Multi-Label and Multilingual News Framing Analysis

Afra Feyza Akyürek[1], **Lei Guo**[2,1], **Randa Elanwar**[3], **Prakash Ishwar**[4,1], **Margrit Betke**[1] and **Derry T. Wijaya**[1]

[1]Department of Computer Science, Boston University
[2]College of Communication, Boston University
[3]Electronics Research Institute, Egypt
[4]Department of Electrical and Computer Engineering, Boston University
{akyurek,guolei,relanwar,pi,betke,wijaya}@bu.edu

Abstract

News framing refers to the practice in which aspects of specific 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 context where there are few/no frame annotations in the target language. We propose a novel method that can leverage elementary resources consisting of a dictionary and few annotations to detect frames in the target 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 work 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 exciting insights on the relationship between different frames of the same problem 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 phenomenon is known as media framing, which is the process of selecting "some aspects of a perceived reality and [making] 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" (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 of 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 understudied 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 if not better than a model trained on the source language and tested on headlines that are translated from the target language to the source. Since our approach requires only simple resources – a dictionary and a few ($\leq$40) annotated examples in the target language

– it is handy for many languages. Moreover, considering the significant improvement gained over the zero-shot transfer, the proposed approach is much more reliable for languages without existing translation technologies or expert annotations.

Due to the subtle nature of framing, it is not uncommon for one news article to involve 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 Network Agenda Setting Model suggests that examining the interrelationships between media elements enables researchers to measure media effects in a more nuanced manner. Note that some frames appear more often than others. 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 by adapting focal loss (Lin et al., 2017) into our multi-label setting. Our multi-label approach allows for the examination of frame co-occurring, 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 revolve around military and economic influence but also largely depend 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 foreign publics (Entman, 2008; Golan and Himelboim, 2016). However, these efforts are not always successful. Editors of international news media serve as the gatekeepers to decisions which may lead to the framing of a given country contrary to how its government intends. In reporting news about a foreign country, news editors and reporters make conscious or unconscious choices to emphasize specific issues, or emphasize certain aspects of a given topic, 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 particular 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 review based on languages spoken in the research team rather than theoretical rationales. This language constraint becomes a more significant 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 greatly benefit the international communication research community.

In NLP, language models have been effectively fine-tuned or used in downstream tasks such as text classification (Dai and Le, 2015; Howard and Ruder, 2018; Radford et al., 2018). Further, the introduction of deep contextual language embedding such ELMO (Peters et al., 2018), which uses bi-directional LSTMs and BERT (Bi-directional Encoder Representations from Transformers) (Devlin et al., 2019), has been another milestone in this line of work. BERT is currently one of the state-of-the-art models in language modeling.

News framing was first brought to the attention of the computational linguistics community by the Media Frames Corpus (Card et al., 2015), which addresses three issues: immigration, tobacco, and

[1]Code and data are available at https://github.com/feyzaakyurek/newsframing

same-sex marriage. Field et al. (2018) analyzes the framing of the U.S. and agenda-setting in Russian news. Our work is similar to (Field et al., 2018) in terms of using nPMI to find essential words. Furthermore, our work advances previous research by leveraging a multilingual language model, facilitating transfer learning in news framing, and relying on parsimonious resources, that is, 50,000 lexical translations vs. ~350 in our case.

The current state-of-the-art model (Liu et al., 2019) for frame detection fine-tunes BERT on frame-annotated English news headlines with the standard multiclass focal loss objective (Lin et al., 2017). Their approach predicts only a single frame, which is insufficient given the multifaceted nature of news framing in which multiple frames often co-occur in the same headline. Indeed, more than a quarter of the Gun Violence Frame Corpus (GVFC) has more than one frame (Liu et al., 2019). In this work, we fine-tune MultiBERT to detect frames in multiple languages' headlines with our multi-label focal loss. Our approach can predict (and be evaluated on) multiple frames for each headline, which is a more complex task 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.

3 Dataset Creation

GVFC is a dataset of news articles from 21 major U.S. news organizations related to U.S. gun violence that contains news headlines and their domain-expert frame annotations (Liu et al., 2019). We extend GVFC to include headlines in other languages by 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 items 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 into German, Turkish, and Arabic respectively by native speakers of the languages. In curating the multilingual datasets, we used the same set of frames as in GVFC.

We then trained two native speaker coders for each language to apply the GVFC codebook protocol for identifying frames and then measured their intercoder reliability (ICR) in annotating a sample of 350, 200, and 210 German, Turkish, and Arabic news headlines, respectively. The coders achieve 92.6%, 98.5%, 78.1% agreement rates in identifying the first frame and 78.9%, 97.9%, 74.3% agreement rates for the second frame for German, Turkish, and Arabic samples. Additionally, Krippendorff's Alpha for the 1st frame and the 2nd frame are 0.89, 0.66; 0.90, 0.74, and 0.69, 0.26 for German, Turkish, and Arabic, respectively.

Once a minimum of 70% agreement was reached, one coder of each language continued to code more headlines. Annotation resulted in a total of 326, 100, and 388 non-duplicate headlines for German, Turkish, and Arabic. The 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, which is in English. As we can observe from the agreement rates, the Arabic data has a relatively weaker ICR, while the Turkish data has the best ICR. As high ICR values imply that two coders consistently categorized the content similarly, they signal a high validity of the coded results. In turn, this is reflected in the performance of our model as it performs the worst in Arabic (Section 5). Nonetheless, the quality of our curated data is substantially higher – the average of Krippendorff's alpha is 0.82 – than contemporaries such as MFC (which is only in English) with an average alpha of less than 0.6 (Card et al., 2015).

4 Model

In this work, we extend the current state-of-the-art model on the GVFC (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, a self-attention layer, and fully-connected layers before being passed to the upper

encoder block. The self-attention layer embodies three matrices called W^Q for the query, W^K for the key, and W^V for the value. Each of these matrices is of size $vocab\_size \times hidden\_size$, and 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. Therefore, the 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. Then it 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 initially inspired by the Cloze task (Taylor, 1953). The contextual representation of the `[CLS]` token encodes the syntactic and semantic constructs of the sequence, and one can fine-tune BERT for various down-stream tasks.

Fine-tuning BERT performs 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, MultiBERT, which is 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 languages. The multilingual pre-training and the utilization of sub-word tokenization allows MultiBERT to represent sequences from any of these 104 languages (Gu et al., 2018) and enables zero-shot classification on any of the languages (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 words possible.

4.1 Multi-label News Frame Detection

For frame detection purposes, we classify news articles into nine frame categories based on their headlines. Devlin et al. (2019) recommends using the embedding generated for the special token called `[CLS]`, which is padded to the beginning of every sentence. All tokens, including `[CLS]` are of length $H = 768$. The representation for `[CLS]` 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. Hence, the only parameters trained from scratch during fine-tuning are those of the classifier layer's, $W \in \mathbb{R}^{HxK}$. Finally, we use Sigmoid activations to obtain nine outputs, each between 0 and 1, which are interpreted as scores for nine classes. During inference, we use the threshold of 0.5 on these scores to binarize the output.

We fine-tune MultiBERT with two different losses: the standard Binary Cross-Entropy loss, and a multi-label variation of the weighted 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,

$$BCE(f) = -\frac{1}{|K|}\sum_{i=1}^{|K|}(y^{(i)} log(\hat{y}^{(i)}) + (1 - y^{(i)}) log(1 - \hat{y}^{(i)}))$$

where predictions are given by

$$\hat{\mathbf{y}} = [\hat{y}^{(1)}, \ldots, \hat{y}^{(|K|)}] = \frac{1}{(1 + exp(-f(\mathbf{x})))}$$

$\mathbf{y} = [y^{(1)}, \ldots, y^{(|K|)}]$ are the gold binary labels and f is BERT with classifier.

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 binary focal loss (Lin et al., 2017). As a reminder, the focal loss for a single sample $\mathbf{x}$ is defined as,

$$FL(f) = -\alpha(1 - p)^2 log(p)$$

where $p = (1-y)(1-\hat{y}) + y\hat{y}$ and $y \in \{0, 1\}$ is the true label, also $\hat{y} = 1/\left(1 + exp(-f(\mathbf{x}))\right) \in \mathbb{R}$, and α is the balancing factor, which is 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 f is the hypothesis e.g., neural network. In the multi-label case, we alter focal loss formulation such that y and $\hat{y}$ become $\mathbf{y} \in \{0,1\}^{|K|}$ and $\hat{\mathbf{y}} \in \mathbb{R}^{|K|}$. Moreover, for α we propose using

$$\alpha = \left[(\alpha_1^{(0)}, \alpha_1^{(1)}), \ldots, (\alpha_k^{(0)}, \alpha_k^{(1)})\right]$$

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 in *F1* scores and prevails it in multi-label accuracy score *EM-2* (Exact Match for two frames) by a significant 11% margin 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 regularization, 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. As our second baseline, we train nine binary MultiBERTs in a one-vs-rest manner.

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 English samples that have more than one correct class by a significant 11% margin, 62% vs. 51% in *EM-2*, while maintaining the same level of F-1 scores as 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 as in 2.1 and 2.3 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. In our few-shot models, we use extra 40 samples from the target language, i.e., DE, AR, or TR, and use the same training configurations as in the initial training, which we describe in Section 5.

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.

Furthermore, since the news framing task is fairly a keyword-driven phenomenon (Field et al., 2018), we developed a set of keywords that occur most frequently in a given frame. To this end, 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. We determine the set of important words based on nPMI by selecting the top 250 words for each frame – that also have nPMI greater than zero – resulting in 358 total words. We, then, use word-to-word translation to code-switch (CS) the English training set with the target language (TL) 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" having both frames "mental illness" and "race/ethnicity". We experiment with using the code-switched data for training in both zero-shot 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 is a way of adapting the model to the target language during training. 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.

	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	0.66	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	**0.53**	**0.72**	**0.64**	**0.39**	**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 (the source) in both directions. We use Google Translate for translation. $X \rightarrow Y$: X translated to Y using Google Translate.

5 Experiments and Results

As input to our models, we follow previous work and rely on news headlines rather than news story content, due to reasons described by Liu et al. (2019). To showcase the gains made on top of a multi-class approach by reformulating the problem as multi-label, we reproduce the method described by Liu et al. (2019) with both English BERT and MultiBERTs (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 we train for ten 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 BERT code repository[2]. As for losses, we experimented with Binary Cross-Entropy and multi-label Focal Loss, as described in Section 4.1.

For evaluation, we follow 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 number of frames (*EM-A*). In Table 1, we also report *Top-2* accuracy, which, for a given sample, computes the top two most confident 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 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 using 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, as well as the label cardinalities in other language datasets, we favor ML Focal loss for multilingual models.

While being a state-of-the-art machine translation tool, Google Translate is the practitioner's handy translation guide, (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 US gun violence. Note that in none of the languages is a sufficient size of news framing training data available; thus, to extend framing analysis to multiple languages, cross-lingual transfer learning is needed.

Firstly, we translate GVFC from English, 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 use EngBERT as well. We experiment with both cased and uncased models and observe that uncased performs better in *DE* and *AR*. Overall, we note that translating test sets to English results in better performance, which is intuitive as the model requires clarity in the language during training. All

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

models in Tables 3 and 2 use the same loss, and MultiBERT experiments always use the cased version, following the authors' recommendation.

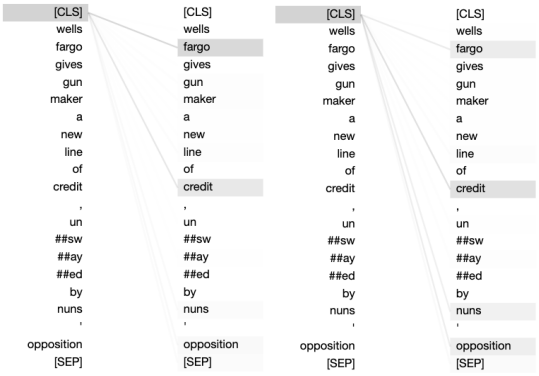
We use 40 target samples of target language, translated to English, and include them in the training set to study few-shot performance. We only train the best performers, primarily based on F1 scores, among (3.2), (3.3) and (3.4), namely the models (3.3), (3.3) and (3.2) for *DE, AR* and *TR*, respectively in (Table 3). For some of the metrics the few-shot performance may drop because the new samples come from a different distribution.

Furthermore, we compare zero-shot and few-shot performances of MultiBERT when trained on original English versus code-switched train sets in Table 2. Both models use the same set of samples; the difference is that in the former, the headlines are in English, whereas in the latter, "important" words are switched with their *TL* translations. In a zero-shot setting, code-switched training (2.2) outperforms English training (2.1) significantly for all three languages (F1-macro and F1-micro scores). Considering the few-shot setting, although the improvement gets smaller, the performance of code-switching is on par if not better 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-labeled cases correctly as well as between common and rare classes. In German, for instance, code-switched few-shot training improves in *F1*-scores from zero-shot but remains around the same in terms of *EM-A*. The reason for that is because the model predicts multi-label cases (*EM-2*) better by 4 percent points, 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 a complete translation of the test set to English. For Turkish, code-switched training beats full translation of the test set into English in a few-shot setting; it results in a comparable performance for German and slightly worse predictions for Arabic. We attribute the overall low performance for Arabic to the relatively small ICR in the annotation process.

6 Analysis

To visualize our multi-label model we use the visualization tool by Vig (2019) in Figure 1. In BERT,



(a) Multi-class Model (b) Multi-label Model

Figure 1: "Wells Fargo gives gun maker a new line of credit, unswayed by nuns' opposition" has *Economic Consequences* as the first frame and *Public Opinion* as the second.

every sequence is padded by a special classification token `[CLS]` from the beginning. Embedding generated for this token is used for classification into 9 classes. Figures 1a and 1b demonstrate the attentions of this token to other tokens in the sequence. Note that the given sample headline has indeed two frames i.e. "Economic Consequences" as the first and "Public Opinion" as the second. However, in a multiclass setup in which the model is configured to produce a single label, it learns to disregard the second frame "Public Opinion" while strongly attending the words "fargo" and "credit" related to for the theme of "Economic Consequences". On the contrary, a multi-label model correctly attends all words that are related to both frames i.e. "fargo", "credit", "nuns" and "opposition" and predicts "Economic Consequences" and "Public Opinion" correctly.

Another interesting observation is related to bias induced by translation. In German, the phrase "schärferes Waffenrecht" means "stricter gun regulation". However, Google Translate translates half of the headlines that include the expression as "stricter/sharper gun rights" which makes the model predict "Gun Rights" rather than "Gun Control" as the frame. A discrepancy like this is widely deceptive and jeopardizes the learning, whether it happens in the training or validation set. However, in code-switched training, one has better control over the translation, as one only translates a manageable number of words. We observe that code-switched training escapes this bias through correctly translated keywords "gun" and "laws" to German. Additionally, we find our models catching several annotation errors such as the headline in Turkish

Code-switch Technique	Unique Switched Words	Total Switched Words	F1-Macro	F1-Micro	EM-1	EM-2	EM-A
Zero-Shot (Train EN, Test DE)	0	0	0.48	0.66	0.47	0.31	0.39
Code-switch Omitted Words	387	2121	0.54	0.70	0.53	0.27	0.40
Code-switch nPMI Words	358	7522	0.53	**0.72**	0.64	**0.39**	**0.52**
Code-switch nPMI + Omitted Words	675	8129	**0.60**	0.70	**0.65**	0.29	0.47

Table 4: Code-switch analysis for German.

"Obama'dan LGBTI bireylerin gittiği bir kulüpte 49 kişiyi öldüren Orlando saldırganı hakkında açıklama" which translates as "Obama gave a statement about the Orlando shooter who killed 49 in an LGBTI club." is annotated as "Politics"." In contrast, the model predicts "Society/Culture" and "Politics", attending to "LGBTI" and "club".

6.1 Code-switching Analysis

In determining the words to code-switch from English to a target language, we mainly considered the metric called nPMI (Section 4.2), which essentially gives the most frequently-used words for each frame. In the English dataset (GVFC), we first list the top 250 words for a given frame based on their nPMI scores and take the union of these across frames, which resulted in a total of 358 case-sensitive words to be dictionary-translated into the target language.

In Table 4, we provide results obtained by using different code-switching methods that use no target language annotations. Note that, since nPMI is a frequency metric, code-switching with nPMI results in this set of words that includes not only frame-indicative words but also a lot of stop words and common words such as "a", "the", "he" or "are". An alternative method, which we called "omitted words" suggests determining important words by omitting a word from the headline and reapplying the trained classifier to the headline with the missing word (similar to Zhong et al. (2019); Ribeiro et al. (2016)). We then compute the drop in the probability as an importance measure for word x_j, $Importance(x_j) = p(y|x_1, \ldots, x_n) - p(y|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$ where y is a true label. The remaining procedure is similar to nPMI, as we determine the set of important words per frame, 45 of them this time, and combine those which resulted in 387 words. Note that this method results in a set of important words that are more disjointed across frames, which in turn makes the words more frame-specific. No common or stop words made it to the top 45 in any of the frames.

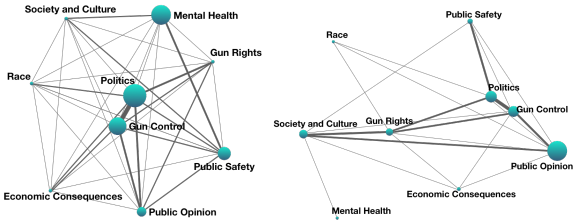
Despite resulting in more sophisticated words, using omitted words to code-switch resulted in more deficient if not on par scores as compared to nPMI – our primary way of doing code-switching. We argue that the reason for nPMI performing better is the much higher number of total words that get translated to the target language. In Table 4, note that using dictionary translations for only 358 unique words results in a total of 7522 words that are in the target language, which is more than 3.5 times what omitted words method yields. The increased amount of words that end up in the target language helped the MultiBERT classifier distinguish frames in the target language better. Note that in the last line of Table 4, including translations for the omitted words results in inconsistent improvement due to negligible size in the increase of the total words that get translated.

Our experiments show that for code-switching purposes, quantity might override quality which may suggest that for code-switching to be effective in multilingual transfer, translations of simpler words can outperform translations of the domain- and task-specific words, making the resources required to leverage knowledge from the source language to target language even more parsimonious.

6.2 Framing Network Analysis

The network visualization software Netdraw (Borgatti, 2002) was used to visualize the two frame networks depicted in Figure 6.2 based on the predictions generated on U.S. and German news articles from the year 2016 to 2018 by best performing models, i.e., uncased English BERT (Table 1) and code-switched model (Table 2) for English and German respectively. While each node represents a frame, each edge represents the number of times the two corresponding frames co-occurred in the news headline. The more central, the more connected the frame is with other frames. The node size was adjusted to reflect the relative frequency of news coverage of the given frame. That is, a frame with a larger node size more frequently occurs in the news coverage.

Several notable patterns emerge by comparing the frame networks in the U.S. and Germany. It appears that the U.S. media highly politicized the



(a) U.S. Frame Network (b) German Frame Network

Figure 2: Comparison of frame association networks in the U.S. and German news.

gun violence issue. The frame "politics" is not only the most salient but also the most central, closely connected with several other frames, reflecting, the sensationalism of the U.S. media landscape. The U.S. media tends to link all aspects of social reality to the political fight between the two parties, a pattern not followed in foreign media.

Another important finding is that while the U.S. media broadly framed the gun violence issue from the perspective of mental health, German media rarely mentions this aspect. Rather than blaming individual shooters, the German press paid more attention to U.S. public opinion manifesting as gun violence protests and the U.S. gun regulations. In other words, compared to the U.S.'s news coverage, foreign media tended to attribute the responsibility to the U.S. government.

In the German news coverage, the close association between the frame "society and culture" and "gun rights" is also noteworthy. Frequently linking the U.S.'s unique culture and people's rights to purchase guns in the news presents the U.S. as a "bizarre" place, which may also lead to a negative perception of the country among Germans.

In conclusion, the two frame networks illustrate how an issue can be framed differently in news media of different countries. Considering that the U.S. and Germany are close allies, it would be exciting to examine how countries with tense relations with the U.S. framed gun violence issues. A large-scale comparative framing study would allow a better understanding of the U.S. global image, which we propose as future work, and our multilingual and multi-label tool would make this type of analysis possible. In general, our approach is practical in looking at how media in different countries frame an international issue.

6.3 Future Work

We want to acknowledge two additional properties of a given headline, which neither this nor the previous works in news framing consider (Card et al., 2015; Liu et al., 2019; Field et al., 2018). First is *relevance*, although rarely, not all headlines that include the specified keywords in Section 3 are actually about U.S. gun violence. Second, an article may be about one particular incident or event related to gun violence, i.e., *episodic*, or it may focus on the issue of gun violence as an ongoing problem, i.e., *thematic*. Moreover, some of the episodic articles may not be tendential enough to have a particular frame. Existing works on framing only includes headlines that are both relevant and have frames, whereas, in reality, 48% of headlines about U.S. gun violence in GVFC do not have a particular frame. Media outlets outside of the U.S. have various rates of tendential articles about gun violence in the U.S. For instance, among the foreign languages we examined, German articles have the highest rate, with 90% of articles having at least one frame. Among Turkish articles that are "relevant" only 10% have a frame. In our evaluations, we only considered headlines that are relevant and have at least one frame. While stressing that determining the frame of an article is the most nuanced task in news framing, addressing the challenges mentioned above is still meaningful and constitutes future work.

7 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 news framing test sets for three new languages, and employ a variant of Focal Loss to account for class imbalance in the data. In conclusion, while accounting for multiple frames per sample, we demonstrate how a cross-lingual analysis of news framing is informative and insightful in developing a global view surrounding the gun violence problem in the U.S.

Acknowledgment This work is supported in part by the U.S. NSF grant 1838193 and DARPA HR001118S0044 (the LwLL program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA and the U.S. Government.

References

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

Stephen Borgatti. 2002. Netdraw: Graph visualization software. *Harvard: Analytic Technologies*.

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.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.

Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087.

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.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

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.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

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*, pages 1135–1144.

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.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

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*.

Ruiqi Zhong, Yanda Chen, Desmond Patton, Charlotte Selous, and Kathleen McKeown. 2019. Detecting and reducing bias in a high stakes domain. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4767–4777.