# Cross-lingual Classification of Crisis-related Tweets Using Machine Translation

**Shareefa Al Amer**[1,2], **Mark Lee**[1], **Phillip Smith**[1]
[1]*School of Computer Science*, University of Birmingham, United Kingdom
[2]*College of Computer Science & Information Technology*, King Faisal University, Saudi Arabia
alamersharifah@gmail.com, {m.g.lee,p.smith.7}@bham.ac.uk

## Abstract

Utilisation of multilingual language models such as mBERT and XLM-RoBERTa has increasingly gained attention in recent work by exploiting the multilingualism of such models in different downstream tasks across different languages. However, performance degradation is expected in transfer learning across languages compared to monolingual performance although it is an acceptable trade-off considering the sparsity of resources and lack of available training data in low-resource languages. In this work, we study the effect of machine translation on the cross-lingual transfer learning in a crisis event classification task. Our experiments include measuring the effect of machine-translating the test data into the source language and vice versa [1]. We evaluated and compared the performance in terms of accuracy and F1-Score. The results show that translating the source data into the target language improves the prediction accuracy by 14.8% and the Weighted Average F1-Score by 19.2% when compared to zero-shot transfer to an unseen language.

## 1 Introduction

We are interested in discovering methods to enhance the detection of emerging crises in social media and the transferability of a classification model fine-tuned on a specific language to other languages. Crisis event detection typically relies on two aspects: (1) the detection of *burstiness* of certain keywords or trends in the timeline, and (2) the classification of the detected *bursts* and whether they indicate the occurrence of a disaster or not. The former aspect can be obtained using unsupervised learning to cluster posts that have certain commonalities such as common keywords, time, and location. However, the latter usually depends on the use of supervised learning algorithms to filter out non-relevant posts since bursts can occur for non-crisis related events such as concerts and media events, political events, and other trends that can interfere with the task of responding to emergencies. It is especially important to develop methods to increase the accuracy of classifying relevant posts to support multilingual settings and therefore help provide better response to emergencies. Although the use of machine translation for cross-lingual transfer learning has shown promising results, there are several drawbacks to the existing work including quality of machine translation, limited parallel data, and structural differences which affect the overall performance of the final model.

In this work, we conducted several experiments to assess the effect of machine translation in bridging language gaps for zero-shot cross-lingual classification of crisis-related tweets. Additionally, we investigated potential limitations on the final predictions. Our study focused on transfer learning between English and Arabic languages by fine-tuning a multilingual pre-trained language model like XLM-R for disaster type classification. Despite the inherent heterogeneity of the two languages, our results surpassed existing benchmarks for more linguistically homogeneous languages such as English and Spanish. Our experiments specifically targeted factors that could influence the performance of existing benchmarks, thereby enabling researchers to address these limitations in future studies. Although our focus is on the crisis events, the approaches can be expanded to other types of events.

The structure of the rest of this paper is as follows. A background about the classification task and relevant knowledge about the different language models is discussed in Section 2. Descrip-

[1]We refer to Arabic as the target language (i.e., language of testing data) and to English as the source language (i.e., language of training data).



*Proceedings of Recent Advances in Natural Language Processing*, pages 22–31
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_003

tion of chosen datasets and the required handling process for our task is shown in Section 3 and 3 respectively. Section 4 explains the experimentation settings to achieve our objective of measuring the effect of machine-translation on the cross-lingual transfer. We discuss our results in details in Section 5 along with an investigation of the possible factors that might have affected the transfer using the machine-translation. We also demonstrate the challenges of cross-lingual transfer learning of the data level and task level in Section 6. While in Section 7 we showcase and compare relevant work, we finally suggest future directions for research in Section 8.

## 2 Background

Classification tasks have gained a significant amount of attention recently. In the domain of event classification, different directions have been pursued including binary classification (i.e., whether the text indicates an event or not), multi-class classification, and multi-label classification. Multi-class classification ranges in granularity from event type to fine-grained aspects of the text content.

We look specifically into cross-lingual classification of social media content for recent contributions in the area. Deep learning models can accommodate the complexity associated with social media data including noise and a lack of structure compared to traditional machine learning algorithms such as SVM, Naïve Bayes, and Random Forests which may suffer from a decline in performance with the increasing complexity of the data (Wang et al., 2021).

Social media posts, especially Tweets, have been very useful in recent years for many tasks and goals including event detection and classification. Among other types of events (e.g., sports, music, political, .etc), disaster detection and classification has a special characteristic: urgency and need of rapid response. Taking advantage of crowd sourced information posted by people in real-time may play a large role to provide timely and proper response. Considering that a specific event can be reported in multiple languages emphasises the need for multilingual and cross-lingual tools that do not discard helpful information just because it is in a different language. Unlike traditional Neural Networks, the introduction of the transformer-based language models such as Generative Pre-trained Transformer (GPT) (Radford et al., 2018) and its successors, and Bidirectional Encoder Representations from Transformers (BERT) and the models built upon it have transformed the area of Natural Language Processing. Following BERT, which is pre-trained on English Wikipedia (2,500M words) and BooksCorpus (800M words), emerged other BERT-based models such RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019) introducing an improvement over BERT's performance since they are pre-trained on more data than the original BERT (Conneau et al., 2020). Particularly, RoBERTa has been pre-trained on 144 GB of data in addition to the 16GB that BERT is pre-trained on while XLNet was trained with over 130GB of textual data including the original 16GB of BERT's.

Following the release of multilingual BERT (mBERT) which is a BERT model trained on Wikipedia text in 104 different languages, other multilingual models such as XLM (Conneau and Lample, 2019) and XLM-RoBERTa (Conneau et al., 2020) have been released as well. Cross-lingual Language Modeling (XLM), like other transformer-based language models, was trained with a masked language modeling (MLM) objective. Additionally, it is trained with a Translation Language Modeling (TLM) which relies on the availability of a dataset with parallel sentences. However, XLM-RoBERTa only uses the MLM objective and trained on a huge corpus of text in 100 languages acquired from the CommonCrawl datasets in a RoBERTa fashion.

Arabic is a widely spoken language, with over 400 million people around the world, according to (UNESCO, 2022). However, there is a scarcity of resources when working on machine learning, especially for domain-specific tasks such as crisis event classification which ignites the need for automated solutions to fill this gap. To address the issue of limited training data for low-resource languages, researchers have employed techniques such as transfer learning (Shi et al., 2022; Yu et al., 2019; Zhang, 2017; Sarioglu Kayi et al., 2021), unsupervised learning (Chauhan et al., 2022; Shi et al., 2022; Sun et al., 2021; Bari et al., 2021), and data augmentation (Maimaiti et al., 2022; Zhou et al., 2022; Şahin and Steedman, 2018). Additionally, initiatives like the Masakhane project (Nekoto et al., 2020) have aimed to build machine translation systems for African languages using collaborative and community-driven approaches. These

efforts strive to make machine learning accessible for low-resource languages. While significant progress has been made, further research and collaboration are essential to enhance the quality and availability of machine translation for low-resource languages.

## 3 Data

We used CrisisNLP[2] (Imran et al., 2016) as an English benchmark dataset for our experiments. CrisisNLP is a widely used hand-labelled Twitter dataset and consists of Tweets collected during ten different disasters including earthquakes, floods, epidemics and other types detailed in Table 1. We grouped similar disasters into a common class, i.e., (*Cyclone, Flood, Hurricane* and *Storm*) were grouped into one broader class called "*Storm*." The total number of samples in the dataset is 20,514 Tweets covering three main types of disasters. On the other hand, we choose Kawarith (Alharbi and Lee, 2021) to be our Arabic dataset used for evaluating the transfer learning of our model. Kawarith contains 12,446 Tweets covering seven different disasters detailed in Table 2. Similar to CrisisNLP, we grouped the classes (*Flood, Rain Storm* and *Storm*) into a broader class called "*Storm*." Since our target is to transfer the model trained on one language to another, we need to keep only common classes existing in both datasets (i.e., *Storm, Epidemic* and *Irrelevant*).

### Pre-Processing

As mentioned in Section 3, we grouped data related to different storm classes into one broader class called "*Storm*." The reasons are: (1) not to confuse the classifier since hurricanes, cyclones and typhoons are all storms that share similar characteristics, they only differ in wind speed and location where they originated (Clements and Casani, 2016), and (2) considering them as different events will cause a loss of data because of the lack of tropical storms in the Arabic dataset. After we refine the two datasets (English and Arabic) to have common classes, we started cleaning the data. Data cleaning included removing non-ASCII and special characters such as (ÛÏ) and (&), removing the URLs, user mentions, retweets, Unicode punctuation, and extra white spaces. We also cleaned the text resulted from removing the hashtags by removing the underscores and separating the words by white space (e.g. under_score becomes under score) and camelCase with (camel case) except if the word is in all uppercase to avoid separating a word into single characters (e.g. UPPERCASE to U P P E R C A S E).

[2]CrisisNLP dataset is released by its authors as tweets text and available to download here: https://crisisnlp.qcri.org/lrec2016/lrec2016.html

## 4 Experiments

We examine the effectiveness of multilingual BERT-based models, specifically the XLM-RoBERTa model, in the cross-lingual transfer learning of a model fine-tuned on a source language and evaluated on an unseen target language for the disaster events classification task. XLM-R has shown considerable improvement over mBERT on many benchmarks (Hu et al., 2020; Ding et al., 2022). Our intention is to measure the machine-translation effect on the prediction performance by translating the target data into the source language (and vice versa) before testing the model. We are aware that the translation of social media text will not be as accurate as translated formal text due to its informality, misspelled words, noise, slangs and dialects. However, translation might provide a working solution for the scarcity of training data in different languages exploiting the abundance of available English data.

Our experiment consists of three parts. In the first part, we fine-tune a multilingual model (XLM-R) on classifying English disaster events and evaluate it on a labeled dataset consisting of original Arabic Tweets (non-translated). The results of this part will give us a benchmark to compare the second part results with and answer our question (i.e., does translation improve the prediction performance when transferring a model to another language?). The second part involves translating the test set into English before evaluating the fine-tuned model. Finally, we translate the source data into the target language and test on the target data (i.e., Arabic). For machine translation, we use Facebook's M2M-100 model (Fan et al., 2021) which translates between a hundred different languages in any direction. Those results will also be compared to monolingual performance of the model on both languages.

## 5 Results

The first set of results found in Table 3 shows the performance of the fine-tuned multilingual model

| Event | Year | Size | Event Type | Mapped Class |
|---|---|---|---|---|
| Nepal Earthquake | 2015 | 3003 | Earthquake | Earthquake |
| Cyclone Pam | 2015 | 2004 | Cyclone | Storm |
| Chile Earthquake | 2014 | 1932 | Earthquake | Earthquake |
| Pakistan Earthquake | 2013 | 1881 | Earthquake | Earthquake |
| India floods | 2014 | 1820 | Flood | Storm |
| Ebola | 2014 | 1774 | Epidemic/Pandemic | Epidemic |
| Pakistan floods | 2014 | 1769 | Flood | Storm |
| California Earthquake | 2014 | 1701 | Earthquake | Earthquake |
| Middle East Resp. Syndrome | 2014 | 1358 | Epidemic/Pandemic | Epidemic |
| Hurricane Odile Mexico | 2014 | 1262 | Hurricane | Storm |

Table 1: Description of CrisisNLP dataset annotated by paid workers. CrisisNLP is an English disaster Tweets dataset.

| Event | Year | Size | Event Type | Mapped Class |
|---|---|---|---|---|
| Hafr Albatin Storm | 2018 | 1615 | Rain storm | Storm |
| Jordan Floods | 2018 | 2000 | Flood | Storm |
| Kuwait rain storm | 2018 | 4100 | Rain storm | Storm |
| Cairo car bomb at cancer hospital | 2019 | 706 | Explosion | Explosion |
| COVID-19 | 2019 | 2005 | Epidemic/Pandemic | Epidemic |
| Egypt Dragon storm and flood | 2020 | 1010 | Storm | Storm |
| Beirut Explosion | 2020 | 1010 | Explosion | Explosion |

Table 2: Description of Kawarith Arabic disaster dataset.

in a monolingual setting (English to English and Arabic to Arabic). In the English setting, the Accuracy and F1-Scores (Average and Micro) are relatively high (96%, 96.2%, and 96% respectively) while they are (91.13%, 91%, and 91.1% respectively) in the Arabic setting. The possible reasons for this 5% decrease might be the number of training samples since the size of the English data was 17K in total while the Arabic was about 5K after cleaning and balancing. The other reason might be that XLM-Roberta was originally pre-trained on more English data than Arabic (Conneau et al., 2020). The motivation of performing a monolingual examination of the model is to set a benchmark for our model after data pre-processing and class manipulation. The original data is labelled for Tweet content whether it is (caution and advice, infrastructure and utilities damage, casualties, etc.). Most of the existing work uses these classes (with minor modifications) for testing their models. However, we use the disaster type labels (Storm, Epidemic, etc.) to classify the Tweets into the type of event. The main purpose is because the two datasets are labelled differently for content, more details about the labelling are found in Section 6. Disaster type labels allow us to first determine whether a Tweet is about a disaster event (relevance), and second to determine what type of disaster is the Tweet talking about. Finer granularity can be adopted later for classifying the type of information provided by the Tweets.

The second set of results can be divided into three parts: (1) the result of evaluating a model fine-tuned on English data and tested on original Arabic data (zero-shot) and (2) the result of evaluating the model on the same Arabic dataset after translating it to English (target-translation). The latter experiment shows an improvement in F1-Score by 8.2% when testing on a translated dataset as compared to the former while (3) the third score is when we translated the training data into the target language (source-translation) which has shown a substantial increase of 19.2% in F1-Score over the zero-shot setting.

Although translating the test set to the source language has increased the accuracy of the classification, however, the result is still not close to the monolingual performance. To explore this limitation, we investigated the potential reasons behind it as follows:

| Setting | Source | Target | Accuracy | M Avg | W Avg |
|---|---|---|---|---|---|
| Monolingual | En | En | 0.960 | 0.962 | 0.960 |
| | Ar | Ar | 0.913 | 0.910 | 0.911 |
| Zero-shot | En | Ar | 0.549 | 0.529 | 0.528 |
| Target translation | En | Ar* | 0.618 | 0.610 | 0.610 |
| Source translation | En** | Ar | **0.697** | **0.645** | **0.720** |
| Target dev data | En | Ar | 0.616 | 0.473 | 0.628 |

* machine-translated to En ** machine-translated to Ar

Table 3: Results of mono-lingual and cross-lingual performance of XLM-Roberta model fine-tuned on English disaster data and evaluated on Arabic data. The Arabic data is translated in English before the test in the second set of results. **En** indicates the use of CrisisNLP dataset while **Ar** indicates the use of Kawarith dataset for Arabic Tweets.

**Quality of translation (Machine Translation Vs. Human Translation):**

Assuming that poor machine-translation might led to loss of accuracy, we employed a human translator to translate a fraction of the test set to English (i.e. 500 samples). If the result is improved, it means that the machine translation does not produce quality data that escalates to the source language data, therefore, the classification will not result in comparable accuracy to the original data.

The result of this check is shown in the first set of scores in Table 4. Although the human translation improved the accuracy by around 5% and the weighted average f1-score has increased by around 4% the difference is still insignificant. We should also note that the size of the test data in this run (i.e., 500) is less than the first experiment (i.e., 5000) which decreased the accuracy score from 0.618 to 0.394 which might imply that if we translate the whole 5000 samples by human the accuracy should improve further as compared to the machine translation. We also noticed a better classification of the event type (floods) since mis-translating it by machine led to poor classification of that class. Figure 1 shows how this class was poorly classified when data was translated by machine.

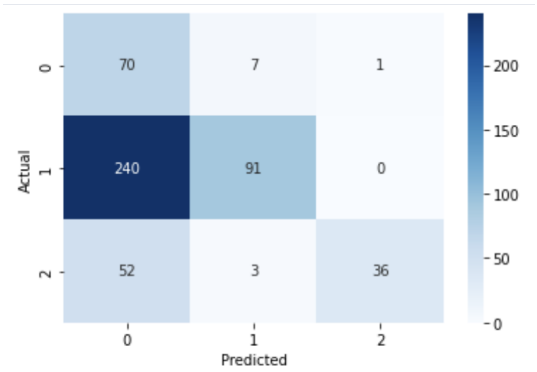
We run a monolingual setting to ensure that the 500 samples of the Arabic data were fairly selected bearing the imbalance of the classes to represent real-world scenarios. The monolingual performance of the model when trained on 80/20 fashion is 0.94, 0.91, and 0.94 for accuracy, macro and weighted average f1-scores respectively as shown in Table 4.

**Human Translation as a Reference**

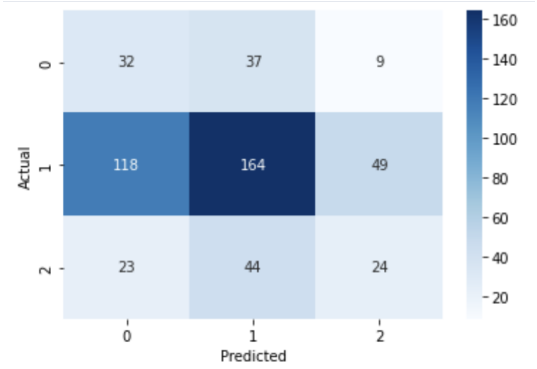
The BLEU score is a widely used metric for measuring translation quality by comparing a machine-translated text to a reference translation (Papineni et al., 2002). It ranges between 0 and 1, with 1 representing a perfect match to the reference translation. In our case, we calculated the BLEU score of the machine translation using human-translated data as a reference to gain insights into the overall performance of the chosen machine translation model. The resulting score was 0.127, which is relatively low but expected, as BLEU primarily focuses on n-gram precision and does not consider semantic or grammatical correctness (Reiter, 2018). A low BLEU score indicates differences in n-grams between the machine translation and the reference translation (i.e., human translation). Similarly, (Ramesh et al., 2020) achieved a notably low BLEU score for English-to-Tamil translation. They argued that the nature of the language contributes to the increased number of n-gram mismatches with the reference translation, despite the translation itself being of good quality. It is important to note that our goal is not to achieve a perfect match with human translation, as that is not the aim of our task. The machine-translated text is not the output of our system; rather, we are using it as a parallel language to train the model.

**Quality of data being translated (normalised data Vs. as-is data):**

On this note, we also employ a linguistic professional to normalise the Arabic Tweets to Modern Standard Arabic (MSA) before translating them by machine. This should give us an idea whether the reason is the poor quality of data found on social media making it hard for the model to generalise




(a) Machine-translation




(b) Human-translation

Figure 1: Confusion Matrix of machine- and human-translated data showing poor classification of class 1 (i.e., storm) when data was translated by machine. 0, 1, and 2 correspond to irrelevant, storm, and epidemic/pandemic, respectively.

to other languages. In other words, the quality of writing and use of various Arabic dialects on social media may affect the model performance.

The second set of results in Table 4 shows a comparison between the two cases with a slight improvement of MSA over the informal Arabic text. Again, there was a drop from 0.549 to 0.398 in accuracy when the test data has decreased from 5K to 500.

**Faults in translators (sequence translation Vs. tokenised data):**

This was an assumption made when we observed that the machine is translating some tweets as a sequence of repeated words such as those samples found in Figure 2 and sometimes when it encounters some characters it stops translating the sequence and moves to the next sequence. Also some words are mistranslated when found in context such as (سيول), Arabic for "floods" is translated to "Seoul" since the Arabic words for floods and Seoul are written in the same way. The machine translates it to "Seoul" whenever it encounters it with a country or city name which is usually the case. Therefore, we wanted to check if the translation quality is affected by the sequence and context. To do so, we tokenised the sequence before translating it to English to check if the translation improves.

The last result in Table 4 shows that lack of context has led to a drop of performance showing the worst scores of all cases.

Overall, drop of performance when transferring the model from English to Arabic as compared to monolingual performance is expected in such scenarios as in the relevant work (Pelicon et al., 2020; Ahmad et al., 2021; Piscitelli et al., 2021; Caselli and Üstün, 2019; Keung et al., 2020) where the accuracy drops when a model is transferred to other languages. In an attempt to improve the performance further, we adopted (Keung et al., 2020)'s approach of using the target language development set instead of the source language data (i.e., English). This led to a very similar effect as the target translation. The last row in Table 3 shows the result of using the target language dev set as an alternative to using the source language dev set.

## 6 Challenges

One of the challenges of cross-lingual transfer learning is the heterogeneity of the source and target data. Even when we acquire disaster datasets in two languages, the way they were labelled can affect the quality of transfer. For instance, CrisisNLP was labelled for information conveyed in the tweet text as discrete labels (e.g., Infrastructure damage, Injured people, etc.) while Kawarith is using multi-label classification where one tweet can have more than one label (e.g., Infrastructure damage AND Injured people). Also, additional labels are found in Kawarith that are not in CrisisNLP such as Opinion and Criticism. Such issues can impose challenges in mapping the labels to the closest possible ones and sometimes discarding some samples. Types of disasters covered in each dataset is also a challenge for disaster type classification. While CrisisNLP contains English data about earthquakes, floods, hurricanes, cyclones and diseases, only two types are in common with the Arabic data which results in discarding the uncommon types when classifying disaster types.

| Data Manipulation | Accuracy | Macro Avg F1-Score | Weighted Avg F1-Score |
|---|---|---|---|
| MT | 0.394 | **0.434** | 0.431 |
| HT | **0.440** | 0.367 | **0.467** |
| MSA | **0.416** | **0.317** | **0.435** |
| As-is | 0.398 | 0.307 | 0.415 |
| Tokenised | 0.318 | 0.321 | 0.324 |
| Monolingual | 0.94 | 0.91 | 0.94 |

Table 4: Evaluating the model on machine-translated test data denoted as (MT), same data translated by a professional human translator denoted as (HT), standardised data to MSA, and un-modified data respectively. Model was trained on around 15K English data and tested on the same 500 Arabic samples manipulated differently.

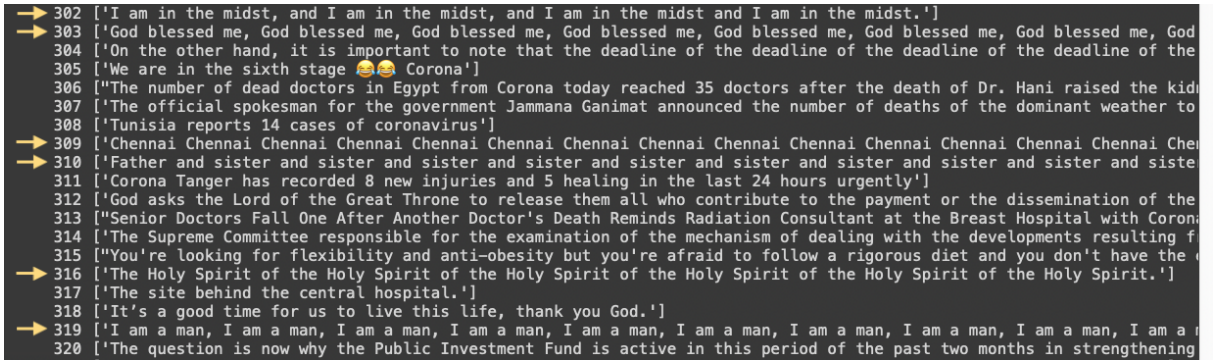



Figure 2: Some results of Machine Translation by Facebook's M2M100 (seq-to-seq) multilingual translation model. The arrows point to the samples that includes repeated words

## 7 Related Work

With the limited work in cross-lingual transfer learning between English and Arabic for the classification task, we needed to set our own benchmark by training the same model on monolingual settings for both languages and also by comparing the transfer to original vs. translated Arabic data. Although the literature lacks comparative work for crosslingual classification between the two particular languages that we are experimenting on, we surveyed the most relevant ones that are either for different languages or different task.

Zero-shot transfer learning from English to Italian has been examined by (Piscitelli et al., 2021) using a Convolutional Neural Network (CNN) exploiting the shared embeddings provided by MUSE (Multilingual Unsupervised and Supervised Embeddings), a Python library. Although the training data was relatively large (45K English Tweets), they achieved a micro-averaged F1-score of 0.52 for Italian when training the model on the English data only. Similarly, (Caselli and Üstün, 2019) investigated the generalisation abilities of mBERT for event detection and classification for Italian and English. Two scenarios were tested: Event detection (i.e. binary classification) and Event detection and classification (i.e. multiclass classification). They experimented with zero-shot learning by training/fine-tuning the model on one language and evaluating it on the other language that it has never been seen in the training. For the zero-shot multiclass scenario, the F1-score was 42.86 when tested on Italian which was improved to 55.38 when the model was fine-tuned with a mixture of data in both English and Italian. A summary of the most relevant work in zero-shot transfer learning is shown in Table 5.

For transferring to Arabic language, (Ahmad et al., 2021) and (Keung et al., 2020) have studied the transfer of mBERT to Arabic language for XNLI task with very close Accuracy in both works. The former has explicitly provided the language syntax to the model to address the challenge of cross-lingual transfer of typologically different languages. Latter work supported the approach of using the target language Dev set for model selection to increase the accuracy and compared the results of both using English dev and target dev. Indeed, using target language Dev set showed improvement over using source language for model selection.

| Authors | F1-Score | Accuracy | Task | Languages |
|---|---|---|---|---|
| (Piscitelli et al., 2021) | 0.52 | - | Classification | English to Italian |
| | 0.70 | - | | English to Spanish |
| (Caselli and Üstün, 2019) | 0.43 | - | Classification | English to Italian |
| (Ahmad et al., 2021) | - | 0.654 | XNLI | English to Arabic |
| (Keung et al., 2020) | - | 0.647 | XNLI | English to Arabic |
| (Pelicon et al., 2020) | 0.52 | - | Sentiment | Slovenian to Croatian |
| **Our work** | **0.72** | **0.697** | **Classification** | **English to Arabic** |

Table 5: Performance scores of relevant work in cross-lingual transfer learning.

A cross-lingual sentiment classification of news documents has been done by (Pelicon et al., 2020) to transfer an mBERT fine-tuned on Slovenian and tested on Croatian without any prior training data in the latter language and achieved an average result of 51.72 F1-Score.

## 8 Conclusions and Future Work

Our study aimed at investigating the impact of using machine translation to leverage the cross-lingual capabilities of multilingual transformer-based models such as XLM-RoBERTa. Specifically, we tested both training data translation and test data translation, in order to mitigate the potential performance loss that can occur when testing on an unseen language. Our findings revealed a considerable improvement in performance, which can be particularly useful for transferring a classifier trained on a resource-rich language to a resource-poor language by translating the same training data into a set of target languages providing an acceptable performance when lacking task data in the target language. However, further research is needed to explore additional approaches that can enhance cross-lingual transfer learning and achieve comparable performance to monolingual models, such as the use of ensemble methods to boost the classification of individual learners. Future work will also include a comparison of different machine-translation models for the same task. Overall, our study highlights the potential of machine translation as a powerful tool for cross-lingual transfer learning, and provides a foundation for future research to further improve the performance of multilingual models on text classification tasks across different languages.

## References


Wasi Uddin Ahmad, Haoran Li, Kai Wei Chang, and Yashar Mehdad. 2021. Syntax-Augmented Multilingual BERT for Cross-Lingual Transfer. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Alaa Alharbi and Mark Lee. 2021. Kawarith: An Arabic Twitter Corpus for Crisis Events. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 42–52.

M. Saiful Bari, Tasnim Mohiuddin, and Shafiq Joty. 2021. UXLA: A Robust Unsupervised Data Augmentation Framework for Zero-Resource Cross-Lingual NLP. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Tommaso Caselli and Ahmet Üstün. 2019. There and Back Again: Cross-lingual Transfer Learning for Event Detection. *CEUR Workshop Proceedings*, 2481.

Shweta Chauhan, Shefali Saxena, and Philemon Daniel. 2022. Enhanced Unsupervised Neural Machine Translation by Cross Lingual Sense Embedding and Filtered Back-Translation for Morphological and Endangered Indic Languages. *Journal of Experimental and Theoretical Artificial Intelligence*.

Bruce W. Clements and Julie Ann P. Casani. 2016. Hurricanes, Typhoons, and Tropical Cyclones. In *Disasters and Public Health: Planning and Response: Second Edition*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32.

Kunbo Ding, Weijie Liu, Yuejian Fang, Weiquan Mao,

Zhe Zhao, Tao Zhu, Haoyan Liu, Rong Tian, and Yiren Chen. 2022. A Simple and Effective Method to Improve Zero-Shot Cross-Lingual Transfer Learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4372–4380.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auliy, and Armand Jouliny. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22:1–48.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multitask Benchmark for Evaluating Cross-lingual Generalization. In *37th International Conference on Machine Learning, ICML 2020*, volume PartF168147-6.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a Lifeline: Human-Annotated Twitter Corpora for NLP of Crisis-Related Messages. *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*.

Phillip Keung, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. 2020. Don't Use English Dev: On the Zero-shot Cross-lingual Evaluation of Contextual Embeddings. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2022. Data Augmentation for Low-Resource Languages NMT Guided by Constrained Sampling. *International Journal of Intelligent Systems*, 37(1).

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Emezue, Bonaventure Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 2002-July.

Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. 2020. Zero-shot Learning for Cross-Lingual News Sentiment Classification. *Applied Sciences*, 10(17).

Sara Piscitelli, Edoardo Arnaudo, and Claudio Rossi. 2021. Multilingual Text Classification from Twitter During Emergencies. *Digest of Technical Papers - IEEE International Conference on Consumer Electronics*, 2021-January.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. *Homology, Homotopy and Applications*.

Akshai Ramesh, Venkatesh Balavadhani Parthasa, Rejwanul Haque, and Andy Way. 2020. Investigating Low-resource Machine Translation for English-to-Tamil. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 118–125, Suzhou, China. Association for Computational Linguistics.

Ehud Reiter. 2018. A Structured Review of the Validity of BLEU. *Association for Computational Linguistics*, 44(3):393–401.

Gözde Gül Şahin and Mark Steedman. 2018. Data Augmentation via Dependency Tree Morphing for Low-Resource Languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen McKeown. 2021. Detecting Urgency Status of Crisis Tweets: A Transfer Learning Approach for Low Resource Languages. *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4693–4703.

Xiayang Shi, Xinyi Liu, Chun Xu, Yuanyuan Huang, Fang Chen, and Shaolin Zhu. 2022. Cross-Lingual Offensive Speech Identification with Transfer Learning for Low-Resource Languages. *Computers and Electrical Engineering*, 101.

Yu Sun, Shaolin Zhu, Chenggang Mi, and Yifan Feng. 2021. Parallel Sentences Mining with Transfer Learning in an Unsupervised Setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 136–142. Association for Computational Linguistics.

UNESCO. 2022. World Arabic Language Day.

Pin Wang, En Fan, and Peng Wang. 2021. Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning. *Pattern Recognition Letters*, 141:61–67.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chongchong Yu, Yunbing Chen, Yueqiao Li, Meng Kang, Shixuan Xu, and Xueer Liu. 2019. Cross-Language End-to-End Speech Recognition Research Based on Transfer Learning for the Low-Resource Tujia Language. *Symmetry*, 11(2).

Yuan Zhang. 2017. *Transfer Learning for Low-resource Natural Language Analysis*. Ph.D. thesis.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. MELM: Data Augmentation with Masked Entity Language Modeling for Low-Resource NER. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1.