.

Cross-Lingual In the cross-lingual setup, we aim to analyse the adaptation to the German language. However, we are not aware of any German datasets which cover crisis events and information types. Therefore, we adopt the GERMAN BASF EXPLOSION (Habdank et al., 2017) and GERMAN FLOODS (Reuter et al., 2015) datasets which represent binary classification tasks about relevancy. We train a classifier on the entire CRISISLEX corpus and map the information type prediction NOT RELATED to the irrelevant class and all other categories to the relevant class. Here, we assume that the SE considers irrelevant and relevant clusters in the embedding space which can boost the relevancy classification. Furthermore, we compare the multilingual variant of MPNET$_{SE}$[4] and the translation[5] to English tweets. We summarize the findings in Table 3 whereby RANDOM corresponds to a randomized classifier. As expected, the comparison of the RANDOM baseline and the MLP classifiers validates the task transfer to the binary classification. Importantly, we experience the effectiveness of CTS in the cross-task transfer by comparing the language model and CTS fine-tuned SE's. The performance improvements with the multilingual variant of MPNET$_{SE}$+CTS further demonstrates the capabilities of fine-tuning in the multi-lingual setting. However, the translated posts outperform the multilingual model in all English model variants. This is an indication of catastrophic forgetting which occurs during training with only English task data. The lack of multi-lingual data in the domain of information type classification is still a challenge.

	Target Corpus		
	CRISISLEX	**HUMAID**	**TRECIS**
CRISISLEX	-	$63.7_{3.8}$ (3.6↑)	$22.9_{2.5}$ (0.0)
HUMAID	$50.4_{5.9}$ (1.4↑)	-	$21.9_{2.2}$ (1.0↓)
TRECIS	$45.8_{4.1}$ (3.2↓)	$56.5_{2.0}$ (0.4↑)	-

Table 2: High-data results of cross-corpus transfer with MPNET$_{SE}$+CTS. For the results, we report the event macro-averaged F_1 (x100%) score with standard deviations. The numbers and arrows in brackets indicate absolute improvements (↑) or degradations (↓) in comparison to MPNET$_{SE}$.

Language	DE	DE → EN
RANDOM	42.2	-
MPNET$_{LM}$	48.8	$\underline{54.6}$
MPNET$_{SE}$+CTS	50.1	**55.1**
MPNET$_{SE}$+ML+CTS	53.4	54.1

Table 3: Results of cross-lingual transfer with MPNET variants trained on CRISISLEX. For the results, we report the event macro-averaged F_1 (x100%) score. The column DE → EN indicates the translation to the English language. The symbol +ML represents the multilingual variant of MPNET which is further trained with CTS.

5 Conclusion

In this work, we investigated the contrastive task-specialization of SE's for the information type classification. The transformation of universal SE's into task-specialized models demonstrates performance gains especially in low-resource setups. Furthermore, we demonstrate the first results and opportunities in cross-corpus, cross-lingual, and cross-task transfer in the focused crisis domain. There are multiple avenues for future research that can improve different aspects with respect to information type classification. Research directions may include but are not limited to: **1)** data augmentation techniques, **2)** retrieval-augmented classification, and **3)** hierarchical contrastive learning.

---

[4] sentence-transformers/
paraphrase-multilingual-mpnet-base-v2
[5] Helsinki-NLP/opus-mt-en-de

Ethical Considerations

Open Source Intelligence (OSINT) has become a significant role for various authorities and NGOs for advancing struggles in global health, human rights, and crisis management (Bernard et al., 2018; Evangelista et al., 2021; Kaufhold, 2021). Following the view of OSINT as a tool, our work pursues the goal to support relief government agencies, organizations, and other stakeholders during ongoing and evolving disaster events. We argue that Natural Language Processing techniques for OSINT and disaster response can have a positive impact on comprehensive situational awareness and in decision-making processes such as coordination of particular services. For example, NLP for social media can enrich the information with the public as co-producers (Li et al., 2018). In contrast, relying on noisy user-generated content as an information source runs the risk of introducing mis- and disinformation. This can cause adverse effects on downstream processing and requires strategies and particular care before the deployment. Furthermore, data privacy issues may arise due to the inherited properties of user-based data. Various anonymization processes should be taken into account for identifying and neutralizing sensitive references (Medlock, 2006).

Limitations

We believe there is much room for improving the contrastive task-specialization method with respect to the multi-class, multi-label, and noisy user-generated text setup. We tested only one variant of language models without considering different transformer encoder sizes or specialized Twitter-based pre-trained models. Furthermore, we did not conduct experiments for a variety of loss functions which may be expanded to triplet loss, cosine similarity, supervised contrastive loss, and the hierarchical variants for higher level information types. We relied for the cross-lingual analysis only on the German language datasets and only conducted experiments for relevancy classification which poses a simplification of information type classification. Future research should consider multi-lingual information type datasets and tasks to comprehensively validate the cross-lingual and cross-task setups in the crisis-related domain. As highlighted in section 5, sophisticated data augmentation methods can further improve the overall classification results but still poses a major challenge for noisy user-generated content. Lastly, recent advanced in the area of instruction-based Large Language Model's (LLM's) should be considered for future research.

References

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM).*

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021a. Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):933–942.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021b. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *15th International Conference on Web and Social Media (ICWSM).*

Rose Bernard, G. Bowsher, C. Milner, P. Boyle, P. Patel, and R. Sullivan. 2018. Intelligence and global health: assessing the role of open source and social media intelligence analysis in infectious disease outbreaks. *Journal of Public Health*, 26(5):509–514.

Cody L. Buntain, Richard McCreadie, and Ian Soboroff. 2021. Incident Streams 2020: TREC-IS in the Time of COVID-19. In *ISCRAM 2021: 18th International Conference on Information Systems for Crisis Response and Management.*

João Rafael Gonçalves Evangelista, Renato José Sassi, Márcio Romero, and Domingos Napolitano. 2021. Systematic Literature Review to Investigate the Application of Open Source Intelligence (OSINT) with Artificial Intelligence. *Journal of Applied Security Research*, 16(3):345–369. Publisher: Routledge _eprint: https://doi.org/10.1080/19361610.2020.1761737.

Matthias Habdank, Nikolai Rodehutskors, and Rainer Koch. 2017. Relevancy assessment of tweets using supervised learning techniques: Mining emergency related tweets for automated relevancy classification. In *2017 4th International conference on information and communication technologies for disaster management (ICT-DM)*, pages 1–8. IEEE.

Marc-André Kaufhold. 2021. *Information Refinement Technologies for Crisis Informatics: User Expectations and Design Principles for Social Media and Mo-*

bile Apps. Springer Fachmedien Wiesbaden, Wiesbaden.

Anna Kruspe, Jens Kersten, and Friederike Klan. 2021. Review article: Detection of actionable tweets in crisis events. *Natural Hazards and Earth System Sciences*, 21(6):1825–1845.

Lifang Li, Qingpeng Zhang, Jun Tian, and Haolin Wang. 2018. Characterizing information propagation patterns in emergencies: A case study with Yiliang Earthquake. *International Journal of Information Management*, 38(1):34–41.

Ben Medlock. 2006. An introduction to NLP-based textual anonymisation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 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 amp; Social Computing*, CSCW '15, page 994–1009, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. 2018. Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *International Journal of Human–Computer Interaction*, 34(4):280–294.

Christian Reuter, Thomas Ludwig, Marc-André Kaufhold, and Volkmar Pipek. 2015. Xhelp: Design of a cross-platform social-media application to support volunteer moderators in disasters. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 4093–4102, New York, NY, USA. Association for Computing Machinery.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 851–860, New York, NY, USA. Association for Computing Machinery.

Philipp Seeberger and Korbinian Riedhammer. 2022. Enhancing crisis-related tweet classification with entity-masked language modeling and multi-task learning. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 70–78, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Xi'ao Su, Ran Wang, and Xinyu Dai. 2022. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679, Dublin, Ireland. Association for Computational Linguistics.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

Ivan Vulić, Iñigo Casanueva, Georgios Spithourakis, Avishek Mondal, Tsung-Hsien Wen, and Paweł Budzianowski. 2022. Multi-label intent detection via contrastive task specialization of sentence encoders. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7544–7559, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ivan Vulić, Pei-Hao Su, Samuel Coope, Daniela Gerz, Paweł Budzianowski, Iñigo Casanueva, Nikola Mrkšić, and Tsung-Hsien Wen. 2021. ConvFiT: Conversational fine-tuning of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1168, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Congcong Wang, Paul Nulty, and David Lillis. 2021. Transformer-based Multi-task Learning for Disaster Tweet Categorisation. In *ISCRAM 2021: 18th International Conference on Information Systems for Crisis Response and Management*.

Matti Wiegmann, Jens Kersten, Friederike Klan, Martin Potthast, and Benno Stein. 2020. Analysis of Detection Models for Disaster-Related Tweets. In *ISCRAM 2020: 17th International Conference on Information Systems for Crisis Response and Management*.

Hamada M. Zahera, Rricha Jalota, Mohamed Ahmed Sherif, and Axel-Cyrille Ngonga Ngomo. 2021. I-aid: Identifying actionable information from disaster-related tweets. *IEEE Access*, 9:118861–118870.

Hongming Zhang, Haoyu Wang, and Dan Roth. 2021a. Zero-shot Label-aware Event Trigger and Argument Classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1331–1340, Online. Association for Computational Linguistics.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip Yu. 2021b. Few-shot intent detection via contrastive pre-training and fine-tuning.

In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1906–1912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.