Multi-domain Emotion Detection using Transfer Learning

Anonymous ACL submission

Abstract

The task of emotion detection in text, particularly in informal and spontaneous messaging, such as email, posts, or tweets, varies in its scope and depth depending upon the requirements of the end application as well as the domain of use. The most popular emotion categories reported in research include the Ekman's or Plutchik's emotion models (Ekman, 1999), (Plutchik, 1984), but often the application domain requires a more specialized emotion categorization, for which there are insufficient annotated datasets available for training. It is additionally complicated by different perceptions and definitions of emotion labels in different domains. The popularity of empathetic systems across a wide range of industries and applications has given rise to the the task of multi-domain emotion detection to increase its adaptability and resiliency across domains. In this paper, we present a generalized approach of emotion detection that can be adapted to any domain and any set of emotion labels with minimal loss in performance. The `multi-domain-emotion` model could be plugged into any emotion detection application without any further training or fine-tuning. We show the zero-shot and few-shot performance of our approach on the publicly available SemEval 2018 dataset and also a new dataset consisting of tweets related to the French elections in 2017. This approach demonstrates good performance in predicting emotion categories previously unseen to the model, including domains different than those on which the model was originally trained. We further propose a few ways to boost the model performance with the availability of a small annotated dataset in the target domain.

1 Introduction

Language is an extremely powerful tool to both express emotion and arouse an emotional response in the audience. Therefore, tools which can effectively analyze the emotional content of text are being used in diverse applications ranging from healthcare (Tivatansakul et al., 2014) and education (Karan et al., 2022) to stock market (Aslam et al., 2022) and political opinion mining (Cabot et al., 2020). But which emotions matter? Clearly, the emotions that may accompany discussions on a new electronic gadget on the market are not quite the same that may arise when comparing political candidates ahead of an election. Depending upon the domain and the context, different sets of emotions may need to be detected.

In recent research, many emotion labeled datasets have been constructed to serve as training data for emotion classification models. Among these datasets, many have emotion label sets which are supersets or subsets of Ekman's or Plutchik's emotion models (Ekman, 1999; Plutchik, 1984). For example, the Cleaned Balanced Emotional Tweets dataset has labels for the six Ekman emotions as well as *love*, *thankfulness*, and *guilt* (Shahraki and Zaiane, 2017), whereas the EmoInt dataset has only four of the six Ekman emotions, leaving out *disgust* and *surprise* (Mohammad and Bravo-Marquez, 2017). As a result, while there is plenty of emotion labeled text data, many of the datasets are incompatible and thus difficult to use for training of a single model. Additionally, when a novel emotion detection problem arises in a domain for which a new label set is more appropriate or desirable and this new label set is not a subset of any existing emotion label set, we face a situation where no training data is available. For such new problems, possible solutions involve curating new datasets with the relevant label set, using semi-supervised or unsupervised techniques, or using zero-shot and few-shot approaches. Existing works in zero-shot emotion detection frame the task as a textual entailment problem (Yin et al., 2019) or utilize the embeddings of the input text and class labels and descriptions for classification (Chen et al., 2022; Zhang et al., 2019). The usage of only the

emotion labels or their definitions from external sources like WordNet does not integrate the understanding of the concept of each emotion label or its underlying intricacies in the application domain. Several works in psychological theories suggest that no emotion definition is universal across domains or people, they are strongly influenced by socio-cultural context and events (Averill, 1980). Such approaches also fail to capture the relationships and inter-dependencies that comprise more complex emotions like *anticipation* and *guilt*.

In this paper, we propose a novel zero-shot approach to emotion detection from text to build a generalized emotion detection model that can be adapted to any unseen domain or target label set. Our method carefully incorporates the interpretations of each label and utilizes their interdependencies to produce more valuable results in the target domain. Following are the steps in our multi-domain transfer learning approach: First, creation of a `multi-domain-emotion` model using a hierarchical structure of preexisting emotion-labeled social media datasets and optimization strategies. Second, development of a weighted linear combination of the outputs of this model to any desired emotion label set. Third (optionally), improve target domain performance by fine-tuning the combination weights and classification thresholds using any in-domain annotated data.

Overall, the contributions of this paper are:

- Development of a generalized emotion detection model for tweets that can be deployed across multiple domains
- A transfer learning method for adaption of the generalized model across unseen applications or domains
- A well-defined methodology to define complex or specialized emotion labels in terms of existing ones
- Multiple ways to boost the zero-shot performance of the model with the availability of in-domain annotated data

2 Related Work

2.1 Emotion Taxonomies

Research on human emotions has led to the development of various ways to dichotomize emotions. Discrete models describe emotions as a set of distinct classes. Notably, Ekman's basic emotions, *joy, sadness, fear, anger, disgust, and surprise* and Plutchik's wheel of emotions, which describes eight basic emotions in pairs of opposites: *joy* and *sadness*, *anger* and *fear*, *trust* and *disgust*, and *surprise* and *anticipation* are popular baselines of much emotion-related research (Ekman, 1999; Plutchik, 1984). Dimensional models like the Circumplex model of affect (Russell, 1980) characterize emotions as regions within a continuous space of emotional response dimensions. With the advancement of research in this field, newer emotion taxonomies specific to the application domains have been developed (Menninghaus et al., 2019; Oberländer et al., 2020). Therefore, the problem of choosing an appropriate taxonomy for an emotion classification task is strongly application dependent.

2.2 Emotion Detection from Text

Emotion detection from text has been a long-standing research problem due to the evolving nature of textual content over various applications and platforms and the complexities of modeling human emotions. Some early approaches include the use of lexicons like WordNet-Affect (Strapparava et al., 2004), NRC (Mohammad and Turney, 2010) or popular machine learning algorithms like Support Vector Machine or Naive Bayes classifiers (Mashal and Asnani, 2017; Hasan et al., 2019). With the availability of large emotion-annotated corpora, large pretrained language models like GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) have become the most powerful tools for this task (Cai and Hao, 2018; Huang et al., 2019; Polignano et al., 2019; Ma et al., 2019; Chiorrini et al., 2021). These models are first pretrained on large, unlabeled text corpora, and then fine-tuned with task-specific annotated data for downstream tasks. We utilize some popular Twitter-specific language models in our approach that have served as a strong baseline for core NLP tasks around social media analysis (Barbieri et al., 2020).

2.3 Zero-Shot Learning

Zero-shot learning entails prediction, at test time, of classes unseen by the model during training, and was first introduced in (Larochelle et al., 2008). Although no training examples of these classes exist, information about these classes are utilized to aid in the classification task. Often in applications of emotion detection tasks there is no training data available or new emotion labels have been developed in the target domain. Recent works in zero-

shot emotion detection formulate the problem using the text entailment approach where each target label is used to create a hypothesis for the model (Yin et al., 2019; Basile et al., 2021). Prompt engineering techniques have also been used to infer the correct emotion label from pre-trained NLI models (Plaza-del Arco et al., 2022). Another category of zero-shot approaches uses sentence embeddings to perform unsupervised or semi-supervised predictions on unlabeled datasets (Chen et al., 2022; Zhang et al., 2019; Olah et al., 2021). The drawback of these approaches are that they have been generalized to perform across all domains and thus perform well only when the target emotion labels match popular definitions. They do not integrate any domain knowledge or understanding of the emotion concepts that may arise in a specialized domain.

3 Methodology

3.1 Problem statement

Our task is to label a tweet x with scores between 0 and 1 for each emotion label in a predefined set of emotions $E = \{e_1, e_2, \ldots e_n\}$. The score for each label $e \in E$ should reflect the confidence that the emotion e is expressed by the author of the tweet x. The set E is dependent on the application and pre-determined by domain experts.

3.2 Approach

Our approach involves producing hierarchical scores for a tweet x over three sentiment categories, the six Ekman emotions, and their fine-grained subcategories defined in (Demszky et al., 2020). The components in the model ensemble can be used to produce these scores without any further training or fine-tuning. To obtain confidence scores over emotions in E, we design a many-to-one mapping from these outputs to the set E, based on domain expertise, the definition of each emotion label and understanding of the relationships in the dimensional models of affect (Plutchik, 1984). As E changes based on the requirements of the application, the first step remains the same, but the mapping from the model outputs to E is updated. We illustrate our emotion model ensemble in Fig.1.

3.3 Datasets and Preprocessing

The following datasets have been used for training and evaluation of our model ensemble:

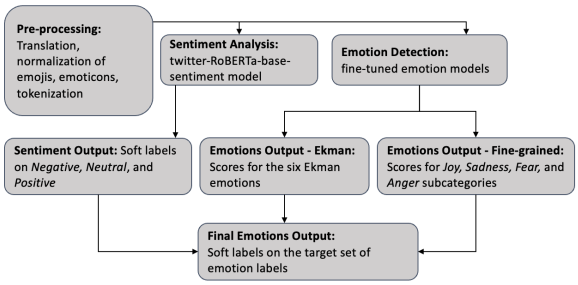

Figure 1: Ensemble Emotion Detection Architecture

Cleaned Balanced Emotional Tweets (CBET) (Shahraki and Zaiane, 2017) is a collection of 81k English tweets that have been collected using a set of hashtags corresponding to the nine emotion labels (*anger, fear, joy, love, sadness, surprise, thankfulness, disgust*, and *guilt*). The dataset has been balanced by utilizing more than one hashtag for each emotion label and finally having an equal number of tweets for each label. We use this dataset to fine-tune a model to predict scores over the six Ekman emotions, removing the annotations for *thankfulness*, *disgust*, and *guilt*. The 56,281 remaining tweets that have at least one nonzero label have been used for fine-tuning. The dataset is split randomly into training (81%), validation (9%), and testing (10%) sets.

GoEmotions (Demszky et al., 2020) is a corpus of 58k English Reddit comments manually annotated with 27 emotion labels or Neutral. The large number of fine-grained emotion labels in this dataset makes it an ideal choice to be used in our task of creating a base emotion model that can be use to build any downstream specialized emotion label set. A series of data curation steps have been carried out while building this dataset to balance the classes and remove the predominant issues usually present in Reddit data (Ferrer et al., 2021). Offensive/adult tokens were removed, and identity and religion terms were masked using predefined lists. Comments that represent gender and ethnic biases were filtered manually. For each text in GoEmotions, a 7-dimensional one-hot vector is created to produce the Ekman output vector. Similarly, for the emotion labels *joy, sadness, fear* and *anger*, we identify their fine-grained outputs using the subcategories prescribed in GoEmotions to produce the training, validation, and testing sets (Table 1) for each lower level emotion model in the hierarchy.

Given an English tweet as input, our system first performs some basic text preprocessing. User-

Model	Training	Validation	Test
joy	17,410	2,219	2,104
sadness	3,263	390	379
fear	726	105	98
anger	5,579	717	726

Table 1: Distribtution of training, validation, and test sets for emotion subcategory models derived from GoEmotions

Model	Output Labels
Sentiment(Sent)	positive, neutral, negative
CBET-Ekman	joy, sadness, fear, anger, disgust, surprise
GE-Ekman	joy, sadness, fear, anger, disgust, surprise
Joy(J)	joy, amusement, approval, excitement, gratitude, love, optimism, relief, pride, admiration, desire, caring
Sadness(S)	sadness, disappointment, embarrassment, grief, remorse
Fear(F)	fear, nervousness
Anger(A)	anger, annoyance, disapproval

Table 2: Set of output labels for each component model

names, retweet IDs and hyperlinks are removed, while emojis are converted to plain text . The preprocessing pipeline is used as a social tokenizer (Baziotis et al., 2017) to remove any hyperlinks, emails, phone numbers, times, dates, and percentages, normalize money values and numbers, annotate any censored or elongated words, and convert complex emoticons to plain text.

3.4 Training and Fine-tuning

For the task of sentiment analysis, we use the twitter-XLM-RoBERTa-base-sentiment model [1] to produce normalized values on the three sentiment categories *negative, neutral*, and *positive*. This model is a RoBERTa base model pre-trained on approximately 198 million tweets and fine-tuned for the task of multilingual sentiment analysis, and achieved a higher performance in comparison to FastText, SVM, and bi-LSTM baselines (Barbieri et al., 2020).

For emotion detection, we use the twitter-RoBERTa-base-emotion pretrained model [2], as a base (Barbieri et al., 2020). We append a dense output layer with a softmax activation function on top of the transformer layer of the pretrained model, with the number of nodes equal to the number of labels in the corresponding dataset. In total, we train six transformer-based models as components to the hierarchical mapping system. First, two models are fine-tuned to output normalized scores on the six Ekman emotions using the CBET Twitter data and GoEmotions Reddit data. We choose to train separate models on Twitter and Reddit data to be able to weigh them in the next step based on the target domain. The remaining four models are fine-tuned to output scores on the subcategories of *joy, sadness, fear, anger*. The fine-tuning setup and metrics for each model are described in Appendix B. To summarize, our emotion classification model ensemble produces scores for each of the fine-grained labels in Table 2. The next section describes how these fine-grained scores are utilized downstream to adapt the model to any new domain.

3.5 Domain-Specific Hierarchical Label Transfer

We map the scores from the model outputs to scores over a desired label set E using a weighted linear combination derived by considering the relatedness of emotions in the Plutchik's wheel of emotions (Plutchik, 1984) and understanding of the label definitions in the target domain. A general set of rules to determine the mapping from the hierarchical emotion model outputs to the any emotion $e \in E$ is as follows:

1. Determine which sentiment categories $S \subseteq Sent$ correspond to emotion e. Usually, this is either *positive* or *negative* (Example: *anger* => *negative*). However, in some cases, an emotion can have *positive* or *negative* sentiment based on the context.

2. The output Ekman scores from the CBET-Ekman and GoEmo-Ekman models have been weighed using a linear combination based on the target domain to produce one output score *EK* for each label.

3. For each sentiment $s \in S$, determine which high-level Ekman emotions corresponding to s, $EK_s \subseteq EK$ have subcategories relevant to emotion e. For example, the output emotion

[1] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

[2] https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion

optimism is *positive*, and the Ekman emotion *joy* has a subcategory *optimism* which is relevant to the output emotion.

4. For each high-level Ekman emotion $ek \in EK_s$, if ek has subcategories, determine which subcategories $sub_{ek} \subseteq Sub_{ek}$ are relevant to emotion e. For example, for the output emotion *optimism*, out of all the *joy* subcategories, the only relevant subcategory is *optimism*.

5. Then, the score of e is

$$\sum_{s \in S}\Big(\sum_{ek \in EK_S}\Big(\sum_{sub_{ek} \in Sub_{ek}}(w_{s,ek,sub_{ek}} (Sent[s] * EK[ek] * Sub_{ek}[sub_{ek}])\Big)\Big)\Big),$$

where $w_{s,ek,sub_{ek}}$ is a weight that can be set to 1, or fine-tuned to maximize a performance metric on a target-domain validation set (if one exists). In other words, the final score for e is a weighted sum of terms, where each term is the product of scores for a sentiment, Ekman emotion, and low-level emotion subcategory triple that is relevant to e. For example, for the output emotion *optimism*, we may have the term $(Sent[positive] * EK[joy] * Joy[optimism])$.

Further, any available in-domain datasets can be used as a validation set for two purposes: 1) find a set of optimal classification thresholds for each emotion label, 2) fine tune the weights of the linear mapping of the emotion scores for a target metric. We fine-tune the classification thresholds by choosing a threshold for each target class to maximize the F1 score on that class over the validation dataset.

We fine-tune the mapping weights by successively applying differential evolution to each individual target label mapping to maximize the F1 score on that label over the validation dataset (Storn and Price, 1997). We fine-tune both the mapping weights and the classification thresholds by first optimizing the weights, and subsequently choosing the thresholds for each label. More details on the label-wise classification thresholds and mapping weights parameters have been listed in Appendix C, along with examples.

The next sections illustrate some applications and evaluation of these general set of rules across two different domains to show their efficacy in producing scores for any new set of emotion labels.

4 Experiments

In this section, we outline the experiments carried out to evaluate our approach on a benchmark emotion dataset which contains a larger label set than the regular Plutchik or Ekman emotions. To further illustrate the adaptability of our method across domains and labels, we conduct a second set of experiments on the French election dataset (Daignan, 2017) which has been annotated with a specialized set of emotion labels. We explain how the `multi-domain-emotion` model has been adapted to these unseen domains and emotion labels. There are several methods available for emotion classification as mentioned in Section 2, but all of them require in-domain training to achieve the SOTA scores. We compare our approach against popular semi-supervised and zero-shot techniques. Our approach stands out as it produces stable performance across any domain with no training data and strong results with the availability of a small in-domain dataset. We perform the below experiments for evaluation:

- Zero-shot mode: Emotion classification on the test set by adapting the model ensemble to the target domain. We also repeat this experiment without the sentiment component in the ensemble to demonstrate its contribution.
- In-domain fine-tuning mode: Use a small subset of available in-domain data to fine-tune the classification thresholds and mapping weights.

4.1 Baselines

We analyze the results of our model against the following baselines:

- Zero-shot textual entailment: Following the work of Yin et al., 2019, we convert each emotion label into the hypothesis: "This text expresses <label>." We use the BART MNLI [3] model to generate entailment and contradiction scores and threshold them to produce binary outputs for each label.
- Zero-shot sentence embeddings: We use SBERT (Reimers and Gurevych, 2019) to obtain input and label embeddings. Tweets are then labeled based on their closeness to the labels in the embedding space using cosine similarity.

[3] `https://huggingface.co/facebook/bart-large-mnli`

Mapping	Output Label
EK[anger] * Sent[negative]	anger
(EK[joy] * J[optimism] * Sent[positive]) + (EK[fear] * F[nervousness] * Sent[negative])	anticipation
EK[disgust] * Sent[negative]	disgust
(EK[fear] * F[fear]) * Sent[negative]	fear
(EK[joy] * J[joy]) * Sent[positive]	joy
(EK[joy] * (J[love] + J[desire] + J[caring])) * Sent[positive]	love
(EK[joy] * J[optimism]) * Sent[positive]	optimism
(EK[fear] * F[nervousness]) * Sent[negative]	pessimism
EK[sadness] * Sent[negative]	sadness
EK[surprise] * max(Sent)	surprise
(EK[joy] * (J[approval] + J[admiration])) * Sent[positive]	trust

Table 3: Mapping of model outputs to SemEval 2018 labels

- Semi-supervised models: We use existing emotion datasets (CBET and GoEmotions) to fine-tune twitter-RoBERTa-base-emotion pretrained models (Barbieri et al., 2020) on the six Ekman labels, and test these models over the label set in the target domain. The outputs for emotions outside of the label set of these models are set to 0.

4.2 SemEval 2018 Task 1e

We choose a popular open source dataset that has been used for multiple emotion labeling tasks: the SemEval 2018 Task 1E-c dataset (Mohammad et al., 2018). Given an input tweet, the goal is to classify it into one of the 11 emotion categories that best represents the emotions of the author. The test dataset contains 3200 English tweets, and 800 tweets have been used to fine-tune the model in the fine-tuning domain as compared to the 7800 tweets available for training in supervised approaches.

We derive a mapping from the output scores of Table 2 to the target label set $E = \{$*anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust*$\}$. The mapping described in Table 3 follows the rules outlined in the previous section, for all target emotions that can be clearly associated to one sentiment. However, when a target label like *surprise* has an ambiguous sentiment, the intuition is to associate it with the most prevalent sentiment in the text and use the mapping $EK[surprise] * max(Sent)$. For example, if $EK[surprise]$ is large and $Sent[positive]$ is the highest of the three sentiment scores, we interpret the *surprise* as positive surprise.

4.3 French Election Dataset

For our next experiment, we use an annotated dataset on the 2017 French presidential election tweets. We note that for this domain, there were no pre-existing available emotion annotated datasets. The experiments have been carried out on the Kaggle dataset (Daignan, 2017), a subset of which were annotated with the set of emotion labels $E = \{$ *anger, embarrassment, admiration, optimism, joy, pride, fear, amusement, positive-other, negative-other*$\}$. It is to be noted that each label was provided with a description and a set of synonymous emotion labels (Appendix A), which further complicates the emotion taxonomy to be used for this task.

The mapping in Table 4 to the destination set E is carried out by domain expertise and the general rules formulated in the previous section. For example, each label in *anger/hate/contempt/disgust* is associated with a *negative* sentiment. Further, for the Ekman emotions *anger* and *disgust*, the only relevant subcategory is *anger*, which results in the final mapping *((EK[anger] * Anger[anger]) + EK[disgust]) * Sentiment[negative]*. The label *positive-other* is associated with a *positive* sentiment and the only positive Ekman emotion *joy*. Additionally, from the label definition, it accumulates scores of all the positive fine-grained emotions that have not been recorded by any other label. Figure 2 illustrates an example tweet from this dataset with its corresponding emotion scores.

5 Results and Analysis

The results of the semi-supervised experiments show how existing emotion datasets can be utilized

Mapping	Output Label
((EK[anger] * A[anger]) + EK[disgust]) * Sent[negative]	anger/contempt/disgust
(EK[sadness] * (S[sadness] + S[embarrassment] + Sent[grief])) * Sent[negative]	embarrassment/guilt
(EK[joy] * (J[admiration] + J[love])) * Sent[positive]	admiration/love
(EK[joy] * (J[optimism])) * Sent[positive]	optimism/hope
(EK[joy] * (J[joy])) * Sent[positive]	joy/happiness
(EK[joy] * (J[pride])) * Sent[positive]	pride
(EK[fear] * (F[fear])) * Sent[negative]	fear/pessimism
(EK[joy] * (J[amusement])) * Sent[positive]	amusement
(EK[joy] * (J[approval] + J[excitement] + J[gratitude] + J[relief] + J[desire] + J[caring])) * Sent[positive]	positive-other
((EK[sadness] * (S[disappointment] + S[remorse])) + (EK[fear] * (F[nervousness])) + (EK[anger] * (A[annoyance] + A[disapproval]))) * Sent[negative]	negative-other

Table 4: Mapping of model outputs to French election labels

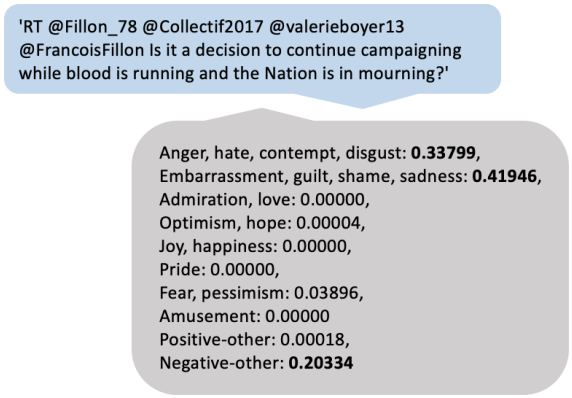

Figure 2: Example tweet from the French election dataset

to predict emotions in a new domain. For SemEval, most of the target labels are present in the GoEmotions and CBET datasets, and the performance on the six Ekman labels is higher than the zero-shot performance of our proposed model (Table 6). This suggests that there is sufficient overlap between the underlying meaning of each emotion label between the datasets to maintain performance. On the other hand, for the French Election dataset, the labels are new or combinations existing ones: some which are Ekman emotions and some which are more fine-grained (Table 7). The performance drops significantly, which suggests that the meanings of these emotions, as interpreted by the annotators, are not consistent with those in the training datasets. To integrate this knowledge, we rely on mappings based on emotion theory and label descriptions. These results show that simply using data from a different domain to predict emotions in a new domain can only be used in applications where the emotion label sets are not entirely novel and have similar definitions across datasets.

The zero-shot experiments on both the datasets demonstrate the adaptability of the emotion model ensemble across any unseen domain (Table 5). Existing zero-shot approaches perform better on SemEval, but fail to maintain the performance on a completely new set of emotion labels in the French election dataset. We can attribute the reason to the fact that the SemEval labels are a direct superset of the popular Ekman emotions. The underlying meaning of each corresponding emotion label is straightforward and thus can be easily detected by textual entailment or sentence embedding methods using a large pre-trained model. The French election labels are much more unusual and are grouped into label categories based on the target domain. For example, the labels *love* and *admiration* can be synonymous in a political influence campaign but not in a general emotion taxonomy. In Tables 6 and 7, we see that semi-supervised and zero-shot methods score high on the *joy* category in the SemEval dataset, but do not score high on the *joy/happiness* category in the French Election dataset, whereas our model maintains relatively stable performance. Our emotion model ensemble carefully integrates specialized label definitions and relationships into the emotion classification task which makes it stand out among general zero-shot classification methods. Further, the addition of the sentiment model to the ensemble improves the scores across all ex-

	SemEval			French Election		
	P	R	F1	P	R	F1
Semi-supervised						
CBET	0.62	0.30	0.41	0.05	0.07	0.06
GoEmotions	0.69	0.33	0.44	0.05	0.08	0.06
CBET + GoEmotions (EK)	0.71	0.35	0.47	0.06	0.09	0.07
Zero-shot						
BART MNLI (TE)	0.37	0.79	**0.50**	0.13	0.86	0.23
SBERT (SB)	0.28	0.78	0.41	0.10	0.65	0.17
Ours	0.59	0.27	0.37	0.32	0.44	0.37
Ours + Sentiment	0.59	0.28	0.38	0.34	0.48	**0.40**
Few-shot						
Ours + Sentiment + fine tune mapping	0.73	0.37	0.49	0.34	0.48	**0.39**
Ours + Sentiment + fine tune threshold	0.47	0.64	0.54	0.29	0.29	0.29
Ours + Sentiment + fine tune both (Ours*)	0.45	0.71	**0.55**	0.34	0.29	0.32
Supervised SOTA	-	-	0.71	-	-	-

Table 5: Evaluation results against baselines on SemEval and French Election dataset. Supervised SOTA results have been obtained from Alhuzali and Ananiadou, 2021. The highest F1 scores in each category are in **bold**.

periments which ascertain that the influence of sentiment is crucial for emotion detection tasks.

For the in-domain fine-tuning mode, although the validation dataset used for SemEval is approximately 12% of the size of the training dataset used in supervised approaches, it boosts the model performance by 44%. For the French election dataset, the ambiguity caused by grouping multiple emotions in one label results in very low inter-annotator agreement and inconsistencies in annotation between the validation and test datasets, which were also provided to us at different times. We believe that with more consistent annotations or sampling fine-tuning data from the same dataset would result in a performance boost similar to SemEval.

	EK	TE	SB	Ours	Ours*
anger	0.61	0.69	0.52	0.62	0.66
anticipation	0	0.26	0.24	0.13	0.26
disgust	0.36	0.70	0.48	0.38	0.64
fear	0.53	0.39	0.36	0.45	0.58
joy	0.8	0.79	0.66	0.32	0.83
love	0	0.55	0.36	0.32	0.50
optimism	0	0.66	0.52	0.09	0.68
pessimism	0	0.30	0.24	0.03	0.21
sadness	0.64	0.65	0.49	0.64	0.64
surprise	0.23	0.11	0.11	0.21	0.19
trust	0	0.14	0.10	0.15	0.11

Table 6: F1 scores across all emotion labels in SemEval

	EK	TE	SB	Ours
anger/cont/disgust	0.17	0.13	0.13	0.23
embarrass/guilt	0.05	0.03	0.04	0.19
admiration/love	0	0.04	0.04	0.15
optimism/hope	0	0.22	0.16	0.30
joy/happiness	0.04	0.04	0.03	0.16
pride	0	0.07	0.07	0.17
fear/pessimism	0.10	0.07	0.06	0.18
amusement	0	0.14	0.14	0.14
positive-other	0	0.56	0.43	0.50
negative-other	0	0.53	0.41	0.50

Table 7: F1 scores across all emotion labels in the French Election dataset

6 Conclusion

We present an emotion classification approach for social media text that can be adapted to any domain regardless of the target set of labels, The model does not require any in-domain training data or fine-tuning steps, although utilizing some in-domain data for fine-tuning can improve its performance. The user has to carefully map the hierarchical fine-grained emotion and sentiment scores accounting for differences in the underlying meanings of emotions between label sets. We have demonstrated the idea with the help of two such mappings to new target label sets. Our experiments indicate that using universal zero-shot models across domains and datasets may not always be sufficient to detect novel target labels, and there are methods of integrating domain knowledge into classification tasks without training or fine-tuning the models.

7 Limitations

Based on our experiments, we see that our approach can be successfully applied to various target domains for English tweets. All the pre-trained models are trained on English and thus would not generalize well to a multilingual setting. Future work would include using multilingual pre-trained models like XLM-RoBERTa and produce emotion annotated training data in non-English languages to build the emotion model ensemble. Additionally, we note that our approach assumes that the user has strong and specific definitions for target labels; the approach depends on the quality of the label mapping as well as the quality of the available fine-tuning data. The annotations on the French Election dataset were carried out by a different group and our results rely on the ground truth provided to us. We also aim to carry out in house annotations by experts to release a publicly available dataset annotated with emotions in the political domain and our `multi-domain-emotion` model which would further enhance our analysis.

Acknowledgements

This work is a part of a funded project but details have been withheld to maintain anonymity. It will be provided as a part of the final paper.

Ethical Considerations

We use multiple Twitter and Reddit datasets to fine-tune our emotion model ensemble. Both these datasets have been cleaned to remove any toxicity, biases and offensive language. The annotated French election dataset cannot be publicly released following the terms and conditions of the project. The data available to us for fine-tuning and evaluation does not contain any personally identifiable data and we do not have any knowledge of the annotators behind creating this dataset. We also utilize multiple pre-trained models which reduces the carbon footprint of training models from scratch. Further, utilization of this transfer learning method for any new domain would not incur any training costs as minimal fine-tuning may be required. However, the results obtained in an unknown domain should be human evaluated before using it for any downstream analytics task.

References

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting Multi-label Emotion Classification as Span-prediction. In *Proceedings of the 16th conference of the European Chapter of the Association for Computational Linguistics*.

Naila Aslam, Furqan Rustam, Ernesto Lee, Patrick Bernard Washington, and Imran Ashraf. 2022. Sentiment analysis and emotion detection on cryptocurrency related Tweets using ensemble LSTM-GRU Model. *IEEE Access*, 10:39313–39324.

James R Averill. 1980. A constructivist view of emotion. In *Theories of emotion*, pages 305–339. Elsevier.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

Angelo Basile, Guillermo Pérez-Torró, and Marc Franco-Salvador. 2021. Probabilistic ensembles of zero-and few-shot learning models for emotion classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 128–137.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The pragmatics behind politics: Modelling metaphor, framing and emotion in political discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488.

Xiaofeng Cai and Zhifeng Hao. 2018. Multi-view and attention-based bi-LSTM for Weibo emotion recognition. In *2018 International Conference on Network, Communication, Computer Engineering*, pages 772–779.

Qi Chen, Wei Wang, Kaizhu Huang, and Frans Coenen. 2022. Zero-shot text classification via knowledge graph embedding for social media data. *IEEE Internet of Things Journal*, 9(12):9205–9213.

Andrea Chiorrini, Claudia Diamantini, Alex Mircoli, and Domenico Potena. 2021. Emotion and sentiment analysis of tweets using BERT. In *EDBT/ICDT Workshops*.

Jean-Michel Daignan. 2017. French presidential election: Extract from twitter about the french election.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Xavier Ferrer, Tom van Nuenen, Jose M Such, and Natalia Criado. 2021. Discovering and Categorising Language Biases in Reddit. In *ICWSM*, pages 140–151.

Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. 2019. Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, 7(1):35–51.

Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. 2019. EmotionX-IDEA: Emotion BERT- an Affectional Model for Conversation. *arXiv preprint arXiv:1908.06264*.

KV Karan, Vedant Bahel, R Ranjana, and T Subha. 2022. Transfer learning approach for analyzing attentiveness of students in an online classroom environment with emotion detection. In *Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2021*, pages 253–261. Springer.

Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. 2008. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3.

Luyao Ma, Long Zhang, Wei Ye, and Wenhui Hu. 2019. PKUSE at SemEval-2019 task 3: emotion detection with emotion-oriented neural attention network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 287–291.

Sonia Xylina Mashal and Kavita Asnani. 2017. Emotion intensity detection for social media data. In *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, pages 155–158. IEEE.

Winfried Menninghaus, Valentin Wagner, Eugen Wassiliwizky, Ines Schindler, Julian Hanich, Thomas Jacobsen, and Stefan Koelsch. 2019. What are aesthetic emotions? *Psychological review*, 126(2):171.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Saif M Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*.

Laura Ana Maria Oberländer, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566.

Justin Olah, Sabyasachee Baruah, Digbalay Bose, and Shrikanth Narayanan. 2021. Cross domain emotion recognition using few shot knowledge transfer. *arXiv preprint arXiv:2110.05021*.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. A comparison of word-embeddings in emotion detection from text using biLSTM, CNN and Self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Ameneh Gholipour Shahraki and Osmar R Zaiane. 2017. Lexical and learning-based emotion mining from text. In *Proceedings of the international conference on computational linguistics and intelligent text processing*, volume 9, pages 24–55.

Rainer Storn and Kenneth Price. 1997. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11:341–359.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. WordNet Affect: An affective extension of wordnet. In *Lrec*, volume 4, page 40. Lisbon, Portugal.

Somchanok Tivatansakul, Michiko Ohkura, Supadchaya Puangpontip, and Tiranee Achalakul. 2014. Emotional healthcare system: Emotion detection by facial expressions using japanese database. In *2014 6th computer science and electronic engineering conference (CEEC)*, pages 41–46. IEEE.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3914–3923.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040.

A Annotation Details

For the emotion classification task, each annotator was presented with the same set of tweets from the French election dataset. Every tweet had to be labelled with one or more emotions expressed by the author. Below is the complete list of emotion labels:

1. Anger, hate, contempt, disgust:
2. Embarrassment, guilt, shame, sadness
3. Admiration, love
4. Optimism, hope
5. Joy, happiness
6. Pride, including national pride
7. Fear, pessimism
8. Amusement
9. Positive-other
10. Negative-other

Three annotators labeled each tweet with one or more emotion labels. The ground truth is considered to be the labels which have at least two annotators agree on them.

B Hyperparameters

To fine-tune the pretrained twitter-RoBERTa-base-emotion models on each of the six training and validation datasets, we use the following settings, chosen in order to stay close to the pretrained weights and also alleviate overfitting to the target domains. We use a binary cross-entropy loss for the task of multi-label classification, an Adam optimizer, an initial learning rate of 1e-6, and a batch size of 16. During each training procedure, we apply early stopping on the validation loss with a patience of 10 epochs to alleviate overfitting by stopping fine-tuning when the validation performance no longer improves. In each case, we choose the model that achieves the lowest validation loss as our final model. We train for 72 epochs on the CBET dataset over the six Ekman emotions, 90 epochs on the GoEmotions dataset over the six Ekman emotions, 66 epochs on the GoEmotions *joy* subcategory dataset, 13 epochs on the GoEmotions *sadness* subcategory dataset, 18 epochs on the GoEmotions *fear* subcategory dataset, and 8 epochs on the GoEmotions *anger* subcategory dataset, in order to achieve these best results in Table 8. Across the six models, the total training procedure converged after approximately 5.5 hours on a single GPU.

Model	**Validation Accuracy**	**Test Accuracy**
CBET-Ekman	0.6558	0.6483
GoEmo-Ekman	0.6966	0.6914
Joy	0.7386	0.7519
Sadness	0.7205	0.7625
Fear	0.9048	0.8878
Anger	0.6541	0.6501

Table 8: Final validation accuracy and final testing accuracy for each of the six fine-tuned twitter-RoBERTa-base-emotion models in our model ensemble

C Fine-Tuning Thresholds and Weights

In the hierarchical label mappings presented in Tables 3 and 4 for the SemEval and French Election datasets, the weights for each term in the linear combinations for each target emotion are by default set to 1. Without any fine-tuning data in the target domain, we let each emotion subcategory have equal weight in determining the value of the target emotion. Additionally, in the evaluation, we let the thresholds for classification of each emotion all be equal to 0.3. However, with the availability of a small in-domain validation dataset, we can improve the classification thresholds as well as the mapping weights. We fine-tune the classification thresholds by choosing a threshold for each target

Target Label	Classification Threshold
anger	0.17
anticipation	0.01
disgust	0.02
fear	0.04
joy	0.01
love	0.02
optimism	0.01
pessimism	0.01
sadness	0.22
surprise	0.34
trust	0.60

Table 9: Label-wise classification thresholds after fine-tuning on the SemEval validation set

class to maximize the F1 score on that class over the validation dataset. For the fine-tuning mode, given the SemEval validation dataset, we obtain the label-wise classification thresholds in Table 9. As shown in Table 5, the performance on SemEval improves, suggesting that there is consistency between the validation and testing data in how strong a signal has to be for a positive classification.

We fine-tune the mapping weights by successively applying differential evolution to each individual target label mapping to maximize the F1 score on that label over the validation dataset (Storn and Price, 1997). The implementation of the differential evolution algorithm for fine-tuning the mapping weights is provided by Scipy[4]. For each target label mapping, we constrain each weight in $[0, 2]$ in the optimization process, and continue iteratively until the improvements in the label-wise F1 scores are sufficiently small. For example, the mapping weights for the emotion *love* in SemEval obtained by this process are as follows: love = EK[joy] * (1.174 * J[love] + 1.465 * J[desire] + 0.751 * J[caring])) * Sent[positive]. We see that the contribution of the subcategory *desire* is the greatest, followed by *love* and then *caring*. Again, as shown in Table 5, the scores of the system on SemEval are improved by this optimization. We fine-tune both the mapping weights and the classification thresholds by first optimizing the weights, and then subsequently choosing the thresholds.

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.differential_evolution.html