Multilingual Emotion Recognition: Discovering the Variations of Lexical Semantics between Languages

Xulang Zhang, Rui Mao, Erik Cambria
School of Computer Science and Engineering
Nanyang Technological University, Singapore
{xulang.zhang,rui.mao,cambria}@ntu.edu.sg;

***Abstract*—The task of multilingual emotion recognition holds significant importance in cross-cultural communication and data mining. While prior research has concentrated on enhancing classification accuracy using state-of-the-art techniques, it has often overlooked a crucial linguistic aspect—the semantic disparities across different languages. This study aims to address this gap by introducing a novel method to identify lexical semantic variations in diverse languages. The detected semantic variation features are subsequently injected into a multilingual emotion recognition model to enhance its performance within a target language. Notably, existing multilingual pre-trained language models are likely biased toward English word meanings, leading to inaccurate emotion predictions in other languages due to the misinterpretation of semantics. Our proposed semantic variation injection method tackles this limitation, resulting in improved accuracy. These findings contribute to the ongoing development of robust and culturally sensitive emotion recognition systems, offering valuable insights for both the linguistics and computational linguistics communities engaged in multilingual research.**

***Index Terms*—multilingual emotion recognition, semantics, computational linguistics**

I. Introduction

There is a growing consensus in the field of psychology and linguistics that emotions can vary systematically in their meaning and experience across cultures and languages [1]–[8]. However, concerns have been raised that the prevalent Multilingual Language Models (XLMs) are not well-versed in representing the cultural variations of emotions [9]. In contrast, XLMs exhibit a proclivity toward anglocentrism, displaying a bias favoring Western norms and preferences rooted in the English language. The work of Havaldar et al. [9] underscores the inadequacy of XLMs in acquiring culturally relevant nuances of emotion. Mao et al. [10] demonstrate the limitations of state-of-the-art machine translation systems, particularly in their propensity to misinterpret English metaphors by adhering strictly to the literal meanings of English expressions when translating into a target language. These systems encounter difficulties in thoroughly unraveling the culturally nuanced and intricate semantic variations inherent in multilingual computation. It is important to address these issues to manifest human-centric values in the domain of Artificial Intelligence [11]. By examining cases of colexification of emotion concepts, i.e., instances where multiple emotions are co-expressed by the same word in a language, Jackson et al. [6] find that the interpretation of associations between emotion concepts varies across different geographical language groups.

For instance, there are more instances of colexification between envy and hate in Austroasiatic languages, while envy and fear are more often colexified in Indo-European languages, meaning that people in the former language group more likely experience envy as a similar emotion to hate, whereas people in the latter group understand it to be more similar to fear. This finding conceptually explains the different interpretations of emotional lexicons in different cultures and languages. We define such differences in semantics as semantic variations. The most extensive multilingual emotion recognition focused on speech [12]–[14]. The very limited works in textual multilingual emotion recognition aimed at developing state-of-the-art algorithms to improve the fitness of data and labels [15]–[17], while overlooking the phenomenon of semantic variations between languages. Given the wide applications of pre-trained language models in diverse tasks, we are motivated to address the existing gap by incorporating semantic variation features into XLMs to enhance multilingual emotion recognition from textual content.

We propose a task-agnostic approach for identifying semantic variations. Subsequently, these identified semantic variation features are integrated with the embeddings of an XLM to alleviate the concern that the semantic representations of non-English languages in XLMs may exhibit bias towards the semantics of their English counterparts. To maintain experimental control over other variables, our study is structured around backbone models, allowing for a performance comparison between instances with and without the injection of semantic variation.

We examine the semantic variation injection method across textual emotion recognition datasets in three languages, namely Spanish, Japanese, and Chinese, utilizing two XMLs. Our approach yields average enhancements of the backbone XLMs by Macro F1 scores of 0.87%, 1.27%, and 1.16% for Spanish, Japanese, and Chinese emotion recognition, respectively. Notably, the observed improvements are more pronounced in Asian languages compared to Spanish, underscoring the greater cultural distinctions from English. The discernible enhancements extend to major emotional classes, highlighting the pervasive impact of semantic variations on diverse emotion interpretations. We further conduct human evaluation to validate the efficacy of the generated semantic variation feature in measuring the degree of lexical semantic variations.

The outcomes prove that our method can accurately identify words exhibiting heightened semantic variations between English and a given target language. The contributions of this work are twofold: (1) We propose a task-agnostic method to inject important linguistic intuitions into multilingual computation; (2) We evaluate the proposed method in emotion recognition tasks across two XLMs and three datasets in Spanish, Japanese, and Chinese.

II. Related Work

A. Semantic Variations

Semantic variation has long been a topic of interest for linguistics, as it provides a peephole into what conceptualizations are universal to human cognition, and what are cultural. Mono-lingual semantic variations were typically explored through the lens of pragmatics, focusing on phenomena such as metaphors [18], [19], and sarcasm [20]. Cross-lingual polysemy and primitive analyses have been conducted for different concept groups, i.e., color metaphors [21], body parts [22], emotions [5], [23], and natural entities [24]. The consensus is that, while some universal patterns could be found, concepts like colors and emotions are at variance among different languages. Early works in computational linguistics have been taking interest in exploring the variations between perceptions of emotions in different languages by examining the semantic space. Romney et al. [25] compare 15 emotion terms in English and Japanese in a semantic space constructed using hand-crafted features. They observe greater distances between the English and Japanese counterparts of three emotion concepts, namely shame, anxious, and bored, indicating a notable difference in emotion perceptions from the two cultures. Jackson et al. [6] aggregate colexifications of concepts from over 2000 languages into a database, and employ a random walk probability procedure to generate colexification networks. Examination of emotion concepts in such networks indicates that emotion concepts show significantly semantic variability among geographically distant language families. With the advancement of neural networks, researchers can more easily quantify semantic distances between words on a much larger scale with the help of word embedding space. However, concerns also arise regarding whether such embedding spaces can capture cultural differences. Havaldar et al. [9] hypothesize that implicit and explicit alignment methods in training XLMs have the undesirable effect of anchoring emotion embeddings to English. By measuring the distances among emotion embeddings of non-English languages in the respective embedding space of monolingual Roberta [26] and XLM-Roberta [27], they demonstrate thatXLMs fail to preserve the embedding space of monolingual non-English language models.

B. Multilingual Emotion Recognition

English emotion recognition from text is a widely studied task [28]–[31]. These works leveraged advanced language models and knowledge bases yielding outstanding performance on monolingual emotion recognition. In the domain of multilingual emotion recognition, the most extensive research focused on emotion detection on speech [12]–[14]. This emphasis is attributed to the distinctive features inherent to speech, including intonation, stress, and speaking speed, which serve as reflective indicators of emotions beyond the confines of textual content. In contrast, the body of research addressing multilingual emotion recognition through text is more limited. Jain et al. [15] proposed a framework for emotion extraction from multilingual text data using the ensemble of machine learning algorithms, e.g., Naïve Bayes and SVM. Becker et al. [16] conducted a comparative analysis to examine the dependability and efficacy of multilingual sentiment analysis in the context of emotion classification through a sequence of experiments. The experimental framework scrutinized divergences in both the original text language and its translated counterparts to elucidate the effects of various languages on the accuracy of emotion classification. Ghosh et al. [17] introduced a multitask learning framework designed to analyze emotions and sentiments in Hindi-English code-mixed texts. This framework was constructed upon a cross-lingual embedding-based Transformer model, using transfer learning from high-resource mono-lingual languages to their low-resource code-mixed counterparts.

However, to the best of our knowledge, the above works did not develop algorithms to address the issue of semantic variants between different languages in the domain of textual emotion recognition.

III. Methodology

A. Preliminary

To alleviate the anglocentric anchoring of XLM, we aim to inject the semantic variations between English and a target language into the embeddings of multilingual LM. Introducing semantic variation features through injection is anticipated to alleviate any semantic biases present in a target language, particularly those aligned with English semantics. Consequently, this is expected to result in improved accuracy in the multilingual emotion recognition tasks. Given that our approach for identifying semantic variations is task-agnostic, it holds the potential for application in various tasks beyond emotion recognition.

To detect such semantic variations, we first train a rotation matrix that maps the semantic space of English embeddings to that of the target language, based on selected anchor words with higher concreteness levels, where the concreteness levels of words are quantified via word concreteness score dictionaries [32], [33]. Our hypothesis is that:

> *In contrast to abstract concepts, the semantics of concrete concepts are expected to exhibit greater similarity across different languages, resulting in fewer semantic variants for concrete concepts.*

Such a hypothesis can be justified by the definition of concrete concepts and neurolinguistic findings [34]. Concrete concepts generally denote specific objects or entities perceptible through the senses, e.g., animals, tools, and natural phenomena. These concepts are typically linked to tangible, physical entities that can be directly perceived or observed. In the context of semantic memory, concrete concepts are distinguished by their connection to perceptual features and are usually acquired through direct sensory experiences. The embodied cognition [35] in the physical world enhances the interpretation of concrete concepts across languages, resulting in fewer semantic variants. This is attributed to the minimal differences in physical environments among different language regions. Kousta et al. [36] suggest that abstract concepts are more emotionally charged than concrete ones, which gives the former a residual advantage when imageability and contextual availability are controlled. Thus, we focus on mitigating cross-lingual semantic variations for abstract concepts.

The selected concrete words are defined as anchor words, which are used for training the rotation matrix. The intuition is that the anchor words are likely semantically consistent across different languages. Thus, the rotation matrix trained upon the anchor words can form a coordinate system, mapping the semantic spaces between two languages. Then, we use the rotation matrix to map the English semantic space to another target language. If a word in the target language's vector space notably deviates from the post-rotated position of its corresponding English word as translated by the rotation matrix, it suggests a semantic variation between the English word and its counterpart in the target language. This incongruity stems from the anticipation that, grounded in the coordinate system defined by concrete concepts, words in different languages are supposed to be aligned in vector space. The observed deviation indicates that the meaning of a word in the target language diverges from the meaning derived through the transformation of its English counterpart.

Figure 1 is a graphical illustration of detecting the semantic variations via a rotation matrix. ◯ and △ are an English word and its counterpart in a target language, respectively. They are supposed to have similar spatial relations to other anchor words (the + and × in the figure) if they have the same meaning in different languages. However, the rotated English word and the corresponding word in the target language are distant from each other in the semantic space, indicating the occurrence of semantic variation. Such a linguistic phenomenon can be found in many languages. For example, "平淡的" is supposed to be the counterpart of "plain", which can sometimes carry a neutral or even positive connotation, implying simplicity and tranquility in Chinese. In Chinese culture, a simple and tranquil life is what many people yearn for. However, in English, "plain" may lean more towards the absence of excitement or embellishment. Taking the connotation of "plain" to explain the meaning of "平淡的" can cause affective interpretation errors.

We can compute the semantic variations of all words using the obtained rotation matrix. In the following emotion recognition task, the semantic variations are used as features to be injected into the XLM. Upon semantic variation feature injection, deep neural networks learn from the adjusted semantic representations in a target language to predict emotional states of given input text, e.g., "joy", "anger", and more.

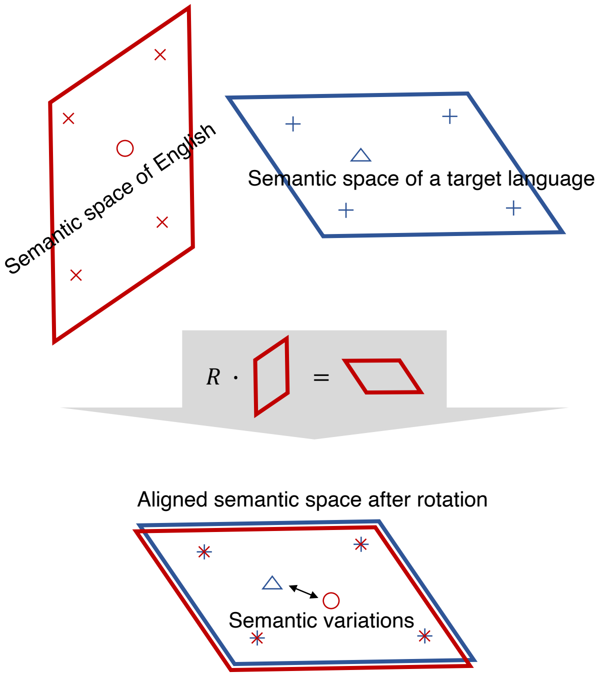
Semantic space of English
Semantic space of a target language
R ·
=
Aligned semantic space after rotation
Semantic variations

Fig. 1. A graphical illustration of detecting the semantic variations via a rotation matrix. + and × denote anchor words from English (red) and a target language (blue), respectively. R denotes a well-trained rotation matrix. ◯, △ and $\longleftrightarrow$ denote an English word, its counterpart in a target language, and their semantic variation, respectively.

B. Anchor Word Selection

To compute the rotation matrix, we compiled a list of concrete concepts as *(English, target language)* word pairs, using two concreteness dictionaries, namely the MRC database [32] and the Concept Abstractness dataset [33]. The MRC database rates the concreteness of 4,292 English words in the range of 100 to 700. We applied min-max normalization to the concreteness scores and selected the words with scores higher than 0.5. We find their best counterparts in the target language by using the word2word lexicon [37]. The Concept Abstractness dataset rates 300K Wikipedia concepts by their degrees of abstractness ranging from 0 to 1. We select the English unigram concepts with scores lower than 0.5, and use the corresponding Wikipedia titles in the target language as their counterparts. We form an anchor word list, containing word pairs in English and a target language.

C. Rotation Matrix Construction

Suppose a word pair in our anchor list is (w_e, w_t), where w_e is an anchor word in English (e); w_t is the counterpart anchor of w_e in a target language (t). We input w_e into the embedding layer of an XLM to obtain its word embedding:

$$v_e = avg(XLMEmb(w_e)), \tag{1}$$

where $avg(\cdot)$ signifies averaging the embeddings of all the wordpieces segmented by the language model. Similarly, we obtain the embedding of w_t by feeding it into the embedding

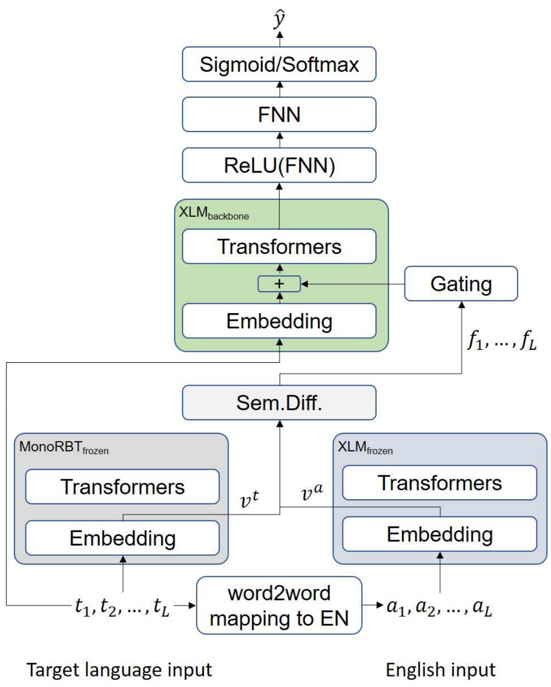

Fig. 2. Our proposed multilingual emotion recognition framework with semantic variation injection. MonoRBT denotes a pretrained monolingual RoBERTa language model. Sem.Diff. denotes semantic difference calculation. XLM denotes a pretrained multilingual language Model. The green XLM is the backbone model for emotion recognition with learnable parameters. The parameters of grey MonoRBT, XLM, and Sem.Diff. are frozen.

layer of a monolingual Roberta language model, pretrained on the target language:

$$v_t = avg(MonoEmb(w_t)). \tag{2}$$

We then train a rotation matrix R such that the English semantic space can be mapped to that of the target language via R. This is achieved by minimizing the Euclidean distance between the transformed English embedding and the target language embedding, i.e., Mean Squared Error (MSE) loss.

$$v'_e = R \cdot v_e, \tag{3}$$

$$\mathcal{L}^R = \frac{\sum_P (v'_{e,p} - v_{t,p})^2}{P}, \tag{4}$$

where $\cdot$ represents the dot product; P denotes the total number of embedding dimensions. Note that in this phase, the parameters in $XLMEmb(\cdot)$ and $MonoEmb(\cdot)$ have been frozen.

D. Emotion Recognition with Semantic Variation Injection

In this phase, we utilize the trained rotation matrix R to produce features for multilingual emotion recognition using XLM. Given an input sequence $t = (t_1, t_2, \ldots, t_L)$ in the target language, we obtain the English counterpart a_l for each word t_l using the word2word lexicon mapping ($w2wMap(\cdot)$).

$$a_l = w2wMap(t_l) \tag{5}$$

Similar to Equations 1 and 2, we use the frozen embedding layers of XLM and monolingual Roberta to compute the embedding of each word pair (a_l, t_l).

$$v^a_l = avg(XLMEmb(a_l)), \tag{6}$$

$$v^t_l = avg(MonoEmb(t_l)). \tag{7}$$

Note that the XLM utilized for calculating variation features in this stage is distinct from the backbone XLM employed in our emotion recognition process.

Next, we use the rotation matrix R to compute the semantic difference (f_l) between the word pair t_l and a_l:

$$f_l = v^t_l - R \cdot v^a_l. \tag{8}$$

In this step, we also freeze the parameters in R to capture consistent semantic differences. The resulting $f1, f_2, \ldots, f_L$ are used as semantic variation features to be injected into our backbone XLM ($XLM^m(\cdot)$) that is used for multilingual emotion recognition. Specifically, we obtain the original embeddings of the input using:

$$V_o = XLMEmb^m(t), \tag{9}$$

where $XLMEmb^m(\cdot)$ represents the embedding layer of the backbone XLM. Assume for the input word t_l, the embeddings of the corresponding wordpieces is $V_{o,l}$. Then, we update the embeddings by adding the respective semantic variation features to the original embeddings through broadcasting ($\mathbf{B}(\cdot)$):

$$V'_{o,l} = V_{o,l} + \mathbf{B}(g \odot f_l), \tag{10}$$

where $\odot$ denotes element-wise multiplication; g is a gating vector that scales the injected features. The values in g is initialized to zeros. This process is repeated for every input word in t to obtain the fully updated embeddings V'_o.

Subsequently, V'_o is fed into the following Transformer layers in the backbone XLM to obtain the sentence representation h, corresponding to the hidden state of the special [cls] token automatically concatenated at the beginning of an input sentence.

$$h = XLMTrans^m(V'_o)_{[cls]}. \tag{11}$$

Then, the sentence representation h is passed on to two layers of feed-forward neural networks (FNN) to obtain the probability distribution of emotion classification, with the first one activated by ReLU [38], and the second by sigmoid (multi-label classification) or softmax (multi-class classification).

$$h' = ReLU(FNN_1(h)) \tag{12}$$

$$\hat{y} = S(FNN_2(h')), \tag{13}$$

where $S(\cdot)$ represents sigmoid or softmax activation depending on the target task setting. For multi-label emotion recognition task, we use Binary Cross Entropy (BCE) as the loss function. For multi-class detection, we use Cross Entropy (CE):

$$\mathcal{L}^{BCE} = -\frac{\sum_N (y_n \log(\hat{y}_n) + (1 - y_n)\log(1 - \hat{y}_n))}{N} \tag{14}$$

$$\mathcal{L}^{CE} = -\frac{\sum_N y_n \log(\hat{y}_n)}{N}, \tag{15}$$

where N is the number of emotion categories.

IV. Experiments

A. Datasets

The Spanish dataset is from the Spanish emotion recognition subtask of SemEval-2018 Task 1 [39], which aims to classify a given Spanish tweet into one, more, or none of 11 given emotion categories, including *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise*, and *trust*. **The Japanese dataset** is derived from the WRIME dataset for emotional intensity estimation [40], which annotates Japanese SNS posts with intensity labels (*none, weak, medium, and strong*) for each of the 8 emotion categories, namely *joy, sadness, anticipation, surprise, anger, fear, disgust*, and *trust*. To align the setting with the Spanish dataset, we reconstruct the dataset for multi-label classification by mapping intensity labels into true (*weak* and above) and false (*none*). **The Chinese dataset** is from the usual subtask of SMP2020-EWECT[1], which labels Weibo posts as one of the 6 emotion categories: *neutral, happy, angry, sad, fear, and surprise*. Details of the datasets are shown in Table I.

TABLE I
Datasets used for experiments.

Dataset	# Train	# Dev	# Test
Spanish	3,561	679	2,854
Japanese	30,000	2,500	2,500
Chinese	27,768	2,000	5,000

B. Backbone Models

To verify the robustness of our method, we test it on two backbone XLMs: **XLM-Roberta-base** (XLM-R) is a multilingual model pretrained on 100 different languages. **XLM-V-base** [41] is a multilingual language model pretrained on 100 different languages with a one million token vocabulary, which allows for more semantically meaningful tokenization. It obtains state-of-the-art performance on a variety of NLP tasks for foundation XLMs.

C. Setup

We use Adam optimizer [42]. When training the rotation matrix R, we set the learning rate to 1e-4 and batch size to 10, and train for 100 epochs with early stopping. The monolingual Robertas are from Huggingface[2][3][4]. For emotion recognition, we set the learning rate to 1e-5. The models are trained with a batch size of 10 for 50 epochs with early stopping based on validation Macro F1 score. Results are averaged from 5 runs. For Chinese and Japanese, we use spaCy[5] to segment input sentences into words that can be looked up in the word2word lexicon to obtain word pairs.

V. Results

A. Multilingual Emotion Classification Evaluation

Tabel II shows the comparison between backbone XLMs and our method on the Spanish dataset. From the results, we can see that by injecting variation features, XLM-R and XLM-V obtain a gain of 1.08% and 0.65% on Macro F1 score respectively, and marginal gains on accuracy. The variation features are able to bring better performance on most of the emotion categories for both XLM-R and XLM-V, with significant improvement in the categories of *anticipation* and *trust* for XLM-R ($\Delta F1 > 2\%$). Our method performs worse than the backbone models on the *surprise* category. We suspect that this is because surprise is inferred from the situations being described instead of textual expressions for a lot of the surprise-labeled samples in this dataset. As such, the word-level variation features are unable to provide appropriate cultural differences in this scenario.

Table III presents the experiment results on the Japanese dataset. It can be observed that our method yields significant performance gain (1.51% F1, 5.41% accuracy) on XLM-R, 1.03% F1 and 0.3% accuracy gains on XLM-V. The variation features are able to improve the performance of the backbone models on almost all emotion categories. Notably, it outperforms the XLM-R by a large margin on *surprise* and *fear*, and the XLM-V on *sadness* and *trust*, showcasing its ability to bridge cultural semantic gaps for different XLMs.

Table IV shows the experiment results on the Chinese dataset. XLM-R is able to obtain an impressive 1.52% F1 gain and 4.79% accuracy with the help of variation features, whereas XLM-V obtains 0.79% F1 gain and 2.10% accuracy gain with variation features. Our method achieves higher F1 scores on all emotion categories for XLM-R. It obtains better or comparable performances for all but the *surprise* category for XLM-V. We hypothesize that this is because, in Chinese, a surprise is often expressed in a hyperbolized way that is similar to fear (discussed in section V-C). As such, the challenge is to distinguish between fear and surprise by gauging the actual severity of the context described. By encompassing this cultural difference, it is possible that the variation features for such expressions end up confusing the model.

Comparing the results from Table II, III, and IV, we can further make noteworthy observation. Firstly, by the extent of improvement over the backbone models, we can deduce that Spanish is semantically more similar to English than Japanese and Chinese, as such less affected by the anglocentric anchoring of XLMs. Therefore, more effort is needed for foundation XLMs to honor the cultural variations in languages that are further away from English in the language families.

Secondly, the gains on XLM-V are consistently lower than those on XLM-R. This is likely because the novel characteristic of XLM-V, i.e., a much larger multilingual vocabulary, allows for an embedding space that preserves more semantic differences of words among languages. However, like XLM-R, it is pretrained with implicit alignment, which encourages anglocentric anchoring [9] and leaves room for improvement

[1] https://github.com/Tencent/TencentPretrain/wiki/SMP2020-EWECT
[2] https://huggingface.co/rinna/japanese-roberta-base
[3] https://huggingface.co/bertin-project/bertin-roberta-base-spanish
[4] https://huggingface.co/hfl/chinese-roberta-wwm-ext
[5] https://spacy.io/usage/linguistic-features

TABLE II

Performance comparison on the Spanish dataset. w/ F stands for with injected variation features, w/o F for without.

Model	Anger		Anticipation		Disgust		Fear		Joy		Love		Optimism		Pessimism		Sadness		Surprise		Trust		Macro	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
XLM-R w/o F	71.32	82.00	43.22	88.67	45.50	84.70	68.13	93.89	**83.09**	89.33	66.22	94.70	42.04	90.04	**45.63**	81.02	**66.97**	84.91	**25.47**	**94.46**	26.46	95.12	53.09	88.99
XLM-R w/ F	**72.41**	**82.34**	**48.29**	**88.69**	**46.24**	**84.72**	**69.96**	**94.24**	82.37	**89.73**	**66.27**	**95.00**	**43.70**	**90.33**	45.56	**81.25**	66.22	**85.09**	24.98	94.45	**29.91**	95.12	**54.17**	**89.18**
XLM-V w/o F	70.34	80.91	44.35	85.82	42.25	84.84	60.64	**91.75**	80.82	88.21	65.93	**94.60**	**43.41**	**89.75**	43.44	**79.26**	62.26	**84.18**	**25.24**	91.68	30.77	94.32	51.76	87.76
XLM-V w/ F	**71.86**	**81.84**	**44.66**	**85.96**	**42.67**	**85.16**	**61.39**	91.44	**81.33**	**88.42**	**66.06**	94.53	43.39	89.73	**45.54**	78.84	**62.53**	83.89	25.08	91.68	**32.00**	**94.74**	**52.41**	**87.84**

TABLE III

Performance comparison on the Japanese dataset. w/ F stands for with injected variation features, w/o F for without.

Model	Joy		Sadness		Anticipation		Surprise		Anger		Fear		Disgust		Trust		Macro	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
XLM-R w/o F	78.78	82.20	64.62	83.88	74.44	80.88	63.34	76.80	34.69	97.44	51.17	80.76	49.68	87.28	37.14	**91.20**	56.73	80.06
XLM-R w/ F	**79.72**	**82.94**	**65.12**	**84.12**	**75.45**	**81.45**	**65.87**	**77.40**	**35.45**	**97.54**	**56.16**	**82.08**	**50.43**	**87.76**	**37.71**	90.44	**58.24**	**85.47**
XLM-V w/o F	79.16	**83.24**	59.69	83.20	74.50	82.20	61.95	76.12	**27.72**	97.08	**56.53**	83.08	51.03	86.72	36.31	91.16	55.86	85.35
XLM-V w/ F	**79.46**	82.76	**62.75**	**83.76**	**74.87**	**82.84**	**63.12**	**77.00**	27.57	**97.24**	56.12	**83.36**	**51.20**	**86.96**	**40.00**	**91.24**	**56.89**	**85.65**

TABLE IV

Performance comparison on the Chinese dataset. w/ F stands for with injected variation features, w/o F for without.

Model	Angry		Sad		Neutral		Happy		Fear		Surprise		Macro	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
XLM-R w/o F	81.21	**86.27**	61.54	56.89	82.60	82.73	78.42	**77.11**	68.60	67.62	62.21	60.96	72.43	71.93
XLM-R w/ F	**81.77**	81.03	**66.87**	**71.56**	**84.19**	**83.13**	**78.53**	73.48	**70.09**	**71.43**	**62.48**	**66.58**	**76.98**	**76.72**
XLM-V w/o F	82.36	81.56	65.58	67.33	82.29	**83.54**	**78.73**	76.92	70.21	72.38	**61.94**	**60.70**	73.52	73.74
XLM-V w/ F	**83.28**	**85.32**	**67.96**	**75.89**	**83.14**	83.43	78.66	**78.88**	**72.06**	**74.29**	60.78	57.22	**74.31**	**75.84**

by semantic variation injection. Additionally, our variation features for both XLMs rely on spaCy to tokenize and construct word pairs for fairness. It is possible that the tokenization of XLM-V captures semantic concepts that are more precise than spaCy, which is important for accurate variation feature calculation, i.e., 気持ち悪い(disgusting) as a full token versus 気持ち(mood) and 悪い(bad) as separate tokens.

B. Human Evaluation for Semantic Variations

To verify whether our method can reliably select words with high semantic variations between languages, i.e., whether f_l in Equation 8 is a proper reflection of cross-lingual semantic variations, we conduct a human evaluation in this section.

First, we construct in total 6 word pair lexicons using XLM-R and XLM-V on the datasets in Table I. For each lexicon, following the steps in Equation 5-8, we record the absolute value of the element-vise average of the semantic difference f_l^i for word pair (t_l, a_l) as its semantic variation score s_l^i for the i-th time t_l appears in the dataset:

$$s_l^i = |\frac{1}{P}\sum_P f_{l,p}^i|, \tag{16}$$

where P denotes the number of embedding dimensions. After processing all samples from a dataset, we compute the overall semantic variation score s_l for word pair (t_l, a_l) as:

$$s_l = \frac{1}{I}\sum_I s_l^i, \tag{17}$$

where I is the total number of appearances of t_l.

We then rank the lexicon of word pairs according to their semantic variation scores and randomly select 100 word pairs from the upper 50% of the abstract lexicons[6] of each dataset for human evaluation. For each lexicon in a target language, we invite three native speakers who have received at least four years of English education in an anglophone country as evaluators. An evaluation instance is formed by randomly selecting two word pairs (t_j, a_j) and (t_k, a_k) that rank at the j-th and k-th place $(j - k > 10)$ in our semantic variation lexicon, and an evaluator which pair they think is more semantically different, t_j and a_j, or t_k and a_k. Evaluators were encouraged to Google example sentences, containing t_j, a_j, t_k, or a_k, respectively, to better understand the semantic variants. We use the label agreed by the major evaluators as the ground truth. If they believe the semantic difference between t_j and a_j is larger, we deem the test instance as correct; otherwise, it is erroneous. We form 200 such test instances (two word pairs) for each language, and measure the results by average accuracy. The evaluators achieved 0.63 Cohen's kappa in the evaluation task. As shown in Table V, we can conclude that the semantic variation feature is effective in measuring the degree of semantic variations between languages, applicable to different XLMs.

TABLE V

Human evaluation results, evaluated by accuracy.

Language	XLM-R lex.	XLM-V lex.
Spanish	69.5	69.3
Japanese	73.3	74.1
Chinese	74.2	75.1

[6] We do not select word pairs from the bottom 50% (the concrete ones) because we aimed to adjust the semantic variations for the abstract ones by using the concrete ones as anchors for training the rotation matrix (see Section III-A).

C. Case Study

In this section, we conduct case studies on the Japanese and Chinese datasets to qualitatively illustrate how our method is able to mitigate the anglocentric anchoring of XLMs. Japanese and Chinese are analyzed because they demonstrated higher semantic variations in Section V-A. Similar to Section V-B, we select the high variation words by comparing the absolute value of the element-vise average of Equation 8 for each word in an input sentence.

Table VI presents the case study on Japanese. In the first instance, the high variation word selected is “ギャップ”, which is a loanword from “gap” in English. Unlike the original word, it describes the unexpected discrepancy between what people think someone should like, and how they actually are. From the predictions, we can observe that XLM-R fails to assign “surprise” to the input sentence, indicating that it is unable to capture the element of unexpectedness in the word “ギャップ”. The second instance means “please no!”, conveying a negative emotion. The high variation word is “勘弁”, whose literal translation is to forgive. XLM-R wrongly labels the sentence as a surprise, suggesting that it is unable to understand the meaning of the word used in this context. On the other hand, our method can infer correctly thanks to the injected variance features, which implicitly provide more word senses to the XLM. In the third instance, the high variation word “残念” can generally be translated as “regrettable”. However, it can also express a subtle distaste for being let down when it is used to describe a situation, which can be detected by our method, but not by XLM-R. The high variation word “楽しみ” in the fourth instance has the word sense of enjoying something, as well as looking forward to something. The latter is generally accompanied by a sense of cheerfulness when not used sarcastically, which XLM-R can neglect in certain instances. In the fifth instance, “やばい” is an emotive word that has no equivalent in English. It can be used to convey strong emotions ranging from positive (awesome) to negative (terrible). Hence, it can pose a challenge for machines to discern their emotional learning according to the context. As shown in Table VI, XLM-V misinterprets “やばい” in this scenario as an expression of shock, whereas our method understands it as an exclamation of amazement.

Table VII presents the case study on Chinese. In the first and fourth instances, “可怕” and “吓人” both translate to “terrifying” and ”scary” in English. However, as mentioned in the previous section, they are often used in a hyperbolic way to express shock. From these two instances, we can infer that without variation features, both XLM-R and XLM-V make the wrong prediction, likely due to the confusion of this hyperbolic usage. In the second instance, XLM-R mislabels the sentence as angry, while our method correctly classifies it as sad. The high variation word in this sentence is “内疚”, which means “guilt”. We suspect this has to do with how the correlations between guilt and other negative emotions are perceived differently for anglophones and Chinese, as similar differences have been studied for other cultural comparisons [43], [44]. The third instance is challenging, as it expresses resiliency and resolution in times of difficulties, leaning towards weak positive. The high variation word “积极” means “optimistic”. It could be that XLM-R lacks the understanding that Chinese culture prizes perseverance over extreme adversity, leading to a negative prediction. The last instance translates to, “She criticize her father for never caring about her and her mother. There is no love in this family.” XLM-V wrongly classifies it as sad instead of angry. The high variation word, “亲情” (love between family members), is likely closer to “moral” in the semantic space compared to its English counterpart (love between family members), since Chinese culture considers it not only as family ties but also as a moral obligation. Hence, XLM-V’s prediction aligns with the anglocentric reading of this sentence, i.e., the mourning over lost of familial love, where as our method is able to detect the culturally appropriate tone, i.e., accusation of failing family duty.

TABLE VI
CASE STUDY ON THE JAPANESE DATASET. THE RELEVANT HIGH VARIATION WORDS IN THE INPUT SENTENCES ARE UNDERLINED.

Content	Labels
ほしのディスコ歌うまいのかよ！！！ギャップ萌！！！	Ground-truth: joy, anticipation, surprise XLM-R w/o F: joy, anticipation XLM-R w/F: joy, anticipation, surprise
ご勘弁！！	Ground-truth: sadness XLM-R w/o F: surprise XLM-R w/F: sadness
スーパーで買ったエビフライは衣が厚くて残念でした。	Ground-truth: sadness, disgust XLM-R w/o F: sadness XLM-R w/F: sadness, disgust
合成音声もそのうち、生身の人間と話すのと変わらないくらいになるのでしょうね。それはそれで楽しみです。	Ground-truth: joy, anticipation XLM-V w/o F: anticipation XLM-V w/F: joy, anticipation
てかフォロワーさん優しすぎるからやばい	Ground-truth: joy, trust XLM-V w/o F: surprise XLM-V w/F: joy

TABLE VII
CASE STUDY ON THE CHINESE DATASET. THE WORDS WITH HIGHEST VARIATION IN THE INPUT SENTENCES ARE UNDERLINED.

Content	Labels
有些为自己现在的心思惶恐，有时候会觉得自己已经自以为地站在了比普通民众高一级的地方，而不由自主地蔑视民众的思想……这种真可怕orz	Ground-truth: surprise XLM-R w/o F: fear XLM-R w/ F: surprise
内疚的感觉真的不是一般的糟特别是别人用心的事情而我竟后知后觉	Ground-truth: sad XLM-R w/o F: angry XLM-R w/ F: sad
即使失败了还是得继续前进。我做不到永远积极向上但至少继续前进。我不会捷径，至少智商的资本已经再也无法显现，现在的我除了天生的伪装乐观就只有一颗被迫冰冻的心，只等梦想的暖阳照进。#我的复旦梦#	Ground-truth: happy XLM-R w/o F: sad XLM-R w/ F: happy
吓人惊了[哆啦A梦汗]666666可怕	Ground-truth: surprise XLM-V w/o F: fear XLM-V w/ F: surprise
她指责父亲从来不关心母亲和她，这个家里没有亲情。	Ground-truth: angry XLM-V w/o F: sad XLM-V w/ F: angry

VI. CONCLUSION

To address the anglocentric anchoring problem in XLMs, we propose a task-agnostic approach that incorporates semantic variations between languages into the XLM embedding space. We illustrate the effectiveness of our proposed method by highlighting improvements in emotion recognition tasks across Spanish, Japanese, and Chinese datasets using two backbone XLMs. We further validate the efficacy of our semantic variation features by human evaluation and case study.

References

[1] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[2] B. Mesquita, M. Boiger, and J. De Leersnyder, "The cultural construction of emotions," *Current opinion in psychology*, vol. 8, pp. 31–36, 2016.

[3] C. Heyes, "Précis of cognitive gadgets: The cultural evolution of thinking," *Behavioral and Brain Sciences*, vol. 42, p. e169, 2019.

[4] L. F. Barrett, "The theory of constructed emotion: an active inference account of interoception and categorization," *Social cognitive and affective neuroscience*, vol. 12, no. 1, pp. 1–23, 2017.

[5] A. Wierzbicka, "Defining emotion concepts," *Cognitive science*, vol. 16, no. 4, pp. 539–581, 1992.

[6] J. C. Jackson, J. Watts, T. R. Henry, J.-M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, and K. A. Lindquist, "Emotion semantics show both cultural variation and universal structure," *Science*, vol. 366, no. 6472, pp. 1517–1522, 2019.

[7] R. Mao, X. Li, M. Ge, and E. Cambria, "MetaPro: A computational metaphor processing model for text pre-processing," *Information Fusion*, vol. 86-87, pp. 30–43, 2022.

[8] R. Mao, X. Li, K. He, M. Ge, and E. Cambria, "MetaPro Online: A computational metaphor processing online system," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2023, pp. 127–135.

[9] S. Havaldar, B. Singhal, S. Rai, L. Liu, S. C. Guntuku, and L. Ungar, "Multilingual language models are not multicultural: A case study in emotion," in *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[10] R. Mao, C. Lin, and F. Guerin, "Word embedding and WordNet based metaphor identification and interpretation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, 2018, pp. 1222–1231.

[11] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 62–69, 2023.

[12] M. Neumann *et al.*, "Cross-lingual and multilingual speech emotion recognition on English and French," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5769–5773.

[13] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6907–6911.

[14] M. Agarla, S. Bianco, L. Celona, P. Napoletano, A. Petrovsky, F. Piccoli, R. Schettini, and I. Shanin, "Semi-supervised cross-lingual speech emotion recognition," *Expert Systems with Applications*, vol. 237, p. 121368, 2024.

[15] V. K. Jain, S. Kumar, and S. L. Fernandes, "Extraction of emotions from multilingual text using intelligent text processing and computational linguistics," *Journal of Computational Science*, vol. 21, p. 316, 2017.

[16] K. Becker, V. P. Moreira, and A. G. dos Santos, "Multilingual emotion classification using supervised learning: Comparative experiments," *Information Processing & Management*, vol. 53, no. 3, p. 684, 2017.

[17] S. Ghosh, A. Priyankar, A. Ekbal, and P. Bhattacharyya, "Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data," *Knowledge-Based Systems*, vol. 260, p. 110182, 2023.

[18] R. Mao and X. Li, "Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 534–13 542.

[19] M. Ge, R. Mao, and E. Cambria, "Explainable metaphor identification inspired by conceptual metaphor theory," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10 681–10 689, 2022.

[20] T. Yue, R. Mao, H. Wang, Z. Hu, and E. Cambria, "KnowleNet: Knowledge fusion network for multimodal sarcasm detection," *Information Fusion*, vol. 100, p. 101921, 2023.

[21] S. Derrig, "Metaphor in the color lexicon," *Papers from the parasession on the lexicon*, pp. 85–96, 1978.

[22] C. H. Brown, "General principles of human anatomical partonomy and speculations on the growth of partonomic nomenclature 1," *American ethnologist*, vol. 3, no. 3, pp. 400–424, 1976.

[23] A. Ogarkova, C. Soriano, and C. Lehr, "Naming feeling: Exploring the equivalence of emotion terms in five european languages," *Dynamicity in emotion concepts*, vol. 27, pp. 253–284, 2012.

[24] H. Youn, L. Sutton, E. Smith, C. Moore, J. F. Wilkins, I. Maddieson, W. Croft, and T. Bhattacharya, "On the universal structure of human lexical semantics," *Proceedings of the National Academy of Sciences*, vol. 113, no. 7, pp. 1766–1771, 2016.

[25] A. K. Romney, C. C. Moore, and C. D. Rusch, "Cultural universals: Measuring the semantic structure of emotion terms in English and Japanese," *Proceedings of the National Academy of Sciences*, vol. 94, no. 10, pp. 5489–5494, 1997.

[26] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[27] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[28] D. Jiang, R. Wei, J. Wen, G. Tu, and E. Cambria, "AutoML-Emo: Automatic knowledge selection using congruent effect for emotion identification in conversations," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1845–1856, 2023.

[29] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "SenticNet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3829–3839.

[30] W. Li, L. Zhu, R. Mao, and E. Cambria, "SKIER: A symbolic knowledge integrated model for conversational emotion recognition," *Proceedings of the AAAI Conference on Arificial Intelligence*, vol. 37, no. 11, pp. 13 121–13 129, 2023.

[31] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1743–1753, 2023.

[32] M. Coltheart, "The MRC psycholinguistic database," *The Quarterly Journal of Experimental Psychology Section A*, vol. 33, p. 497, 1981.

[33] E. Rabinovich, B. Sznajder, A. Spector, I. Shnayderman, R. Aharonov, D. Konopnicki, and N. Slonim, "Learning concept abstractness using weak supervision," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4854–4859.

[34] S. J. Crutch and E. K. Warrington, "Abstract and concrete concepts have structurally different representational frameworks," *Brain*, vol. 128, no. 3, pp. 615–627, 2005.

[35] R. Mao, K. Du, Y. Ma, L. Zhu, and E. Cambria, "Discovering the cognition behind language: Financial metaphor analysis with MetaPro," in *2023 IEEE International Conference on Data Mining (ICDM)*, 2023.

[36] S.-T. Kousta, G. Vigliocco, D. P. Vinson, M. Andrews, and E. Del Campo, "The representation of abstract words: why emotion matters." *Journal of Experimental Psychology: General*, vol. 140, no. 1, p. 14, 2011.

[37] Y. J. Choe, K. Park, and D. Kim, "word2word: A collection of bilingual lexicons for 3,564 language pairs," in *Proceedings of the 12th International Conference on Language Resources and Evaluation*, 2020.

[38] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," *arXiv preprint arXiv:1803.08375*, 2018.

[39] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "Semeval-2018 task 1: Affect in tweets," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 1–17.

[40] H. Suzuki, Y. Miyauchi, K. Akiyama, T. Kajiwara, T. Ninomiya, N. Takemura, Y. Nakashima, and H. Nagahara, "A Japanese dataset for subjective and objective sentiment polarity classification in micro blog domain," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 7022–7028.

[41] D. Liang, H. Gonen, Y. Mao, R. Hou, N. Goyal, M. Ghazvininejad, L. Zettlemoyer, and M. Khabsa, “XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models,” *arXiv preprint arXiv:2301.10472*, 2023.

[42] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[43] T. Stompe, G. Ortwein-Swoboda, H. R. Chaudhry, A. Friedmann, T. Wenzel, and H. Schanda, “Guilt and depression: a cross-cultural comparative study,” *Psychopathology*, vol. 34, no. 6, pp. 289–298, 2001.

[44] G. G. Bear, X. Uribe-Zarain, M. A. Manning, and K. Shiomi, “Shame, guilt, blaming, and anger: Differences between children in Japan and the US,” *Motivation and Emotion*, vol. 33, pp. 229–238, 2009.