ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095597

USING EMOTION EMBEDDINGS TO TRANSFER KNOWLEDGE BETWEEN EMOTIONS, LANGUAGES, AND ANNOTATION FORMATS

Georgios Chochlakis[1,2] *Gireesh Mahajan*[3] *Sabyasachee Baruah*[1,2]
Keith Burghardt[2] *Kristina Lerman*[2] *Shrikanth Narayanan*[1,2]

[1] Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA 90089, USA
[2] Information Science Institute, University of Southern California, Marina del Rey, CA 90292, USA
[3] Microsoft Cognitive Services, Redmond, WA 98052, USA

ABSTRACT

The need for emotional inference from text continues to diversify as more and more disciplines integrate emotions into their theories and applications. These needs include inferring different emotion types, handling multiple languages, and different annotation formats. A shared model between different configurations would enable the sharing of knowledge and a decrease in training costs, and would simplify the process of deploying emotion recognition models in novel environments. In this work, we study how we can build a single model that can transition between these different configurations by leveraging multilingual models and *Demux*, a transformer-based model whose input includes the emotions of interest, enabling us to dynamically change the emotions predicted by the model. *Demux* also produces emotion embeddings, and performing operations on them allows us to transition to clusters of emotions by pooling the embeddings of each cluster. We show that *Demux* can simultaneously transfer knowledge in a zero-shot manner to a new language, to a novel annotation format and to unseen emotions. Code is available at `https://github.com/gchochla/Demux-MEmo`.[1]

***Index Terms*—** Multilingual emotion recognition, Zero-shot, Emotion clusters

1. INTRODUCTION

Human experience is permeated by emotions. They can guide our attention and influence our information consumption, beliefs, and our interactions [10, 26]. Deep learning has enabled us to extract affective constructs from natural language [7, 5], allowing emotion recognition from text at scale [13]. Nevertheless, the need for better performance across various metrics of interest still exists.

When inferring emotions from text, earlier approaches have utilized emotion lexicons [21]. These struggle in more realistic settings, because, for example, they do not handle context, like negation. On the other hand, while modern efforts relying on deep learning achieve better performance [7, 3], these data-driven models have to contend with a multitude of biases, such as annotation biases in the data used to train the models.

The needs for emotional inference from text have also diversified. First, the domain of interest can vary greatly between applications, ranging from everyday dialogues [17] to tweets [19]. Secondly, it is desirable for the models to be able to handle multiple languages [19], such as when studying perceptions and reactions to international news stories. Finally, the emotions of interest can differ between applications, and perhaps even the annotation scheme might not be similar. For example, in this work, we analyzed tweets annotated for clusters of emotions, where emotions that co-occur frequently were grouped, in contrast to single emotions in other settings. Hence, transfer learning is hindered by the mismatch.

In this work, our main goal is to achieve transfer of emotion recognition between annotation formats, emotions and languages. Our experimental design examines this step-by-step. First, we leverage pretrained multilingual language models [2, 9] to enable knowledge transfer between languages. We also use and extend *Demux* [7], a model that incorporates the labels in its input space to achieve the final classification. Emotions are then embedded in the same space as the language tokens. Emotion word embeddings can facilitate transfer between emotions, as shown in [7], so we study whether this can also be achieved in a zero-shot manner. Lastly, we examine how an extension of *Demux* to clusters can transfer knowledge between different annotation formats by directly performing operations on label embeddings [18]. Our contributions include the following:

- We show that multilingual emotion recognition models can be competitive with or even outperform monolingual baselines, and that knowledge can be transferred to new languages.
- We demonstrate that *Demux* can inherently transfer knowledge to emotions it has not been trained with.
- We illustrate that operations on the contextual emotion embeddings of *Demux* can successfully achieve transfer to novel annotation formats in a zero-shot manner. To the best of our knowledge, we are the first to study this setting.
- We show that *Demux* can be critical for flexible emotion recognition in a dynamic environment with ever-changing inference needs, such as the addition and subtraction of emotion types, changes in language, and alterations in the annotation format, e.g., the clustering of different emotions.

2. RELATED WORK

2.1. Emotion Recognition

Earlier works utilized Bag-of-Words algorithms driven by emotion lexicons. For instance, LIWC [21] is a lexicon that is widely used to perform word counting, while DDR [12] extends lexicon-based methods from word counting to computing similarities between words. More recently, deep learning has enabled more accurate extraction of emotion signals from text. Initial efforts have treated the task as single-label, and use a threshold to transform into the desired multi-label output [15]. LSTMs have been widely used for the task [11, 3] e.g., for SemEval 2018 Task 1 [19], where some also

[1]Funded in part by DARPA under contract HR001121C0168

used features from affective lexicons. More recently, Transformers [25] have dominated the field. *Demux* and MEmo [7], state of the art models in SemEval 2018 Task 1 E-c, prompt BERT-based [9] models in different ways, by including all emotions in the input or [MASK] tokens in language prompts, respectively. They also employ an intra-group correlation loss to further improve performance. Transformers have also been used with other architectures [27].

2.2. Multilingual Models & Emotion Recognition

Multilingual transformers attempt to model many languages simultaneously. Normalized sampling from each language is used so that low-resource languages are not severely hindered, which we also adopt. This is achieved, given $\alpha \in [0, 1]$, by transforming the frequency p_l of each language as $p'_l \leftarrow p_l^\alpha$ and renormalizing to create the new sampling distribution [16]. Note that as α decreases, the distribution becomes more balanced, achieving parity at $\alpha = 0$.

BERT [9] and XLM [16] require preprocessing of languages that do not use spaces to delimit words. Both are trained on 100 languages on Wikipedia, and use $\alpha = 0.7,\ 0.5$ respectively. XLM also uses language embeddings, and incorporates Translation Language Modeling (TLM) as a pretraining technique, which requires parallel data. XLM-R [8] handles all languages without preprocessing. It decreases α to 0.3 and switches to the CommonCrawl dataset, which has a more balanced language distribution. It also disposes of language embeddings and TLM. XLM-T [2] extends XLM-R by finetuning it on tweets to perform multilingual sentiment analysis, as does XLM-EMO [4] for emotion recognition on four emotions.

2.3. Zero-shot Emotion Recognition

Very few works explicitly study zero-shot emotion recognition in text with transformer-based models [28, 22]. They do so by formulating the problem as a *Natural Language Inference* problem, i.e., by creating a different prompt per emotion of interest, and classifying whether each prompt follows from the input sequence (entailment) or not (contradiction). This requires running the model once per emotion, creating a bottleneck for classification. Most similar to ours are earlier approaches that used semantic similarity with emotion word embeddings to classify in a zero-shot manner [24].

3. METHODOLOGY

We present the technical details of interest for *Demux* and our simple extension for it to handle clusters of emotions. Let $E = \{e_i : i \in [n]\}$ be the set of emotions and $C = \{C_i : i \in [m]\}$ be some clustering of E s.t. $n \geq m$, $\cup_{i\in[m]} C_i = [n]$ and $\cap_{i\in[m]} C_i = \emptyset$.

3.1. Demux

Let x be an input text sequence. *Demux* constructs $x' =$ "$e_1, e_2, \ldots$ or e_n?" and use a LM L with its corresponding tokenizer T:

$$\begin{aligned}\tilde{x} = T(x', x) = ([CLS], t_{1,1}, \ldots, t_{1,N_1}, \ldots, \\ t_{n,1}, \ldots, t_{n,N_n}, [SEP], x_1, \ldots, x_l),\end{aligned} \tag{1}$$

where x_i are the tokens from x, $t_{i,j}$ the j-th subtoken of e_i, and [SEP] and [CLS] are special tokens of T. $\tilde{x}$ is propagated through L to get $\hat{x} = L(\tilde{x})$, where $\hat{x}$ contains one output embedding corresponding to each input token. We denote the output embedding corresponding to $t_{i,j}$ as $\hat{t}_{i,j} \in \mathbb{R}^d$, where d is the feature dimension of L. Finally, *Demux* averages the embeddings of each emotion's subtokens, and predicts using a 2-layer neural network mapping embeddings to scalars, NN : $\mathbb{R}^d \to \mathbb{R}$, followed by sigmoid σ:

$$\forall i \in [n], \quad p(e_i|x) = \sigma(\text{NN}(\frac{\sum_{j=1}^{N_i} \hat{t}_{i,j}}{N_i})). \tag{2}$$

Notice that the same NN is applied to all emotions. For emotion clusters, we modify x' to contain all emotions from all clusters. After the forward pass through L, we instead aggregate across all emotions of a cluster instead of a single emotion and predict, for each cluster:

$$\forall i \in [m], \quad p(C_i|x) = \sigma(\text{NN}(\frac{\sum_{j\in C_i}\sum_{k=1}^{N_j} \hat{t}_{j,k}}{\sum_{j\in C_i} N_j})). \tag{3}$$

Moreover, when using multilingual models, we keep all emotions in English to retain the same prompt, x', across all languages.

3.2. Correlation-aware Regularization

To provide extra supervision to the model and enhance its correlation awareness between emotions, *Demux* includes a label-correlation regularization loss. This loss takes into account the ground-truth labels for each example in its formulation. Therefore, the emotions are split into two groups, the present and the absent emotions based on annotations y, $\mathcal{P}$ and $\mathcal{N}$ respectively. Intra-group relationships are regularized, meaning we only pick pairs of emotions when they are both in $\mathcal{P}$ or both in $\mathcal{N}$. The formulation is:

$$\begin{aligned}\mathcal{L}_{L,\text{intra}}(y, \hat{y}) = \frac{1}{2}\Bigg[\frac{1}{|\mathcal{N}|^2 - |\mathcal{N}|}\sum_{(i,j)\in\mathcal{N}^2}^{i>j} e^{\hat{y}_i + \hat{y}_j} + \\ \frac{1}{|\mathcal{P}|^2 - |\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}^2}^{i>j} e^{-\hat{y}_i - \hat{y}_j}\Bigg],\end{aligned} \tag{4}$$

where $\hat{y}$ is the prediction of the model and subscripts indicate indexing. In this manner, we decrease the distance of pairs of emotions when they have the same gold labels. The denominators simply average the terms. The final loss is a convex combination of the classification and the regularization loss, dictated by hyperparameter c:

$$\mathcal{L} = (1 - c)\mathcal{L}_{BCE} + c\mathcal{L}_{L,\text{intra}}. \tag{5}$$

4. EXPERIMENTS

4.1. Datasets

We use the publicly available SemEval 2018 Task 1 E-c [19], and private data containing tweets from the French elections of 2017.

SemEval 2018 Task 1 E-c (SemEval E-c) contains tweets annotated for 11 emotions in a multilabel setting, namely *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise,* and *trust*, in *English, Arabic,* and *Spanish*. The cardinalities are 6838 training, 886 development and 3259 testing for English, 2278 training, 585 development and 1518 testing for Arabic, and 3561 training, 679 development and 2854 testing for Spanish. We have also machine-translated the English subset of the tweets into French.

The French election dataset contains mostly French tweets annotated for 10 clusters of emotions in a multilabel setting, namely *admiration-love, amusement-sarcasm, anger-hate-contempt-disgust, embarrassment-guilt-shame-sadness, fear-pessimism, joy-happiness, optimism-hope, pride (including national pride),* any *other positive*,

Emotion Cluster	Support in FrE-A	Support in FrE-B
Admiration	589 (14.4%)	51 (1.1%)
Sarcasm	395 (9.6%)	233 (5.1%)
Anger	595 (14.5%)	279 (6.1%)
Embarrassment	182 (4.4%)	56 (1.2%)
Fear	271 (6.6%)	116 (2.5%)
Joy	169 (4.1%)	40 (0.9%)
Optimism	501 (12.2%)	350 (7.7%)
Pride	539 (13.1%)	92 (2%)
Positive-other	632 (15.4%)	1268 (27.7%)
Negative-other	542 (13.2%)	1426 (31.2%)
All tweets	4102 (100%)	4574 (100%)

Table 1. Support per emotion cluster for our French election datasets.

and any *other negative* emotions. The formulation dictates that if one emotion in the cluster is present, the cluster is considered present. Only cluster-wise annotations are requested. We have two separate subsets, collected by third parties using keywords related to prominent politicians, agendas and concerns of the French elections, and annotated independently also by third parties. The first ("FrE-A") contains 4102 tweets that we randomly split into training, development and test sets with ratios 8:1:1. The second subset ("FrE-B") contains 4574 tweet that we also randomly split with ratios 8:1:1. In Table 1, we present the support for each class. Annotations for FrE-B are a lot sparser.

4.2. Implementation Details

We use Python (v3.7.4), PyTorch (v1.11.0) and the *transformers* library (v4.19.2). We use NVIDIA GeForce GTX 1080 Ti. Given the difference in writing formality on Twitter, we use Twitter-based in addition to general-purpose models. For the former, XLM-T [2] is our multilingual model, BERTweet [20] for English, RoBERTuito [23] for Spanish, AraBERT [1] for Arabic, and BERTweetFR [14] for French. For the latter, we use multilingual BERT [9] as our multilingual model, BERT for English and Arabic, and BETO [6] for Spanish. We retain the hyperparameters used in [7], such as the learning rate and its warmup, early stopping, batch size, the convex combination coefficient c (α in [7]), and the text preprocessor. During multilingual finetuning on SemEval E-c, we sample from different languages with equal probability ($\alpha = 0$ in aforementioned normalization in Section 2.2). We evaluate using the F1 score for individual emotions, and micro F1, macro F1 and Jaccard score (JS) otherwise. We also used Micro F1 for early stopping in FrE-A because we found JS to be unreliable due to the label distribution.

4.3. Knowledge Transfer between Languages

Multilingual Training and Evaluation First, we examine how feasible and competitive it is to use multilingual emotion recognition models trained and evaluated on a mixture of languages. To establish this, we first consider training monolingual and multilingual models, pretrained either on tweets or general-purpose text. For multilingual models, we also consider how training and/or evaluating on multiple languages fares with training and/or evaluating on a single language. For example, we compare performance on the Arabic subset when training and performing early stopping solely on Arabic, with the performance when training on all languages simultaneously, but still performing early stopping on Arabic, and other combinations.

We find that training and evaluating on all languages in SemEval E-c has either only slightly negative or strongly positive influence on accuracy for higher-resource languages. Our dev set results are presented in Table 2. It becomes immediately obvious that models trained on Twitter comfortably outperform their general-purpose alternatives (first and the second row, and third and final row). We also observe that using a dev set of multiple languages not only does not hurt performance, but actually achieves positive knowledge transfer for Spanish, achieving the best JS overall. The monolingual alternatives perform favorably in English and Arabic, with the increase in the former being relatively minor. Overall, since our ultimate goal is to transfer to French tweets, and given the competent or superior performance for Latin-based languages, we do adopt the multilingual model trained and evaluated on multilingual data for pretraining.

Transfer to New Languages In trying to establish how emotion recognition knowledge is transferred to new languages, we conducted experiments with one SemEval E-c language left out during training. Results are shown in Table 3. We notice a drop in performance, with all models performing roughly equivalently across all metrics on the new language notwithstanding original performance. In detail, all metrics drop by around 25% in Arabic, $< 22\%$ for Spanish, while the drop in English ranges from 18% to 32%. Nonetheless, emotion recognition in the new language occurs at a competent level, picking up clear signals despite the noise from the language switch, rendering multilingual emotion recognition models capable of being used with new languages.

4.4. Knowledge Transfer to New Emotions

We also assess *Demux*'s ability to perform zero-shot emotion recognition by excluding emotions from SemEval E-c, one at a time. We can predict unseen emotions since the final classifier maps embeddings to probabilities, agnostic to specific emotions. We choose *anger*, *joy*, *pessimism*, and *trust* to capture the change in accuracy across a wide spectrum of performance levels. In particular, *joy* and *trust* are the highest and lowest performing emotions, whereas *anger* and *pessimism* have relatively high and low scores, respectively. Results are presented in Table 4. We notice a decrease in performance. However, the model can still predict these unseen emotions, especially those with which the original model was competent at.

To remedy the decrease in performance, we experimented with freezing word embeddings in an effort to retain the relationships between emotions. We examine two alternatives, freezing the word embedding layer altogether, and freezing only the emotion word embeddings (including the novel emotions). Results are also presented in Table 4. We find this decreases performance for the model, indicating it already captured relationships across emotions in its input.

4.5. Knowledge Transfer to New Annotation Format

Finally, we evaluate if the model can successfully transfer knowledge to a new annotation format. In particular, we transfer from SemEval E-c to FrE-a and FrE-B. The former contains tweets in English, Spanish and Arabic, and annotations for emotions. The latter contain French tweets annotated for clusters. Additionally, the emotions of the latter are not a subset of the former. Therefore, we expect to observe compounding effects from the multiple changes in the setting. To address the language switch, we also use the French translations to train BERTweetFr as a monolingual parallel to XLM-T.

We also show two simple baselines, one that predicts the most frequent label for each emotion, and another that predicts uniformly randomly. Moreover, our experiments include an alternative to aver-

Setting				JS En	JS Es	JS Ar
	Monolingual models			61.0±0.3	53.2±0.5	55.1±0.3
Twitter-based	Monolingual models			**62.0**±0.4	55.6±0.3	**61.6**±0.5
	Multilingual models	w/ Multilingual Training	& Evaluation	58.4±1.0	50.3±0.4	49.2±0.0
Twitter-based	Multilingual models			61.6±0.1	56.8±1.0	56.5±0.6
Twitter-based	Multilingual models	w/ Multilingual Training		61.3±0.4	56.5±0.6	57.9±1.0
Twitter-based	Multilingual models	w/ Multilingual Training	& Evaluation	61.3±0.2	**58.0**±0.7	57.7±0.4

Table 2. Comparing Jaccard scores in *SemEval 2018 Task 1 E-c*. The variables we consider are: monolingual or multilingual models, Twitter-based or general-purpose models, monolingual or multilingual training, where the training set is comprised of one or a mixture of all languages, and monolingual or multilingual evaluation, where the evaluation set is comprised of one or a mixture of all languages.

Train langs			Eval langs	SemEval 2018 Task 1 E-c Mic-F1	Mac-F1	JS
En	Es		Ar	55.8±0.9	42.4±2.7	43.3±1.3
En	Es	Ar	Ar	70.2±0.8	57.4±1.4	57.9±1.0
En		Ar	Es	55.9±0.7	45.1±1.6	44.0±0.9
En	Es	Ar	Es	65.1±0.6	54.7±0.8	56.5±0.6
	Es	Ar	En	54.6±0.6	47.0±1.3	41.9±0.7
En	Es	Ar	En	72.3±0.3	57.6±2.0	61.3±0.4

Table 3. Leave-one-language-out experiments.

ZS	Frozen	SemEval 2018 Task 1 E-c F1 Anger	Joy	Pessimism	Trust
✗	-	78.6±0.3	86.0±0.3	40.5±1.5	9.9±3.9
✓	-	42.5±11.8	57.8±3.2	11.7±0.9	3.6±3.5
✓	Words	25.1±18.2	51.5±7.3	15.1±12.9	1.7±2.0
✓	Emos	21.5±20.8	36.2±14.3	4.3±4.3	3.7±4.6

Table 4. Zero-shot performance of unseen emotions.

aging emotion embeddings with *Demux*, where we instead select the maximum predicted probability across emotions in a cluster.

Results are presented in Table 5. We consider three different settings for each model and each dataset in order to study how training on SemEval E-c (*Pretrained*) and training on the French election datasets (*Finetuned*) affects performance. In the first row per model and dataset, we see the zero-shot performance on the dataset. The second shows the performance of a model only trained on the dataset. The last row shows performance from a model first trained on SemEval E-c and then on the corresponding dataset.

For FrE-A, zero-shot performance of XLM-T is close to either finetuned performance, indicating the model can successfully transfer despite the complete change of environment. Performance increases for both when training on French data, and when SemEval E-c is also included. The monolingual model performs favorably only when fully supervised, which indicates that translations are not ideal for zero-shot transfer, and overall proves the ability of multilingual models to transfer knowledge to a new language.

Performance is evidently better in FrE-B. Again, XLM-T performs better than random and favorably to BERTweetFr in the zero-shot setting. Finetuning improvements are significant for both models. Lastly, the alternative pooling performs worse than our proposed one. This speaks to the subjectivity of the annotations, allowing operations on emotion embeddings but not the predictions themselves.

		Pretrained	Finetuned	Mic-F1	Mac-F1	JS
				FrE-A		
		Most frequent		0	0	24.9
		Uni. Random		17.6±0.4	18.0±0.4	10.0±0.2
XLM-T		✓	✗	21.0±1.2	15.8±2.0	26.6±2.2
XLM-T		✗	✓	28.9±1.4	25.7±0.7	**29.8**±1.7
XLM-T		✓	✓	**30.5**±2.5	**26.5**±5.2	29.4±2.3
BERT TweetFr		✓	✗	24.8±0.7	18.5±1.4	15.2±0.6
BERT TweetFr		✗	✓	27.0±12.5	24.9±12.8	29.6±2.6
BERT TweetFr		✓	✓	**34.3**±1.1	**31.4**±1.7	**31.8**±1.0
				FrE-B		
		Most frequent		0	0	35.0
		Uni. Random		15.0±0.7	12.3±0.7	8.3±0.4
XLM-T		✓	✗	18.3±3.7	11.5±1.9	29.4±4.6
XLM-T		✗	✓	55.3±6.5	19.3±9.9	55.3±6.5
XLM-T		✓	✓	**62.7**±0.9	**44.4**±2.8	**59.3**±0.8
XLM-T	*	✓	✗	15.9±2.6	11.3±1.0	14.8±3.3
XLM-T	*	✓	✓	61.9±1.2	39.6±9.6	58.4±0.8
BERT TweetFr		✓	✗	27.9±1.9	15.9±1.1	15.3±1.4
BERT TweetFr		✗	✓	63.8±1.4	29.7±7.6	60.2±1.7
BERT TweetFr		✓	✓	**66.9**±0.5	**40.3**±6.0	**63.2**±1.3

Table 5. Performance in French election data when the models pretrain on *SemEval E-c* and/or finetune on the corresponding dataset. *: maximum probability instead of pooling embeddings in Demux.

5. CONCLUSION

In this work, we study how to transfer emotion recognition knowledge to different languages, different emotions, and different annotation formats. We find that multilingual models have the capacity to transfer that kind of knowledge sufficiently well. In order to transfer knowledge between emotions, we leverage *Demux*'s transferability between emotions through word-level associations. We see that the model also inherently performs zero-shot emotion recognition without the need for further changes. Finally, we modify *Demux* to perform aggregation operations on its label embeddings, and show this can transfer knowledge to novel annotation formats, such as clusters of emotions, even in conjunction with the presence of novel emotions and in a different language. We show that multilingual models pretrained on other languages perform favorably in the zero-shot setting to native models pretrained on machine translations.

6. REFERENCES

[1] ANTOUN, W., BALY, F., AND HAJJ, H. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020* (2020), p. 9.

[2] BARBIERI, F., ESPINOSA-ANKE, L., AND CAMACHO-COLLADOS, J. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *Proceedings of the LREC, Marseille, France* (2022), 20–25.

[3] BAZIOTIS, C., ATHANASIOU, N., CHRONOPOULOU, A., KOLOVOU, A., PARASKEVOPOULOS, G., ELLINAS, N., NARAYANAN, S., AND POTAMIANOS, A. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658* (2018).

[4] BIANCHI, F., NOZZA, D., AND HOVY, D. Xlm-emo: Multilingual emotion prediction in social media text. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis* (2022), pp. 195–203.

[5] CALVO, R. A., D'MELLO, S., GRATCH, J. M., AND KAPPAS, A. *The Oxford handbook of affective computing*. Oxford Library of Psychology, 2015.

[6] CAÑETE, J., CHAPERON, G., FUENTES, R., HO, J.-H., KANG, H., AND PÉREZ, J. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020* (2020).

[7] CHOCHLAKIS, G., MAHAJAN, G., BARUAH, S., BURGHARDT, K., LERMAN, K., AND NARAYANAN, S. Leveraging label correlations in a multi-label setting: A case study in emotion. *arXiv preprint arXiv:2210.15842* (2022).

[8] CONNEAU, A., KHANDELWAL, K., GOYAL, N., CHAUDHARY, V., WENZEK, G., GUZMÁN, F., GRAVE, E., OTT, M., ZETTLEMOYER, L., AND STOYANOV, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[9] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] DUKES, D., ABRAMS, K., ADOLPHS, R., AHMED, M. E., BEATTY, A., BERRIDGE, K. C., BROOMHALL, S., BROSCH, T., CAMPOS, J. J., CLAY, Z., ET AL. The rise of affectivism. *Nature human behaviour 5*, 7 (2021), 816–820.

[11] FELBO, B., MISLOVE, A., SØGAARD, A., RAHWAN, I., AND LEHMANN, S. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017).

[12] GARTEN, J., HOOVER, J., JOHNSON, K. M., BOGHRATI, R., ISKIWITCH, C., AND DEHGHANI, M. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods 50*, 1 (2018), 344–361.

[13] GUO, S., BURGHARDT, K., RAO, A., AND LERMAN, K. Emotion regulation and dynamics of moral concerns during the early covid-19 pandemic. *arXiv preprint arXiv:2203.03608* (2022).

[14] GUO, Y., RENNARD, V., XYPOLOPOULOS, C., AND VAZIRGIANNIS, M. Bertweetfr: Domain adaptation of pre-trained language models for french tweets. *arXiv preprint arXiv:2109.10234* (2021).

[15] HE, H., AND XIA, R. Joint binary neural network for multi-label learning with applications to emotion classification. In *CCF International Conference on Natural Language Processing and Chinese Computing* (2018), Springer, pp. 250–259.

[16] LAMPLE, G., AND CONNEAU, A. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291* (2019).

[17] LI, Y., SU, H., SHEN, X., LI, W., CAO, Z., AND NIU, S. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957* (2017).

[18] MIKOLOV, T., YIH, W.-T., AND ZWEIG, G. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (2013), pp. 746–751.

[19] MOHAMMAD, S., BRAVO-MARQUEZ, F., SALAMEH, M., AND KIRITCHENKO, S. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation* (2018), pp. 1–17.

[20] NGUYEN, D. Q., VU, T., AND NGUYEN, A. T. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200* (2020).

[21] PENNEBAKER, J. W., FRANCIS, M. E., AND BOOTH, R. J. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates 71*, 2001 (2001), 2001.

[22] PLAZA-DEL ARCO, F. M., MARTÍN-VALDIVIA, M.-T., AND KLINGER, R. Natural language inference prompts for zero-shot emotion classification in text across corpora. *arXiv preprint arXiv:2209.06701* (2022).

[23] PÉREZ, J. M., FURMAN, D. A., ALEMANY, L. A., AND LUQUE, F. Robertuito: a pre-trained language model for social media text in spanish, 2021.

[24] SAPPADLA, P. V., NAM, J., MENCÍA, E. L., AND FÜRNKRANZ, J. Using semantic similarity for multi-label zero-shot classification of text documents. In *ESANN* (2016).

[25] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

[26] WAHL-JORGENSEN, K. *Emotions, media and politics*. John Wiley & Sons, 2019.

[27] XU, P., LIU, Z., WINATA, G. I., LIN, Z., AND FUNG, P. Emograph: Capturing emotion correlations using graph networks. *arXiv preprint arXiv:2008.09378* (2020).

[28] YIN, W., HAY, J., AND ROTH, D. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 3914–3923.