

# Multilingual and Multilabel Emotion Recognition using Virtual Adversarial Training

Vikram Gupta

ShareChat, India

vikramgupta@sharechat.co

## Abstract

Virtual Adversarial Training (VAT) has been effective in learning robust models under supervised and semi-supervised settings for both computer vision and NLP tasks. However, the efficacy of VAT for multilingual and multilabel text classification has not been explored before. In this work, we explore VAT for multilabel emotion recognition with a focus on leveraging unlabelled data from different languages to improve the model performance. We perform extensive semi-supervised experiments on SemEval2018 multilabel and multilingual emotion recognition dataset and show performance gains of **6.2%** (Arabic), **3.8%** (Spanish) and **1.8%** (English) over supervised learning with same amount of labelled data (10% of training data). We also improve the existing state-of-the-art by **7%**, **4.5%** and **1%** (Jaccard Index) for Spanish, Arabic and English respectively and perform probing experiments for understanding the impact of different layers of the contextual models.

## 1 Introduction

Emotion recognition is an active and crucial area of research, especially for social media platforms. Understanding the emotional state of the users from textual data forms an important problem as it helps in discovering signs of fear, anxiety, bullying, hatred etc. and maintaining the emotional health of the people and platform. With the advent of deep neural networks and contextual models, text understanding has advanced dramatically by leveraging huge amount of unlabelled data freely available on web. However, even with these advancements, annotating emotion categories is expensive and time consuming as emotion categories are highly correlated and subjective in nature and can co-occur in the same text. Psychological studies suggest that emotions like "anger" and "sadness" are correlated and co-occur more frequently than "anger" and "happiness" (Plutchik, 1980). In a multilingual

setup, the annotation becomes even more challenging as annotator team are expected to be familiar with different languages and culture for understanding the emotions accurately. Imbalance in availability of the data across languages further creates problems, especially in case of resource impoverished languages. In this work, we investigate the following key points; *a) Can unlabelled data from other languages improve recognition performance of target language and help in reducing requirement of labelled data? b) Efficacy of VAT for multilingual and multilabel setup.*

To address the aforementioned questions, we focus our experiments towards *semi-supervised* learning in a *multilingual* and *multilabel* emotion classification framework. We formulate semi-supervised Virtual Adversarial Training (VAT) (Miyato et al., 2018) for multilabel emotion classification using contextual models and perform extensive experiments to demonstrate that unlabelled data from other languages  $L_{ul} = \{L_1, L_2, \dots, L_n\}$  improves the classification on the target language  $L_{tgt}$ . We obtain competitive performance by reducing the amount of annotated data demonstrating cross-language learning. To effectively leverage the multilingual content, we use multilingual contextual models for representing the text across languages. We also evaluate monolingual contextual models to understand the performance differences between multilingual and monolingual models and explore the advantages of *domain-adaptive* and *task-adaptive* pretraining of models for our task and observe substantial gains.

We perform extensive experiments on the SemEval2018 (Affect in Tweets: Task E-c<sup>1</sup>) dataset (Mohammad et al., 2018) which contains tweets from Twitter annotated with 11 emotion categories across three languages - English, Spanish and Arabic and demonstrate the effectiveness of semi-supervised learning across languages. To the

<sup>1</sup><https://competitions.codalab.org/competitions/17751>

best of our knowledge, our study is the first one to explore semi-supervised adversarial learning across different languages for multilabel classification. In summary, the main contributions of our work are the following:

- We explore Virtual Adversarial Training (VAT) for semi-supervised multilabel classification on multilingual corpus
- Experiments demonstrating **6.2%**, **3.8%** and **1.8%** improvements (Jaccard Index) on Arabic, Spanish and English by leveraging unlabelled data of other languages while using 10% of labelled samples.
- Improve state-of-the-art of multilabel emotion recognition by **7%**, **4.5%** and **1%** (Jaccard Index) for Spanish, Arabic and English respectively.
- Experiments showcasing the advantages of *domain-adaptive* and *task-adaptive* pre-training

## 2 Related Work

Semi-supervised learning is an important paradigm for tackling the scarcity of labelled data as it marries the advantages of supervised and unsupervised learning by leveraging the information hidden in large amount of unlabelled data along with small amount of labelled data (Yang et al., 2021), (Van Engelen and Hoos, 2020). Early approaches used self-training for leveraging the model’s own predictions on unlabelled data to obtain additional information during training (Yarowsky, 1995) (McClosky et al., 2006). Clark et al. (2018) proposed cross-view training (CVT) for various tasks like chunking, dependency parsing, machine translation and reported state-of-the-art results. CVT forces the model to make consistent predictions when using the full input or partial input. Ladder networks (Laine and Aila, 2016), Mean Teacher networks (Tarvainen and Valpola, 2017) are another way for semi-supervised learning where temporal and model-weights are ensembled. Another popular direction towards semi-supervised learning is adversarial training where the data point is perturbed with random or carefully tuned perturbations to create an adversarial sample. The model is then encouraged to maintain consistent predictions for the original sample and the adversarial sample. Adversarial training

was initially explored for developing secure and robust models (Goodfellow et al., 2014), (Xiao et al., 2018), (Saadatpanah et al., 2020) to prevent attacks. Miyato et al. (2016), Cheng et al. (2019), Zhu et al. (2019) showed that adversarial training can improve both robustness and generalization for classification tasks, machine translation and GLUE benchmark respectively. Miyato et al. (2016), Sachan et al. (2019), Miyato et al. (2018) then applied the adversarial training for semi-supervised image and text classification showing substantial improvements.

Emotion recognition is an important problem and has received lot of attention from the community (Yadollahi et al., 2017), (Sailunaz et al., 2018). The taxonomies of emotions suggested by Plutchik wheel of emotions (Plutchik, 1980) and (Ekman, 1984) have been used by the majority of the previous work in emotion recognition. Emotion recognition can be formulated as a *multiclass* problem (Scherer and Wallbott, 1994), (Mohammad, 2012) or a *multilabel* problem (Mohammad et al., 2018), (Demszky et al., 2020). In the *multiclass* formulation, the objective is to identify the presence of one of the emotion from the taxonomy whereas in a *multilabel* setting, more than one emotion can be present in the text instance. Binary relevance approach (Godbole and Sarawagi, 2004) is another way to break multilabel problem into multiple binary classification problems. However, this approach does not model the co-relation between emotions. Seq2Seq approaches (Yang et al., 2018), (Huang et al., 2021) solve this problem by modelling the relationship between emotions by inferring emotion in an incremental manner. An interesting direction for handling data scarcity in emotion recognition is to use distant supervision by exploiting emojis (Felbo et al., 2017), hashtags (Mohammad, 2012) or pretraining emotion specific embeddings and language models (Barbieri et al., 2021), (Ghosh et al., 2017).

With the emergence of contextual models like BERT (Devlin et al., 2018), Roberta (Liu et al., 2019) etc., the field of NLP and text classification has been revolutionized as these models are able to learn efficient representations from a huge corpus of unlabelled data across different languages and domains (Hassan et al., 2021), (Barbieri et al., 2021). Social media content contains linguistic errors, idiosyncratic styles, spelling mistakes, grammatical inconsistency, slangs, hashtags,

emoticons etc. (Barbieri et al., 2018), (Derczynski et al., 2013) due to which off-the-shelf contextual models may not be optimum. We use *language-adaptive*, *domain-adaptive* and *task-adaptive* pre-training which has shown performance gains (Peters et al., 2019), (Gururangan et al., 2020), (Barbieri et al., 2021), (Howard and Ruder, 2018), (Lee et al., 2020) for different tasks and domains.

### 3 Methodology

We consider the task of multilabel emotion classification, where given a text  $t \in T$  and  $t = \{w_1, w_2, \dots, w_l\}$ , we predict the presence of  $y$  emotion categories denoted by  $\{1, 2, \dots, y\}$ .  $T$  represents the corpus of all the sentences across the different languages and  $w_i$  represent the tokens in the sentence. We leverage contextual models as feature extractors  $\phi: t_i \rightarrow x_i$ , where  $x_i \in \mathbb{R}^d$  and  $d$  is the dimension of the text representations and train a classifier over these representations.

#### 3.1 Virtual Adversarial Training (VAT)

Virtual Adversarial Training (VAT) (Miyato et al., 2018) is a regularization method for learning robust representations by encouraging the models to produce similar outputs for the input data points and local perturbations. VAT creates the adversary by perturbing the input in the direction which maximizes the change in the output of the model. Since VAT does not require labels it is well suited for semi-supervised applications. Consider  $x \in \mathbb{R}^d$  as the  $d$  dimensional representation of the text and  $y$  as the ground truth. Objective function of VAT ( $L_{vadv}$ ) is represented as,

$$L_{vadv}(x, \theta) := D[p(y|x, \hat{\theta}), p(y|x + r_{vadv}, \theta)] \quad (1)$$

where,

$$r_{vadv} := \arg \max D[p(y|x, \hat{\theta}), p(y|x + r, \theta)] \quad (2)$$

and  $\|r\|_2 < \epsilon$  and  $r_{vadv} \in \mathbb{R}^d$ .  $D[p, p']$  measures the divergence between the two probability distributions and  $r_{vadv}$  is the *virtual adversarial perturbation* that maximizes this divergence. In order to leverage the unlabelled data, the predictions from the current estimate of the model  $\hat{\theta}$  are used as the target. However, it is not possible to exactly compute  $r_{vadv}$  by a closed form solution or linear approximation as gradient  $g$  (Equation 4) with respect to  $r$  is always zero at  $r = 0$ . Miyato et al. (2018) propose fast approximation method to formulate

$r_{adv}$  as:

$$r_{vadv} \approx \epsilon \frac{g}{\|g\|_2}, \quad (3)$$

where,

$$g = \nabla_r D[p(y|x, \hat{\theta}), p(y|x + r, \hat{\theta})] \quad (4)$$

and  $r = \epsilon * q$ , where  $q$  is a randomly sampled unit vector. With this approximation, we can use back-propagation to compute the gradients  $g$  in Equation 4. The overall training objective,  $L_{VAT}$  becomes:

$$L_{VAT} = L_{ce} + \alpha * L_{vadv} \quad (5)$$

where  $L_{ce}$  is the multiclass classification loss and  $L_{adv}$  is the adversarial loss.  $\alpha$  is the balancing hyperparameter between the two losses.

#### 3.2 Multilabel Virtual Adversarial Training (m1VAT)

We explore VAT for multilingual contextual models and multilabel classification. For computer vision tasks, perturbing the raw pixel values to generate adversarial examples is intuitive as the input space is continuous. However, contextual models use the indices of the words as input which are not present in the continuous domain and thus perturbing them is not optimal. Perturbing an index  $k$  of a word  $w_k$  to  $k + r_{vadv}$  would not result in a word closer to  $w_k$ . To overcome this problem, instead of perturbing the input, we perturb the intermediate layer of the contextual models which form a continuous representation space and allows us to use VAT with contextual models. Similar strategy for text classification was also explored by Miyato et al. (2016). For modelling multilabel classification, we measure the divergence of multilabel outputs by Mean Square Error (MSE),

$$L_{vadv}(x, \theta) := MSE[p(y|x, \hat{\theta}), p(y|x + r_{vadv}, \theta)] \quad (6)$$

MSE is calculated over the logits normalized by sigmoid. This is important as the outputs in case of multilabel classification are not probability distributions across classes which renders the usage of KL-Divergence incompatible for this scenario. We also experiment by treating the probability for each emotion separately but our results demonstrate the effectiveness of Mean Square Error (MSE) for our task (Table 4). The overall training objective,  $L_{m1VAT}$  is:

$$L_{m1VAT} = L_{bce} + \alpha * L_{vadv} \quad (7)$$

where,  $L_{bce}$  is the multilabel binary cross entropy loss. We represent the text instances using monolingual/multilingual contextual representations.

### 3.3 Multilingual Semi-Supervised Setup

**mlVAT:** For each target language  $L_{tgt}$ , we randomly select a percentage of samples from the training set of this language and use them as labelled examples for training. We use the remaining data of the same language and the complete dataset of the other languages  $L_{ul}$  as the unlabelled set. Each training batch is created by maintaining a ratio between labelled and unlabelled examples for stable training. For the labelled set, both multilabel classification loss  $L_{bce}$  and adversarial loss  $L_{vadv}$  is applied. For the unlabelled examples, only the adversarial loss  $L_{vadv}$  is used.

**Sup:** We also train supervised classifiers (Sup) by using the same amount of labelled data for target language  $L_{tgt}$ . Supervised classifiers (Sup) act as baseline and help in measuring the gains obtained by semi-supervised learning. We vary the ratio of sampled labelled examples as 10%, 25%, 50% and 100% to study the progression of our framework across different amount of labelled data of the target language.

### 3.4 Multilingual Representation

For leveraging cross-learning between multiple languages in a semi-supervised setup, we experiment with different multilingual models. We experiment with off-the-shelf multilingual BERT, mBERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2019) models which have been trained with corpus from multiple languages. Since we are performing emotion recognition on multilingual tweets, we evaluate the *domain-adaptive* multilingual model XLM-Tw (Barbieri et al., 2021) trained using a 198M tweet corpus across 30 languages over the XLM-R checkpoint. For exploring the effect of *task-adaptive* pretraining, we evaluate XLM-Tw-S, which is finetuned for sentiment analysis over tweets which is arguably a task related to emotion recognition.

### 3.5 Monolingual Representation

We also experiment with monolingual models trained over the corpus from the same language for comparison with multilingual models and setting up the baselines for each language: English BERT (E-BERT) (Devlin et al., 2018) for English, BeToBERT (Cañete et al., 2020) for Spanish and AraBERT (Antoun et al., 2020) for Arabic. We experiment with and without finetuning the representations to evaluate the performance of these

representations out-of-the box and finetuning over our task.

%	Method	JI	MiF1	MaF1
10	Sup	44.05	57.86	40.91
	mlVAT	<b>46.79</b>	<b>60.36</b>	<b>44.41</b>
25	Sup	49.69	62.80	44.19
	mlVAT	<b>51.08</b>	<b>63.96</b>	<b>47.31</b>
50	Sup	53.95	66.26	48.57
	mlVAT	<b>55.11</b>	<b>66.79</b>	<b>52.52</b>
100	Sup	55.78	67.41	50.12
	mlVAT	<b>57.31</b>	<b>68.18</b>	<b>52.15</b>

Table 1: mlVAT and Supervised (Sup) experiments on **Arabic** across different ratios of labelled examples

%	Method	JI	MiF1	MaF1
10	Sup	54.15	66.33	48.94
	mlVAT	<b>55.15</b>	<b>67.01</b>	<b>50.57</b>
25	Sup	55.11	66.99	47.83
	mlVAT	<b>56.54</b>	<b>68.52</b>	51.18
50	Sup	57.20	69.14	<b>54.14</b>
	mlVAT	<b>58.67</b>	<b>70.03</b>	<b>51.55</b>
100	Sup	59.78	71.19	53.43
	mlVAT	<b>60.70</b>	<b>71.90</b>	<b>56.10</b>

Table 2: mlVAT and Supervised (Sup) experiments on **English** across different ratios of labelled examples

%	Method	JI	MiF1	MaF1
10	Sup	44.36	53.17	38.28
	mlVAT	<b>46.05</b>	<b>54.83</b>	<b>42.49</b>
25	Sup	<b>52.89</b>	<b>61.30</b>	48.99
	mlVAT	52.05	60.17	<b>49.15</b>
50	Sup	55.17	63.20	51.70
	mlVAT	<b>55.70</b>	<b>63.39</b>	<b>54.19</b>
100	Sup	<b>57.04</b>	<b>65.31</b>	51.53
	mlVAT	56.89	64.89	<b>51.77</b>

Table 3: mlVAT and Supervised (Sup) experiments on **Spanish** across different ratios of labelled examples

### 3.6 Dataset and Evaluation

We evaluate on the SemEval2018 dataset (Affect in Tweets: Task E-c) (Mohammad et al., 2018) dataset. The dataset consists of tweets scraped from twitter in English, Spanish and Arabic. Each tweet is annotated with the presence of 11 emotions *anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise and trust*. Some tweets are neutral and do not have the presence of any emotion. The dataset has 3 splits - train, dev and test



(Table 15). Following Mohammad et al. (2018), we measure the multilabel accuracy using Jaccard Index (JI), Macro F1 (MaF1) and Micro F1 (MiF1) scores (Chinchor, 1992) over the test set of these languages.

#### 4 Semi-Supervised Experiments

We select a percentage (10%, 25%, 50%, 100%) of the data from the target language as labelled data and use the remaining data from same language along with data of other languages as the unlabelled data. In Table 1 for Arabic, we see that by using 10%, 25%, 50% and 100% of the labelled data, mLVAT improves upon the results of training over the same amount of supervised data by **6.2%**, **2.8%**, **2.2%** and **2.7%** (Jaccard Index; JI) respectively. Similar improvements are also observed on the micro F1 (MiF1) and macro F1 (MaF1). It is interesting to note that by using only 50% of the labelled data with unlabelled data, we are able to match the performance of supervised learning with 100% of the data for Spanish. This shows that mLVAT is able to leverage the unlabelled data of Spanish and English for improving the performance over Arabic language.

Similar observations on English can be made from Table 2 also where we notice an improvement of **1.8%**, **2.6%**, **2.6%** and **2%** on the Jaccard Index and proportional improvements on other metrics also. For English also, we note that by using 10% of labelled data, mLVAT is able to improve on supervised results with 25% of the data. For Spanish, mLVAT helps for the 10% and 50% split as reported in Table 3 but is not able to improve all the metrics for the other splits. Overall, for majority of the languages and splits, we see that by adding unlabelled data, mLVAT improves upon the performance over supervised learning consistently and helps in decreasing the requirements for annotated data.

**Frozen backbone:** We perform semi-supervised experiments with frozen backbone to investigate the effect of mLVAT on the backbone and classification head. We repeat similar experiments as in previous sections for Spanish and English, but freeze the backbone and only train the classification head. From the Figure 1, we can observe that mLVAT consistently improves the performance for both languages over all the splits. This demonstrates that the performance gains are backbone-agnostic allowing for application of mLVAT on other back-

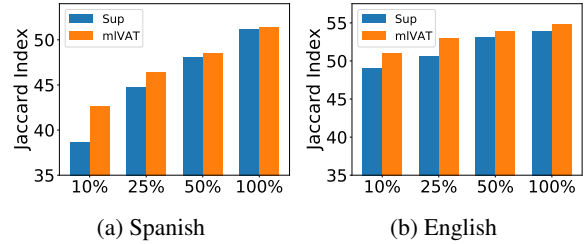


Figure 1: Comparison of Jaccard Index for English and Spanish across different ratio of labelled examples with *frozen backbone*. mLVAT (orange) improves upon supervised settings (Sup) (blue) for both languages.

bones also.

Loss	JI	MiF1	MaF1
mLVAT	55.2	67.1	50.6
mLVAT (w/o sig)	50.7	63.5	41.6
KLDivergence	21.9	35.9	34.1

Table 4: Comparison of loss functions on English with 10% labelled data

**Loss Function:** We evaluate Mean Square Loss (mLVAT), MSE without sigmoid and KL-divergence (KLDivergence) loss in Table 4. MSE in presence of sigmoid shows superior performance than the other loss functions. The superior performance can be attributed to the normalization of the logits which encourages more stable activations and training. For experimenting with KL-divergence, we interpreted the normalized logits as probabilities but observed substantially poorer performance. We used English language with 10% of labelled examples and XLM-Tw model for these experiments.

Ratio	1	2	3	4	5
JI	55.1	54.4	<b>55.2</b>	53.6	52.9
MiF1	66.9	66.4	<b>67.0</b>	65.8	65.3
MaF1	50.0	<b>50.9</b>	50.8	50.5	47.0

Table 5: Comparison of batch ratios on English with 10% labelled data

**Unlabelled Batch Ratio:** In Table 5, we study the impact of ratio of the batch size of the unlabelled examples while keeping the batch size of the labelled data fixed. At higher ratios, the adversarial loss overpowers the supervised learning resulting in a performance drop. However, for the lower ratios, we did not observe a consistent trend.

**Epsilon:** We study the impact of epsilon ( $\epsilon$ ) on the performance in Table 6. Higher values create more aggressive adversarial samples with high pertur-

$\epsilon$	0.1	0.25	0.5	0.75	1
JI	54.9	54.9	<b>55.2</b>	54.7	54.6
MiF1	66.7	66.8	<b>67.0</b>	66.6	66.8
MaF1	50.4	50.3	<b>50.8</b>	50.3	49.9

Table 6: Comparison of epsilon ( $\epsilon$ ) values on English with 10% labelled data

bation while lower values may create insufficient perturbation. From our empirical experiments, we note that 0.5 works better than the other values and we use this for all our semi-supervised experiments.

## 5 Domain and Task Adaptive Pretraining

In this section, we perform supervised learning experiments with frozen and finetuned representations by using the labelled data of each language for evaluating the performance of *domain-adaptive*, *task-adaptive*, monolingual and multilingual contextual models. In Table 8, 9 and 7, we present the results for different monolingual and multilingual contextual models for the three languages with frozen backbones. We use English BERT (E-BERT), BetoBERT and AraBERT as monolingual models for English, Spanish and Arabic respectively. We note that for all the languages, mBERT performs substantially poorer than the monolingual contextual models of the respective languages. However, XLM-R which is another multilingual model performs competitive with the monolingual models which is not surprising as XLM-R has shown improvements over mBERT in other language tasks also (Conneau et al., 2019).

We further evaluate *Domain-adaptive* (XLM-Tw) and *Task-adaptive* (XLM-Tw-S) versions of the XLM-R multilingual model and observe substantial improvements. XLM-Tw-S improves the Jaccard Index (JI) by **5.5%**, **6.5%** and **8.4%** for Arabic, English and Spanish respectively, highlighting the advantages of *task-specific* pretraining for contextual models. XLM-Tw also improves upon XLM-R for all the languages reiterating the importance of pretraining the contextual models with domain specific data.

We study the impact of finetuning the monolingual and best performing multilingual model on our task to compare the capabilities of multilingual models with monolingual after finetuning on the task. We notice that finetuning bridges the gap to some extent but still the domain adaptive multilingual XLM-Tw works better than the finetuned

Model	JI	MiF1	MaF1
XLM-Tw-S	<b>52.0</b>	<b>64.4</b>	<b>47.9</b>
XLM-Tw	49.3	62.2	47.1
XLM-R	45.2	58.4	42.9
mBERT	37.5	51.2	36.2
AraBERT	46.4	59.7	43.7

Table 7: Performance of pretrained models on Arabic

Model	JI	MiF1	MaF1
XLM-Tw-S	<b>53.9</b>	<b>66.2</b>	<b>47.8</b>
XLM-Tw	50.6	63.5	45.9
XLM-R	48.6	61.9	45.7
mBERT	44.7	57.6	39.2
E-BERT	48.2	61.4	42.9

Table 8: Performance of pretrained models on English

monolingual models for all the languages as shown in Table 10, 11 and 12. For English, the improvement is relatively moderate but for Spanish and Arabic, XLM-Tw demonstrates substantial gains.

### 5.1 Comparison with existing methods

**English:** Alhuzali and Ananiadou (2021) (SpanEmo) use sentences along with emotion categories as input to the contextual model and use label correlation aware loss (LCA) to model correlation among emotions classes. LVC-Seq2Emo (Huang et al., 2019) propose a latent variable chain transformation and use it with sequence to emotion for modelling correlation between emotions. BinC (Jabreel and Moreno, 2019) transform the multilabel classification problem into binary classification problems and train a recurrent neural network over this transformed setting. (Baziotis et al., 2018) (NTUA) used a Bi-LSTM architecture with self-attention models over word2vec trained on large collection of twitter tweets and were winner of the task. Huang et al. (2021) trained a sequence to emotion (Seq2Emo) encoder where the text is encoded using a bi-directional recurrent network and emotions are predicted by the decoder in an iterative fashion. Seq2Emo architecture allows for understanding the correlation between emotions. Yu et al. (2018) (DATN) use sentiments to improve emotion classification using bi-directional LSTM. Meisheri and Dey (2018) (TCS) uses SVM on manually engineered features.

**Spanish:** Mulki et al. (2018) (TW-StAR) used binary relevance transformation strategy over tweet

Model	JI	MiF1	MaF1
XLM-Tw-S	<b>51.1</b>	<b>60.0</b>	<b>48.8</b>
XLM-Tw	47.1	56.6	42.7
XLM-R	42.9	51.9	39.8
mBERT	37.0	44.8	31.2
BetoBERT	41.3	50.3	37.0

Table 9: Performance of pretrained models on Spanish

features while [González et al. \(2018\)](#) (ELiRF) explored preprocessing and adapted the tokeniser for Spanish tweets. MILAB was the wining entry in the SemEval2018 task. [Hassan et al. \(2021\)](#) (CER) finetuned the Spanish BERT representations (BetoBERT).

**Arabic:** For Arabic, [Samy et al. \(2018\)](#) (CA-GRU) extract contextual information from the tweets and uses them as context for emotion recognition using RNNs. [Hassan et al. \(2021\)](#) (CER) finetuned BERT representations. [Alswaidan and Menai \(2020\)](#) (HEF) proposed hybrid neural network using different embeddings. [Badaro et al. \(2018\)](#) (EMA) used preprocessing techniques like normalisation, stemming etc.

Overall, our results improve upon the existing approaches on Jaccard Index(JI) by **7%** for Spanish, **4.5%** for Arabic and around **1%** for English and setup a new state-of-the-art for all the three languages highlighting the efficacy of semi-supervised learning and *domain-adaptive* multilingual models.

## 5.2 Crosslingual Experiments

We combine data of all the three languages and train a combined model and test this model on the test set of each language. We notice that the combined model improves upon the performance of individual models for Arabic and Spanish (Table 13) while the performance of English is comparable.

In Table 14, we perform crosslingual experiments to evaluate the performance of a model trained on one language on another language. It is interesting to note that for Arabic and Spanish, the cross lingual performance is competitive with performance using some of the pretrained networks which is encouraging. We also observe that English demonstrates better crosslingual capability than Arabic and Spanish. A possible reason might be the large size of the English training dataset.

Model	JI	MiF1	MaF1
mLVAT	<b>60.7</b>	<b>71.9</b>	56.1
XLM-Tw	59.8	71.2	53.4
E-BERT	59.1	70.4	53.3
SpanEmo	60.1	71.3	<b>57.8</b>
LVC-Seq2Emo	59.2	70.9	-
BinC	59.0	69.2	56.4
NTUA	58.8	70.1	52.8
Seq2Emo	58.7	70.1	51.9
DATN	58.3	-	54.4
TCS	58.2	69.3	53.0

Table 10: Results on English

Model	JI	MiF1	MaF1
mLVAT	56.9	64.9	51.8
XLM-Tw	<b>57.0</b>	<b>65.3</b>	51.5
BetoBERT	52.7	60.8	48.7
SpanEmo	53.2	64.1	53.2
CER	52.4	-	<b>53.7</b>
MILAB	46.9	55.8	40.7
ELiRF	45.8	53.5	44.0
TW-StAR	43.8	52.0	39.2

Table 11: Results on Spanish

## 6 Probing Experiments

We perform experiments to evaluate the contribution of different layers of the XLM-Tw-S model. We extract representation of the tokens of a sentence from a particular layer of the contextual model and take an average across tokens for obtaining the representation of the sentence. We train a classifier over these sentence representations and report the results. From Figure 2, we note that higher layers provide better performance for all the three languages showing that the higher-order contextual information is useful for understanding the emotions in the text. Refer Appendix A for detailed results. Similar to [Tenney et al. \(2019\)](#), we also compute the improvement due to incrementally adding more layers to the previous layers and calculate the expected layer:

$$E_{\Delta}[l] = \frac{\sum_{l=1}^L l * \Delta^{(l)}}{\sum_{l=1}^L \Delta^{(l)}} \quad (8)$$

where,  $\Delta^{(l)}$  is the change in the Jaccard Index metric when adding layer  $l$  to the previous layers. We start from layer 0 and incrementally add higher layers for representing the tokens of the sentence

Model	JI	MiF1	MaF1
m1VAT	<b>57.3</b>	<b>68.2</b>	<b>52.2</b>
XLM-Tw	55.8	67.4	50.1
AraBERT	54.3	65.9	49.0
SpanEmo	54.8	66.6	52.1
CA-GRU	53.2	64.8	49.5
CER	52.9	-	48.9
HEF	51.2	63.1	50.2
EMA	48.9	61.8	46.1

Table 12: Results on Arabic

Language	JI	MiF1	MaF1
EN	59.4	70.6	55.7
ES	57.8	65.8	56.6
AR	57.8	68.6	55.5

Table 13: Experiments on the combination of languages

followed by averaging for representing the whole sentence. The expected layer for English, Spanish and Arabic computes to 6.9, 6.2 and 6.8 respectively showing that higher layers are useful for the task. This analysis is helpful to understand the improvement achieved by adding layers to the previous layers. For all the three languages, we obtain the best results on using the average of all the layers for representing the sentences which shows that different layers encapsulate complementary information about emotions.

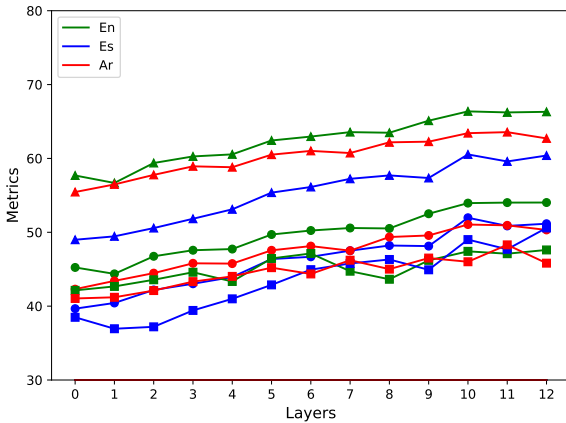


Figure 2: Performance metrics across different layers for XLM-Tw-S. Circle, Triangle and Square represent JI, MiF1 and MaF1 respectively.

Train/Eval	JI	MiF1	MaF1
Es → En	42.9	54.6	42.6
Ar → En	39.2	51.7	42.0
En → Ar	49.7	62.3	45.3
Es → Ar	46.4	57.2	44.9
En → Es	44.6	55.9	41.5
Ar → Es	40.0	51.1	41.9

Table 14: Crosslingual experiments between the languages

Language	Train	Dev	Test
English	6838	886	3259
Arabic	2278	585	1518
Spanish	3561	679	2854

Table 15: SemEval2018 dataset statistics

## 7 Training Details

We finetune the contextual models following huggingface<sup>2</sup> with a batch size of 8, learning rate of  $2e-5$  and weight decay of 0.01 using AdamW optimizer for 30 epochs. The classifier is a two layered neural network with 768 hidden dimensions and 11 output dimensions with 0.1 dropout. For m1VAT experiments, the number of examples sampled from the unlabelled set for each batch are 24,  $\epsilon$  and  $\alpha$  are set to 0.5 and 1 using cross validation. We apply sigmoid over the logits and train using binary cross entropy loss. We use validation set for finding optimal hyperparameters and evaluate on the test set using combination of training and validation set for training.

## 8 Conclusion

In this work, we explored semi-supervised learning using Virtual Adversarial Training (VAT) for multi-label emotion classification in a multilingual setup and showed performance improvement by leveraging unlabelled data from different languages. We used Mean Square Error (MSE) as the divergence measure for leveraging VAT for multilabel emotion classification. We also evaluated the performance of monolingual, multilingual and *domain-adaptive* and *task-adaptive* multilingual contextual models across three languages - English, Spanish and Arabic for multilabel and multilingual emotion recognition and obtained state-of-the-art results. We also performed probing experiments for understanding the impact of different layers of contextual models.

<sup>2</sup><https://huggingface.co/>



## 9 Broader Impact and Discussion of Ethics

In recent years, deep learning approaches have played an important role in state-of-the-art natural language processing systems. However, obtaining labelled data for training these models is expensive and time consuming, especially for multilingual and multilabel scenarios. In such case, multilingual semi-supervised and unsupervised techniques can play a pivotal role. Our work introduces a semisupervised way for detecting and understanding textual data across multiple languages. Our methods could be used in sensitive contexts such as legal or healthcare settings, and it is essential that any work using our probe method undertake extensive quality assurance and robustness testing before using it in their setting. The datasets used in our work do not contain any sensitive information to the best of our knowledge.

## References

- Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. *arXiv preprint arXiv:2101.10038*.
- Nourah Alswaidan and Mohamed El Bachir Menai. 2020. Hybrid feature model for emotion recognition in arabic text. *IEEE Access*, 8:37843–37854.
- Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.
- Gilbert Badaro, Obeida El Jundi, Alaa Khaddaj, Alaa Maarouf, Raslan Kain, Hazem Hajj, and Wassim El-Hajj. 2018. [EMA at SemEval-2018 task 1: Emotion mining for Arabic](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana. Association for Computational Linguistics.
- Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33.
- Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2021. A Multilingual Language Model Toolkit for Twitter. In *arXiv preprint arXiv:2104.12250*.
- Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. Ntua-slp at semeval-2018 task 1: Predicting affective content in tweets with deep attentive rnns and transfer learning. *arXiv preprint arXiv:1804.06658*.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PMLADC at ICLR 2020*.
- Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. *arXiv preprint arXiv:1906.02443*.
- Nancy Chinchor. 1992. [MUC-4 evaluation metrics](#). In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.
- Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. 2018. Semi-supervised sequence modeling with cross-view training. *arXiv preprint arXiv:1809.08370*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.
- Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the international conference recent advances in natural language processing ranlp 2013*, pages 198–206.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Paul Ekman. 1984. Expression and the nature of emotion. *Approaches to emotion*, 3(19):344.
- Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. [Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

- Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. [Affect-LM: A neural language model for customizable affective text generation](#). Vancouver, Canada. Association for Computational Linguistics.
- Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative methods for multi-labeled classification. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 22–30. Springer.
- José-Ángel González, Lluís-F. Hurtado, and Ferran Pla. 2018. [ELiRF-UPV at SemEval-2018 tasks 1 and 3: Affect and irony detection in tweets](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 565–569, New Orleans, Louisiana. Association for Computational Linguistics.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.
- Sabit Hassan, Shaden Shaar, and Kareem Darwish. 2021. Cross-lingual emotion detection. *arXiv preprint arXiv:2106.06017*.
- Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, Lili Mou, and Osmar R Zaiane. 2021. Seq2emo: A sequence to multi-label emotion classification model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4717–4724.
- Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, and Osmar R Zaiane. 2019. Seq2emo for multi-label emotion classification based on latent variable chains transformation. *arXiv preprint arXiv:1911.02147*.
- Mohammed Jabreel and Antonio Moreno. 2019. A deep learning-based approach for multi-label emotion classification in tweets. *Applied Sciences*, 9(6):1123.
- Samuli Laine and Timo Aila. 2016. Temporal ensemble for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159.
- Hardik Meisheri and Lipika Dey. 2018. [TCS research at SemEval-2018 task 1: Learning robust representations using multi-attention architecture](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 291–299, New Orleans, Louisiana. Association for Computational Linguistics.
- Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.
- Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.
- Saif Mohammad. 2012. [#emotional tweets](#). In *\*SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.
- Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. [SemEval-2018 task 1: Affect in tweets](#). New Orleans, Louisiana. Association for Computational Linguistics.
- Hala Mulki, Chedi Bechikh Ali, Hatem Haddad, and Ismail Babaoğlu. 2018. [Tw-StAR at SemEval-2018 task 1: Preprocessing impact on multi-label emotion classification](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 167–171, New Orleans, Louisiana. Association for Computational Linguistics.
- Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.
- Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

- Parsa Saadatpanah, Ali Shafahi, and Tom Goldstein. 2020. Adversarial attacks on copyright detection systems. In *International Conference on Machine Learning*, pages 8307–8315. PMLR.
- Devendra Singh Sachan, Manzil Zaheer, and Ruslan Salakhutdinov. 2019. Revisiting lstm networks for semi-supervised text classification via mixed objective function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6940–6948.
- Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26.
- Ahmed E Samy, Samhaa R El-Beltagy, and Ehab Hassanien. 2018. A context integrated model for multi-label emotion detection. *Procedia computer science*, 142:61–71.
- Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.
- Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovered the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.
- Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440.
- Chaowei Xiao, Ruizhi Deng, Bo Li, Fisher Yu, Mingyan Liu, and Dawn Song. 2018. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 217–234.
- Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. 2017. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33.
- Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. **SGM: Sequence generation model for multi-label classification**. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics.
- Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. 2021. A survey on deep semi-supervised learning. *arXiv preprint arXiv:2103.00550*.
- David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.
- Jianfei Yu, Luis Marujo, Jing Jiang, Pradeep Karuri, and William Brendel. 2018. Improving multi-label emotion classification via sentiment classification with dual attention transfer network. *ACL*.
- Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freeb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.

## A Probing Experiments

Layers	JI	MiF1	MaF1
Layer 0	45.24	57.69	42.14
Layer 1	44.37	56.68	42.67
Layer 2	46.75	59.38	43.57
Layer 3	47.57	60.27	44.59
Layer 4	47.73	60.55	43.34
Layer 5	49.70	62.43	46.46
Layer 6	50.24	62.96	47.13
Layer 7	50.58	63.56	44.73
Layer 8	50.52	63.48	43.64
Layer 9	52.51	65.11	46.20
Layer 10	53.95	66.37	47.42
Layer 11	54.02	66.23	47.10
Layer 12	54.03	66.31	47.61

Table 16: Comparison of layer performance for English using XLM-Tw-S model

Layers	JI	MiF1	MaF1
Layer 0	39.66	48.99	38.49
Layer 1	40.43	49.45	36.94
Layer 2	42.19	50.57	37.20
Layer 3	43.03	51.83	39.42
Layer 4	43.94	53.11	40.99
Layer 5	46.38	55.37	42.88
Layer 6	46.68	56.13	44.94
Layer 7	47.51	57.24	45.78
Layer 8	48.21	57.70	46.32
Layer 9	48.13	57.35	44.92
Layer 10	51.97	60.54	49.01
Layer 11	50.86	59.59	47.70
Layer 12	51.16	60.39	50.59

Table 17: Comparison of layer performance for Spanish using XLM-Tw-S model

Layers	JI	MiF1	MaF1
Layer 0	42.30	55.45	41.04
Layer 1	43.42	56.48	41.20
Layer 2	44.47	57.77	42.11
Layer 3	45.80	58.93	43.32
Layer 4	45.76	58.81	44.03
Layer 5	47.56	60.51	45.21
Layer 6	48.13	61.02	44.36
Layer 7	47.51	60.73	46.22
Layer 8	49.36	62.18	45.01
Layer 9	49.57	62.27	46.51
Layer 10	51.05	63.42	46.01
Layer 11	50.94	63.56	48.30
Layer 12	50.32	62.71	45.82

Table 18: Comparison of layer performance for Arabic using XLM-Tw-S model

In Table 16, 17 and 18, we report the performance of each layer of frozen XLM-Tw-S model. We extract the layer representation of each token of the sentence and average them for representing the sentence. For all the languages, we note that the higher layers show superior performance.