Multi-label and Multilingual News Framing Analysis

Afra Feyza Akyürek Lei Guo Randa Elanwar

Prakash Ishwar Margrit Betke Derry Wijaya

Boston University
akyurek@bu.edu

Abstract

News framing refers to the strategy in which aspects of certain issues are highlighted in the news to promote a particular interpretation. In NLP, although recent works have studied framing in English news, few have studied how the analysis can be extended to other languages and in a multi-label setting. In this work, we explore multilingual transfer learning to detect multiple frames from just the news headline in a genuinely low-resource setting where there are few/no frame annotations in the target language. We propose a novel method that can leverage very basic resources consisting of a dictionary and few annotations in a target language to detect frames in the language. Our method performs comparably or better than translating the entire target language headline to the source language for which we have annotated data. This opens up an exciting new capability of scaling up frame analysis to many languages, even those without existing translation technologies. Lastly, we apply our method to detect frames on the issue of U.S. gun violence in multiple languages and obtain interesting insights on the relationship between different frames of the same issue across different countries with different languages.

1 Introduction

The worldwide image of the United States has dropped precipitously during the past few years (Wike et al., 2018). Among other factors, the increasing number of gun violence incidents appears to affect the U.S. reputation abroad. Whenever a fatal mass shooting happens, it often attracts significant international news attention. While the domestic U.S. news media often links gun violence to individual shooters' mental illness (DeFoster and Swalve, 2018; Liu et al., 2019), foreign media may attribute it to U.S. gun policy and its gun culture e.g. (Atkinson, 2019). This is known as media framing, which refers to the process of selecting "some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item" (Entman, 1993). When foreign media frame the gun violence issue in a way to depict the U.S. as an unsafe and undesired place, it erodes the country's "soft power" —the ability to attract (Nye Jr, 2004). Evaluating how different countries frame the U.S. gun violence issue will enrich our understanding of the U.S. soft power in particular and international relations in general. In this work, we develop a multilingual approach to automatically detect frames in news coverage of different languages, thus facilitating the analysis on how different countries with different languages frame a particular issue. Aside from enabling this, understanding of foreign public opinion regarding a certain issue or nation, a multilingual approach is essential in media framing analysis, as it is also an under-studied problem in many parts of the world.

Given frame annotated news headlines of a particular topic in a source language (e.g., English), our approach uses word-to-word translation to translate keywords that are indicative of the frames in these headlines to a target language. Then, we fine-tune a state-of-the-art multilingual language model MultiBERT (Devlin et al., 2019) to detect frames on these "code-switched" headlines combined with a few annotated headlines from the target language. The translated keywords and a few-shot examples act as anchors to adapt MultiBERT to detect frames in the target language. This approach performs comparably or better than a model trained on the source language and tested on headlines that have been translated from the target language to source. Since our approach requires only simple resources: a dictionary and a few ($\leq$40) annotated examples in the target language, it can

be applied to many languages even those without existing translation technologies.

Due to the subtle nature of framing, it is not uncommon that one news article involves more than one message. Communication researchers have suggested that the association of different constructs such as issues and frames in the news will influence how the audience associate these elements, thus determining how they perceive the world (Guo and McCombs, 2015). The so-called Network Agenda Setting Model suggests that examining the interrelationships between media elements enables researchers to measure media effects in a more nuanced manner. In this work, we formulate our frame detection model to allow for multi-label frame detection while also addressing the imbalance in the frame distribution – since some frames are used more often in the news than others – by adapting focal loss (Lin et al., 2017) to our multi-label setting. Our multi-label approach allows for the examination of frame co-occurrence, or "associative frames" (Schultz et al., 2012), across the news articles.

Overall, the contribution of this work are manifold: (1) we devise a novel code-switch few-shot scheme to train a frame detection model for any language, (2) we extend the formulation of the frame classification problem and focal loss to a multi-label setting, allowing the model to predict multiple frames for each instance, (3) we use our multilingual multi-label frame detection model to detect frames in news headlines pertaining to U.S. gun violence issue in multiple countries and languages and obtain interesting insights on how other countries view the gun violence issue in the U.S. and how frames are related across news articles in different countries with different languages.[1]

2 Background and Related Work

Today's international politics not only resolves around military and economic influence, but also largely depends on a country's soft power (Nye Jr, 2004). For each nation, constructing a positive country image to the outside world is crucial to ensure its international competitiveness in this global information society (Buhmann and Ingenhoff, 2015). In this light, more and more governments have realized the importance of public diplomacy, making great efforts to promote their countries' values and perspectives to the foreign public (Entman, 2008; Golan and Himelboim, 2016). However, these efforts are not always successful. Editors of foreign news media serve as the gatekeepers to decisions which may lead to the framing of a given country contrary to what its government intends. In reporting news about a foreign country, news editors and reporters make conscious or unconscious choices to emphasize certain issues, or emphasize certain aspects of a given issue, which may alter the country's image in the minds of their audience. A multilingual approach is essential to analyze media framing in different parts of the world, which will shed light on foreign public opinion regarding a certain nation.

Communication researchers often rely on manual content analysis to examine media framing in news outlets of different languages (H. De Vreese, 2001). One critique for this type of study is that researchers tend to decide countries for analysis based on languages spoken in the research team rather than theoretical rationales. This language constraint becomes a greater challenge in this increasingly globalized media landscape; capturing a holistic picture of international communication would require the analysis of news coverage in a larger number of languages. Arguably, an automatic, multilingual approach of framing analysis would largely benefit international communication research community.

The current state-of-the-art model for frame detection (Liu et al., 2019) fine-tunes BERT on frame annotated English news headlines with the regular multiclass focal loss (Lin et al., 2017). It predicts one frame for each headline. Since a headline can be annotated with more than one frame, their model is only evaluated on the first annotated frame. In this work, we fine-tune MultiBERT to detect frames in multiple languages' headlines with our multi-label focal loss. Contrary to their work, our approach is able to predict (and be evaluated on) multiple frames for each headline while being comparable to their work in terms of the average F1 performance. Similar to their work, we detect frames on news headlines as they provide the most direct clue to the potential influence of the news coverage (Liu et al., 2019).

3 Dataset Creation

We extend the Gun Violence Frame Corpus (GVC) (Liu et al., 2019) to other languages. GVFC is a dataset of news articles from 21 major U.S. news

[1] Code and data will be made available.

organizations related to U.S. gun violence that contains news headlines and their domain-expert frame annotations. To extend GVFC to news headlines in other languages, following their process of curating GVFC, we first drew our sample of news articles from German-, Turkish-, and Arabic-speaking news websites, using Crimson Hexagon's ForSight social media analytics platform (Hexagon, 2018), retrieving articles that had at least one keyword in their headlines from the following list of words: {"gun," "firearm," "NRA," "2nd amendment," "second amendment," "AR15," "assault weapon," "rifle," "Brady act," "Brady bill," "mass shooting"} that have been translated to German, Turkish, and Arabic words respectively by native speakers of the languages.

We then train two annotators for each language to apply the GVFC codebook protocol for annotation and measure their agreement on how to apply the codes on a sample of 350, 200, and 210 German, Turkish, and Arabic headlines, respectively. The coders achieve 92.6%, 98.5%, 78.1% agreement rates on the first frame and 78.9%, 97.9%, 74.3% agreement rates on the second frame on these German, Turkish, and Arabic samples. One coder of each language continues to code more headlines from the retrieved articles, resulting in a total of 326, 100, and 388 non-duplicate headlines for German, Turkish, and Arabic that are annotated as relevant to the issue of U.S. gun violence and with their frames. Average number of labels, i.e. label cardinalities, per headline are 1.4, 1.5, and 1.5, for German, Turkish and Arabic whereas it's 1.3 in GVFC. As we can observe from the agreement rates, the Arabic data has a relatively poorer intercoder reliability (ICR) while the Turkish data has the best ICR. As high ICR values implies that two coders consistently categorized the content similarly, they signal a high validity of the coded results. In turn, this is reflected on the performance of our model on these datasets. Since there are potentially erroneous annotations in the Arabic data, our model performs the worst on this data (section 5).

4 Model

In this work, we extend the current state-of-the-art model on GVFC dataset (Liu et al., 2019), which predicts only the first frame, into a multi-label approach and evaluate it across multiple languages. As previous work has showcased that BERT surpasses LSTM and GRU-based architectures, we shift our focus in this work from architecture optimization to scalability of news framing analysis across multiple languages in a multi-label setting.

BERT relies on multiple stacks of the Transformer's encoder blocks (Devlin et al., 2019), (Vaswani et al., 2017) to learn vector representations of sentences. A single encoder block is composed of a self-attention layer followed by a fully-connected layer. When a sentence – a sequence of tokens – is fed into the encoder, it passes through an embedding layer, then through the self-attention and fully-connected layers before being passed to the upper encoder block. The self-attention layer embodies three matrices called W^Q for query, W^K for key and W^V for value. Each of these matrices are of size $vocab\_size \times hidden\_size$, thus each token in the vocabulary has its corresponding q, k and v vectors. Representations for each token are contextualized, namely, the representation of a token is the weighted average of all representations in the sequence. Thus vector representation for token x_i is given by

$$\text{vec\_rep}(x_i) = \sum_{j \in S} v_j \, \text{Softmax}(q_i \cdot k_j / \sqrt{d})$$

where d is the size of the key vectors in W^K and S is the set of all tokens in the same sequence as x_i, including x_i.

BERT adds a special token for classification `[CLS]` at the beginning of each sequence and learns the representation of this token and other tokens in the sequence by training on Wikipedia corpus for two language tasks: next sentence prediction and Masked Language Model (MLM), which was originally inspired by the Cloze task (Taylor, 1953). The contextual representation of the `[CLS]` token encodes the syntactic and semantic constructs of the sequence and can be fine-tuned for various down-stream tasks.

Fine-tuning BERT has been shown to perform well on new tasks even with small datasets which can be attributed to the data-efficient deep attention mechanism (Devlin et al., 2019), (Vinyals et al., 2015). The knowledge encoded within the vector representations of the tokens through pre-training also helps the classifier with the language understanding part of the task, reducing the need for a larger dataset. Finally, a multilingual version of pre-trained BERT, which are trained on the entire Wikipedia dumps of 104 languages with the largest Wikipedia, has recently been released making it an excellent candidate for scaling to multiple lan-

guages. Thanks to multilingual pre-training and the utilization of sub-word tokenization that facilitates scaling to multiple languages and knowledge sharing (Gu et al., 2018), MultiBERT can represent sequences from any of these 104 languages that enables zero-shot classification on them (i.e., train on one language and test on another).

In our case, since reproducing the effort put in GVFC which was created by highly qualified journalism students in other languages is prohibitive, employing a cross-lingual model such as MultiBERT renders scaling to other languages possible.

4.1 Multi-label News Frame Detection

One useful property of BERT is that it is designed for deployability in various down-stream tasks such as next sentence prediction, question answering and text classification. For our frame detection purposes, we would like to classify news articles into 9 frame categories based on their headlines. Devlin et al. (2019) recommend using the embedding generated for the special token called `[CLS]` which is padded to the beginning of every sentence. As with every token, `[CLS]` is also of length $H = 768$ and it's representation is generated by attending every word in the sequence. We modify BERT by appending to it a fully connected layer which acts as a classifier taking in the embedding generated for `[CLS]` after 12 layers of encoders and mapping it into $K = 9$ output neurons. Thus the only parameters trained from scratch during fine-tuning are those of the classifier layer's, $W \in \mathcal{R}^{HxK}$. Finally, we use Sigmoid activations to obtain 9 outputs between 0 and 1 which are interpreted as scores for each of the 9 classes. During inference, we use the canonical threshold of 0.5 on these scores.

We fine-tune MultiBERT with two different losses: the standard Binary Cross-Entropy loss, and a multi-label variation of the focal loss (Lin et al., 2017). We compute the Binary Cross-Entropy (BCE) loss, also named as Sigmoid Cross-Entropy loss, for a single sample $\mathbf{x}$ as $BCE(f) = -\frac{1}{|K|}\sum_{i=1}^{|K|}\left(y^{(i)}log(\hat{y}^{(i)}) + (1-y^{(i)})log(1-\hat{y}^{(i)})\right)$ where predictions given by $\hat{\mathbf{y}} = [\hat{y}^{(1)}, \ldots, \hat{y}^{(|K|)}] = 1/\left(1 + exp(-f(\mathbf{x}))\right)$, $\mathbf{y} = [y^{(1)}, \ldots, y^{(|K|)}]$ are the gold binary labels and f is BERT followed by the classification layer.

Considering the high degree of class imbalance in the GVFC dataset which deteriorates within the multilingual datasets we developed, we adopt a multi-label variation of focal loss which was originally proposed for the binary classification task by Lin et al. (2017). Focal loss for a single sample $\mathbf{x}$ is defined as $FL(g) = -\alpha(1-p)^2 log(p)$ where $p = (1-y)\cdot(1-\hat{y}) + y\cdot\hat{y}$ and $y \in \{0,1\}$ is the true label, also $\hat{y} = 1/\left(1 + exp(-g(\mathbf{x}))\right) \in \mathbb{R}$, α is the balancing factor, usually normalized inverse class frequency – hence the smaller the class, the higher the α and vice versa, which balances the importance of each class' examples – while g is the hypothesis e.g. linear hypothesis. In our case, we alter focal loss formulation such that y and $\hat{y}$ are replaced by $\mathbf{y} \in \{0,1\}^{|K|}$ and $\hat{\mathbf{y}} \in \mathbb{R}^{|K|}$. Also we use the weights α as $\alpha = \left[(\alpha_1^{(0)}, \alpha_1^{(1)}), \ldots, (\alpha_k^{(0)}, \alpha_k^{(1)})\right]$ where $\alpha_k^{(j)}$ is the normalized inverse frequency of the event $y^k = j$ where $j \in \{0,1\}$. In other words, we interpret each class as two classes either 0 or 1 and compute inverse class frequencies for all $2 * |K|$ classes and normalize them such that $\sum_{k\in K}\sum_{j\in\{0,1\}} \alpha_k^{(j)} = 1$. We observe that this loss matches BCE and in F1-scores and prevails them in multi-label accuracy as in Table 1.

We use two Binary Relevance approaches based on Naïve Bayes and MultiBERT respectively as our baselines. Naïve Bayes is a standard baseline for text classification which leverages Bayes theorem and utilizes word frequencies as features (McCallum et al., 1998). For regularizaton, we apply add-1 smoothing. The standard configuration for Naïve Bayes is multi-class. One intuitive technique of tailoring Naïve Bayes into a multi-label problem is called Binary Relevance (BR). BR is the method of training $|K|$ one-vs-rest classifiers independently for each of class $k \in K$ on the same dataset. Moreover, we do BR using MultiBERT by training 9 binary MultiBERTs in a one-vs-rest manner as our second baseline.

Among the potential weaknesses of BR, it is criticized for ignoring label correlations and incapability of handling class imbalance effectively the most (Zhang et al., 2018). Plus, modeling complexity is linear in number of classes which is inefficient when it comes to models like BERT which entail immense spatial and computational costs.

4.2 Multilingual Models

GVFC dataset is composed of 1300 relevant samples for the issue of Gun Violence and is only available in English. For cross-lingual transfer, MultiBERT with multi-label Focal loss provides the highest accuracy within multi-labeled English samples (samples that have more than one true class) by a

10% margin while maintaining the same level of macro and micro F-1 scores given in Table 1.

Firstly, we explore zero-shot and few-shot performances of our MultiBERT model with Focal loss which is trained on the English dataset provided in the 1th and 3rd lines of Table 2. We use German (DE), Arabic (AR) and Turkish (TR) as our target languages to explore the cross-lingual performance of our model to a variety of languages for which we have some validation set but not train set. German is the closest language to our source language English, Turkish is from a different language family while still using the Latin alphabet and finally Arabic is far from English in both senses. In our few-shot model we use extra 40 samples from the target language, i.e. DE, AR or TR and use the same training configurations as in the original training which are described in Section 5.

Furthermore, since the task in hand, news framing, is fairly a keyword-driven phenomena (Field et al., 2018), we developed a set of keywords that are the most helpful in labeling. In order to determine the most salient words for frame classes, we utilize the metric called *normalized pointwise-mutual information* (nPMI) which was suggested by Field et al. (2018). nPMI score for a given frame F and word w is $I(F, w) = log\frac{P(w|F)}{P(w)}$. Both $P(w)$ and $P(w|F)$ are estimated from the training corpus. After determining the set of most important words based on nPMI (we select the top 250 words for each frame that also have nPMI scores greater than zero, resulting in 358 of them total), we use word-to-word translation to code-switch (CS) the English training set with the target language solely for these words. In other words, we replace all utterances of "important" words with it's TL dictionary translation. For instance, a sample headline in the training set that was code-switched with German becomes

```
Florida Schütze ein troubled
loner mit Weiß supremacist
Bindungen.
```

which originally was "Florida shooter a troubled loner with white supremacist ties" which is annotated as both "mental illness" and "race/ethnicity" frames. We experiment with using the code-switched data for training in both zero-shot, namely using no examples from the target language datasets, and few-shot, using 40 target language examples. Models based on code-switched training are indicated with CS_{TL} for target language (TL) in Table 2. Code-switched translation can be viewed as a way of adapting the model to the target language during training. In fact, we observed significant improvements or comparable results both in zero-shot and few-shot settings over the model that was trained on the original English data as demonstrated in Table 2 for all three languages. Furthermore, we explore the effect of translation direction for the news frame detection task using Google Translate in Table 3.

Model (Loss)	F1-Macro	F1-Micro	EM-1	EM-2	Top-2	EM-A
MULTICLASS						
EngBERT (Liu et al., 2019)	0.77	0.83	0.86	N/A	0.93	0.83
MultiBERT	0.73	0.79	0.82	N/A	0.89	0.79
MULTI-LABEL						
BR w/ Naïve Bayes	0.58	0.65	0.58	0.29	0.68	0.51
BR w/ MultiBERT (Binary Focal)	0.74	0.82	0.69	0.58	0.87	0.66
EngBERT (ML Focal)	0.76	0.82	0.71	0.62	0.94	0.69
MultiBERT (ML Focal)	0.76	0.82	0.71	0.62	0.92	0.69
MultiBERT (BCE Loss)	0.76	0.82	0.79	0.51	0.91	0.72

Table 1: English results. Multiclass models consider only the first frame correct and are evaluated accordingly. EM-1, EM-2, EM-A, Top-2: See Section 5. ML: Multi-Label, BR: Binary Relevance.

5 Experiments and Results

As input to our models we follow the previous work and rely on the news headlines rather than contents as input due to reasons motivated in (Liu et al., 2019). In order to showcase the gains made on top of a multi-class approach by re-formulating the problem as multi-label, we reproduce the method described in Liu et al. (2019) with both English BERT and MultiBERTs in Table 1. In our implementations involving BERT we use Adam optimizer with a learning rate of 0.02, a maximum sequence length of 128 and train for 10 epochs.

In Table 1, we include experiments that use different configurations of BERT, such as uncased English BERT (EngBERT) and cased Multilingual BERT (MultiBERT) with two different loss functions. Casing decisions were based on previous work (Liu et al., 2019) and recommendations in the code repository[2] for BERT, respectively. As of losses, we experimented with Binary Cross Entropy and multi-label Focal loss which are described in Section 4.1.

With regards to evaluation, we follow the recent work and report macro and micro-averaged F1-scores (Wu et al., 2019) as well as exact match (EM) for samples which have single frames (EM-1), two frames (EM-2) and any kind of frames (EM-A). In Table 1 we also report Top-2 accuracy which, for a given sample, computes the top two most confi-

[2]https://github.com/google-research/bert

	DE					AR					TR				
Model	**F1-Macro**	**F1-Micro**	**EM-1**	**EM-2**	**EM-A**	**F1-Macro**	**F1-Micro**	**EM-1**	**EM-2**	**EM-A**	**F1-Macro**	**F1-Micro**	**EM-1**	**EM-2**	**EM-A**
Zero-shot															
(2.1) Train EN, Test TL	0.48	0.65	0.47	0.31	0.39	0.37	0.39	0.38	0.04	0.24	0.50	0.77	0.76	0.29	0.53
(2.2) Train $CS_{TL}(EN)$, Test TL	**0.56**	**0.73**	**0.70**	**0.35**	**0.52**	**0.42**	**0.46**	**0.39**	**0.06**	**0.26**	**0.57**	**0.82**	**0.86**	**0.39**	**0.63**
Few-shot (40 TL samples)															
(2.3) Train EN, Test TL	**0.66**	0.75	0.52	0.37	0.44	0.48	0.54	**0.41**	0.17	**0.31**	0.77	0.89	0.67	0.73	0.70
(2.4) Train $CS_{TL}(EN)$, Test TL	0.64	**0.76**	**0.59**	**0.43**	**0.51**	**0.53**	**0.58**	0.35	**0.19**	0.29	**0.84**	**0.92**	**0.80**	**0.73**	**0.77**

Table 2: Comparison of pure-English training and code-switched training in zero-shot and few-shot settings. CS: Code-Switched. EN: English. TL: Target Language (DE, AR or TR). $CS_Y(X)$: Code-switch X with Y. Underlying models are MultiBERT with ML Focal loss.

	DE					AR					TR				
Setup	**F1-Macro**	**F1-Micro**	**EM-1**	**EM-2**	**EM-A**	**F1-Macro**	**F1-Micro**	**EM-1**	**EM-2**	**EM-A**	**F1-Macro**	**F1-Micro**	**EM-1**	**EM-2**	**EM-A**
Train: $EN \rightarrow TL$. Test: TL															
(3.1) MultiBERT	0.59	0.72	0.67	0.33	0.50	0.45	0.49	0.36	0.11	0.26	0.69	0.88	0.82	0.65	0.74
Train: EN. Test: $TL \rightarrow EN$															
(3.2) MultiBERT	0.65	0.75	0.72	0.42	0.58	0.50	0.54	0.42	0.10	0.29	0.59	0.84	0.71	0.57	0.64
(3.3) EngBERT Uncased	0.63	0.78	0.75	0.44	0.60	0.52	0.55	0.48	0.13	0.34	0.48	0.78	0.73	0.43	0.58
(3.4) EngBERT Cased	0.53	0.75	0.74	0.41	0.58	0.51	0.54	0.46	0.11	0.32	0.54	0.86	0.75	0.63	0.69
(3.5) Few-shot w/ the best among (3.2), (3.3), (3.4)	0.61	0.79	0.62	0.50	0.56	0.62	0.66	0.48	0.29	0.40	0.70	0.84	0.63	0.57	0.60

Table 3: Exploring the effect of translation between target languages and English (source) in both directions. Google Translate is used for translation. $X \rightarrow Y$: X translated to Y using Google Translate.

dent predictions for each model based on the scores for each frame after the last activation layer, and checks whether those comprise the first frame. We report this metric to demonstrate that by switching from a multi-class model to a multi-label one, we retain accuracy for the first frame while providing more predictive power with multiple labels.

Note that, to be able to accommodate multiple languages, we favor a multilingual language model. Results in Table 1 show that for our application there is only an insignificant drop in the predictive power from EngBERT to MultiBERT that are both based on multi-label Focal loss (ML Focal). Moreover, Focal loss results in higher accuracy in EM-2 while maintaining as high F1-scores to canonical BCE Loss. Considering the purposes of this paper and as well as the label cardinalities in other languages datasets, we favor ML Focal loss for the remainder of the models.

Google Translate, while being the practitioner's handy translation guide, is a state-of-the-art machine translation tool (Edunov et al., 2018). In Table 3, we explore the effect of the direction of translation to detect frames in German (DE), Arabic (AR) and Turkish (TR) headlines about U.S. gun violence. Note that in neither of the languages, a sufficient size of training data is available. To extend framing analysis to multiple languages, one needs to employ cross-lingual transfer learning.

Firstly, we translate the training set in English from GVFC, to target language $TL \in \{DE, AR, TR\}$, train MultiBERT with ML Focal loss and test on the TL. Secondly, we use the English training set as is and translate target test sets to English. This latter setup lets us to use EngBERT as well. We experiment with both cased and uncased models and observe that uncased performs better in DE and AR. Overall, we observe that translating test sets to English results in better performance which is intuitive as the model requires clarity in language during training the most. All models in Tables 3 and 2 use the same loss and MultiBERT experiments always use the cased version, following the authors' recommendation (Devlin et al., 2019).

Moreover, we use 40 target samples, translated to English, and include them in the training set to inquire few-shot performance on the task. Using 40 new samples from target language, we only train the best performers, primarily based on F1 scores, among (3.2), (3.3) and (3.4) as translating tests to English perform better overall.Thus the few-shot models for DE, AR and TR are models (3.3), (3.3) and (3.2), respectively whose results are given in line (3.5) of Table 3. Note that for some metrics in few-shot, the performance may drop as the few samples come from a different distribution.

Furthermore, in Table 2 we compare zero-shot and few-shot performances of MultiBERT when trained on original English versus on code-switched train sets. Both models use the same set of samples, the difference is in the former a headline in English where in the latter "important" words are switched with their TL translations. In a zero-shot setting, in terms of F1-macro and F1-micro scores, code-switched training (2.2) outperforms English training (2.1) significantly for all three languages. Considering the few-shot setting, we con-

tinue to observe a consistent improvement for all three languages, see (2.3, 2.4). Note that the comparisons we make are primarily based on F1-scores as the model's capability might shift from predicting single-label cases correctly to predicting more multi-labelled cases correctly. For instance in German, code-switched training improves in terms of F1-scores from zero-shot to few-shot but remains around the same in terms of EM-A. This is because model gets to predict multi-label cases (EM-2) better by a 8% margin, see (2.2), (2.4) in Table 2.

Notably, considering Tables 2 and 3 together, a simple word-to-word translation for as little as 358 words, improves frame detection performance drastically even to the level of complete translation of the test set to English using Google Translate. For Turkish, code-switched training beats complete translation of test set into English in few-shot setting, it results in a comparable performance in German and slightly worse predictions in Arabic. We attribute the overall low performance in Arabic to inter-rater reliability in annotation process and the lowest label cardinality among all languages.

6 Conclusion

In this work we present a novel code-switch model for the task of automatic cross-lingual news frame detection and show that it matches the performance of full translation if not overrides. Moreover, we leverage an existing dataset by making use of multiple labels, create benchmark test sets for three new languages and employ a variation of Focal loss to account for class imbalance in the data. In conclusion, we demonstrate how cross-lingual analysis of news-framing phenomena, while accounting for multiple frames per sample, could be informative and insightful in developing a global view surrounding gun violence issue in the U.S.

References

Claire Atkinson. 2019. Americans are crazy: Foreign journalists grapple with covering u.s. mass shootings.

Alexander Buhmann and Diana Ingenhoff. 2015. The 4d model of the country image: An integrative approach from the perspective of communication management. *International Communication Gazette*, 77(1):102–124.

Ruth DeFoster and Natashia Swalve. 2018. Guns, culture or mental health? framing mass shootings as a public health crisis. *Health communication*, 33(10):1211–1222.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

Robert M Entman. 2008. Theorizing mediated public diplomacy: The us case. *The International Journal of Press/Politics*, 13(2):87–102.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*.

Guy J Golan and Itai Himelboim. 2016. Can world system theory predict news flow on twitter? the case of government-sponsored broadcasting. *Information, Communication & Society*, 19(8):1150–1170.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Lei Guo and Maxwell McCombs. 2015. *The power of information networks: New directions for agenda setting*. Routledge, New York and London.

Holli A. Semetko Claes H. De Vreese, Jochen Peter. 2001. Framing politics at the launch of the euro: A cross-national comparative study of frames in the news. *Political communication*, 18(2):107–122.

Crimson Hexagon. 2018. ForSight social media analytics platform, Last accessed on November 1, 2018.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Joseph S Nye Jr. 2004. *Soft power: The means to success in world politics*. Public affairs.

Friederike Schultz, Jan Kleinnijenhuis, Dirk Oegema, Sonja Utz, and Wouter Van Atteveldt. 2012. Strategic framing in the bp crisis: A semantic network analysis of associative frames. *Public Relations Review*, 38(1):97–107.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Ł ukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2773–2781. Curran Associates, Inc.

Richard Wike, Bruce Stokes, Jacob Poushter, Laura Silver, Janell Fetterolf, and Kat Devlin. 2018. America's international image continues to suffer. *Pew Research Center, October*, 1.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label classification. *arXiv preprint arXiv:1909.04176*.

Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. 2018. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2):191–202.