Proceedings of the OSACT 2022 Workshop @LREC2022, pages 71–78
Marseille, 20 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Classifying Arabic Crisis Tweets using Data Selection and Pre-trained Language Models

Alaa Alharbi, Mark Lee
University of Birmingham, UK
Taibah University, KSA
{axa1314, m.g.lee}@bham.ac.uk
alaharbi@taibahu.edu.sa

Abstract

User-generated Social Media (SM) content has been explored as a valuable and accessible source of data about crises to enhance situational awareness and support humanitarian response efforts. However, the timely extraction of crisis-related SM messages is challenging as it involves processing large quantities of noisy data in real-time. Supervised machine learning methods have been successfully applied to this task but such approaches require human-labelled data, which are unlikely to be available from novel and emerging crises. Supervised machine learning algorithms trained on labelled data from past events did not usually perform well when classifying a new disaster due to data variations across events. Using the BERT embeddings, we propose and investigate an instance distance-based data selection approach for adaptation to improve classifiers' performance under a domain shift. The K-nearest neighbours algorithm selects a subset of multi-event training data that is most similar to the target event. Results show that fine-tuning a BERT model on a selected subset of data to classify crisis tweets outperforms a model that has been fine-tuned on all available source data. We demonstrated that our approach generally works better than the self-training adaptation method. Combing the self-training with our proposed classifier does not enhance the performance.

Keywords: Crisis Detection, Domain Adaptation, Data Selection, Self-training, BERT

1. Introduction

In the last decade, Social Media (SM) content has been explored by Natural Language Processing (NLP) and data mining researchers as a valuable and accessible source of data. Many of these studies have investigated the problem of mining social media (notably microblogging sites such as Twitter) to extract emergency events. They have demonstrated that SM posts contain actionable and time-critical information that can be leveraged to respond effectively to disasters (Olteanu et al., 2015; Imran et al., 2016). The automatic extraction of crises from SM as they unfold can enhance situational awareness and support humanitarian response efforts.

While SM event detection is challenging because of noisy language, several studies have demonstrated the possibility of identifying crisis-related data from Twitter and categorising these into different information types using conventional supervised Machine Learning (ML) techniques (Imran et al., 2013; Rudra et al., 2015; Rudra et al., 2018; Singh et al., 2017) and Deep Neural Networks (DNNs) (Caragea et al., 2016; Nguyen et al., 2017; Neppalli et al., 2018; Alharbi and Lee, 2019; Kozlowski et al., 2020). Annotated training datasets are unlikely to be available in real-time from current events since obtaining sufficient human-labelled data is time-consuming and labour intensive. The unavailability of training data from newly occurred crises and the growing generation of SM content challenge the timely processing of crisis data by the emergency responders. Researchers proposed to use data from historical events. Several studies have shown that DNNs generalise better than conventional ML approaches (Nguyen et al., 2017; Neppalli et al., 2018; Alharbi and Lee, 2019). However, DNNs are still challenged by out-of-distribution learning when classifying unseen crises – especially if they are trained on data from cross-type crises due to data variations across such events. Out-of-distribution (covariate shift) refers to the different probability distributions of input features across the training (source) and test (target) data.

Textual data varies across SM events in two main aspects: topic and language. Messages from two events of the same type can discuss various topics emerging from the event properties and its distinct aspects such as time, cause, related entities and impact. Such topic variations across events can lead to substantially different feature distributions. Events on SM are discussed with varying levels of formality, in various dialects and languages, resulting in more challenges to handle the out-of-distribution problem when learning from cross-event historical data. Supervised classifiers' performance drops on test data if it does not follow the training set distribution as many ML algorithms assume (Ramponi and Plank, 2020).

Recent works on Domain Adaptation (DA) show that training on a domain similar to the target data results in performance gains for various NLP tasks. Ruder et al. (2017) explored the performance of several domain similarity metrics on different text representations for data selection in the sentiment analysis context. The authors also proposed a subset-level data selection approach that outperforms instance-level selection. In the same vein, Guo et al. (2020) studied differ-

ent domain distance measures and presented a bandit-based multi-source domain adaptation model for sentiment classification. Ma et al. (2019) presented a domain adaptation method based on data selection and curriculum learning to fine-tune BERT models for text classification. Leveraging pre-trained Language Model (LM) representations, Aharoni and Goldberg (2020) proposed data selection approaches for multi-domain Machine Translation (MT) using cosine similarity in embedding space.

This study aims at improving the cross-crisis (cross-domain) classification tasks using DA methods. We adopt a data selection approach to train a classifier on the best matching data for the target emergency event instead of using all multi-event source data. Training a classifier using examples that are dissimilar to the target data can adversely affect the model performance. As pre-trained contextualised LMs have achieved state-of-the-art results in various NLP tasks, we exploit their representation to propose a data selection approach based on the document similarity in the embedding space. The selected data has been used to fine-tune a Bidirectional Encoder Representation from Transformer (BERT) model to identify crisis-related posts from Twitter and classify the messages into different information types. We took advantage of transfer learning and used BERT as a classifier. Fine-tuning a model pre-trained on large data eliminates the need for massive training examples that are required to train a DNN from scratch. The presented adaptation strategy is unsupervised as it does not require labelled data from the target domain. It can also be adopted during the early hours of a crisis when small unlabelled data is available from the target event. To the best of our knowledge, we are the first to adopt the data selection with pre-trained LMs in the crisis detection domain. This work is also the first to investigate DA approaches to perform cross-domain crisis Arabic Twitter classification.

2. Related Work

Recently, researchers have adopted DA approaches to improve the generalisation of supervised models trained on past crisis data to classify unseen new crises. They showed that learning from both the labelled source and unlabelled target data is better than learning only from source labelled data. Several studies adopted an unsupervised DA approach using self-training (Li et al., 2015; Li et al., 2017; Li et al., 2018a; Li et al., 2018b; Li et al., 2021). Earlier work showed that an iterative self-training improved the performance of a NB classifier (Li et al., 2015; Li et al., 2017; Li et al., 2018a; Li et al., 2018b). Li et al. (2018b) demonstrated that the self-training strategy outperformed a feature-based correlation alignment method. Mazloom et al. (2018) proposed a hybrid feature-instance DA method using matrix factorisation and the k-nearest neighbours algorithm to learn a NB classifier for the target event. The work was extended by Mazloom et al. (2019) who combined the feature-instance approach with the self-training algorithm presented by Li et al. (2017). Li et al. (2021) used self-training with CNN and BERT models and highlighted that self-training improved the performance of the Deep Learning (DL) models. For retraining the base classifier, they used a soft-labelling strategy.

Alam et al. (2018) proposed an approach based on adversarial training and graph embeddings in a single DL framework. The adversarial training minimises the distribution shift across domains, whereas graph-based learning encodes similarity between source and target instances. Krishnan et al. (2020) created a multi-task domain adversarial attention network based on a shared Bi-LSTM layer to filter Twitter posts for crisis analytics under domain shift. The tasks are relevancy, priority level, sentiment and factoid detection. Chen et al. (2020) used a BERT-based adversarial model to classify tweets gathered during a disaster into different information categories. ALRashdi and O'Keefe (2019) proposed to use a distant supervision-based framework to label the data from emerging disasters. The pseudo-labelled target data is then used with labelled data from past crises of a similar type to train a classifier. Li and Caragea (2020) explored the use of the domain reconstruction classification approach on disaster tweets, which aims at reducing the covariate shift between source and target data distributions using an autoencoder. The authors showed that this approach outperformed the DA method proposed by Alam et al. (2018). To contribute to this line of research, we have adopted a selection-based DA approach that leverages pre-trained LMs. Our approach is an unsupervised DA as it does not require any labelled instances from the target event and can be utilised during the early hours of a disaster when small unlabelled data is available from the target event. We also explored whether combining the selection method with the self-training DA approach improves the performance. The evaluation was conducted on two crisis-related tasks as described in the next sections.

3. Methodology

3.1. Domain Definition and Tasks Description

In the NLP literature, the notion of domain typically refers to a specific corpus that differs from other domains in the topic, genre, style, etc. (Ramponi and Plank, 2020). In this work, a domain is defined as a dataset that has been collected from SM for a specific crisis. Hence, each crisis data represents a distinct domain. A crisis is a real-world emergency event that occurs at a particular time and location and is characterised by a main topic representing its type (flood, shooting, etc.). In this research, experiments will be performed using Arabic SM data which poses an additional challenge as they include several dialects that can vary across events based on their geographical lo-

cations. We consider two main SM crisis-related tasks as follows.

3.1.1. Relevancy Detection

This task aims at identifying crisis messages from SM by classifying them as relevant (on-topic) or irrelevant (off-topic). Twitter crisis datasets are usually collected by tracking specific relevant keywords and hashtags. Such a process captures lots of relevant data but can also include some irrelevant posts with various distribution across crises. Unrelated posts include advertisements, jokes, political views, unrelated personal messages or posts related to other disasters. Such posts usually exploit trending hashtags to be more visible. Other off-topic tweets were crawled due to the keywords' ambiguity. Relevancy detection is challenging because of the unbalanced datasets.

3.1.2. Information Classification

The task categorises relevant messages into one or more information categories that support situational awareness, such as infrastructure damage, caution, etc. This task is modelled as either a multi-class or multi-label problem. In this work, we used data that has been annotated using a multi-label scheme.

3.2. Multi-source Data Selection for SM Crisis Classification

This work proposes an unsupervised multi-source data selection approach using the K-nearest neighbours algorithm. The strategy aims at building a good model for target data by leveraging labelled data from several related source domains and unlabelled data from a target domain. We hypothesise that fine-tuning LMs on the most similar data can produce more accurate results on crisis classification tasks than using all multi-source training data.

We mimic the real scenario and assume that we are given labelled data from several historical emergency events (multi-source datasets) and only unlabelled data from an emerging crisis, representing the source and target domains, respectively. The goal is to find an optimal set of training data from a multi-source domain that enhances the model generalisation for the target data. Finding this set can be achieved by identifying the examples from training source data that are similar (as close as possible) to the target domain. Hence, we consider an instance-level data selection strategy using cosine similarity in the embedding space of the pre-trained LMs. The selected instances are expected to have similar feature distribution to the target domain of interest.

In our approach, we selected the data only from the crisis-related messages. Due to the unbalanced dataset, we used all off-topic posts for the relevancy detection task. Messages labelled irrelevant to a particular crisis but related to other emergency events were excluded from the off-topic set as depicted in Figure 1. For example, the Jordanian flood data contain two tweets about the Kuwait flood. Those posts were labelled irrelevant to Jordan flood because the annotation was performed per event level. We filtered out such posts from the off-topic training set. Therefore, the classifier will be trained to learn whether or not a post represents a crisis, as the aim is to build a model for cross-event crisis detection.

We leveraged the contextualised text representations produced by pre-trained LMs. In this work, we used Sentence-BERT (S-BERT) (Reimers and Gurevych, 2019). The authors created S-BERT by fine-tuning BERT using Siamese network architecture to produce fixed-sized sentence embeddings that can be compared using similarity measures. As there is no monolingual Arabic S-BERT, we used the multilingual S-BERT (Reimers and Gurevych, 2020). We used the K-nearest neighbours algorithm to select the K most similar instances for each example in the target set based on the cosine similarity on the S-BERT embedding space. The selected data was used to fine-tune a BERT model for classifying target tweets as outlined in Algorithm 1.

Algorithm 1: Multi-source instance-level data selection

Input:
S_L: $\{ S_1 \cup S_2 \cup \cdots \cup S_n \}$ Labelled source domain examples from n historical crisis data
T_U: Unlabelled target domain data for a new crisis
SET $trainset$ to [ ]
S_R = S_L[label=relevant]
S_I = S_L[label=irrelevant]
Remove duplicates in T_U
Encode data in S_R using S-BERT
FOR EACH instance in T_U **DO:**
  Encode instance using S-BERT
  Select the nearest k instances S_k from S_R
  $trainset.append(S_k)$
END LOOP
Remove duplicates in $trainset$
$trainset.append(S_I)$
Output: $trainset$ to **Fine-tune** a BERT model M for classifying T_U

3.3. Self-training

We investigated the effect of combining the data selection approach with self-training. Self-training is a semi-supervised learning approach that learns a base classifier from source data and then uses that classifier to label the unseen target data. The classifier is then retrained utilising the source and pseudo-labelled target data. The self-training continues for a fixed number of iterations or until convergence. For retraining, we used a hard-labelled approach, in which most confidently classified instances are added with their predicted labels (e.g. 0 or 1) to the training set in subsequent training iterations. Figure 1 illustrates the DA framework that combines data selection and self-training.

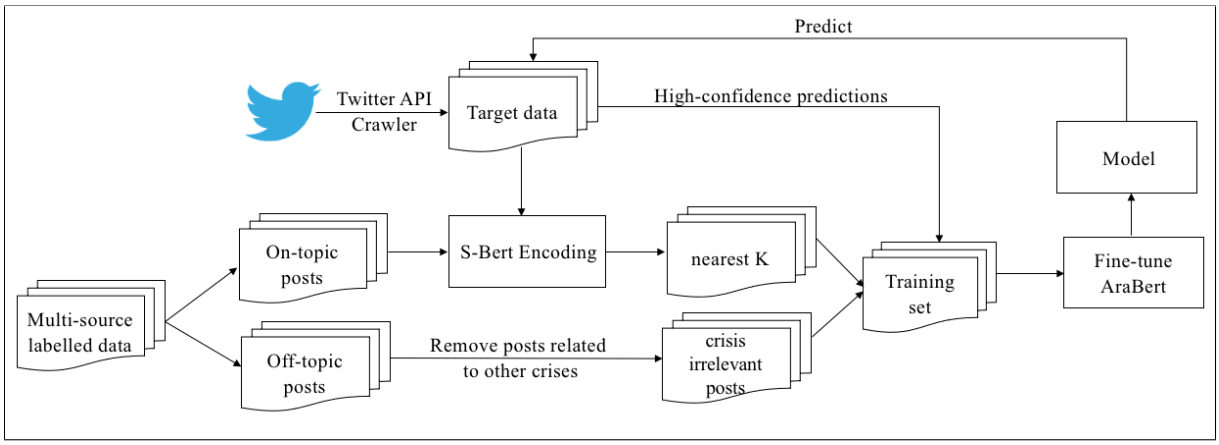

Figure 1: The DA approach with data selection and self-training.

4. Experimental Setup

For data selection, instances are encoded using the distiluse-base-multilingual-cased-v1 multilingual model[1] that supports 15 languages, including Arabic. We experimented with different values of K when choosing the nearest neighbours (3, 5 and 10). We evaluated the binary classification model on each selected set. The information type classifier was assessed on the last chosen dataset (K=10) as some classes can be under-represented in the smaller training set. AraBERT (Antoun et al., 2020) (trained on news corpus) was used as a classification model as it slightly outperforms other Arabic BERT variants (trained on Twitter corpora) on crisis tasks. We fine-tune AraBERT using the same parameters and text pre-processing steps introduced by Alharbi and Lee (2021). The batch size and the number of epochs were set to 32 and 3, respectively. For reproducibility, the random seed was set to 1. For self-training, we set the confidence threshold and the number of iterations to 0.99 and 2, respectively.

We experimented with different source and target crisis pairs using the leave-one-out strategy. In other words, the evaluation was performed by choosing one target event as the test set and the rest events for training. For self-training, the evaluation was performed using 3-folds cross-validation. The target data was split into three parts: one for testing and the rest (two-thirds) for adaptation. We report the average score. We used the weighted F1 and macro F1 to evaluate the models' performance on the relevancy detection task as the dataset is unbalanced. For information classification, we used the accuracy (Godbole and Sarawagi, 2004; Sorower, 2010) and macro F1 score.

The current study used the Kawarith (Alharbi and Lee, 2021) corpus, an Arabic crisis-related Twitter dataset, to evaluate our proposed DA approach. Table 1 shows the number of relevant and irrelevant messages per event in the dataset, while Table 2 presents the distribution of information types (for relevant posts) per event. We manually removed those messages related to other crises from the off-topic set in the source data. We found 30 such tweets in the dataset, most of which (18 posts) were in the Dragon storm data. We did not handle the imbalanced class distribution in this work.

In the following section, we will present the results of classifiers with the data selection method and the data selection with self-training. We will compare the performance of the two proposed models with the same classifier fine-tuned using the entirety of the historical source data (baseline-1). Besides that, we compare them against the self-training model (baseline-2) as self-training with BERT shows promising results in one of the most recent works on English crisis detection (Li et al., 2021). It is worth emphasising that we used a hard-labelling self-training strategy, while Li et al. (2021) adopted soft-labelling.

5. Results and Discussion

Table 3 presents the results of the proposed models and baselines. We found that fine-tuning BERT on the most similar data (BERT-DS models) improves performance in all cases over BERT(all) despite using smaller training data. On average, choosing different values of K results in slightly different performance. BERT-DS(k=10) achieved the best scores in four out of seven cases. We compared the best model with the baseline. BERT-DS(k=10) improved the average weighted F1 and macro F1 over BERT(all) by 3.67% and 4.57%, respectively. The improvement in weighted F1 scores ranges from 1% to 7.53, whereas the macro F1 improved by values ranging from 1.38 to 9.93. We observed that the macro F1 was enhanced by 9.93%, 6% and 5.79% when classifying KF, HF and CD, respectively. The pronounced enhanced performance for classifying CD emphasised the usefulness of DA based on data selection, as the features differ substantially between COVID-19 data and other crises.

[1] https://www.sbert.net/docs/pretrained_models.html

Crisis	Relevant posts	Irrelevant posts	Total
Jordan floods (JF)	1882	118	2000
Kuwait floods (KF)	3701	399	4100
Hafr Albatin floods (HF)	978	637	1615
Cairo bombing (CB)	700	6	706
COVID-19 (CD)	1782	223	2005
Dragon storms (DS)	701	309	1010
Beirut explosion (BE)	833	177	1010

Table 1: Distribution of tweets by relevancy in the Kawarith dataset

Label	JF	KF	HF	CB	DS	BE
Affected individuals & help	331	414	83	138	70	186
Infrastructure & utilities damage	39	271	100	17	105	64
Caution, preparations & other crisis updates	268	980	475	214	252	170
Emotional support, prayers & supplications	709	816	202	222	120	277
Opinions & criticism	604	1355	189	181	221	198

Table 2: Distribution of tweets by information types in the Kawarith dataset

The self-training (BERT-ST(all)) model achieved higher results than BERT(all) in four cases and comparable results in two cases. The BERT-ST(all) model improved the weighted F1 by 12.18% in the COVID-19 case. On average, it enhanced the weighted F1 and macro F1 by 1.82% and 2.27%, respectively. BERT(all) worked better than BERT-ST(all) for the HF data. The reason was that many irrelevant tweets from HF were misclassified and added to the classifier in the next iteration as ground truth data, which degraded the performance.

The DS models generally worked better than the BERT-ST(all) model. The BERT-ST(all) achieved comparable scores in two cases. However, the ST worked better for the CD data. It improved the weighted F1 by 5.54% over the best DS model. The self-training works well when there is a significant feature distribution gap between the source and target, shifting the weights gradually towards the target data.

Finally, we explore whether self-training enhanced the performance of the DS models. To do that, we combine the self-training strategy with the best DS model BERT-DS(K=10). We found that BERT-DS(K=10)+ST achieved the highest scores on CD data. It enhanced the weighted F1 and macro F1 by 12.53% and 10.36%, respectively. However, BERT-DS(K=10) produced a higher performance for KF and HF. Otherwise, they achieved comparable results in two cases. Hence, adding the pseudo-labels to the training data does not constantly improve the performance. We recommend using self-training on the relevancy detection task when the target event is very different from the source data, as in the case of COVID-19 and other disasters.

For the relevancy detection task, we excluded the crisis messages from the off-topic set to reduce the false negatives (crisis messages classified as not crisis). However, we still need to handle the irrelevant messages detected as relevant because they are about another crisis. For example, we found instances related to Covid-19 were classified as on-topic in the DS data because examples of CD were chosen as the most similar data and were added to the selected set with their positive labels. This results in false positives. We can solve this by following the crisis detection task with a message type classification to filter out such posts. We left this for future work.

Regarding the information category classification, we set the K value to 10. This task has been assessed separately, i.e. we suppose that we managed to filter out all irrelevant posts and need to categorise the crisis-related tweets into pre-defined information types. Table 4 displays the results of our experiments. When training the model on the chosen data, we obtained further improvements in macro F1 scores ranging from 1.17% to 6.73% absolute gains. Overall, data selection improved the performance in most cases. Similarly, BERT-ST(all) worked generally better than the BERT(all) model. The average of all scores showed that BERT-ST(all) and BERT-SD achieved comparable results, while BERT-SD(K=10)+ST resulted in lower performance on this task. We generally recommend using the data selection DA for information type classification as the self-training could damage the performance as in the KF event, which failed to detect many cases related to the 'affected individuals' class.

6. Conclusion

This paper proposed a selection-based multi-source domain adaptation approach to identify crisis Twitter messages for new events. Data was selected using the cosine similarity metric in the embedding that has been generated from transformer-based models. Selecting a subset of data that is semantically similar to the target

Target Data	Model	No. of human-labelled examples	Weighted F1	Mac. F1
JF	BERT(all)	10418	93.26	73.52
	BERT-DS(K=3)	3901	94.08	73.89
	BERT-DS(K=5)	4724	94.19	76.32
	BERT-DS(K=10)	6082	94.45	76.14
	BERT-ST(all)	10418	94.85	75.95
	BERT-DS(K=10)+BERT-ST	6082	**95.06**	**77.83**
KF	BERT(all)	8317	87.40	72.39
	BERT-DS(K=3)	3552	90.95	78.34
	BERT-DS(K=5)	4227	90.50	76.83
	BERT-DS(K=10)	5329	93.69	**82.32**
	BERT-ST(all)	8317	**95.83**	77.51
	BERT-DS(K=10)+BERT-ST	5329	**95.18**	77.94
HF	BERT(all)	10801	76.39	78.36
	BERT-DS(K=3)	3107	**85.25**	**84.54**
	BERT-DS(K=5)	3850	80.34	81.69
	BERT-DS(K=10)	5137	82.40	83.63
	BERT-ST(all)	5137	71.07	67.99
	BERT-DS(K=10)+BERT-ST	10801	73.86	85.25
CB	BERT(all)	11710	96.51	55.43
	BERT-DS(K=3)	2783	97.63	56.58
	BERT-DS(K=5)	3169	**98.21**	**65.39**
	BERT-DS(K=10)	3966	97.94	58.01
	BERT-ST(all)	3966	96.96	61.52
	BERT-DS(K=10)+BERT-ST	11710	97.29	63.46
CD	BERT(all)	10412	64.11	49.88
	BERT-DS(K=3)	4071	69.56	52.95
	BERT-DS(K=5)	5044	69.20	53.82
	BERT-DS(K=10)	6484	71.64	55.67
	BERT-ST(all)	6484	76.29	61.21
	BERT-DS(K=10)+BERT-ST	10412	**84.17**	**66.03**
DS	BERT(all)	11424	77.71	71.24
	BERT-DS(K=3)	3115	**84.55**	**80.81**
	BERT-DS(K=5)	3809	81.36	76.31
	BERT-DS(K=10)	5092	78.70	72.62
	BERT-ST(all)	11424	77.60	71.13
	BERT-DS(K=10)+BERT-ST	5092	79.45	73.66
BE	BERT(all)	11414	86.72	78.1
	BERT-DS(K=3)	3019	87.8	79.04
	BERT-DS(K=5)	3628	**90.54**	**82.76**
	BERT-DS(K=10)	4824	89.72	81.75
	BERT-ST(all)	11414	87.77	77.49
	BERT-DS(K=10)+BERT-ST	4824	89.01	79.32

Table 3: The weighted F1 and macro F1 for the relevancy detection task. DS and ST refer to the data selection and self-training techniques, respectively. The keyword (all) indicates training on the whole labelled data.

for fine-tuning BERT LMs showed promising results and outperformed two baselines: training on all data and self-training. We suppose that using monolingual Arabic S-BERT for data representation may achieve better results, so we left that for future work. We think that our instance-level DA approach is useful during the early hours of a crisis when no large unlabelled data is available from an emerging disaster.

7. Bibliographical References

Aharoni, R. and Goldberg, Y. (2020). Unsupervised domain clusters in pretrained language models. *arXiv preprint arXiv:2004.02105*.

Alam, F., Joty, S., and Imran, M. (2018). Domain adaptation with adversarial training and graph embeddings. *arXiv preprint arXiv:1805.05151*.

Alharbi, A. and Lee, M. (2019). Crisis detection from arabic tweets. In *Proceedings of the 3rd workshop*

Target Data	Model	No. of human-labelled examples	Accuracy	Mac. F1
JF	BERT(all)	6913	82	75.97
	BERT-DS(K=10)	3743	87.44	81.65
	BERT-ST(all)	6913	**88.78**	**82.42**
	BERT-DS(K=10)+BERT-ST	3743	87.36	79.29
KF	BERT(all)	5094	79.01	77.32
	BERT-DS(K=10)	3148	**81.23**	**78.86**
	BERT-ST(all)	5094	73.09	68.84
	BERT-DS(K=10)+BERT-ST	3148	71.45	56.55
HF	BERT(all)	7817	84.27	81.56
	BERT-DS(K=10)	3397	83.73	82.98
	BERT-ST(all)	7817	**85.18**	**82.97**
	BERT-DS(K=10)+BERT-ST	3397	83.35	79.65
CB	BERT(all)	8095	77.05	71.36
	BERT-DS(K=10)	1851	77.48	72.53
	BERT-ST(all)	8095	78.20	72.58
	BERT-DS(K=10)+BERT-ST	1851	**79.09**	**74.92**
DS	BERT(all)	8094	78.14	75.08
	BERT-DS(K=10)	3023	**81.88**	**81.18**
	BERT-ST(all)	8094	79.87	81.06
	BERT-DS(K=10)+BERT-ST	3023	**81.83**	**81.46**
BE	BERT(all)	7962	78.14	75.08
	BERT-DS(K=10)	2694	79.89	80.41
	BERT-ST(all)	7962	**83.44**	**77.72**
	BERT-DS(K=10)+BERT-ST	2694	76.70	65.01

Table 4: The accuracy and macro F1 for the information classification task. DS and ST refer to the data selection and self-training techniques, respectively. The keyword (all) indicates training on the whole labelled data.

on arabic corpus linguistics, pages 72–79.

Alharbi, A. and Lee, M. (2021). Kawarith: an arabic twitter corpus for crisis events. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 42–52.

ALRashdi, R. and O'Keefe, S. (2019). Robust domain adaptation approach for tweet classification for crisis response. In *International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning*, pages 124–134. Springer.

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Caragea, C., Silvescu, A., and Tapia, A. H. (2016). Identifying informative messages in disaster events using convolutional neural networks. In *International conference on information systems for crisis response and management*, pages 137–147.

Chen, Q., Wang, W., Huang, K., De, S., and Coenen, F. (2020). Adversarial domain adaptation for crisis data classification on social media. In *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, pages 282–287. IEEE.

Godbole, S. and Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 22–30. Springer.

Guo, H., Pasunuru, R., and Bansal, M. (2020). Multi-source domain adaptation for text classification via distancenet-bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7830–7838.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1021–1024. ACM.

Imran, M., Mitra, P., and Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.

Kozlowski, D., Lannelongue, E., Saudemont, F., Benamara, F., Mari, A., Moriceau, V., and Boumadane, A. (2020). A three-level classification of french tweets in ecological crises. *Information Processing & Management*, 57(5):102284.

Krishnan, J., Purohit, H., and Rangwala, H. (2020). Unsupervised and interpretable domain adaptation to rapidly filter tweets for emergency services. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*,

pages 409–416. IEEE.

Li, X. and Caragea, D. (2020). Domain adaptation with reconstruction for disaster tweet classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1561–1564.

Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C., Squicciarini, A. C., and Tapia, A. H. (2015). Twitter mining for disaster response: A domain adaptation approach. In *ISCRAM*.

Li, H., Caragea, D., and Caragea, C. (2017). Towards practical usage of a domain adaptation algorithm in the early hours of a disaster. In *ISCRAM*.

Li, H., Caragea, D., Caragea, C., and Herndon, N. (2018a). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1):16–27.

Li, H., Sopova, O., Caragea, D., and Caragea, C. (2018b). Domain adaptation for crisis data using correlation alignment and self-training. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 10(4):1–20.

Li, H., Caragea, D., and Caragea, C. (2021). Combining self-training with deep learning for disaster tweet classification. In *The 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021)*.

Ma, X., Xu, P., Wang, Z., Nallapati, R., and Xiang, B. (2019). Domain adaptation with bert-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83.

Mazloom, R., Li, H., Caragea, D., Caragea, C., and Imran, M. (2018). Classification of twitter disaster data using a hybrid feature-instance adaptation approach. In *Proceedings of the 15th Annual Conference for Information Systems for Crisis Response and Management (ISCRAM)*.

Mazloom, R., Li, H., Caragea, D., Caragea, C., and Imran, M. (2019). A hybrid domain adaptation approach for identifying crisis-relevant tweets. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 11(2):1–19.

Neppalli, V. K., Caragea, C., and Caragea, D. (2018). Deep neural networks versus naïve bayes classifiers for identifying informative tweets during disasters.

Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., and Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. In *Eleventh International AAAI Conference on Web and Social Media*.

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*, pages 994–1009. ACM.

Ramponi, A. and Plank, B. (2020). Neural unsupervised domain adaptation in nlp—a survey. *arXiv preprint arXiv:2006.00632*.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Ruder, S., Ghaffari, P., and Breslin, J. G. (2017). Data selection strategies for multi-domain sentiment analysis. *arXiv preprint arXiv:1702.02426*.

Rudra, K., Ghosh, S., Ganguly, N., Goyal, P., and Ghosh, S. (2015). Extracting situational information from microblogs during disaster events: a classification-summarization approach. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 583–592.

Rudra, K., Ganguly, N., Goyal, P., and Ghosh, S. (2018). Extracting and summarizing situational information from the twitter social media during disasters. *ACM Transactions on the Web (TWEB)*, 12(3):1–35.

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., and Kapoor, K. K. (2017). Event classification and location prediction from tweets during disasters. *Annals of Operations Research*, pages 1–21.

Sorower, M. S. (2010). A literature survey on algorithms for multi-label learning.