Domain Adaptation for Arabic Crisis Response

Reem ALRashdi
University of Ha'il, Ha'il, KSA
University of York, York, UK
reem.alreshede@uoh.edu.sa
rmma502@york.ac.uk

Simon O'Keefe
University of York, York, UK
simon.okeefe@york.ac.uk

Abstract

Deep learning algorithms can identify related tweets to reduce the information overload that prevents humanitarian organisations from using valuable Twitter posts. However, they rely heavily on human-labelled data, which are unavailable for emerging crises. Because each crisis has its own features, such as location, time and social media response, current models are known to suffer from generalising to unseen disaster events when pre-trained on past ones.Tweet classifiers for low-resource languages like Arabic has the additional issue of limited labelled data duplicates caused by the absence of good language resources. Thus, we propose a novel domain adaptation approach that does not rely on human-labelled data to automatically label tweets from emerging Arabic crisis events to be used to train a model along with available human-labelled data. We evaluate our work on data from seven 2018–2020 Arabic events from different crisis types (flood, explosion, virus and storm). Results show that our method outperforms self-training in classifying crisis-related tweets in real-time scenarios.

1 Introduction

Arabic represents the world's fifth most spoken language and Arabic language users are the fastest-growing language group on the web (Lane, 2019). In February 2011, protestors in Egypt used Twitter as their main communication platform (Tufekci and Wilson, 2012). This emphasises that Twitter is an important and rich source of real-time and useful information during crises in Arabic countries. People share their statuses and post information about injured or dead people and infrastructural damage (Vieweg, 2012). They also tweet to ask for help or to offer help to others. Although humanitarian organisations could use these information to significantly improve crisis response with regard to reducing human and financial losses, they do not due to the information overload issue (George et al., 2021). To solve this problem, deep learning algorithms have been utilised to identify Arabic tweets from unseen crises to support disaster management and enhance situational awareness in the Middle East (Adel and Wang, 2020; Alharbi and Lee, 2021). However, they did not consider the domain-shift between source and target tweets posted during these events, which prevents the models from reaching a good generalisation level. As a result, semi-supervised approaches that automatically generate new labelled training data from an unlabelled corpus to reduce the gaps between the two domains are desirable.

Distant supervision has been applied to automatically generate new labelled training data for event extraction task (Chen et al., 2017; Zeng et al., 2018). Moreover, semi-supervised domain adaptation techniques have been successfully adopted to incorporate unlabelled target data to labelled source data to reduce the domain-shift between the two domains. Our work here is motivated by the success of applying distant supervision and domain adaptation methods to high-resource English-language tweets presented in our previous works (ALRashdi and O'Keefe, 2019; Alrashdi and O'Keefe, 2020). However and unlike English, Arabic is considered a low-resource language, with several notable issues highlighted in the crisis literature. First is the lack of labelled Arabic tweets for crisis response (Adel and Wang, 2020). Second, the lack of good supporting resources for Arabic, such as external knowledge bases or language dictionaries (Alharbi and Lee, 2019). Finally, Arabic tweets are informal and regional in nature, and Arabic regions have unique dialects which differ in syntax, phonology and morphology (Chiang et al., 2006).

In this paper, we propose an adaptive domain adaptation method from our previous work for English crisis response in (Alrashdi and O'Keefe, 2020) to overcome all these challenges for Arabic

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 249 - 259

crisis response. Our work, here, aims at minimising the domain shift between the target and the source Arabic tweets. We use a distant supervision-based framework to label the unlabelled target data (pseudo-labelling), whereby an initial keyword list is established using clusters from past events. The most related keywords are then selected using a statistical method. The selected keyword list is then expanded by employing distant supervision via an external source (Almaany[1]), and those tweets with a bigram of keywords are labelled as positive tweets, while tweets with none of the keywords are labelled as negative tweets. The generated labelled data is then mixed with the available source data to train a new target model. Unlike self-training in (Win and Aung, 2018; Li, 2021), our method does not replicate the label noise that exists in the current dataset. In addition, crisis data that cannot be detected using existing keyword alert systems, as in (Sakaki et al., 2010), will be detected by our method because of the new crisis keywords derived from Almaany. To the best of our knowledge, this is the first attempt to use distant supervision under the umbrella of domain adaptation techniques to classify unseen crisis-related Arabic data from current events. The experimental results show that the proposed method can be seen as a robust approach to classifying unseen Arabic tweets from an emerging event regardless of the crisis types used to create the keyword list. Furthermore, it extends our framework's abilities from our prior work to automatically label data from low-resource languages with limited capabilities.

2 Related work

Distant supervision (DS). Recent NLP studies have shown the effectiveness of using DS to generate training data via external sources. The researchers in (Chen et al., 2017) employ DS to automatically generate a large-scale dataset using a linguistic knowledge base (FrameNet) for event extraction tasks, where triggers and arguments are extracted from Wikipedia data. Zeng et al. (2018) argue that detecting key argument is enough for determining the event type for event extraction tasks. They extract the most related arguments that best describe the event from existing structured knowledge (FreeBase). However, we use an Arabic dictionary (Almaany) for Arabic ill-formed texts, tweets, based on the existence of essential keywords in the synonyms of a related form.

Domain adaptation (DA). Li et al. (2018b) introduce a semi-supervised DA approach that does not require limited labelled data from the target domain. They use a pre-trained model on one crisis dataset to classify tweets from an emerging event – to be added to the training data in the retrained stage. Their iterative self-training method shows good results, particularly when classifying tweets related to a specific crisis. This method outperforms expectation-maximisation when combined with naive Bayes (Li et al., 2018a). Self-training has been also combined with deep learning models and findings indicate that using unlabelled target data resulted in better adaptation performance (Li et al., 2021). Alharbi and Lee (2022) preform similar study by applying data selection with pre-trained learning models on tweets related to Arabic crises. Another work extends domain adaptation with adversarial training to include a graph-based semi-supervised learning (Alam et al., 2018). F1 score on only two datasets (Queensland Floods and Nepal Earthquake) improves the performance with 5%–7% absolute gain.

To contribute to this line of research, we propose an adaptive yet novel semi-supervised DA that uses DS to give pseudo-labels to unlabelled data from target event to be then incorporated to labeled source data from past disasters to build a robust Arabic crisis-related classifier. We compare our method to the widely used labeling technique in the literature, self-training. We also explore using keyword sets from different crisis type to the target event.

3 Proposed Method

The method consists of two stages as described in algorithm 1.

3.1 Distant supervision-based labelling framework

The proposed labelling framework is described by the steps shown in Figure 1.

Step one: Creating the initial keyword list. We use K-means to classify several Arabic corpora from different events. K-means has been successfully applied to different Arabic Twitter data (Sangaiah et al., 2019; Saeed et al., 2022). For cluster optimisation, elbow method was uncertain for our data because the results shown in the figures are not clear. Because of

[1]available on: https://www.almaany.com

Algorithm 1 Robust domain adaptation approach with pseudo-labelled target data.

1. Given: Clusters of tweets related to several crisis events from different time intervals and locations (CLS); manually Labelled tweets of source data (MLS); unlabelled tweets from target domain (UT) retrieved using Twitter API and publicly available tweet IDs; and manually labelled test data from target domain (MLTT).
2. DS-based labelling stage: Use our framework to label UT based on CLS and employing distant supervision via external knowledge base (giving them pseudo-labels).
3. Adaptation stage: Build a target model using MLS with the pseudo-labelled data from the target domain.
4. Evaluate the model on MLTT.

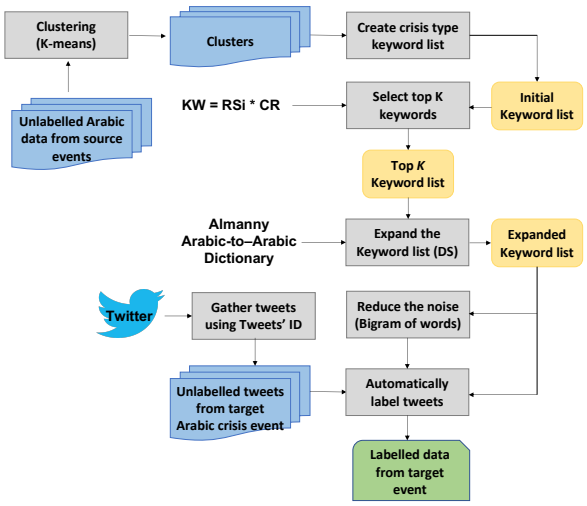

Figure 1: The proposed labelling framework.

that, we use silhouette score measurements to determine the optimal number of clusters to apply K-means to the data for every crisis event from the unlabelled corpora. After that, we assign profiles as labels for each cluster. The reason behind labelling the clusters is that assigning profiles that describe the tweets within the clusters is another way to decide whether the cluster is related to the crisis and informative. To do so, we follow the centroid approach: we pick the centre data point of each cluster to extract the cluster's features. This approach is suitable for our work because the variance within the clusters is slight, and the centre data point of the cluster is the closest one to represent it. Our data are similar in that the tweets are all posted during a crisis and different in providing information about it. Therefore, other approaches can be misleading for our data. To ensure the effectiveness of the centroid approach, we select the closest three data points instead of one. We then extract the features for each cluster and use them to assign profiles from these data points. Our data represents many crisis topics, including: advertisements; political opinions; irrelevant to the crisis; emotional support; infrastructural and utility damage; dead, injured or affected people; and providing help and caution advice. For example, the closest three data points for cluster #3 in the Beirut Explosion corpus (3,445 tweets) are: "حصيلة قتلى انفجار ميناء # بيروت ترتفع إلى ٥٣١. https://t.co/Hbah7vFFKi.https://t.co/HLYVLF7zco" , "أشبه ما يكون بما حدث في هيروشيما وناجاساكي أعداد كبيرة من القتلى والجرحى #انفجار_بيروت # انفجار_مرفأ_بيروت" and "اوضح فيديو للانفجار الهائل الذي هز العاصمة اللبنانية # بيروت الانفجار نتج عنه دمار كبير وانباء عن سقوط عشرات الجرحى https://t.co/uwD2JWpyF4". These tweets contain the words "قتلى", which mean (dead people), "انفجار" (explosion) and "جرحى" (injured people). It is obvious that the most represented tweets of the cluster talk about dead and injured people during the Beirut Explosion incident. As a result, this topic is assigned to cluster #3. After assigning topics to clusters, we divide the clusters into two classes: related and informative and irrelevant or not informative. In particular, infrastructure and utility damages, dead, injured or affected people, and providing help and caution advice are classified as related and informative. On the other hand, advertisements, political opinions, and emotional support are labelled as irrelevant or not informative. While doing this, we observe that all the crisis events have a cluster with a vast number of tweets advertising for specific products or services. We decide to label the tweets expressing political opinions and emotional support as not informative because the information in the posted tweets offers no benefits to humanitarian organisations. Finally, the initial keyword list is created based on the chosen clusters (related and informative) from different collections

related to the same crisis type. We stem each word to its root by utilising ISIRI Stemmer, as in (Al-Horaibi et al., 2017; Abuaiadah et al., 2017), to avoid word redundancy and reduce the amount of linguistically similar words. We also use NLTK libraries to remove stop words such as "في، من، هذا", hashtags such as "#انفجار", places such as "حفر الباطن" and useless Twitter-specific words such as "RT" and "via" from the initial keyword list. To conduct fair experiments, at this point, we eliminate the test event data.

Step two: Selecting top K keywords. The top K keywords are then chosen based on an intrinsic filtering method. To select the top K keyword list for Arabic crisis events, we calculate the keyword (KW) value, inspired by (Chen et al., 2017), for each keyword in the initial keyword list. In a tweet, a word that describes a given crisis type can be a verb, a noun or an adverb. For instance, for the Floods crisis type, the top K keyword list contains "غرق", "سيل",and "مطر", which have the highest KW values compared to the other words in the initial Floods list. Intuitively, a word describing a crisis type appears more than other words in the related tweets. In addition, if the same word appears in both related and unrelated tweets, it has a low probability to be a keyword of this crisis type. Thus, KW is calculated as follows:

$$RS_i = \frac{Count(W_i, CT)}{Count(CT)} \quad (1)$$

$$CR_i = \log \frac{3}{(Count(CTC_i)} \quad (2)$$

$$KW_i = RS_i * CR_i, \quad (3)$$

where RS_i (role saliency) represents the saliency of *i*-th keyword to identify a specific word of a given crisis type, $Count(W_i, CT)$ is the number of a word W_i that occurs in all the tweets related to the crisis type *CT*, and *Count(CT)* is the count of times that all words occur in all the tweets related to the crisis type. CR_i (crisis relevance) represents the ability of the *i*-th keyword to distinguish between the tweets related to the crisis type and irrelevant tweets, and $Count(CTC_i)$ equals 1 if the *i*-th keyword occurs only in the related tweets and 2 if the *i*-th keyword occurs in both related and irrelevant tweets. We compute KW_i for all the words in the initial keyword list from step one and sort them according to their KW values to select the top K keywords for a given crisis type. Table 1 shows that crisis-related and flood-related words have higher KW values than the unrelated ones. Other statistical methods such as pointwise mutual information (PMI; (Church and Hanks, 1990)) or term frequency-inverse document frequency (TF-IDF; (Jones, 1972)) have not been used here for solid reasons. Calculating PMI for positive and negative examples to give the final PMI score is not a fair metric in our case because of the imbalanced data problem in Kawarith dataset. On the other hand, this problem does not affect our formula as *Count(CT)* accounts for the total number of words in the related tweets only. TF-IDF is not suitable in our case because IDF has more impact on the final result than TF; in our case, they should be equally important since tweets are short and full of noise. If we used TF-IDF on our data, rare words such as misspelled words would have higher TF-IDF than essential keywords. Additionally, an important keyword may appear in both related and not related tweets. For instance, in earthquake crisis-type data, the word "earthquake" may appear very frequently in related earthquake event tweets but only once or twice in unrelated earthquake event tweets. On the other hand, our method does not discard the impact of word frequency if the word appears in both related and unrelated tweets.

Ranking	Keyword	KW Value
1	مطر	0.00371
2	غرق	0.00130
32	كسر	0.00065
98	رادار	0.00019

Table 1: KW values of some words from the initial Floods keyword list.

Step three: Applying distant supervision. The list containing top K keywords is then expanded to include similar semantic words from the Almaany Arabic-to-Arabic dictionary. Almaany is an online dictionary that provides corresponding meanings with similar semantic words for each term in Arabic and has been widely used by in Arabic researches (Touahri and Mazroui, 2021; Al-Matham and Al-Khalifa, 2021). We retrieve all the synonyms provided by Almaany for each crisis keyword if the corresponding meaning of the top keyword is related to the crisis type. For example, the top keyword "سيل" exists in the Almaany dictionary but with two corresponding meanings based on the shape and the signs of the word: " سَيْل " and " سيْل".The meaning of " سيْل" is the water of the

rain that rushes over the earth's surface, whereas "سيّل " refers to converting material from a solid state to a liquid state. According to their meanings, " سَيْل " is related to the Floods crisis type, but " سيل" is not. Thus, all the synonyms associated with " سَيْل ", such as " فيضان " and " طوفان " can be mapped to " سَيْل ", which is a crisis keyword gathered from the first step and selected in the second step as one of the top K keywords based on its high KW value. In other words, if one of the top crisis type keywords exists in the Almaany dictionary and its meaning relates to a given crisis type (Floods or Explosion), then distant supervision assumes that all the synonyms related to the given word express that crisis type. As a result, the number of keywords increases in the final list. For instance, the number of keywords rises from 10 to 78 in the keyword list for the Floods crisis type. This list contains two types of keywords: strong keywords (top K keywords) and weak keywords (extracted from Almaany). If a word exists in the top K keywords and is a synonym associated with another top K keyword at the same time, then we consider it a strong keyword. Weak keywords may bring noise to the data, which we try to reduce in step five. As a result, 7 final keyword lists are generated according to the test event and the crisis type of the test event.

Step four: Gathering unlabelled tweets from prior crisis events. These tweets are obtained using Twitter API by their IDs provided by an Arabic twitter corpus (Kawarith) (Alharbi and Lee, 2021).

Step five: Noise reduction. We filter the unlabelled corpus gathered from step four after deleting duplicated and non-Arabic tweets by applying a specific lexical feature (bigram of keywords). After cleaning the unlabelled tweets, only the examples with two keywords from the final keyword list remain. This step reduces the effect of a powerful hashtag when the hashtag without the "#" symbol is one of the keywords. For example, if we use "#كورونا" as one of our hashtags in the previous step, and "كورونا" is one of the keywords in the final keyword list, then tweets like "ناس خايفين من # كورونا و ناس تشل و تحط خطبات و ملكات برويد برويد ان شاء الله لاحقين ماهو ذا الشغل خطبات و ملكات بالله عندكم و الا بس عندنا ؟؟" will not be selected for the Covid'19 event. On the other hand, the tweet "@RT @masrawy: #عاجل مصرع ٥ إصابة ٥١ آخرين الداخلية تكشف تفاصيل انفجار معهد الأورام" will be selected for the Cairo Bombing event because of the appearance of at least two keywords from the final Explosion keyword list: "إصابة" (derived from "اصاب" and "انفجار" (derived from "فجّر" in this case. This process also eliminates several tweets that contain only one weak keyword expanded from Almaany, which decreases most of the noise caused by step three. For instance, the tweet "@3ashoouur: إن شاء الله العاصفه الجايه نكون محبوسين انا و إنتي في بيت واحد " will not be chosen for the Dragon Storm event since " عاصفه" is a weak keyword derived from Almaany using "عصف" which is associated with one of the top K keywords for the Floods crisis type, "اعصار".

Step six: Labelling the corpus as related and not related examples. A collection of data from the new crisis event is automatically generated by labelling tweets from step five as relevant (positive) examples and tweets with no keywords from the expanded keyword list as not related (negative) examples. For instance, the tweet " RT @ww6223ww6: بالتوفيق بإذن الله لابناء العم في انتخابات الغرفة التجارية في فئة الصناعيين و فئة التجار #حفر_الباطن" will be labeled as not related because of the absence of keywords from the final Floods crisis-type list.

3.2 Adaptation stage

We add the pseudo-labelled target data created in the first stage to the available manually labelled source data from the same crisis type as the target crisis (from Kawarith) to build a new target model to classify the unseen tweets from the emerging event. Pseudo-labelled target data generated by our distant supervision-based framework provides new keywords than those existed in the source data. Adding these data to the manually labelled tweets brings target-related features to the training data, including location and crisis nature. By mixing the source and target data in training the target model, we increase the ability of the target classifier to identify related target tweets, including any type of information during the target event lifetime (Sit et al., 2019). For example, tweets containing advice, warnings and alerts start to appear at the beginning of the event onset and decrease thereafter while tweets containing reports on damage and affected individuals reach their peak in the middle of the disaster.

4 Experiments

To determine the effectiveness of using pseudo-labelled target data generated by our framework in domain adaptation settings, we compare two labelling with three adaptation methods. To automatically give labels to the unlabelled target data we apply Distant Supervision (DS) – using our distant supervision-based framework; and Self-Labelling (SelfL; (Li et al., 2018b)) – using a pre-trained model on MLS.

To incorporate target labelled data, we use three adaptation methods: Target Model (TM) – building a model following the source architecture as described in the above section; Finetuning (FT) – modifying all the weights of the pre-trained model using the pseudo-labelled or self-labelled target data; and Feature extraction (FX) – treating the pre-trained model as a feature extractor. Here, we only train a linear classifier using pseudo-labelled or self-labelled data on the top of the extracted features.

As a result, we compare 8 classifiers on 14 settings (keyword sets from the same or different crisis type of the target event - both from Kawarith, as shown in Table 2): (1) SL-LT, supervised learning model trained on MLTT (upper limit); (2) SL-LS, supervised learning model pre-trained on MLS (lower limit); (3) DS-TM; (4) SelfL-TM; (5) DS-FX; (6) SelfL-FX; (7) DS-FT; and (8) SelfL-FT. All the models are tested on MLTT. To train SL-LT, we split MLTT into training (70%) and testing sets (30%). The same testing set is then used to evaluate all the models on the given events. We consider the lower limit model to be our baseline, while the upper limit model is our ideal case.

We follow (Alharbi and Lee, 2021) in cleaning Arabic input tweets. We substitute hyperlinks with the Arabic word "رابط", which means HTTP address or URL. Similarly, we replace user mentions with "مستخدم", hashtags with "هاشتاق", and numbers with "رقم". Four types of letter normalizations are performed: (1) "إ ، آ، أ", the different forms of *alef* are normalized to " ا "; (2) "ى ،ئ", forms of *elaf maqsora*, to "ي ", (3) "ؤ", a form of *waw*, to " و"; and (4) *ta marboutah* "ة ، ه" to "ه".We also eliminate stop words, special characters, punctuation, Twitter-specific words such as "RT", elongation, emojis, non-Arabic characters, diacritics and short vowels. We use ConvBiLSTM (Tam et al., 2021) as the tweet classifier which contains two sub-models: the CNN model for feature extraction and the BiLSTM model for interpreting the features across time steps in both directions. We define a sequential model and add various layers to it. The first is the embedding layer, which represents fastText Arabic embedding as it has been pre-trained using Arabic Wikipedia articles and outperforms other embeddings in Arabic text classification (DHARMA et al., 2022; Habib et al., 2021). The pre-trained embedding has been also fine-tuned in our work using tweets from Kawarith. The embedding layer converts tweets into numerical values and feature embedding. Feature embedding is then fed into the CNN layer with 64 filters and max pooling of size 4. The output of the CNN layer (reduced dimensions of features) is received by the BiLSTM layer with 100 neurons, followed by dropout layers with a rate of 0.5 for regulating the network. The final dense layer is the output layer with two cells representing categories along with a sigmoid activation function to produce classification results. To obtain the best parameter for our model, we utilise Adam as an optimiser and binary cross-entropy loss and set the maximum length to 100. In the end, our model with 25 epochs and a batch size of 32 yields better results. And due to the stochastic nature of the learning algorithm, we repeat every experiment 30 times and take the mean as the final score.

Setting	Keyword Set	Target Set
S1	Explosion	Cairo Bombing
S2	Explosion	Beirut Explosion
S3	Floods	Jordan Floods
S4	Floods	Kuwait Floods
S5	Floods	Hafer-albatin Floods
S6	Floods	Covid'19
S7	Explosion	Covid'19
S8	Floods	Dragon Storm
S9	Explosion	Dragon Storm
S10	Floods	Cairo Bombing
S11	Floods	Beurit Explosion
S12	Explosion	Jordan Floods
S13	Explosion	Kuwait Floods
S14	Explosion	Hafer-albatin Floods

Table 2: Source, keywords and target set for each setting (S) in our experiments.

5 Results and Discussion

Results from the first column in Table 3 show that SL-LS can be useful when classifying target Arabic

S/M	SL-LS	DS-TM	SelfL-TM	DS-FX	SelfL-FX	DS-FT	SelfL-FT	SL-LT
S1	0.753	**0.833**	0.608	0.683	0.784	0.628	0.795	**0.945**
S2	0.768	**0.831**	0.589	0.618	0.584	0.635	0.592	**0.881**
S3	0.798	**0.822**	0.687	0.804	0.647	0.803	0.625	**0.924**
S4	0.746	0.803	0.653	0.708	**0.819**	0.725	0.802	**0.929**
S5	0.717	0.747	**0.757**	0.754	0.679	0.754	0.670	**0.839**
S6	0.744	**0.846**	0.741	**0.850**	0.757	0.842	0.757	**0.954**
S7	0.744	**0.831**	0.741	0.730	0.757	0.729	0.757	**0.954**
S8	0.658	**0.741**	0.560	**0.742**	0.647	0.725	0.640	**0.852**
S9	0.658	**0.734**	0.560	0.651	0.647	0.612	0.640	**0.852**
S10	0.753	**0.843**	0.608	0.694	0.784	0.689	0.795	**0.945**
S11	0.768	**0.771**	0.589	0.682	0.584	0.687	0.592	**0.881**
S12	0.798	**0.810**	0.687	0.640	0.647	0.644	0.625	**0.924**
S13	0.746	0.767	0.653	0.719	**0.819**	0.753	0.802	**0.929**
S14	0.717	0.737	**0.757**	0.505	0.679	0.532	0.670	**0.839**

Table 3: Results in F1 score for 8 models tested on 5 crisis events from the same crisis type and 9 crisis events from different crisis type as the keywords set. Note that S is the setting and M is the model. Best results are in bold.

data. F1 scores for most settings are above 0.70, except for settings 8 and 9 (0.658), which represent the same target data (Dragon Storm). This outcome suggests that crisis data from other crisis types of the target event can be used to train a model for identifying Arabic tweets for crisis response. This result is consistent with prior studies (Nguyen et al., 2017; Li et al., 2018a). On the other hand, Dragon Storm in settings 8 and 9 does not share any of the common features, such as crisis type, location, occurrence time or dialects, with the source events or the keyword sets. This is not the case for the Covid'19 event, since dialects used to post tweets about Covid'19 have been used in the data of the source event, including Saudi and Kuwaiti. This observation clarifies the gap in F1 scores between Dragon Storm and Covid'19 (0.658 < 0.744).

5.1 Keyword and target sets share crisis types

From Table 3, we find out that at least one of the domain adaptation models outperforms SL-LS in all the settings. The highest scores are recorded by DS-TM for all the settings except settings 4 (SelfL-FX) and 5 (SelfL-TM). In contrast, it is clear that DA techniques are not always better than SL-LS. For example, SelfL-FX causes the Beirut Explosion model's performance to decrease by 18%, while SelfL-FT causes the Hafer-albatin Floods model's performance to fall by 4%. This is based on the level of similarity between source and target data and the nature of the adaptation methods. In FX, the high-level features of the source data are transferred to the target data; in FT, more specific target features are incorporated through changing the weights of some layers. Having said that, the Beirut Explosion data differs from the source data even with the existence of another explosion event (Cairo Explosion). The Cairo and Beirut Explosion data are written in different dialects and have dissimilar characteristics: Cairo Explosion was a terrorist act, whereas Beirut Explosion was caused by mismanagement on the part of the Lebanese government. On the other hand, the two Floods events in the source data used to train the model make the Hafer-albatin Floods data very similar. To summarise, DS-TM can be seen as the best general approach among the other 5 domain adaptation classifiers – regardless of the similarity between source and target domains – as it reports the best results in 3 out of 5 settings and a very minor gap compared to the best score in the other two (< 1%). An interesting finding, from columns 2 and 3 for settings (1-5) in Table 3, is that DS performs better as a labelling method than SelfL when TM is used as an adaptation method in 4 out of 5 settings. For setting 5, SelfL-TM is better than DS-TM with a gap of 1% in model performance. However, it is clear from the results that DS-TM always improves the performance by an average of 5.5%. In contrast, SelfL-TM causes a decline in performance for 4 out of 5 target events (average of 12.2%). The model performance when FX is used to adopt pseudo-labelled target outperforms that with self-labelled data in 3 settings (2, 3 and 5). The same scenario is

replicated for the last adaptation method, finetuning (FT). These outcomes suggest that the impact of the labelling method is greater than the impact of the adaptation method when pre-trained models are used due to the nature of the labeling method. DS produces pseudo-labelled target data with important keywords extracted from the keyword set with the same type and new keywords derived from Almaany. This can be very useful if the test set includes these initial or derived keywords. However, if the source and target data are alike in terms of having similar event features (e.g., location, infrastructure damage, people response and dialects), then SelfL can produce accurate self-labelled target data. On review, we observe that 5 out of the 10 top keywords are present in tweets from setting 3, the Jordan Floods incident. Additionally, 62.5% (50 out of 80) of the expanded keyword list occur in the target data. This increases the ability of the DS labelling method to accurately label tweets from this event to the extent that building a target model along with the source data performs better than other models. In setting 5, SelfL-TM outperforms other domain adaptation methods. The reason behind this result is that Hafer-albatin is very similar to the other two Floods events, especially Kuwait Floods. Hafer-albatin and Kuwait are proximal locations and share dialects. Another reason is that the incident data contain 5 out of the top 10 Floods keywords, yet the percentage of the expanded keywords from Almaany is low (38%). Although SelfL-TM should report better results for Kuwait floods than DS-TM because of the similarity level with Hafer-albatin Floods and the small number of common top keywords (3 out of 10), it does not. This can be explained by the nature of the Arabic language, any root word in Arabic has more than 10 shapes regardless of the language signs. This increases the ability of our framework to retrieve more related tweets where most of the expanded keywords occurring in the target data are shapes from root words such as " حذر " from "تحذير، حذر، يحذرون، يحذر ". This represents a significant advantage in using our framework to automatically label Arabic crisis tweets from emerging events. We also note that, in setting 2, both labelling methods cause a substantial drop in model performance when FT or FX is used as the adaptation method, unlike in the other settings. This is because of the high level of divergence between the source and target domains – to the extent that using a pre-trained model in the domain adaptation method always inhibits model performance.

5.2 Keyword and target sets from different crisis types

As stated in column 2 for settings (6-14) from Table 3, and as expected, DS-TM results slightly decrease when using crisis data from different crisis types as the target data to create the keyword set. We find that the number of the shared top or expanded keywords occurring in the target data decreases. Evidently, when the number of shared keywords decreases, the performance of DS labelling method also declines. However, this is not the case in settings 1 and 10. Our results are better in classifying the Cairo Explosion data when the Floods keyword set is used in place of the Explosion keyword set. This is because the number of the top Floods keywords exist in tweets related to Cairo Explosion event is higher than that of the top Explosion keywords (6 > 5). The high divergence level between the Cairo and Beirut Explosion data helps in producing such an outcome. For the Kuwait Floods event, the performance of DS-TM drops from 0.803 to 0.767 in F1 score. It is worth noting that the top keyword list changes from the previous list and does not include " حذر", which gives DS-TM an advantage in the previous section. For the Covid'19 and Dragon Storm events, Table 3 shows that the results of DS-TM change when using different crisis types to build the keyword sets for Floods and Explosion- settings 6 to 9. It seems that the framework with the Floods keyword set generates better pseudo-labelled data from Covid'19 and Dragon Storm than with the Explosion keyword set. This is definitely caused by the number of shared top or expanded keywords. The Dragon Storm data includes 6 top keywords and 55% of the expanded keywords from the Floods keyword set. On the other hand, only 2 top keywords and 16% of the expanded keywords are shared with the Explosion keyword set. The performance of our standalone model supports this finding: for example, its F1 score for tweets related to Covid'19 in setting 6 is higher than in setting 7. This is because setting 6 uses the Floods keyword set, while setting 7 uses the Explosion keyword set. Based on these observations, we can posit that Arabic tweets from an event of any crisis type can be used to generate keyword sets for any emerging disaster. However,

the performance of DS-TM can be improved by using crisis data from the same or similar crisis type to establish the initial keyword list for the given emerging Arabic event. We note that using tweets from different crisis types to pre-train a model to classify target events presents several problems. The main issue is that keywords from related tweets in the source data can be remarkable keywords in the irrelevant target data. An example of this case is setting 9, where the Explosion crisis type included in the source data features terrorism-associated words due to the nature of bombings and explosions, while unrelated tweets from the Dragon Storm target event contain these words due to the crisis locations (Palestine and Syria), where people often post about terrorist acts. Using DS to automatically label the Arabic target corpus – before merging with the manually labelled source tweets to build DS-TM – dramatically reduces this problem. DS-TM does not use models pre-trained on source data, and the DS labels the tweet as related and informative only if it contains two keywords from the expanded keyword set; it is rare to find two terrorist words in one tweet posted during the Dragon Storm crisis. Thus, the DS outperforms SelfL in the three adaptation methods. Another issue is that the number of shared top or expanded keywords can be reduced when tweets from crisis events belonging to different crisis types to the target data are used to generate the keyword sets. This is the case in settings 11, 12, 13 and 14. Although this issue restricts the capacity of DS to produce good target pseudo-labelled data, the best reported domain adaptation model for setting 11 is DS-TM. This is because of the divergence level between the source and target events, which leads SelfL to produce noisy self-labelled data related to Beirut Explosion incident. In contrast, DS-TM does not outperform SelfL-FX for the Kuwait Floods event – even in setting 13 with the increased number of common top keywords. Nevertheless, this number is still too small (4 > 3) to change the performance of DS-TM. We also observe that DS-TM remains the best reported domain adaptation model for the Jordan Floods disaster in setting 12. Here, the length and content of the keyword set change when using incidents from another crisis type. Although the number decreases, the list becomes richer by including words with multiple shapes present in the Jordan Floods data: "انقاذ" and "كارثة". This is because of the powerful nature of the Arabic language in having multiple shapes on one root as discussed above. For setting 14 (Hafer-albatin), the number of common keywords decreases from 5 to 2, with no words with multiple shapes like "صوت". Thus, SelfL-TM produces the best results among the 6 domain adaptation models. In general, DS-TM is the most robust tweet classifier among all the mentioned domain adaptation models. In all cases, it improves model performance after incorporating the pseudo-labelled data, unlike the alternatives.

The last column in Table 3 show that the best recorded DA models for all settings are very far from the results for the upper limit, LT. One possible explanation is that the source data are collected from events from various crisis types. In general, therefore, the results of these Arabic domain adaptation models show much room for improvement.

6 Conclusion

We introduced a domain adaptation method to automatically label Arabic tweets from emerging disasters. Our goal is to overcome the issues of low-resource languages in applying solutions to domain shifts between source and target data. We use clusters instead of manually labelled tweets along with Almaany to extend the initial keyword list. Results showed that our method always improves the model performance (average of 3.7% absolute gain in F1 score) if the keyword sets share the crisis type of the target events.We also ran experiments to use keyword sets from different crisis types to the target incident. As a result, we found out that our framework can classify unseen tweets from a given disaster using a keyword set from different disasters and DS-TM always improves model performance (average of 5.5% absolute gain in F1 score). To this end, we can say that that DS-TM represents robust models to classify tweets from emerging events for languages with limited resources. It also expands our approach's ability to use corpora from other crisis types of the target data to create keyword sets that suit the situation of Arabic tweets. We hope that leveraging automatically labelled data will accelerate the current research on classifying Arabic tweets in crisis response. In the future, we want to extend our method to other low-resource languages like Spanish. We also believe that tweets share features with ill-formed texts, which points to the potential of our method to identify specific events, behaviors or feelings expressed on other communication platforms.

References

Diab Abuaiadah, Dileep Rajendran, and Mustafa Jarrar. 2017. Clustering arabic tweets for sentiment analysis. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 449–456. IEEE.

Ghadah Adel and Yuping Wang. 2020. Detecting and classifying humanitarian crisis in arabic tweets. In *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 269–274. IEEE.

Lamia Al-Horaibi, M Badruddin Khan, and L Al-Horaibi Muhammad Badruddin Khan. 2017. Sentiment analysis of arabic tweets using semantic resources. *Int. J. Comput. Inf. Sci*, 13(1).

Rawan N Al-Matham and Hend S Al-Khalifa. 2021. Synoextractor: a novel pipeline for arabic synonym extraction using word2vec word embeddings. *Complexity*, 2021.

Firoj Alam, Shafiq Joty, and Muhammad Imran. 2018. Domain adaptation with adversarial training and graph embeddings. *arXiv preprint arXiv:1805.05151*.

Alaa Alharbi and Mark Lee. 2019. Crisis detection from arabic tweets. In *Proceedings of the 3rd workshop on arabic corpus linguistics*, pages 72–79.

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

Alaa Alharbi and Mark Lee. 2022. Classifying arabic crisis tweets using data selection and pre-trained language models. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 71–78.

Reem Alrashdi and Simon O'Keefe. 2020. Automatic labeling of tweets for crisis response using distant supervision. In *Companion Proceedings of the Web Conference 2020*, pages 418–425.

Reem ALRashdi and Simon O'Keefe. 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.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–419.

David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing arabic dialects. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 369–376.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

EDDY MUNTINA DHARMA, FORD LUMBAN GAOL, HARCO LESLIE HENDRIC SPITS WARNARS, and BENFANO SOEWITO. 2022. The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. *Journal of Theoretical and Applied Information Technology*, 100(2).

Yasmeen George, Shanika Karunasekera, Aaron Harwood, and Kwan Hui Lim. 2021. Real-time spatio-temporal event detection on geotagged social media. *Journal of Big Data*, 8(1):1–28.

Maria Habib, Mohammad Faris, Alaa Alomari, and Hossam Faris. 2021. Altibbivec: A word embedding model for medical and health applications in the arabic language. *IEEE Access*, 9:133875–133888.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

James Lane. 2019. The 10 most spoken languages in the world. *Babbel Magazine*, 6.

Hongmin Li. 2021. *Domain adaptation approaches for classifying social media crisis data*. Kansas State University.

Hongmin Li, Doina Caragea, and Cornelia Caragea. 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)*.

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

Hongmin Li, Oleksandra Sopova, Doina Caragea, and Cornelia Caragea. 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.

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 *Eleventh international AAAI conference on web and social media*.

Radwa MK Saeed, Sherine Rady, and Tarek F Gharib. 2022. An ensemble approach for spam detection in arabic opinion texts. *Journal of King*

Saud University-Computer and Information Sciences, 34(1):1407–1416.

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*, pages 851–860.

Arun Kumar Sangaiah, Ahmed E Fakhry, Mohamed Abdel-Basset, and Ibrahim El-henawy. 2019. Arabic text clustering using improved clustering algorithms with dimensionality reduction. *Cluster Computing*, 22(2):4535–4549.

Muhammed Ali Sit, Caglar Koylu, and Ibrahim Demir. 2019. Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of hurricane irma. *International Journal of Digital Earth*.

Sakirin Tam, Rachid Ben Said, and Ö Özgür Tanriöver. 2021. A convbilstm deep learning model-based approach for twitter sentiment classification. *IEEE Access*, 9:41283–41293.

Ibtissam Touahri and Azzeddine Mazroui. 2021. Deep analysis of an arabic sentiment classification system based on lexical resource expansion and custom approaches building. *International Journal of Speech Technology*, 24(1):109–126.

Zeynep Tufekci and Christopher Wilson. 2012. Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of communication*, 62(2):363–379.

Sarah Elizabeth Vieweg. 2012. *Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications*. Ph.D. thesis, University of Colorado at Boulder.

Si Si Mar Win and Than Nwe Aung. 2018. *Automated text annotation for social media data during natural disasters*. Ph.D. thesis, MERAL Portal.

Ying Zeng, Yansong Feng, Rong Ma, Zheng Wang, Rui Yan, Chongde Shi, and Dongyan Zhao. 2018. Scale up event extraction learning via automatic training data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.