Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup

Jishnu Ray Chowdhury[1], **Cornelia Caragea**[1], **and Doina Caragea**[2]
[1]Computer Science, University of Illinois at Chicago
[2]Computer Science, Kansas State University
jraych2@uic.edu, cornelia@uic.edu, dcaragea@ksu.edu

Abstract

Distinguishing informative and actionable messages from a social media platform like Twitter is critical for facilitating disaster management. For this purpose, we compile a multilingual dataset of over 130K samples for multi-label classification of disaster-related tweets. We present a masking-based loss function for partially labeled samples and demonstrate the effectiveness of Manifold Mixup in the text domain. Our main model is based on Multilingual BERT, which we further improve with Manifold Mixup. We show that our model generalizes to unseen disasters in the test set. Furthermore, we analyze the capability of our model for zero-shot generalization to new languages. Our code, dataset, and other resources are available on Github.[1]

1 Introduction

In times of disaster, affected individuals often turn to social media platforms, such as Twitter or Facebook, to express their feelings generated by a disaster, update friends and relatives on their status, request help or supplies, or report useful information to the disaster response teams. Response organizations can use social media to increase situational awareness by providing information about disaster status, ongoing rescue operations, and disaster warnings (Palen and Hughes, 2018). However, the low entry-barrier of social media platforms, where everybody can post their own "news" in real-time, leads to information overload, making it hard for users to find relevant and useful information (Reuter et al., 2018). Thus, it is crucial to filter out the non-informative messages, and to distinguish among different categories of informative messages to ensure that a message reaches its target users. In turn, this can help facilitate disaster response and increase situational awareness.

Towards this goal, in recent years, many works have focused on disaster-related tweet classification (Alam et al., 2018b; Mazloom et al., 2018; Nguyen et al., 2017; Li et al., 2017; Neppalli et al., 2018; Caragea et al., 2016, 2011). However, most of these works have focused on the classification of English tweets only, with a few notable exceptions (Musaev and Pu, 2017; Khare et al., 2018; Lorini et al., 2019; Torres et al., 2019). We stress that there are a lot of disaster-prone non-English-speaking countries, which could benefit from a multilingual classifier that can be used in real-time to identify useful information on social media. Furthermore, there is a lack of a large scale standard multilingual disaster-related dataset for multi-label classification with diverse disaster types. Against this background and needs, we make the following contributions:

1. We aggregate existing datasets into a large disaster dataset using a new annotation scheme. Furthermore, by utilizing a class-mask (elaborated in Section 4.1), we make use of both binary-classification data and multi-class classification data in the same training phase.

2. We explore Manifold Mixup (Verma et al., 2019) in the natural language-based disaster domain. Manifold Mixup is a regularization technique originally introduced in computer vision tasks.

3. We employ Multilingual BERT (Devlin et al., 2019) to train multilingual classifiers. We demonstrate its generalization on unseen disasters and its zero-shot transfer-ability to languages not present in the training data.

[1]https://github.com/JRC1995/Multilingual-BERT-Disaster

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 292–298

2 Related Work

There are numerous prior works on disaster-related tweet classification. For example, Imran et al. (2013; 2015) focus on classifying and extracting actionable information from disaster-related tweets, assuming that sufficient labeled tweets from the ongoing disaster are available for model training. Later, Imran et al. (2016b) explore real-time classification of tweets from a target disaster using models trained on past disasters. Nguyen et al. (2017) introduce a Convolutional Neural Network that performs robustly even on out-of-event data during inference. Other works explore domain-adaptation that uses labeled tweets from past disasters and unlabeled tweets from an ongoing disaster (Li et al., 2018; Alam et al., 2018a). Kruspe (2019) take a few-shot learning approach, in which a disaster-specific model is trained using only a few (around 10) examples for disaster-related tweet detection. In contrast, we train a universal model on diverse disaster types for fine-grained classification and show that it performs remarkably well on unseen disaster types without further training (specifically, it achieves zero-shot generalization to unseen events). Wang and Lillis (2019) classify actionable tweets using ELMo contextual word embeddings, whereas Ma (2019) uses a monolingual BERT-based model for disaster-related tweet classification. In contrast, we work with a multilingual model, which we compare with multiple baselines, and augment with Manifold Mixup.

Regarding cross-lingual approaches, Dittrich and Lucas (2014) present a real-time application tool for multilingual tweet classification and disaster detection. However, this tool requires a long training phase with tweets from specific areas for robust detection, and its multilingual classifier filters messages based on shallow matching of pre-selected keywords (and their translations). Musaev and Pu (2017) construct a multilingual model for tweet classification using multilingual Wikipedia articles as knowledge repository. Khare et al. (2018) also take into account cross-lingual capabilities, however, this is limited to the fixed few number of languages that are present in their annotated training data and do not generalize to new languages without further training. M-BERT overcomes these shortcomings. Similar to us, Lorini et al. (2019) use multi-lingual word embeddings for cross-lingual classification, but they use non-contextual embeddings. Torres et al. (2019) use contextualized word embeddings for cross-lingual analysis, but only on limited samples ($8K$) and only for two languages (English and Spanish).

A few recent works (Pires et al., 2019; K et al., 2020) also demonstrate the strong cross-lingual and zero-shot transfer capabilities of M-BERT, but not in the disaster domain.

3 Aggregated Dataset

To prepare our large multilingual dataset, we aggregated several resources from CrisisNLP,[2] together with two resources from CrisisLex.[3] Specifically, we used Resource #1 (Imran et al., 2016a), Resource #4 (Nguyen et al., 2017), Resource #5 (Alam et al., 2018c), and Resource #7 (Alam et al., 2018a) from CrisisNLP, and CrisisLexT6 (Olteanu et al., 2014) and CrisisLexT26 (Olteanu et al., 2015) from CrisisLex. The *original classes* in each resource, together with the mapping to the *new classes* included in our data set, can be seen in Table 1. Some examples from the dataset are shown in Table 2. For the dataset construction, the following classes were included:

1. **Casualties and Damage (C & D):** This class consists of tweets related to affected individuals, displaced people, building collapse, rescue operations, infrastructure and utilities damage, needs of affected people, missing or trapped people, and other topics related to situational awareness and disaster response.
2. **Donation and Volunteering (D & V):** This class consists of tweets related to donations, volunteering requests, and other needs and requests targeted to individuals following the disaster and/or supporting the victims.
3. **Caution and Advice (C & A):** This class consists of tweets recommending caution, expressing warnings, or providing advice regarding the crisis situation. Such tweets are useful for the affected individuals.
4. **Informative (I):** This is a general class, which includes: tweets belonging to any of the above three classes; tweets with niche categories that do not fit into the above classes; tweets with more vague classes (e.g., "*other useful information*"); and tweets originally labeled with only binary classes such as *relevant* or *informative*.

[2] https://crisisnlp.qcri.org/
[3] https://crisislex.org/data-collections.html

Original Class	*New Class*	*Original Class*	*New Class*	*Original Class*	*New Class*
CrisisNLP Resource #1				**CrisisNLP Resource #5**	
Other relevant info.	I	Disease signs, symptoms	C & D, I	Other relevant info.	I
Displaced people	C & D, I	Affected People	C & D, I	Affected individuals	C & D, I
Needs of those affected	C & D, I	Prevention	C & A, I	Injured or dead	C & D, I
Donations of money	D & V, I	Death Reports	C & D, I	Vehicle damage	C & D, I
Not related to crisis	N	Disease Transmission	I	Infrastructure & util.	C & D, I
Infrastructure	C & D, I	Treatment	I	Volun. & Donation	D & V, I
Shelter and supplies	D & V, I	Displaced people & evac.	C & D, I	Missing or found	C & D, I
Other relevant	I	Other Useful Info.	I	**CrisisNLP Resource #7**	
Injured and dead	C & D, I	Money	D & V, I	Relevant	I
Volunteer or Prof. services	D & V, I	Caution & Advice	C & A, I	Not relevant	N
Sympathy & emotional	N	Humanitarian Aid	D & V, I	**CrisisLexT6**	
Infrastructure & util.	C & D, I	People missing or found	C & D, I	on-topic	I
Donations supp. & volun.	D & V, I	Response Efforts	C & D, I	off-topic	N
Not related or irrelevant	N	Urgent Needs	D & V, I	**CrisisLexT26**	
Requests for Help/Needs	D & V, I	Not Informative	N	Affected individuals	C & D, I
Praying	N	**CrisisNLP Resource #4**		Not applicable	N
Missing, trapped, found	C & D, I	Other Useful Info.	I	Donations & volun.	D & V, I
Not Relevant	N	Not related or irrelevant	N	Sympathy & support	N
Informative	I	Affected Individuals	C & D, I	Caution and advice	C & A, I
Injured or dead people	C & D, I	Sympathy and support	N	Infrastructure & util.	C & D, I
Infrastructure damage	C & D, I	Donations and volunteering	D & V, I	Other Useful Info.	I
Personal, sympathy, support	N				

Table 1: Overview of mappings between the original classes and the new classes.

Examples	Original label	New label
Another typhoon named internationally as #BOPHA will hit #Southern #Mindanao. It will be named #Pablo in RP. Oh noooooes!!	Other Useful Info.	Informative
#RescuePH Rescue pls family trapped at Blk64 Lot2 Phase2 Dela Costa Homes V Burgos Montalban Rizal.Family of 4w/2 children. ...”	Affected individuals	Casualties & Damage
Methods of prevention of Coronavirus: Use a tissue when coughing or sneezing, cover your mouth and nose with it, and then get rid of it	Prevention	Caution & Advice

Table 2: Examples from the aggregated dataset with the original and new label.

5. **Non-Informative (N):** This class consists of all the tweets that are not included in the Informative class.

Some of the above classes (for example, Casualties and Damage) are very broad and could be broken down into more specific classes. However, keeping them broad simplifies the aggregation of different annotation schemes and prevents the formation of multiple fine-grained but sparse classes.

During aggregation, we treat the first four classes as mutually exclusive (they are also mutually exclusive with the Non-Informative class). We filter out duplicate tweets. For duplicates from different resources that were originally associated with more than one mutually exclusive classes, we keep only the first class, based on the order in which classes are listed above.

Statistics about the final dataset with respect to the number of tweets per class and per language are shown in Table 3.

Number of Tweets per Class				
C & D	D & V	C & A	I	N
16,235	9,125	3,634	79,473	54,947
Number of Tweets per Language				
English	Spanish	Italian	French	Others
123,406	4,724	1,581	666	4043

Table 3: Samples per class and per language.

4 Methods

4.1 Classification Approach

In general, all of our models use a sentence encoder to map a tweet to a single vector sentence representation. The vector is then fed to multiple binary classifiers. Specifically, we train four classifiers. One classifier distinguishes between *Informative* and *Non-Informative* classes, while the other three classifiers correspond to the remaining three classes: *Casualties and Damage*, *Caution and Advise*, and *Donations and Volunteers*, respectively

(each classifier predicts whether a tweet belongs to a particular class or not).

We should note that there are many tweets belonging to the *Informative* class, which originally only had binary classes (informative/non-informative or relevant/non-relevant). While those tweets may also belong to one of the more fine-grained classes, their class could not be determined, if it was not available in the original resources. In other words, many of the samples in the dataset are **partially labeled** (where the binary "informative" or "Non-Informative" class is present but the other fine-grained class information is absent). However, ignoring all partially labeled tweets would result in removing nearly half of the data. In order to get the benefit from the binary-classification-only data while also enabling the same model to work on multi-label classification we devise a label masking strategy. Precisely, the mask is used to ensure that the loss signal is only propagated from classes which are annotated. The strategy is discussed in further details below.

By default, we use the negative class for the three fine-grained categories as dummy ground-truth for such cases. We then mask out (i.e., zero out) the loss from the dummy ground truth cases during training. For masking the loss from dummy ground truth, we use a class mask m_{ij} (i.e., a mask for the j^{th} class and the i^{th} sample), where m_{ij} is 0 if the actual j^{th} class ground truth is not present for the i^{th} sample, otherwise it is 1. Overall, we use binary cross entropy for each of the classifiers with the class masks and class weights. The loss function can be formalized as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{j=1}^{K} m_{ij}\cdot(c_j y_{ij}\cdot\log P_\theta(y_{ij}|x_i) + (1-y_{ij})\cdot\log(1-P_\theta(y_{ij}|x_i))) \quad (1)$$

where K is the number of classes, N is the number of samples, c_j is the class weight for the j^{th} class, x_i is the i^{th} tweet string, θ represents the model parameters, and $P_\theta(y_{ij}|x_i)$ is the model prediction for the i^{th} tweet string and the j^{th} class. We use class weights to handle class imbalance. We consider the cost of filtering out an important and urgent tweet to be higher than the cost of including a non-informative tweet. This is why we bias our model towards recall by using class-weights of value ≥ 1 for the positive classes. We use a class weight of 1 for the *Informative* versus *Non-Informative* classes (as they are fairly balanced, with already a small bias towards the positive class). For the fine-grained classes, we use the following formula to find the class weights:

$$c_i = \frac{\max_j(\{count(class_j)|class_j \in C\})}{count(class_i)} \quad (2)$$

where C = {*'Non-Informative'*, *'Casualties & Damage'*, *'Donation & Volunteering'*, *'Caution & Advice'*}. We should note here that the loss function does not take the positive classes as mutually exclusive since, in principle, a single tweet could have multiple classes (for example, a tweet could have both 'Caution and Advice' and 'Casualties and Damage').[4]

4.2 Sentence Encoders

As we focus on supervised learning from large data, we use some standard text classification models, such as FastText (Joulin et al., 2017; Mikolov et al., 2018), CNN (Kim, 2014), XML-CNN (Liu et al., 2017), and BiLSTM (Adhikari et al., 2019) as baseline sentence encoders. We compare them with M-BERT (Devlin et al., 2019) encoders.

Manifold Mixup. We also adopt Manifold Mixup (Verma et al., 2019) in our main model (M-BERT). Mixup (Zhang et al., 2018) was originally introduced in the image classification domain as a data augmentation based regularization technique. The original technique augments data by linearly interpolating two different input data samples and their associated classes. In effect, this helps make a model more robust by inducing a linear behavior in-between training samples. Guo et al. (2019) show that Mixup both at the level of word embeddings and at the level of sentence embeddings (output of sentence encoder) is effective for text classification. Manifold Mixup is a more recent variant of the original input Mixup, where the hidden states of two different data samples, along with their associated classes, are linearly interpolated. To do this, a mixup ratio λ is sampled from a $Beta$ distribution, as: $\lambda \sim Beta(\alpha, \alpha)$.

Next, a hidden layer l is randomly chosen for mix up. Let h_i^l be the randomly chosen l^{th} hidden

[4]Even though the classifiers are not mutually exclusive, the annotated classes (excluding the more general informative class) are kept mutually exclusive because there were not many multi-label annotations in the original data and most tweets tend to belong to only one of the specific classes. One could also use mutually exclusive classifiers with the given data.

Model	$\mathbf{F_1}$ (mean, std)	$\mathbf{F_1}$ (mean, std)
	Meteor (802)	Cyclone (2,473)
FastText	73.79 ± 0.55	81.81 ± 0.67
FastText$_{hier}$	74.61 ± 1.19	81.59 ± 0.50
CNN	66.83 ± 2.16	82.40 ± 0.48
XML-CNN	74.34 ± 2.51	82.04 ± 1.02
BiLSTM	72.40 ± 2.47	82.65 ± 0.88
M-BERT	83.63 ± 0.78	83.99 ± 0.47
+Word Mixup	83.02 ± 0.95	$\mathbf{85.48 \pm 0.64}^*$
+Sentence Mixup	82.62 ± 0.55	83.78 ± 0.73
+Mixup	$\mathbf{84.90 \pm 0.84}^*$	$85.09 \pm 0.86^*$
	Flood (684)	Mixed (10,000)
FastText	75.53 ± 0.54	82.91 ± 0.05
FastText$_{hier}$	76.21 ± 0.81	84.09 ± 0.24
CNN	72.78 ± 3.18	83.70 ± 0.12
XML-CNN	77.03 ± 1.26	84.56 ± 0.32
BiLSTM	74.35 ± 0.25	84.33 ± 0.15
M-BERT	78.51 ± 0.93	86.77 ± 0.18
+Word Mixup	$79.18 \pm 0.95^*$	$87.31 \pm 0.29^*$
+Sentence Mixup	$\mathbf{79.85 \pm 0.50}$	$87.32 \pm 0.20^*$
+Manifold Mixup	79.36 ± 0.79	$\mathbf{87.39 \pm 0.23}^*$

Table 4: $\mathbf{F_1}$ scores on four test datasets (English Only). * means that the difference from M-BERT is statistically significant.

layer output from the i^{th} tweet sample, and let h^l_j be the hidden layer output from the j^{th} sample. The two outputs can be mixed up as follows:

$$\tilde{h^l}_i = \lambda \cdot h^l_i + (1 - \lambda) \cdot h^l_j \quad (3)$$

where $\tilde{h^l}_i$ is the augmented (mixed-up) hidden state. We use the same λ to mix the hidden states of the tweet samples i and j, and also the corresponding ground truth classes and class masks for each class k included in our dataset:

$$\tilde{y}_{ik} = \lambda \cdot y_{ik} + (1 - \lambda) \cdot y_{jk} \quad (4)$$

$$\tilde{m}_{ik} = \lambda \cdot m_{ik} + (1 - \lambda) \cdot m_{jk} \quad (5)$$

where, $\tilde{y}_{ik}$ and $\tilde{m}_{ik}$ are the corresponding mixed-up class and class-mask, respectively. The augmented class-masks can be intuitively thought of as indicating to what extent the loss following the corresponding augmented ground truth class should be considered. If the major fraction of the mixed up class is a dummy class, then the corresponding augmented class-mask should have a low value. We also compare Word Mixup and Sentence Mixup. Word Mixup and Sentence Mixup can be considered as special cases of Manifold Mixup where the mixup is applied on only a specific layer. In case of word mixup, it is the first embedding layer, and in case of Sentence Mixup it is the final layer output of the sentence encoder.

Model	$\mathbf{F_1}$ (mean, std)	$\mathbf{F_1}$ (mean, std)
	Meteor (930)	Cyclone (2,558)
M-BERT	81.39 ± 1.42	84.60 ± 1.06
+Word Mixup	80.16 ± 3.09	84.47 ± 0.81
+Sentence Mixup	80.97 ± 1.65	84.87 ± 0.71
+Manifold Mixup	$\mathbf{81.73 \pm 0.78}$	$\mathbf{85.15 \pm 0.79}^*$
	Flood (768)	Mixed (10,000)
M-BERT	79.24 ± 1.28	86.63 ± 0.22
+Word Mixup	$78.32 \pm 0.81^*$	87.10 ± 0.19
+Sentence Mixup	78.77 ± 0.69	86.98 ± 0.18
+Manifold Mixup	$\mathbf{79.84 \pm 0.68}^*$	$\mathbf{87.44 \pm 0.11}^*$

Table 5: $\mathbf{F_1}$ scores on four test datasets (Full Dataset). * means that the difference from M-BERT is statistically significant.

5 Experiments and Results

5.1 Experimental Setup

We use four datasets for testing: Russia Meteor, Cyclone Pam, Philippines Flood, and Mixed disasters. To demonstrate the generalization capabilities of our models, we ensured that the first three datasets are from disasters that are absent in the training set. For M-BERT-based models, we use a mini batch size of 32, a learning rate of 10^{-3} for non-BERT parameters, and a fine-tuning rate of 2×10^{-5} for M-BERT parameters. We set the parameter α of the $Beta$ distribution for the Mixup equation to 2. We run each model five times and report the mean and standard deviation of the results obtained in the 5 runs. For the other models, we import the parameter settings from their corresponding paper and then perform light manual tuning. The exact hyperparameters are available on Github.[5]. For significance testing, we used the paired t-test ($p \leq 0.05$) (Dror et al., 2018) Note that the CNN baseline is also similar to the model used by Nguyen et al. (2017) which was demonstrated to be a strong performer in disaster-related classification.

5.2 Results

In Table 4, we show the results on English only samples, and in Table 5, we show the results on the full multilingual test sets. As can be seen in Table 4, the M-BERT outperforms all the non-BERT baseline models. Using Manifold Mixup consistently increases the performance of base M-BERT in all cases, often also working better than Word Mixup and Sentence Mixup, especially for the multilingual setting (see Table 5). Manifold Mixup

[5] https://github.com/JRC1995/Multilingual-BERT-Disaster

Language	Samples	$\mathbf{F_1}$ (mean, std)
French (zero shot)	666	81.33 ± 0.77
Italian (zero shot)	1,581	75.44 ± 0.67
Spanish (zero shot)	4,724	85.26 ± 0.37

Table 6: $\mathbf{F_1}$ scores of M-BERT + Manifold Mixup.

C & D	D & V	C & A	I
79.8 ± 0.5	77.5 ± 1	70.3 ± 1	90.9 ± 0.2

Table 7: Per-class $\mathbf{F_1}$ of M-BERT+Manifold Mixup on Multilingual Mixed Disasters.

either outperforms or is very close to the other Mixup techniques. Table 6 shows the results of the cross-lingual experiments with M-BERT and Manifold Mixup for French, Italian, and Spanish languages, respectively, in a zero shot setting (Pires et al., 2019), where no tweets in the test language are included in the training set.

As can be seen from the table, the zero shot F1-score on Spanish is 85.25% (which is comparable to the best results in the previous experiments), despite the fact that no Spanish tweets were included in the training. The zero shot F1-scores on French and Italian are 81.33% and 75.44%, respectively. These results show that the M-BERT+Manifold Mixup model has good generalization capability in the new language (zero shot) setting. Thus, we can conclude that our M-BERT+Manifold Mixup model has great capability to generalize to a disaster in a new language (unseen in the training set) as long as the language is one of the 104 languages on which M-BERT was pre-trained. This is a strong result given that disasters can happen in countries with limited resources for automated classification of social media information.

In Table 7, we check the binary classification performance of M-BERT+Manifold Mixup for each class. As we can see, our model achieves an F_1 above 90% for the binary classification task of distinguishing whether a tweet is informative or not. Interestingly, it does not perform too poorly on Caution & Advice either despite having very limited samples for this class in the training set.

6 Conclusion

We present a way to aggregate prior disaster-related resources to compile a large scale tweet dataset for multi-label classification utilizing both multi-class classes and binary classes. We motivate the use of M-BERT for disaster-related tweet classification and we demonstrate its strong performance on unseen disasters and languages. We also motivate the use of Manifold Mixup for further improvement. In the future, it would be interesting to explore weak suervision and other data augmentation techniques to improve models' robustness further.

Acknowledgments

We thank NSF and Amazon Web Services for support from grants IIS-1741345 and IIS-1912887, which supported the research and the computation in this study. We also thank NSF for support from the grants IIS-1903963, and CMMI-1541155. We are grateful to our anonymous reviewers and the ACL student mentor Valerio Basile for their constructive feedback.

References

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. Rethinking complex neural network architectures for document classification. In *ACL*.

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018a. Domain adaptation with adversarial training and graph embeddings. In *ACL*.

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018b. Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. In *AAAI Conference on Web and Social Media*.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018c. Crisismmd: Multimodal twitter datasets from natural disasters. In *ICWSM*.

Cornelia Caragea, Nathan J. McNeese, Anuj R. Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, Andrea H. Tapia, C. Lee Giles, Bernard J. Jansen, and John Yen. 2011. Classifying text messages for the haiti earthquake. In *8th Proceedings of ISCRAM*.

Cornelia Caragea, Adrian Silvescu, and Andrea H. Tapia. 2016. Identifying informative messages in disaster events using convolutional neural networks. In *Proceedings of ISCRAM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*.

André Dittrich and Christian Lucas. 2014. Is this twitter event a disaster?

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Comp. Surveys (CSUR)*.

Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. 2013. Practical extraction of disaster-relevant information from social media. In *WWW*.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016a. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *LREC*.

Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. 2016b. Enabling rapid classification of social media communications during crises. *IJISCRAM*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *ACL*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual {bert}: An empirical study. In *ICLR*.

Prashant Khare, Grégoire Burel, Diana Maynard, and Harith Alani. 2018. Cross-lingual classification of crisis data. In *ISWC*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*.

Anna Kruspe. 2019. Few-shot tweet detection in emerging disaster events. *arXiv preprint arXiv:1910.02290*.

Hongmin Li, Doina Caragea, and Cornelia Caragea. 2017. Towards practical usage of a domain adaptation algorithm in the early hours of a disaster. In *ISCRAM*.

Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018. Disaster response aided by tweet classification with a domain adaptation approach. *JCCM*.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *ACM SIGIR*.

Valerio Lorini, Carlos Castillo, Francesco Dottori, Milan Kalas, Domenico Nappo, and Peter Salamon. 2019. Integrating social media into a pan-european flood awareness system: A multilingual approach. In *ISCRAM*.

Guoqin Ma. 2019. Tweets classification with bert in the field of disaster management.

Reza Mazloom, HongMin Li, Doina Caragea, Muhammad Imran, and Cornelia Caragea. 2018. Classification of twitter disaster data using a hybrid feature-instance adaptation approach. In *ISCRAM*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *LREC*.

A. Musaev and C. Pu. 2017. Towards multilingual automated classification systems. In *ICDCS*.

Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. 2018. Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In *ISCRAM*.

Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2017. Robust classification of crisis-related data on social networks using convolutional neural networks. In *ICWSM*.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *AAAI Conference on Weblogs and Social Media*.

Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to expect when the unexpected happens: Social media communications across crises. In *CSCW*.

Leysia Palen and Amanda L. Hughes. 2018. *Social Media in Disaster Communication*, pages 497–518. Springer International Publishing, Cham.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. 2018. Social media in crisis management: An evaluation and analysis of crisis informatics research. *Int. J. of HCI*, 34(4):280–294.

Torres, Carmen Vaca, and Johnny. 2019. Cross-lingual perspectives about crisis-related conversations on twitter. In *WWW*.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *ICML*.

Congcong Wang and David Lillis. 2019. Classification for crisis-related tweets leveraging word embeddings and data augmentation.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *ICLR*.