Adopting Ensemble Learning for Cross-lingual Classification of Crisis-related Text On Social Media

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

Cross-lingual classification poses a significant challenge in Natural Language Processing (NLP), especially when dealing with languages with scarce training data. This paper delves into the adaptation of ensemble learning to address this challenge, specifically for disaster-related social media texts. Initially, we employ Machine Translation to generate a parallel corpus in the target language to mitigate the issue of data scarcity and foster a robust training environment. Following this, we implement the bagging ensemble technique, integrating multiple classifiers into a cohesive model that demonstrates enhanced performance over individual classifiers. Our experimental results reveal significant improvements in adapting models for Arabic, utilising only English training data and markedly outperforming models intended for linguistically similar languages to English, with our ensemble model achieving an accuracy and F1 score of 0.78 when tested on original Arabic data. This research makes a substantial contribution to the field of cross-lingual classification, establishing a new benchmark for enhancing the effectiveness of language transfer in linguistically challenging scenarios.

1 Introduction

Cross-lingual transfer learning, which involves transferring models from one language to another or from one task to another, has gained significant attention in the field of natural language processing. This approach is particularly valuable in scenarios where task data in the target language is scarce, posing a limitation to training machine learning models for specific tasks such as classification. In such cases, machine translation has emerged as an effective solution to bridge the language gap, enabling acceptable performance in transfer learning (Ji et al., 2024; Huang et al., 2021).

Furthermore, the use of ensemble techniques in the context of cross-lingual classification has shown promise in achieving better performance and generalisation across multiple languages. Ensemble models, by combining multiple base models, can effectively capture diverse aspects of the data and mitigate the impact of language variations, thereby enhancing the robustness of cross-lingual classification systems.

The significance of transfer learning lies in its ability to utilise the wealth of data available in high-resource languages to benefit low-resource languages, thus enabling access to various NLP tasks. Data augmentation techniques, like Machine Translation, serve to amplify this effect by artificially expanding the dataset in the target language, which allows for a richer and more diverse linguistic feature set that models can learn from, leading to improved performance and reliability in cross-lingual applications.

The potential of ensemble learning in addressing the challenges of cross-lingual classification cannot be overstated. By leveraging the strengths of multiple learners, ensemble learning introduces a level of diversity that single models alone cannot achieve, significantly enhancing performance and generalisation capabilities across languages. This diversity is particularly crucial in cross-lingual scenarios, where linguistic and semantic disparities between languages can pose substantial barriers to effective model transfer. Ensemble methods can mitigate these barriers by combining predictions from multiple models, thereby reducing the risk of misclassification due to language-specific nuances or translation inaccuracies. Moreover, ensemble learning can adaptively focus on difficult-to-classify instances, ensuring that the aggregated model is not only more accurate but also more robust to the variability inherent in cross-lingual data. Consequently, the application of ensemble learning in cross-lingual classification opens up new avenues for building more resilient and adaptive NLP systems that can better serve the needs of a

Proceedings of the The Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024), pages 159–165

linguistically diverse world.

In addressing the challenge of cross-lingual transfer learning in situations where training data is non-existent, our work introduces an effective approach that significantly improves the efficacy of model transfer to the Arabic language, a language markedly different in structure and lexicon from English. A key aspect of our contribution is investigating viable strategies, including the integration of machine translation with a bagging ensemble approach, for classifying disaster-related social media posts in Arabic using solely English data. This technique demonstrates potential for broad application across various languages and domains, offering a solution for scenarios with limited data availability in the target language, provided there's access to extensive data in a resource-rich language.

2 Related Work

With growing interest in cross-lingual text classification, the challenge persists due to linguistic variations and data scarcity across languages. Ensemble classification models, employing multiple weak classifiers and combining their predictions through consistency functions like voting, have been widely used in monolingual tasks but less explored in cross-lingual contexts. Techniques such as bagging, AdaBoost, random forest, and gradient boosting have shown promise across various domains (Dong et al., 2020). Among the limited literature on cross-lingual applications, *Funnelling* and its advanced iteration, Generalised Funnelling (*GFun*), stand out for incorporating calibrated posterior probabilities and additional feature vectors to enhance classification performance on multilingual datasets. However, they assume the availability of training data in all target languages (Esuli et al., 2019; Moreo et al., 2021).

Earlier studies, such as those by (Kilimci and Akyokus, 2018) and (Bashmal and Alzeer, 2021), demonstrate the effectiveness of ensemble models in monolingual settings, suggesting potential for cross-lingual adaptation. Beyond ensemble models, research has explored leveraging linguistic similarities through character-based embeddings, joint training, and embedding alignment to address cross-lingual text classification. Techniques such as instance-weighting have also been employed to assign larger weights to source instances sharing common features with target samples during training. This approach aims to utilise resource-rich data while accommodating the specifics of the target language (Li et al., 2021).

While the foundation for cross-lingual text classification is robust, marked by a variety of methodologies from ensemble learning to linguistic feature exploitation, challenges remain in data availability, computational demands, and language diversity. Our work contributes to this ongoing effort by adapting an ensemble approach designed to enhance the effectiveness of cross-lingual classification, especially for under-resourced languages. This approach seeks to build on the existing body of research, pushing the boundaries of what is achievable in the realm of cross-lingual text classification.

3 Proposed Methodology

3.1 Problem Formulation

The challenge of accurately classifying disaster-related social media texts across multiple languages is paramount for effective emergency response, yet is significantly hindered by the lack of training data for various languages. This scarcity affects the development of effective cross-lingual classification models, especially for languages with minimal resources. We aim to tackle this issue by focusing on the cross-lingual classification of disaster-related texts within the context of languages that are underrepresented in training datasets.

Addressing the data disparity between high-resource and low-resource languages, which are often spoken by communities most affected by disasters, is crucial. The goal of this study is to utilise the abundant data from high-resource languages to improve the classification accuracy of texts in low-resource languages. In doing so, we aspire to enhance global disaster response efforts by ensuring that critical information reaches all linguistic groups, thereby overcoming language barriers that could potentially hinder timely and effective disaster management.

3.2 Model Overview

Our proposed model, depicted in Figures 1 and 2, addresses the challenge of cross-lingual classification of disaster-related social media texts through a structured methodology comprising four main components: Data Collection and Translation, Bootstrapping, Ensemble Model Learning, and Testing the Ensemble. The process begins with the acquisition of disaster-related texts from a high-resource language, namely English, followed by their trans-

lation into the target language, Arabic, to mitigate data scarcity. This is complemented by a bootstrapping phase that employs a bagging approach to split the dataset into separate subsets.

At the core of our methodology is the Ensemble Model Learning component, which utilises three classifiers to construct a robust model. This approach benefits from the diversity of data and significantly reduces the risks of overfitting. The ensemble model is subsequently tested with separate target language data to evaluate its effectiveness and applicability in real-world disaster scenarios. Through this integrated approach, which combines machine translation, iterative learning, and ensemble learning, we aim to enhance the classification of disaster-related texts across languages, thereby improving global disaster response capabilities.

4 Experiments

4.1 Data

We utilise the CrisisNLP dataset (Imran et al., 2016), with over 17,000 English X posts covering a range of disaster types such as earthquakes, floods, and diseases. This dataset was translated into Arabic using Google Translate to create a parallel corpus. To assess the model's performance in Arabic, we used the Kawarith dataset (Alharbi and Lee, 2021) which consists of 5,000 Arabic X posts with similar disaster classifications. This setup allows for effective cross-lingual model training and testing.

To prepare the data, we consolidated storm-related classes into a single "Storm" category to align both datasets and simplify classification. We ended up with three classes that are common to both the training and testing datasets, namely storm, disease, and irrelevant. Our preprocessing included removing non-ASCII characters, URLs, mentions, and normalising text (removing extra spaces, handling hashtags). This ensured clean, uniform datasets for our cross-lingual classification experiments.

4.2 Machine Translation Model Selection

To choose a suitable Machine Translation (MT) model for translating the data, we evaluated the performance of three open-source MT systems that support a wide range of languages, including Arabic, by calculating BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) scores. We sampled 1,000 posts from the English data and employed a human translator to obtain the reference translation. The same sample was then translated using three different MT models: Google Translate, Facebook's M2M, and MarianMT. While the differences in performance were not substantial, this evaluation assisted us in making the MT decision. We acknowledge that these metrics measure how closely the MT translations align with human translations, which is less critical in our case since the translation is for classifier consumption.

The results of this study are presented in Table 1. While the observed BLEU scores may appear low, they do not necessarily indicate poor model performance given the complexity of the task. Translating social media content is particularly challenging in machine translation, and achieving high BLEU scores in this domain is more difficult than in more formal types of text (Sabtan et al., 2021).

Notably, the observed METEOR scores are higher than the BLEU scores for the same models and languages. This could be attributed to METEOR's more comprehensive assessment of translation quality, including synonymy and sentence structure, which might be more forgiving than BLEU's strict n-gram matching approach, especially in the context of social media text.

4.3 Evaluation Metrics

For both the individual models and the ensemble model, these evaluation metrics are calculated based on their predictions on a separate test dataset. The evaluation includes calculating accuracy and F1 scores, including weighted, micro, and macro F1. The accuracy metric provides an overview of the model's overall performance, while the F1 scores give insights into the model's precision and recall for each class and their overall performance.

The ensemble model's evaluation involves combining the predictions of the individual classifiers using a voting mechanism. The predictions made by each individual classifier are considered, and the final prediction for each instance is determined by the class with the majority vote. The ensemble model's accuracy and F1 scores are then computed based on these aggregated predictions.

4.4 Experimental Settings

We employed the "xlm-roberta-base" (Conneau et al., 2020) as the base classifier for our ensembles with the same hyper-parameters, differing only in the data being handled. The tokenizer was configured with truncation enabled and a maximum to-

	BLEU	METEOR
Google Translate	**0.144**	**0.354**
Facebook M2M	0.097	0.283
MarianMT	0.081	0.236

Table 1: BLEU and METEOR scores calculated for each Machine Translation model translating the 1K sample to Arabic using human translation as the reference.

ken length set to the longest instance. Padding was also used to ensure uniform input lengths during training. The model architecture was loaded using the 'AutoModelFor-SequenceClassification.from-pretrained' method, which adapted the pre-trained XLM-RoBERTa to our specific task with three classes.

For our model training, we chose a batch size of 64 to balance computational efficiency and gradient stability, and we limited training to 10 epochs to optimise exposure to the dataset while preventing overfitting. The learning rate was set to 1e-5, chosen through experimentation to ensure fast convergence without overshooting, and weight decay was applied at a rate of 0.01 as a regularisation measure to enhance generalisation. To manage resources effectively, we saved the model at the end of each epoch but limited storage to only the latest model checkpoint, avoiding the resource strain of multiple checkpoints. This ensures training efficiency, model performance, and computational resource management.

4.5 Results

4.5.1 Experiment 1

In the first experiment, we fine-tuned two separate XLM-RoBERTa models for classifying parallel datasets (CrisisNLP and translated CrisisNLP), with each model being exposed to data in one language during training. These individual classifiers were then combined to create an ensemble model using a voting function to determine predictions for test instances. The best-performing model for each language was selected for predicting these instances. Subsequently, the ensemble model aggregated the predictions from the individual classifiers to generate the final prediction through a voting mechanism, as illustrated in Figure 1.

The performance of the individual models in the monolingual setting is presented in Table 2, where one XLM-RoBERTa model was fine-tuned on the original English data and another on the Arabic translation of the same data. Results are reported using accuracy, macro F1, and weighted F1 scores. The ensemble, combining predictions from both models and generating final predictions for the original Arabic test data through voting, achieved an accuracy of 0.75, with macro and weighted F1 scores of 0.63 and 0.74, respectively. This marks a significant improvement over our previous benchmark, achieving a 0.72 weighted average F1 score with machine-translated source data for training. These results underscore the effectiveness of the ensemble technique in a cross-lingual context, indicating the potential for further improvement by exploring alternative ensemble approaches.

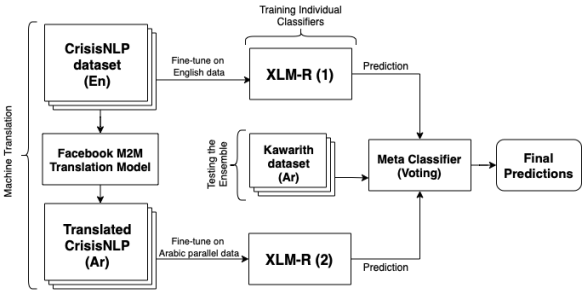

Figure 1: The approach of ensembling two individual classifiers trained on parallel (machine-translated) data in English and Arabic.

4.5.2 Experiment 2

Building on the successes of the first experiment, which already marked improvements over previous experiments and existing baselines as discussed in Section 4.5.1, our second experiment aimed at further enhancing performance by altering the architecture of the ensemble model. We introduced a joint training approach, leveraging an ensemble of three base classifiers. Each classifier was trained on a unique segment of the data, ensuring complete data separation among the models. This was achieved by initially merging the parallel datasets and then dividing this combined dataset into three segments using a bagging technique. The classifiers were then trained independently on

	Data	Accuracy	Macro F1	Micro F1	W Avg F1
Classifier 1	CrisisNLP (En)	0.96	0.96	0.95	0.96
Classifier 2	CrisisNLP (Ar)	0.94	0.94	0.94	0.94
Ensemble	Kawarith (test)	0.75	0.63	0.74	0.74

Table 2: Performance of individual XLM-R classifiers on monolingual data. The CrisisNLP (Ar) dataset represents the machine-translated Arabic data, while Kawarith is an original Arabic dataset used for testing the ensemble. The last set of results showcases the voting ensemble of both individual classifiers when evaluated on the Kawarith dataset.

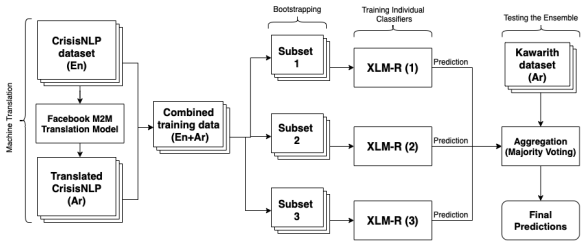

Figure 2: The bagging approach used in this experiment involves splitting the combined parallel data into three distinct subsets. Each subset is utilised to train a different instance of the XLM-R classifier.

these segments, with the best-performing model for each segment chosen for test sample prediction. A majority voting mechanism was subsequently employed for the final prediction, showcasing the ensemble's combined strength in making accurate cross-lingual classifications, as depicted in Figure 2.

This experiment aimed to maximise the ensemble's effectiveness by combining parallel training data, thereby exposing the models to both languages. The goal was to leverage the collective strength of the ensemble in adeptly handling the linguistic diversity presented by the datasets. The performance of the individual classifiers is presented in Table 3, with all three models achieving an accuracy and F1 score of approximately 0.93 on homogeneous data, underscoring their consistency. When evaluated on the original Arabic data (Kawarith), the ensemble of the three models demonstrated substantial improvement over the results of the first experiment, achieving an accuracy and F1 score of 0.78. Remarkably, this performance sets a new benchmark, surpassing existing efforts in similar cross-lingual classification challenges and underlining the potency of our ensemble approach in achieving state-of-the-art results in cross-lingual contexts.

4.5.3 Joint Training

To comprehensively assess the method's efficacy, we conducted benchmark comparisons by training a classifier on combined English and its Arabic translated datasets. This benchmark allowed a direct evaluation of the ensemble strategy's benefits over single-classifier approaches, crucial for understanding the impact of using multiple classifiers together in a cross-lingual context. Our findings show that while individual classifiers in the ensemble may perform less effectively than a singular classifier on homogeneous data, the ensemble as a whole surpasses the single classifier's performance on Arabic test data (zero-shot), highlighting the ensemble's superior handling of linguistic diversity. These results, summarised in Table 4, showcase the ensemble's ability to outperform despite the individual weaknesses of its components, demonstrating its strength in cross-lingual classification.

5 Discussion

The presence of data imbalance posed a significant challenge and had a noticeable impact on the model's performance. To address this issue, we made a trade-off between better classification and the potential loss of data. As a mitigation strategy, we introduced an additional layer to calculate class weights, accounting for the class imbalance. The class weight calculation function was designed to dynamically assign weights based on the distribution of instances in each class. If the data is already balanced, the function assigns equal weights to all classes. However, for classes with fewer instances, the function assigns higher weights, effectively prioritising those classes during training. By incorporating this mechanism, we aimed to balance the training process and alleviate the negative impact of data imbalance on the model's performance. This approach is particularly useful in scenarios where the dataset is heavily imbalanced, as it allows the model to focus more on the underrepre-

	Data	Accuracy	Macro F1	Micro F1	W Avg F1
Classifier 1	CrisisNLP (En+Ar*)	0.928	0.929	0.928	0.927
Classifier 2	CrisisNLP (En+Ar*)	0.933	0.934	0.933	0.932
Classifier 3	CrisisNLP (En+Ar*)	0.936	0.938	0.936	0.935
Ensemble	Kawarith (test)	0.78	0.70	0.78	0.78

Table 3: Performance of individual XLM-R classifiers on distinct subsets of the data. The last row showcases the voting ensemble of the three individual classifiers when evaluated on the Kawarith dataset. *The CrisisNLP (Ar) dataset represents the machine-translated Arabic data.

	Accuracy	Macro F1	Weighted Avg F1
Bagging Ensemble	0.78	0.70	0.78
Joint Training	0.69	0.70	0.70
Joint Training (homo)	0.95	0.95	0.95

Table 4: Performance comparison between bagging ensemble and an individual classifier trained on the same combined dataset. Last row shows the performance of the classifier when tested on the same training data (test portion).

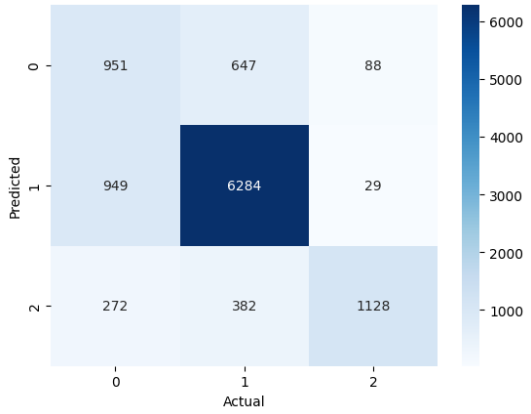

Figure 3: The confusion matrix of the ensemble classification results achieved through majority voting of three classifiers. 0, 1, and 2 correspond to irrelevant, storm, and disease classes, respectively.

sented classes, leading to improved generalisation and performance across all classes.

With a closer look at the individual scores, we notice that the mis-classification of class 0 (i.e., irrelevant) has affected the macro-average F1 resulting in a score of 0.70. However, the model performs relatively well in classifying other classes. The diverse nature of the irrelevant posts makes it challenging for the model to accurately classify them, and they are often mis-classified into other classes, mostly class 1 (storm). Figure 3 displays the confusion matrix, which provides a visualisation of the classification performance for each individual class.

6 Conclusion and future work

In this work, we have presented a practical solution for transferring models across languages when confronted with limited or nonexistent training data. Our experimentation involved the application of a bagging ensemble technique, with each experiment employing a distinct approach. By combining training data from both English and its Arabic translation, and partitioning (bagging) this combined dataset into separate splits, we observed a noteworthy enhancement in prediction performance compared to existing methodologies. Looking ahead, our future work will explore alternative ensemble approaches to tackle the same challenge. Additionally, extending the scope of our approach to a wider set of languages and tasks holds promising potential for further advancement.

References

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.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Proceedings of the Workshop ACL 2005*.

Laila Bashmal and Daliyah H. Alzeer. 2021. ArSarcasm Shared Task: An Ensemble BERT Model for Sarcasm Detection in Arabic Tweets. In *WANLP 2021 - 6th*

Arabic Natural Language Processing Workshop, Proceedings of the Workshop.

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.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A Survey on Ensemble Learning.

Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2019. Funnelling: A New Ensemble Method for Heterogeneous Transfer Learning and Its Application to Cross-lingual Text Classification. *ACM Transactions on Information Systems*, 37(3).

Kuan Hao Huang, Wasi Uddin Ahmad, Nanyun Peng, and Kai Wei Chang. 2021. Improving Zero-Shot Cross-Lingual Transfer Learning via Robust Training. In *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*.

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*.

Shaoxiong Ji, Timothee Mickus, Vincent Segonne, and Jörg Tiedemann. 2024. Can Machine Translation Bridge Multilingual Pretraining and Cross-lingual Transfer Learning?

Zeynep H. Kilimci and Selim Akyokus. 2018. Deep Learning- and Word Embedding-based Heterogeneous Classifier Ensembles for Text Classification. *Complexity*, 2018.

Irene Li, Prithviraj Sen, Huaiyu Zhu, Yunyao Li, and Dragomir Radev. 2021. Improving Cross-lingual Text Classification with Zero-shot Instance-Weighting.

Alejandro Moreo, Andrea Pedrotti, and Fabrizio Sebastiani. 2021. Generalized Funnelling: Ensemble Learning and Heterogeneous Document Embeddings for Cross-lingual Text Classification. In *CEUR Workshop Proceedings*, volume 2947.

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.

Yasser Muhammad Naguib Sabtan, Mohamed Saad Mahmoud Hussein, Hamza Ethelb, and Abdulfattah Omar. 2021. An Evaluation of the Accuracy of the Machine Translation Systems of Social Media Language. *International Journal of Advanced Computer Science and Applications*, 12(7).