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. The task is additionally complicated by social and cultural factors that make certain words and expressions emotionally charged in one context but entirely neutral in another. In this paper, we present a generalized approach of transfer learning for emotion detection that can be adapted to any domain and any set of classification labels. We show the performance improvements that could be achieved by fine-tuning our approach with limited annotated data from the target domain. 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. Furthermore, the system output can be easily adapted by end users to detect additional emotion categories. Lastly, we present an evaluation of this method on the publicly available SemEval 2018 Task 1e-c dataset and also a new annotated dataset consisting of tweets related to the French elections in 2017 (Daignan, 2017).

1 Introduction

It is now widely acknowledged that internet social media are powerful platforms for launching wide-reaching influence campaigns related to important events such as elections, pandemics, armed conflicts, as well as commercial interests. The main objectives of such campaigns is to manipulate public opinion in a particular way: to favor or oppose a political candidate, to accept or resist vaccination, to justify an aggression, etc. To achieve their objectives, the campaigns send messages that push a specific agenda, using language, imagery, and topics that are likely to be persuasive to their target audiences. One powerful device is the use of language that both expresses emotion and arouses an emotional response in the audience. But which emotions matter? Clearly, the emotions that may accompany discussions 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 (CBET) 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 be a superset or subset of any existing emotion label set, we face a situation where no training data is available for some labels. For such new problems, possible solutions involve curating new datasets with the relevant label set, using semi-supervised or unsupervised approaches,

or framing the emotion classification task in such a way that no training data is needed. In this paper, we propose a generalized approach of transfer learning with multiple steps. First, neural models are trained on sentiment analysis and emotion detection tasks using a variety of preexisting emotion-labeled social media data. Second, the outputs of these models are combined and mapped to the desired emotion labels by a weighted linear combination derived by considering the relatedness of emotions. Third (optionally), given target domain data, the linear combination weights or classification thresholds are fine-tuned to improve target domain performance.

Overall, the contributions of this paper are:

- A generalized approach for emotion detection across domains.
- A zero-shot transfer learning method for novel or specialized emotion label sets for which there is no in-domain training data.
- A few shot fine-tuning when limited in-domain training data is available.

2 Background

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*, is the baseline of much emotion-related research (Ekman, 1999). Another prominent model is 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* (Plutchik, 1984). This wheel can be used to compose more complex emotions by varying the emotion intensities. Dimensional models characterize emotions as regions within a continuous space of emotional response dimensions. For example, the Circumplex model of affect (Russell, 1980) specifies the dimensions *valence* and *arousal*, and interprets 28 emotional states in terms of these dimensions. In some related models, a third dimension of *dominance* is added (Russell and Mehrabian, 1977). The Plutchik's wheel of emotions also depicts *valence* and *arousal* on the two axes of the wheel. In general, the problem of choosing an appropriate taxonomy for an emotion classification task is dependent on domain and end-use.

2.2 Pre-trained Language Models

Large pretrained language models (PLMs) like GPT (Radford et al., 2018), BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) have achieved state of the art performance in various NLP tasks like text classification (Sun et al., 2019), (Munikar et al., 2019), summarization (Miller, 2019) and machine translation (Zhu et al., 2020). They are mostly the highest scorers on the GLUE (Wang et al., 2018), SQUAD (Rajpurkar et al., 2016) and MultiNLI (Williams et al., 2017) benchmarks. These models are first pretrained on large, unlabeled text corpora, and then fine-tuned with task-specific annotated data for various downstream tasks. Some model architectures were adapted to short and spontaneous texts such as tweets and social media comments, by pretraining on Twitter corpora. Models like BERTweet (Nguyen et al., 2020) and XLM-T (Barbieri et al., 2021) are popular Twitter-specific language models. TweetEval serves as a strong baseline for the seven 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 at training time, and was first introduced in (**?**). Although no training examples of these classes exist, information about these classes is utilized to aid in the classification. In the task of emotion detection, it is possible that the application domain calls for the prediction of emotion classes for which there is no training data. These emotions can be more fine-grained than what is available in the training data (e.g., different types of *anger*), or they may be emotions that do not correspond to any of the labels of the training data. Additionally, domain-specific emotion classes may arise as the application and its requirements evolve. In the emotion detection approach described in this paper, we build upon the idea of zero-shot learning, in that we must predict emotion classes unseen during training. However, instead of using supplementary information to aid in this classification, we use a predefined hierarchical mapping from seen emotion classes to unseen emotion classes based on descriptions of these emotions.

3 Related Work

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 to the task are lexicon-based. Popular emotion lexicons include WordNet-Affect (Strapparava et al., 2004), NRC Emotion Lexicon (Mohammad and Turney, 2010), EmoSenticSpace (Poria et al., 2014), DepecheMood (Staiano and Guerini, 2014). These lexicons consist of words annotated with emotion labels or scores, and various rule-based or machine learning algorithms have been developed to utilize lexicons to classify emotions in sentences and documents (Bandhakavi et al., 2017), (Tzacheva et al., 2019), (Bravo-Marquez et al., 2019), (Kušen et al., 2017), (Seal et al., 2020). Mac Kim et al. (2010) and Zad and Finlayson (2020) use lexicons and dimensionality reduction techniques for unsupervised emotion detection from text. The major drawback of these methods is the focus on individual words resulting in the lack of context incorporation. Additionally, the use of a specific lexicon limits the number of available annotated keywords and emotion labels.

Several supervised machine learning approaches have been developed using a combination of datasets collected from Twitter, Reddit, blogs, and news articles, with curated features such as unigrams, bigrams, lexicon labels, hashtags, and emoticons. The most popular algorithms are the Support Vector Machine or Naive Bayes classifiers, which have achieved accuracy scores of over 80% in some emotion classification tasks (Alm et al., 2005), (Hasan et al., 2014), (Wikarsa and Thahir, 2015), (Mashal and Asnani, 2017), (Alotaibi, 2019), (Hasan et al., 2019). The lack of a consistent emotion taxonomy make these methods inadequate when used across domains.

With the recent availability of large emotion-annotated corpora, word embeddings and deep learning approaches were applied to emotion detection to incorporate contextual information. CNN, LSTM and BERT models became the most powerful tools (Cai and Hao, 2018), (Huang et al., 2019), (Polignano et al., 2019), (Ma et al., 2019), (Chiorrini et al., 2021). The recent works of Fei et al. (2020), He and Xia (2018), Alhuzali and Ananiadou (2021) aim to integrate label dependencies in multi-label emotion detection by modeling them in the loss function.

4 Methodology

4.1 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 experts in the application domain.

4.2 Approach

Our approach involves producing hierarchical scores for a tweet x over three sentiment categories, the six Ekman emotions, and their fine-grained sub-categories defined in (Demszky et al., 2020). To obtain confidence scores over emotions in E, we design a many-to-one mapping from the model outputs to the set E, based on domain knowledge and the understanding of the categorical and dimensional models of affect (Russell, 1980), (Plutchik, 1984). This mapping can be applied without any training data for emotions in E, but can be fine-tuned to improve performance if there is existing data for E in the target application domain. 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.

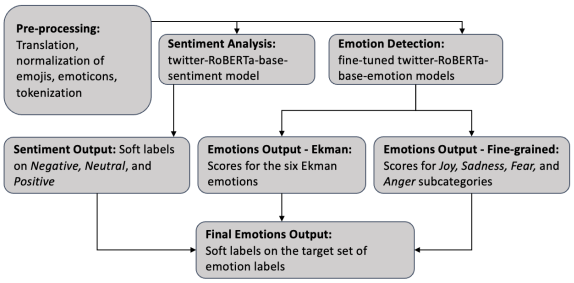

Figure 1: Ensemble Emotion Detection Architecture

4.3 Datasets and Preprocessing

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

Model	Output Labels
Sentiment(Sent)	positive, neutral, negative
CBET-Ekman	joy, sadness, fear, anger, disgust, surprise
GoEmo-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 1: Set of output labels for each component model

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 rich taxonomy of emotions has been identified after recent works ascertained how the Ekman or Plutchik labels are insufficient to label the complex emotions expressed by facial expressions, speech and other gestures (Cowen et al., 2019). Human feedback was incorporated to identify additional labels during the annotation process. The emotions can be grouped into positive, negative, ambiguous and neutral sentiment labels, or the six Ekman emotions (Ekman, 1999). The large number of fine-grained emotion labels in this dataset makes it an ideal choice to be used in our task of creating more generalized or specialized labels based on the domain. A series of data curation steps have been carried out to 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. The dataset was also balanced to limit the number of samples for each emotion. Consistent inter-rater agreement scores were achieved across most of the emotion labels, with emotion frequency being directly correlated to the agreement score. We use the subcategories of *joy, sadness, fear* and *anger* as prescribed in GoEmotions to produce training, testing and validation datasets for each lower level emotion model in the hierarchy (Table 2).

Emotion	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 2: Sizes of training, validation, and testing sets for emotion subcategory datasets derived from the GoEmotions

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

4.4 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

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

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 the task of 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 both Twitter and Reddit data so that, in the subsequent mapping step, we can weigh them based on the target domain of the application. The remaining four models are fine-tuned to output scores on the subcategories of *joy, sadness, fear, anger*. The fine-tuning details and results for each model are described in Appendix B. To summarize, our emotion classification model ensemble produces scores for each of the fine-grained labels as outlined in Table 1. The next section describes how these scores are utilized downstream to adapt our model to a new domain.

4.5 Domain-Specific Hierarchical Label Transfer

For a desired label set E, we map the scores from the model ensemble to scores on the new set, using a weighted linear combination derived by considering the relatedness of emotions, as in Plutchik's wheel of emotions (Plutchik, 1984), where the eight primary strong emotions are associated with weaker ones such as *contempt* and *optimism*. For simplicity, let *EK* be a model with the six Ekman output labels from Table 1 with scores for each emotion equal to a weighted linear combination of the scores from CBET-Ekman and GoEmo-Ekman. A general set of rules to determine the mapping from the 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*; for example, the emotion *anger* is *negative*. However, in some cases, an emotion can have *positive* and *negative* sentiments in different contexts.

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

3. 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*.

4. 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 $(Sent[positive] * EK[joy] * Joy[optimism])$. We provide examples of specific emotion mappings in the experiments.

Section 5 outlines some examples of label transfer that we adopted to map the hierarchical outputs to different sets of emotion labels that may be used for transfer learning. In each experiment, we additionally use an ablation study to show the significance of applying sentiment scores in addition to the hierarchical emotion scores to determine each label score.

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

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

5 Experiments

In this section, we discuss the results of our emotion classification model on a benchmark emotion dataset. 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). Both are unseen domains for the model as no samples from these datasets were used during the training phase. There are several methods available for emotion classification as mentioned in Section 3, but all of them require in-domain training to achieve the SOTA scores. Our approach stands out as it produces competitive scores with no available in-domain training data, and thus is an important baseline for transfer learning of emotions across domains.

5.1 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. 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 around 7k English tweets, and none of this data has been used to train or fine-tune our emotion model ensemble. We derive a mapping from the output scores of Table 1 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.

In general, the model scores higher for target emotions that are more closely related to the Ekman emotions as well as those that have more testing examples. 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, with respect to a target metric 2) find an optimal set of weights for each component in the linear mapping to E, with respect to a target metric. After the first zero-shot evaluation, we utilize a small subset of in-domain data to fine-tune the model weights and classification thresholds. Although fine-tuning has been performed using only the validation dataset, which is approximately 12% of the size of the original training dataset, it produces strong results and is successful in adapting the system to the SemEval target domain (Table 4). The micro-average F1 scores and AUC scores in the ablation study (Table 5) show the consistent performance of the zero-shot method and also the relevance of the sentiment layer as a crucial step in emotion detection.

5.2 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

Emotion	F1	Support
anger	0.66	1101
anticipation	0.26	425
disgust	0.64	1099
fear	0.58	485
joy	0.83	1442
love	0.5	516
optimism	0.68	1143
pessimism	0.21	375
sadness	0.64	960
surprise	0.19	170
trust	0.11	153

Table 4: Classification report on SemEval 2018 Task 1e test dataset with fine-tuning on the validation dataset

Model	F1	AUC
Emotion model ensemble	0.55	0.83
- In-domain fine-tuning	0.38	0.74
- Sentiment layer	0.37	0.72

Table 5: Ablation study shows the influence of hierarchical layers on SemEval 2018 evaluation; F1 denotes micro-average F1 score

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*$\}$. Every label was also provided with a description and a set of synonymous emotion labels (Appendix A). Due to the ambiguity caused by grouping multiple emotions in one label, the inter-annotator agreement across all labels is very low and there are inconsistencies in annotation guidelines between validation and test datasets. In spite of several of these issues in this dataset, our model adapts to the unknown domain using very little or no fine-tuning data.

The mapping of the output scores from the emotion model ensemble to the destination set E is carried out by the understanding of the label definitions in the target domain and the general rules formulated in the previous section. For example, the label *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]*. Figure 2 illustrates an example output produced by our system on a tweet from this dataset.

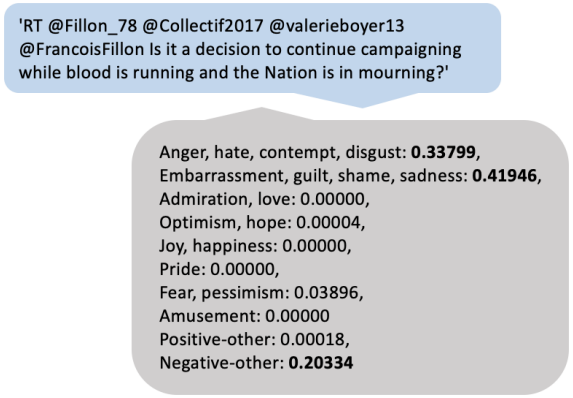

Figure 2: Example tweet from the French election dataset

In Table 7, we show the per-class F1 scores of our system on the French election annotated dataset after fine-tuning on a 1000 tweet validation set. In general, we see that the model scores higher for target emotions such as *positive-other* and *negative-other* which have more examples, and lower for target emotion such as *joy/happiness* which have fewer examples. The scores in this domain are lower than those achieved than for the SemEval 2018 Task 1e dataset, reflecting the lack in reliability of annotation, but are improvements over baseline models with little to no training data.

6 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

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 6: Mapping of model outputs to French election labels

Emotion	F1	Support
anger/contempt/disgust	0.22	520
embarrassment/guilt	0.20	114
admiration/love	0.16	118
optimism/hope	0.38	711
joy/happiness	0.18	94
pride	0.25	192
fear/pessimism	0.17	222
amusement	0.16	455
positive-other	0.49	2572
negative-other	0.49	2779

Table 7: Classification report on French election dataset with fine-tuning on a validation set

which would further enable us to produce stronger results and analysis.

7 Conclusion

We present an approach for the task of emotion detection from social media text, and an off-the-shelf emotion classification ensemble that can be adapted in 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 target domain validation data for fine-tuning can improve performance within that domain. The user has to carefully map the hierarchical fine-grained emotion and sentiment scores available from the model to their required set of labels. We have demonstrated the idea with the help of two such mappings to datasets in various domains and with various target label sets that the model has not seen before.

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 classi-

fication as span-prediction. *arXiv preprint arXiv:2101.10038*.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586.

Fahad Mazaed Alotaibi. 2019. Classifying text-based emotions using logistic regression.

Anil Bandhakavi, Nirmalie Wiratunga, Deepak Padmanabhan, and Stewart Massie. 2017. Lexicon based feature extraction for emotion text classification. *Pattern recognition letters*, 93:133–142.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2021. Xlm-t: A multilingual language model toolkit for twitter. *arXiv preprint arXiv:2104.12250*.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

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)*, Vancouver, Canada. Association for Computational Linguistics.

Felipe Bravo-Marquez, Eibe Frank, Bernhard Pfahringer, and Saif M. Mohammad. 2019. AffectiveTweets: a Weka package for analyzing affect in tweets. *Journal of Machine Learning Research*, 20(92):1–6.

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 (NCCE 2018)*, pages 772–779. Atlantis Press.

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

Alan Cowen, Disa Sauter, Jessica L Tracy, and Dacher Keltner. 2019. Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1):69–90.

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. *arXiv preprint arXiv:2005.00547*.

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

Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7692–7699.

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, Emmanuel Agu, and Elke Rundensteiner. 2014. Using hashtags as labels for supervised learning of emotions in twitter messages. In *ACM SIGKDD workshop on health informatics, New York, USA*, volume 34, page 100.

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.

Huihui He and Rui Xia. 2018. Joint binary neural network for multi-label learning with applications to emotion classification. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 250–259. Springer.

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*.

Ema Kušen, Giuseppe Cascavilla, Kathrin Figl, Mauro Conti, and Mark Strembeck. 2017. Identifying emotions in social media: comparison of word-emotion lexicons. In *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pages 132–137. IEEE.

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*.

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.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70.

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.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

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*.

Manish Munikar, Sushil Shakya, and Aakash Shrestha. 2019. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5. IEEE.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

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.

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014. Emosenticspace: A novel framework for affective common-sense reasoning. 69(1):108–123.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

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

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Dibyendu Seal, Uttam K Roy, and Rohini Basak. 2020. Sentence-level emotion detection from text based on semantic rules. In *Information and Communication Technology for Sustainable Development*, pages 423–430. Springer.

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.

Jacopo Staiano and Marco Guerini. 2014. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprint arXiv:1405.1605*.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *Lrec*, volume 4, page 40. Lisbon, Portugal.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Angelina Tzacheva, Jaishree Ranganathan, and Sai Yesawy Mylavarapu. 2019. Actionable pattern discovery for tweet emotions. In *International Conference on Applied Human Factors and Ergonomics*, pages 46–57. Springer.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Liza Wikarsa and Sherly Novianti Thahir. 2015. A text mining application of emotion classifications of twitter's users using naive bayes method. In *2015 1st International Conference on Wireless and Telematics (ICWT)*, pages 1–6. IEEE.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Samira Zad and Mark Finlayson. 2020. Systematic evaluation of a framework for unsupervised emotion recognition for narrative text. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 26–37, Online. Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.

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, incl. 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