Adversarial Learning for Robust Emergency Need Discovery in Low Resource Settings

Developing technologies to discover emergency needs in low resource settings is vital for effectively providing aid during disastrous events. In emergency scenarios with limited time and resources, humans may not be able to quickly scan incoming texts and SOSs. NLP models might help with identifying, classifying, and prioritizing distress signals. In low resource and time-sensitive settings, supervised data for training such models is sparse and human annotators might be hard to find. Furthermore, distributions of needs might not be consistent across different emergency scenarios, and populations in varying emergency scenarios may use distinct vocabulary or phrases to express the same need. In turn, applying models across multiple emergency scenarios might be disadvantageous.

Inspired by prior work that uses adversarial learning to overcome domain- and dataset-specific biases, artifacts or distributions [Ganin et al., 2016, Belinkov et al., 2019], we apply adversarial learning to the task of discovering emergency needs in low resource settings. When training a classifier to predict whether and which type of emergency need is expressed in a text, we force our model to predict which disaster occurred. Adversarial learning, implemented through a gradient reversal, penalizes our model when correctly predicting the disaster that occurred. We hypothesize that this may force our networks to generalize well across different disaster scenarios.

Baseline: For each emergency need $n \in \mathcal{N}$, a pre-defined set of possible needs, we train a binary classifier to predict whether n is expressed in sentence s. Each binary classifier consists of a Bi-LSTM encoder $g(s)$ that maps each sentence s to a vector representation v_s, and a MLP $f_n(v_s)$ that predicts whether n is expressed in s. To deal with large class imbalances due to that fact that most texts do not express an emergency need, we weight our loss function, specifically cross-entropy, based on the class imbalance of the training set. Our loss function for each binary classifier is $\mathcal{L}_n = \mathcal{L}(f_n(v_s), y)$, where y is a boolean indicating whether emergency need n is expressed in s.

Applying Adversarial Learning: Since each emergency situation may have different distributions of emergency needs and the needs may be expressed differently in different situations, applying these binary classifiers across events may not work well. During adversarial training, we additionally feed v_s to a new MLP $f_{situation}$ that predicts which disastrous event e occurred. We modify the loss function of our network to become $\mathcal{L} = \mathcal{L}_n + \lambda\mathcal{L}_{\text{Adv}}$, where $\mathcal{L}_{\text{Adv}} = \mathcal{L}(f_{situation}(\lambda_{enc}GRL(g(s))), e)$. λ and λ_{enc} respectively control the weight of the adversarial loss function and the gradient reversal to $g(s)$.

Experiments and Data: We use tweets associated with 8 emergency situations in the past ten years, that were internally annotated with the 11 emergency needs described in the LORELEI Situation Frame task [Christianson et al., 2018] using the EASL framework [Sakaguchi and Van Durme, 2018]. To test our hypothesis, we use a leave-one-out setup where we train our baseline & adversarially-trained binary classifiers on all but one disaster event and test on the held-out event. We repeat this process for all 8 events collected. Table 1 reports the difference in F1 measure between the best performing adversarial model and the baseline model for each emergency need and disaster event. In 42 of the 88 settings, we see no difference (in F1 score) between the best performing adversarial model vs the baseline. In one case, the best performing adversarial model does slightly worse than the baseline, and in the remaining 43 examples, the best adversarial model outperforms the baseline in F1. Inspecting our results, the binary classifiers that predict whether a tweet mentions a *"search"* or *"med"* need achieve the same F1 score as the baseline in all but one case. While each emergency need is often expressed in less than 50% of the tweets, it should be noted that these two needs each appear in less than 10% of the training examples. For the *"regimechange"* need during the 2015 Paris Attack, we notice a 10+ absolute F1 improvement. Upon inspection, the model correctly predicted that a significantly smaller number of tweets represented a *"regimechange"* need compared to the baseline model's predictions. In **extensions** to this work, we will evaluate on a held-out test of new disaster events, and explore other training techniques and experimental setups.

	violence	med	search	food	utils	infra	water	shelter	regimechange	evac	terrorism
2011 NabroEruption	0.12	0.00	0.00	0.00	0.28	0.00	0.13	0.00	0.00	0.14	0.44
2011 EastAfricaDroughts	1.14	0.00	0.00	0.01	0.94	0.03	0.06	0.03	0.00	0.00	0.00
2013 Iran Earthquake	0.00	0.02	0.00	2.37	0.00	0.72	0.00	0.63	4.34	0.00	0.20
2013 India Cyclone	1.50	0.00	0.00	0.00	0.10	0.00	0.00	0.60	0.00	0.42	0.00
2013 EgyptCoupD'état	0.01	0.00	0.00	0.00	0.68	0.72	1.88	0.00	0.10	-0.06	0.00
2014 Turkey Flash Floods	0.00	0.00	0.00	0.00	0.27	0.02	1.68	0.18	0.00	3.54	0.00
2015 Paris Attacks	0.01	0.00	0.00	0.34	0.63	1.00	0.00	0.00	10.02	0.58	0.24
2016 OromoProtest	0.00	0.00	0.00	0.00	0.34	0.00	0.19	0.00	2.66	0.08	0.08

Table 1: Each row indicates the held-out event and each column represents the emergency need predicted. Numbers represent the difference in F1 between the best performing model for each setting and the corresponding baseline binary classifier.

		violence	med	search	food	utils	infra	water	shelter	regimechange	evac	terrorism
accuracy	2016 OromoProtest	0.00	0.00	0.00	0.00	5.61	0.00	0.45	0.65	2.54	0.04	0.00
	2011 NabroEruption	1.06	0.00	0.00	0.00	3.87	0.00	0.97	0.00	0.00	0.39	1.79
	2013 Iran Earthquake	0.00	0.04	0.00	0.08	0.04	0.04	0.00	0.96	9.25	1.08	1.95
	2013 India Cyclone	3.16	0.00	0.00	0.00	0.77	0.00	0.00	3.41	0.00	3.20	0.00
	2015 Paris Attacks	0.04	0.00	0.00	0.61	0.08	1.63	0.00	0.00	6.33	0.08	8.64
	2014 Turkey Flash Floods	0.22	0.00	0.00	0.00	0.89	0.06	0.39	8.42	0.00	0.00	0.00
	2011 EastAfricaDroughts	2.53	0.00	0.00	0.09	1.29	0.04	0.04	1.11	0.00	0.04	0.00
	2013 EgyptCoupD'état	0.12	0.00	0.00	0.00	0.00	0.12	0.08	0.00	0.08	-0.04	0.49
precision	2016 OromoProtest	0.00	0.00	0.00	0.00	4.47	0.00	2.25	2.89	14.01	0.11	0.05
	2011 NabroEruption	0.10	0.00	0.00	0.00	4.77	0.00	0.11	0.00	0.00	0.43	0.35
	2013 Iran Earthquake	0.00	0.02	0.00	5.88	8.33	0.28	0.00	0.63	15.40	0.31	0.20
	2013 India Cyclone	5.22	0.00	0.00	0.00	0.11	0.00	0.00	0.62	0.00	0.60	0.00
	2015 Paris Attacks	0.01	0.00	0.00	4.18	0.85	0.96	0.00	0.00	19.31	0.93	1.00
	2014 Turkey Flash Floods	0.39	0.00	0.00	0.00	0.19	0.40	6.94	1.68	0.00	0.00	0.00
	2011 EastAfricaDroughts	1.78	0.00	0.00	1.39	1.08	0.11	0.15	0.88	0.00	0.00	0.00
	2013 EgyptCoupD'état	0.01	0.00	0.00	0.00	0.41	0.97	2.02	0.00	4.09	-0.20	2.75
recall	2016 OromoProtest	0.00	0.00	0.00	0.00	0.39	0.00	0.00	0.00	0.00	0.00	0.36
	2011 NabroEruption	0.00	0.00	0.00	0.00	0.00	0.00	4.84	0.00	0.00	0.00	0.37
	2013 Iran Earthquake	0.00	3.91	0.00	1.36	0.00	1.25	0.00	0.64	0.00	0.00	0.00
	2013 India Cyclone	0.81	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2015 Paris Attacks	13.73	0.00	0.00	0.35	0.50	1.55	0.00	0.00	0.00	0.43	-0.32
	2014 Turkey Flash Floods	0.00	0.00	0.00	0.00	0.85	0.00	0.90	0.00	0.00	2.20	0.00
	2011 EastAfricaDroughts	22.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2013 EgyptCoupD'état	0.00	0.00	0.00	0.00	0.84	0.57	1.04	0.00	0.00	0.00	0.00

Table 2: Each row indicates the held-out event and each column represents the emergency need predicted. Numbers represent the difference in accuracy between the best performing model for each setting and the corresponding baseline binary classifier. The first column indicates the difference in which metric is reported. Note that for a given disaster event and SF type, the numbers do not correspond to the same model.

References

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 256–262, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-1028. URL `https://www.aclweb.org/anthology/S19-1028`.

Caitlin Christianson, Jason Duncan, and Boyan Onyshkevych. Overview of the darpa lorelei program. *Machine Translation*, 32(1-2):3–9, June 2018. ISSN 0922-6567. doi: 10.1007/s10590-017-9212-4. URL `https://doi.org/10.1007/s10590-017-9212-4`.

Keisuke Sakaguchi and Benjamin Van Durme. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P18-1020`.