Attention Realignment and Pseudo-Labelling for Interpretable Cross-Lingual Classification of Crisis Tweets

Jitin Krishnan
Department of Computer Science
George Mason University
Fairfax, VA
jkrishn2@gmu.edu

Hemant Purohit
Department of Information Sciences & Technology
George Mason University
Fairfax, VA
hpurohit@gmu.edu

Huzefa Rangwala
Department of Computer Science
George Mason University
Fairfax, VA
rangwala@gmu.edu

ABSTRACT

State-of-the-art models for cross-lingual language understanding such as XLM-R [7] have shown great performance on benchmark data sets. However, they typically require some fine-tuning or customization to adapt to downstream NLP tasks for a domain. In this work, we study unsupervised cross-lingual text classification task in the context of crisis domain, where rapidly filtering relevant data regardless of language is critical to improve situational awareness of emergency services. Specifically, we address two research questions: a) Can a custom neural network model over XLM-R trained only in English for such classification task transfer knowledge to multilingual data and vice-versa? b) By employing an attention mechanism, does the model attend to words relevant to the task regardless of the language? To this goal, we present an attention realignment mechanism that utilizes a parallel language classifier to minimize any linguistic differences between the source and target languages. Additionally, we pseudo-label the tweets from the target language which is then augmented with the tweets in the source language for retraining the model. We conduct experiments using Twitter posts (tweets) labelled as a 'request' in the open source data set by Appen[1], consisting of multilingual tweets for crisis response. Experimental results show that attention realignment and pseudo-labelling improve the performance of unsupervised cross-lingual classification. We also present an interpretability analysis by evaluating the performance of attention layers on original versus translated messages.

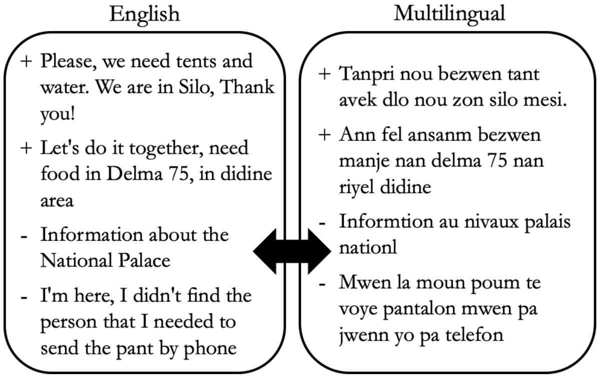
KEYWORDS

Social Media, Crisis Management, Text Classification, Unsupervised Cross-Lingual Adaptation, Interpretability

ACM Reference Format:
Jitin Krishnan, Hemant Purohit, and Huzefa Rangwala. 2020. Attention Realignment and Pseudo-Labelling for Interpretable Cross-Lingual Classification of Crisis Tweets. In *Proceedings of KDD Workshop on Knowledge-infused Mining and Learning (KiML'20).* , 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social media platforms such as Twitter provide valuable information to aid emergency response organizations in gaining real-time situational awareness during the sudden onset of crisis situations [4]. Extracting critical information about affected individuals, infrastructure damage, medical emergencies, or food and shelter needs can help emergency managers make time-critical decisions and allocate resources efficiently [15, 21, 22, 30, 31, 36]. Researchers have designed numerous classification models to help towards this humanitarian goal of converting real-time social media streams into actionable knowledge [1, 22, 26, 28, 29]. Recently, with the advent of multilingual models such as multilingual BERT [9] and XLM [20], researchers have started adopting them to multilingual disaster tweets [6, 25]. Since XLM-R [7] has been shown to be the most superior model in cross-lingual language understanding, we restrict our work to this model to explore the aspects of cross-lingual transfer of knowledge and interpretability.



+ Please, we need tents and water. We are in Silo, Thank you!
+ Let's do it together, need food in Delma 75, in didine area
- Information about the National Palace
- I'm here, I didn't find the person that I needed to send the pant by phone

+ Tanpri nou bezwen tant avek dlo nou zon silo mesi.
+ Ann fel ansanm bezwen manje nan delma 75 nan riyel didine
- Informtion au nivaux palais nationl
- Mwen la moun poum te voye pantalon mwen pa jwenn yo pa telefon

Figure 1: Problem: Unsupervised cross-lingual tweet classification, e.g., train a model using English tweets, predict labels for Multilingual tweets, and vice-versa.

In this work, we address two questions. First is to examine whether XLM-R is effective in capturing multilingual knowledge by constructing a custom model over it to analyze if a model trained using English-only tweets will generalize to multilingual data and vice-versa. Social media streams are generally different from other text, given the user-generated content. For example, tweets are usually short with possibly errors and ambiguity in the behavioral expressions. These properties in turn make the classification task or extracting representations a bit more challenging. Second question

[1] https://appen.com/datasets/combined-disaster-response-data/

In M. Gaur, A. Jaimes, F. Ozcan, S. Shah, A. Sheth, B. Srivastava, Proceedings of the Workshop on Knowledge-infused Mining and Learning (KDD-KiML 2020). San Diego, California, USA, August 24, 2020.
KiML'20, August 24, 2020, San Diego, California, USA,

https://doi.org/10.1145/nnnnnnn.nnnnnnn

is to examine whether word translations will be equally attended by the attention layers. For instance, the words with higher attention weights in a sentence in Haitian Creole such as "*Tanpri nou bezwen tant avek dlo nou zon silo mesi*" should align with the words in its corresponding translated tweet in English "*Please, we need tents and water. We are in Silo, Thank you!*". Our core idea is that if '*dlo*' in the Haitian tweet has a higher weight, so should its English translation '*water*'. This word-level language agnostic property can promote machine learning models to be more interpretable. This also brings several benefits to downstream tasks such as knowledge graph construction using keywords extracted from tweets. In situations where data is available only in one language, this similarity in attention would still allow us to extract relevant phrases in cross-lingual settings. To the best of our knowledge in crisis analytics domain, aligning attention in cross-lingual setting is not attempted before. In this work, we focus our classification experiments only to tweets containing '*request*' intent, which will be expanded to other behaviors, tasks, and datasets in the future.

Contributions: We propose a novel attention realignment method which promotes the task classifier to be more language agnostic, which in turn tests the effectiveness of multilingual knowledge capture of XLM-R model for crisis tweets; and a pseudo-labelling procedure to further enhance the model's generalizability. Furher, incorporating the attention-based mechanism allows us to perform an interpretability analysis on the model, by comparing how words are attended in the original versus translated tweets.

2 RELATED WORK AND BACKGROUND

There are numerous prior works (*c.f.* surveys [4, 14]) that focus specifically on disaster related data to perform classification and other rapid assessments during an onset of a new disaster event. Crisis period is an important but challenging situation, where collecting labeled data during an ongoing event is very expensive. This problem led to several works on domain adaptation techniques in which machine learning models can learn and generalize to unseen crisis event [3, 10, 18, 23]. In the context of crisis data, Nguyen et al. [28] designed a convolutional neural network model which does not require any feature engineering and Alam et al. [1] designed a CNN architecture with adversarial training on graph embeddings. Krishnan et al. [19] showed that sharing a common layer for multiple tasks can improve performance of tasks with limited labels.

In multilingual or cross-lingual direction, many works [8, 17] tried to align word embeddings (such as fastText [27]) from different languages into the same space so that a word and its translations have the same vector. These models are superseded by models such as multilingual BERT [9] and XLM-R [7] that produce contextual embeddings which can be pretrained using several languages together to achieve impressive performance gains on multilingual use-cases.

Attention mechanism [2, 24] is one of the most widely used methods in deep learning that can construct a context vector by weighing on the entire input sequence which improves over previous sequence-to-sequence models [13, 34, 35]. As the model produces weights associated with each word in a sentence, this allows for evaluating interpretability by comparing the words that are given priority in original versus translated tweets.

With more and more machine learning systems being adopted by diverse application domains, transparency in decision-making inevitably becomes an essential criteria, especially in high-risk scenarios [12] where trust is of utmost importance. With deep neural networks, including natural language systems, shown to be easily fooled [16], there has been many promising ideas that empower machine learning systems with the ability to explain their predictions [5, 32]. Gilpin et al. [11] presents a survey of interpretability in machine learning, which provides a taxonomy of research that addresses various aspects of this problem. Similar to the work by Ross et al. [33], we employ an attention-based approach to evaluate model interpretability applied to the crisis-domain.

3 METHODOLOGY

3.1 Problem Statement: Unsupervised Cross-Lingual Crisis Tweet Classification

Consider tweets in language A and their corresponding translated tweets in language B. The task of unsupervised cross-lingual classification is to train a classifier using the data only from the source language and predict the labels for the data in the target language. This experimental set up is usually represented as $A \longrightarrow B$ for training a model using A and testing on B or $A \longrightarrow B$ for training a model using B and testing on A. X refers to the data and Y refers to the ground truth labels. The multilingual dataset used in our experiments consists of original multilingual (*ml*) tweets and their translated (*en*) tweets in English. To summarize:

Experiment A ($en \longrightarrow ml$):

Input: X_{en}, Y_{en}, X_{ml}
Output: $Y_{ml}^{pred} \longleftarrow predict(X_{ml})$

Experiment B ($ml \longrightarrow en$):

Input: X_{ml}, Y_{ml}, X_{en}
Output: $Y_{en}^{pred} \longleftarrow predict(X_{en})$

3.2 Overview

In the following sections, we propose two methodologies to enhance cross-lingual classification: 1) Attention Realignment and 2) Pseudo-Labelling. Attention realignment utilizes a language classifier which is trained in parallel to realign the attention layer of the task classifier such that the weights are more geared towards task-specific words regardless of the language. Pseudo-Labelling further enhances the classifier by adding high quality seeds from the target language that are pseudo-labelled by the task classifier.

3.3 Attention Realignment by Parallel Language Classifier

As depicted in Fig 2, model on the left side is the task classifier and the model on the right side is a language classifier that is trained in parallel. The purpose of this language classifier is to pick up aspects that is missed by the XLM-R model. This could be tweet-specific, crisis-specific, or other linguistic nuances that can separate original tweets and translated tweets. Note that semantically, translated words are expected to have similar XLM-R representations.

Attention realignment is a mechanism we introduce to promote the task classifier to be more language independent. The main idea is that the words that are given higher attention in a language

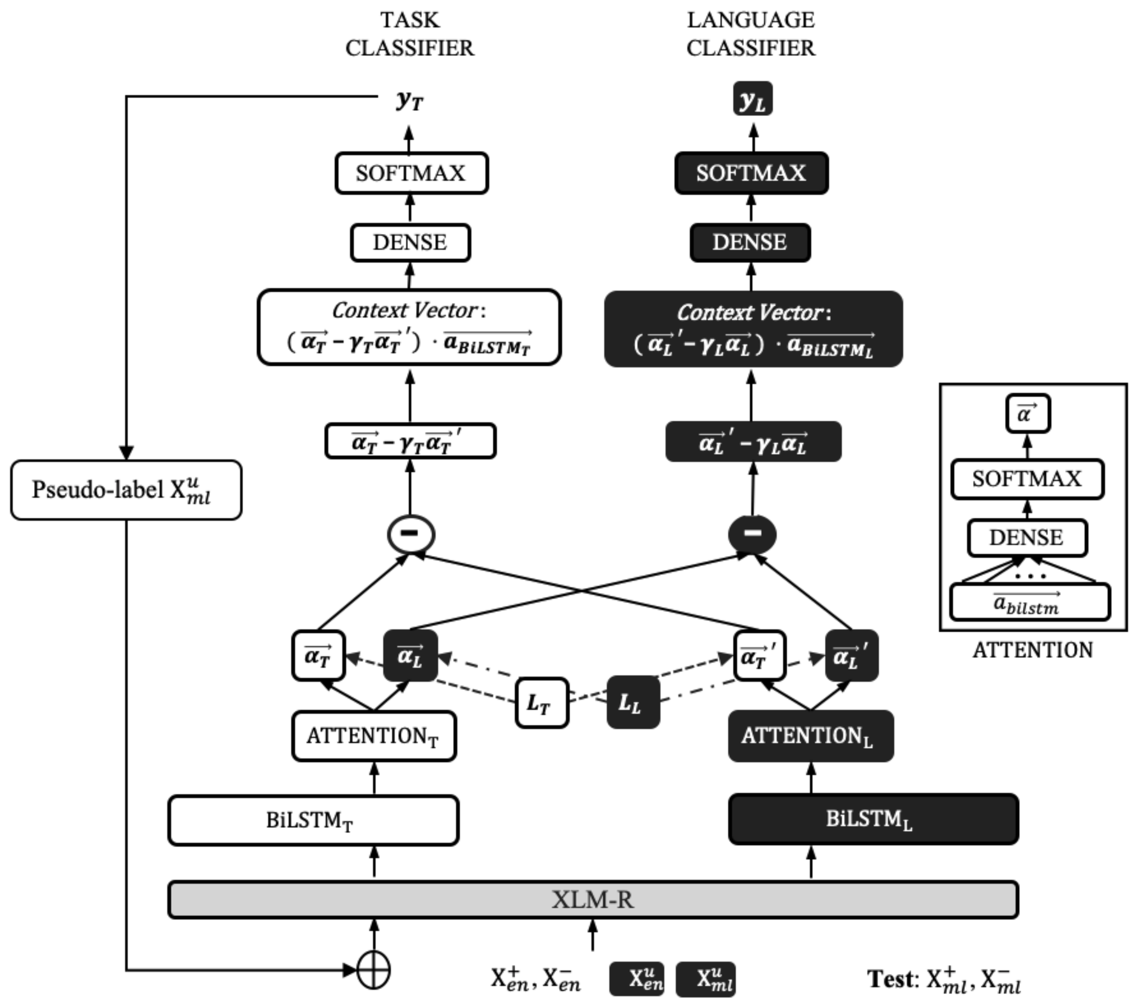

TASK CLASSIFIER
LANGUAGE CLASSIFIER
y_T
y_L
SOFTMAX
SOFTMAX
DENSE
DENSE
Context Vector: $(\overrightarrow{\alpha_T} - \gamma_T \overrightarrow{\alpha_T}') \cdot \overrightarrow{a_{BiLSTM_T}}$
Context Vector: $(\overrightarrow{\alpha_L}' - \gamma_L \overrightarrow{\alpha_L}) \cdot \overrightarrow{a_{BiLSTM_L}}$
$\overrightarrow{\alpha_T} - \gamma_T \overrightarrow{\alpha_T}'$
$\overrightarrow{\alpha_L}' - \gamma_L \overrightarrow{\alpha_L}$
Pseudo-label X^u_{ml}
$\overrightarrow{\alpha_T}$
$\overrightarrow{\alpha_L}$
L_T
L_L
$\overrightarrow{\alpha_T}'$
$\overrightarrow{\alpha_L}'$
ATTENTION$_T$
ATTENTION$_L$
BiLSTM$_T$
BiLSTM$_L$
XLM-R
X^+_{en}, X^-_{en}
X^u_{en}
X^u_{ml}
Test: X^+_{ml}, X^-_{ml}
$\overrightarrow{\alpha}$
SOFTMAX
DENSE
$\overrightarrow{a_{bilstm}}$
ATTENTION

Figure 2: Attention Realignment with Pseudo-Labelling over XLM-R model

Notation	Definition
en	Tweets translated to English ('message' column in the dataset)
ml	Multilingual Tweets ('original' column in the dataset)
α	Attention Layer
T	A component that uses Task-specific data. i.e., + and − 'Request' tweets
L	A component that uses Language-specific data. i.e., *en* and *ml* tweets
a_{BiLSTM}	Activation from the BiLSTM layer
β, γ, ζ	Hyperparameters

Table 1: Notations

classifier should be less important in a task classifier. For example, '*dlo*' in Haitian and '*water*' in English should have the same vector representation in language agnostic models; while the sentence structure, grammar, and other nuances can vary. We enforce this rule by constructing two operations:

(1) **Attention Difference**: When a sentence goes through model M1, it also goes through model M2. For the same sentence, this returns two attention layer weights: one from the task classifier ($\overrightarrow{\alpha_T}$) and the other from the language classifier ($\overrightarrow{\alpha_T}'$). Directly subtracting $\overrightarrow{\alpha_T}'$ from $\overrightarrow{\alpha_T}$ poses two issues: 1) we do not know whether they are comparable and 2) $\overrightarrow{\alpha_T}'$ may have negative values. A simple solution to this is to normalize bothe vectors and clip $\overrightarrow{\alpha_T}'$ such that it is between 0 and 1. Thus, an attention subtraction step is as follows:

$$\frac{\overrightarrow{\alpha_T}}{\|\overrightarrow{\alpha_T}\|} - \gamma_T \; clip\left(\frac{\overrightarrow{\alpha_T}'}{\|\overrightarrow{\alpha_T}'\|}, 0, 1\right) \tag{1}$$

where γ_T is a hyperparameter to tune the amount of subtraction needed for the task classifier. Similarly, for the language classifier,

$$\frac{\overrightarrow{\alpha_L}'}{\|\overrightarrow{\alpha_L}'\|} - \gamma_L \, clip\left(\frac{\overrightarrow{\alpha_L}}{\|\overrightarrow{\alpha_L}\|}, 0, 1\right) \tag{2}$$

(2) **Attention Loss**: Along with attention difference, the model can also be trained by inserting an additional loss function term that penalizes the similarity between the attention weights from the two classifiers. We use the Frobenius norm.

$$L_{At} = \|\overrightarrow{\alpha_T}^T \; \overrightarrow{\alpha_T}'\|_F^2 \tag{3}$$

$$L_{Al} = \|\overrightarrow{\alpha_L}^T \; \overrightarrow{\alpha_L}'\|_F^2 \tag{4}$$

for task and language respectively. Resulting final loss function of joint training will be:

$$L(\theta) = \zeta_T\Big(CE_T + \beta_T L_{At}\Big) + \zeta_L\Big(CE_L + \beta_L L_{Al}\Big) \tag{5}$$

where β is the hyperparameter to tune the attention loss weight, ζ is the hyperparameter to tune the joint training loss, and CE denotes the binary cross entropy loss,

$$CE = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y_i} + (1 - y_i)\log(1 - \hat{y_i})] \tag{6}$$

It is important to note that the Frobenius norm is **not** simply between the attention weights of the two models but rather between the attention weights produced by the two models on the same input tweet. For example, for a given tweet, the task classifier attends more to task-specific words and the language classifier attends to language-specific words. So the mechanism makes sure that they are distinct.

3.4 Pseudo-Labelling

To enhance the model further, we pseudo-label the data in the target language. For example, if we are training a model using the English tweets, we use the original tweets before translation for pseudo-labelling. The idea is simply to gather high-quality seeds from the target to retrain the model. Note that, we still do not use any target labels here; still following the unsupervised goal. Thus, for retraining model M1 for $en \longrightarrow ml$, the new dataset would consist of: X_{en}^{+} and $X_{ml}^{pseudo+}$ as positive examples and X_{en}^{-} and $X_{ml}^{pseudo-}$ as negative examples.

3.5 XLM-R Usage

The recommended feature usage of XLM-R[2] is either by fine-tuning to the task or by aggregating features from all the 25 layers. We employ the later to extract the multilingual embeddings for the tweets.

4 DATASET & EXPERIMENTAL SETUP

	Train	Validation	Test
Positive	3554	418	496
Negative	17473	2152	2128

Table 2: Dataset Statistics for both *en* amd *ml*

T_x	30
Deep Learning Library	Keras
Optimizer	Adam [$lr = 0.005$, $beta_1 = 0.9$, $beta_2 = 0.999$, $decay = 0.01$]
Maximum Epoch	100
Dropout	0.2
Early Stopping Patience	10
Batch Size	32
ζ_T	1
ζ_L	0.1
$\beta_T, \beta_L, \gamma_T, \gamma_L$	0.01

Table 3: Implementation Details

We use the open source dataset from Appen[3] consisting of multilingual crisis response tweets. The dataset statistics for tweets with 'request' behavior labels is shown in Table 2. For all the experiments, the dataset is balanced for each split.

Each experiment is denoted as $A \longrightarrow B$, where A is the data that is used to train the model and B is the data that is used for testing the model. For example, $en \longrightarrow ml$ means we train the model using English tweets and test on multilingual tweets.

Models are implemented in Keras and the details are shown in table 3. Hyperparameters β_T, β_L, γ_T, and γ_L are not exhaustively tuned; we leave this exploration for future work.

	Baseline	Model M1	Model M2
$en \longrightarrow ml$	59.98 (80.57)	62.53 (77.02)	**66.79** (82.39)
$ml \longrightarrow en$	60.93 (70.07)	65.69 (63.50)	**70.95** (73.84)

Table 4: Performance Comparison (Accuracy in %) for $Source \longrightarrow Target$ ($Source \longrightarrow Source$).
Baseline = XLMR + BiLSTM + Attention.
Model M1 = Baseline + Attention Realignment.
Model M2 = Model M1 + Pseudo-Labelling.

5 RESULTS & DISCUSSION

Table 4 shows the cross-lingual performance comparison of all the models. The three models are described below:

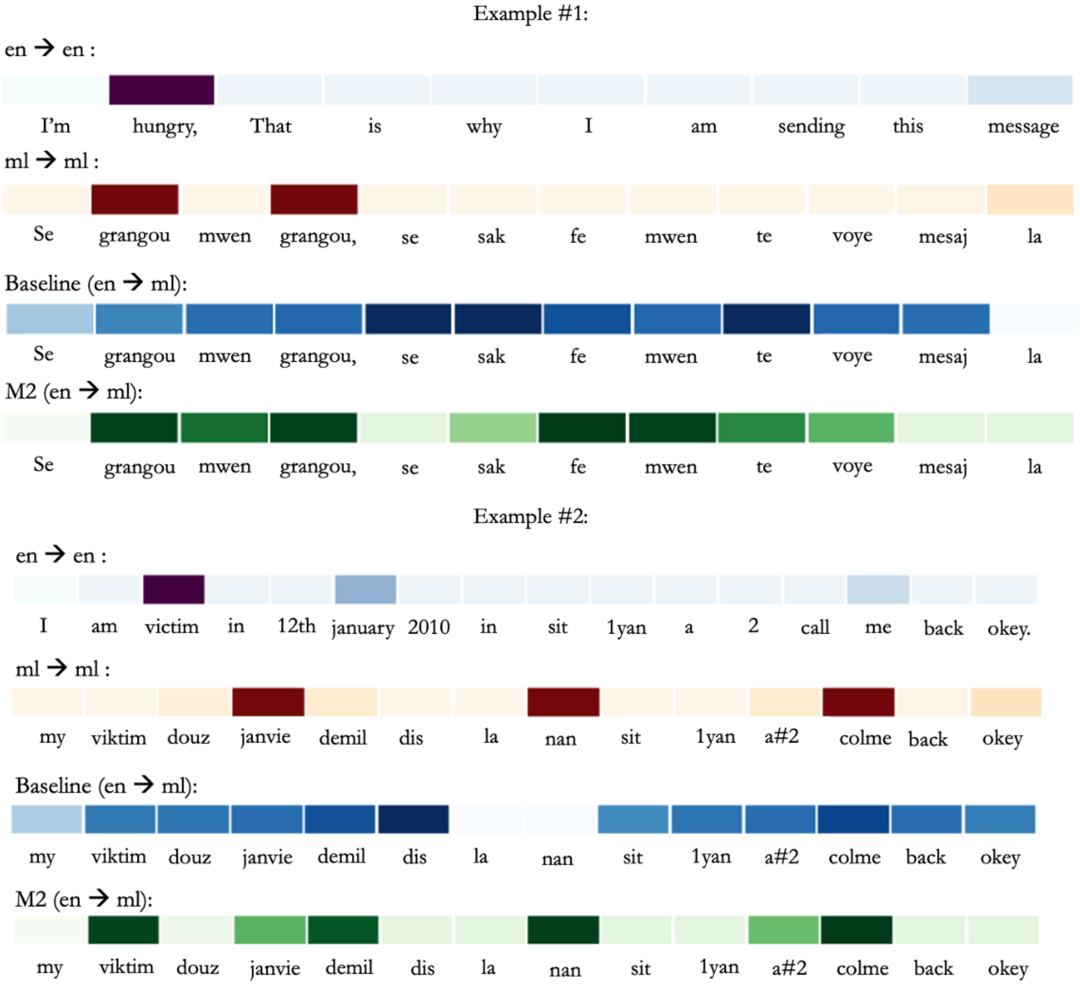
(1) **Baseline**: The baseline model consists of embeddings retrieved from XLM-R trained over BiLSTMs and Attention layers. This is a traditional sequence (text) classifier enhanced with attention mechanism. Activations from the BiLSTM layers are weighed by the attention layer to construct the context vector which is then passed through a dense layer and softmax function to produce the classification output.

(2) **Model M1**: Adding attention realignment to the baseline model produces model M1. Attention realignment is achieved through a language classifier which is trained in parallel with the goal to make the task classifier more language agnostic.

[2] https://github.com/facebookresearch/XLM

[3] https://appen.com/datasets/combined-disaster-response-data/

Figure 3: Attention visualization example for '*request*' tweets: words and their attention weights for two tweets in Haitian Creole and its translation in English (darker the shade, higher the attention).



The attention weights for both task and language classifiers are manipulated by each other during training by a process of subtraction (attention difference) as well a loss component (attention loss). See section 3.3.

(3) **Model M2**: Adding the pseudo-labelling procedure to model M1 produces model M2. Using Model M1 which is trained to be language agnostic, tweets from the target languages are pseudo-labelled. High quality seeds are selected (using Model M1 p>0.7) and augmented to the original training dataset to retrain the task classifier.

Results show that, for cross-lingual evaluation on *en* → *ml*, model M1 outperforms the baseline by **+4.3%** and model M2 outperforms by **+11.4%**. On *ml* → *en*, model M1 outperforms the baseline by **+7.8%** and model M2 outperforms by **+16.5%**. This shows that both models are effective in cross-lingual crisis tweet classification. An interesting observation to note is that using attention realignment alone decreased the classification performance in the same language, which is brought back up by pseudo-labelling. These scores are shown in brackets in table 4. A deeper investigation in this direction on various other tasks can shed more light on the impact of realignment mechanism.

5.1 Interpretability: Attention Visualization

We follow a similar attention architecture shown in [18]. The context vector is constructed as a result of dot product between the attention weights and word activations. This represents the interpretable layer in our architecture. The attention weights represent the importance of each word in the classification process. Two examples are shown in figure 3. In the first example, both *en* → *en* and *ml* → *ml* give attention to the word '*hungry*' (i.e., '*grangou*' in Haitian Creole). Note that these two are results from the models that are trained in the same language in which they are tested; thus, expecting an ideal performance. For the baseline model in the cross-lingual set-up *en* → *ml*, although it correctly predicts the label, the attention weights are more spread apart. In model M2 with attention realignment and pseudo-labelling, although with some spread, the attention weights are shifted more toward '*grangou*'. Similarly in example 2, the attention weights in the baseline model are more spread apart. Cross-lingual performance of model M2 aligns more with $en \rightarrow en$ and $ml \rightarrow ml$. These examples show the importance of having interpretability as a key criterion in cross-lingual crisis tweet classification problems; which can also be used for downstream tasks such as extracting relevant keywords for knowledge graph construction.

6 CONCLUSION

We presented a novel approach for unsupervised cross-lingual crisis tweet classification problem using a combination of attention realignment mechanism and a pseudo-labelling procedure (over the state-of-the-art multilingual model XLM-R) to promote the task classifier to be more language agnostic. Performance evaluation showed that both models M1 and M2 outperformed the baseline by +4.3% and +11.4% respectively for cross-lingual text classification from English to Multilingual. We also presented an interpretability analysis by comparing the attention layers of the models. It shows the importance of incorporating a word-level language agnostic characteristic in the learning process, when training data is available only in one language. Performing extensive hyperparameter tuning and expanding the idea to other tasks (including cross-task/multi-task) are left as future work. We also plan another direction for future work as to incorporate the human-engineered knowledge from the multilingual knowledge graphs such as BabelNet in our model architecture that could improve the learning of similar concepts across languages critical to the crisis response agencies.

Reproducibility: Source code is available available at: https://github.com/jitinkrishnan/Cross-Lingual-Crisis-Tweet-Classification

7 ACKNOWLEDGEMENT

Authors would like to thank U.S. National Science Foundation grants IIS-1815459 and IIS-1657379 for partially supporting this research.

REFERENCES

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

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[3] John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*. 120–128.

[4] Carlos Castillo. 2016. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*. 2172–2180.

[6] Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. 2020. Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. 292–298.

[7] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[8] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *arXiv preprint arXiv:1710.04087* (2017).

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Yaroslav Ganin and Victor Lempitsky. 2014. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* (2014).

[11] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[12] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[13] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[14] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47, 4 (2015), 1–38.

[15] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. *arXiv preprint arXiv:1605.05894* (2016).

[16] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328* (2017).

[17] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. *arXiv preprint arXiv:1804.07745* (2018).

[18] Jitin Krishnan, Hemant Purohit, and Huzefa Rangwala. 2020. Diversity-Based Generalization for Neural Unsupervised Text Classification under Domain Shift. *https://arxiv.org/pdf/2002.10937.pdf* (2020).

[19] Jitin Krishnan, Hemant Purohit, and Huzefa Rangwala. 2020. Unsupervised and Interpretable Domain Adaptation to Rapidly Filter Social Web Data for Emergency Services. *arXiv preprint arXiv:2003.04991* (2020).

[20] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* (2019).

[21] Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2013. Real-time disease surveillance using twitter data: demonstration on flu and cancer. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1474–1477.

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

[23] Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[24] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[25] Guoqin Ma. 2019. Tweets Classification with BERT in the Field of Disaster Management. https://pdfs.semanticscholar.org/d226/185fa1e14118d746cf0b04dc5be8f545ec24.pdf.

[26] Reza Mazloom, Hongmin Li, Doina Caragea, Cornelia Caragea, and Muhammad Imran. 2019. A Hybrid Domain Adaptation Approach for Identifying Crisis-Relevant Tweets. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 11, 2 (2019), 1–19.

[27] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[28] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. 2016. Rapid classification of crisis-related data on social networks using convolutional neural networks. *arXiv preprint arXiv:1608.03902* (2016).

[29] Ferda Ofli, Patrick Meier, Muhammad Imran, Carlos Castillo, Devis Tuia, Nicolas Rey, Julien Briant, Pauline Millet, Friedrich Reinhard, Matthew Parkan, et al. 2016. Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big data* 4, 1 (2016), 47–59.

[30] Bahman Pedrood and Hemant Purohit. 2018. Mining help intent on twitter during disasters via transfer learning with sparse coding. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 141–153.

[31] Hemant Purohit, Carlos Castillo, Fernando Diaz, Amit Sheth, and Patrick Meier. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday* 19, 1 (Dec. 2013).

[32] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[33] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717* (2017).

[34] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[35] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[36] István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1619–1629.