Meta-learning for Classifying Previously Unseen Data Source into Previously Unseen Emotional Categories

Gaël Guibon, Matthieu Labeau, Hélène Flamein, Luce Lefeuvre, Chloé Clavel

Meta-learning for Classifying Previously Unseen Data Source into Previously Unseen Emotional Categories

Gaël Guibon[1,2], Matthieu Labeau[1], Hélène Flamein[2], Luce Lefeuvre[2], and Chloé Clavel[1]

[1]LTCI, Télécom-Paris, Institut Polytechnique de Paris
[2]Direction Innovation & Recherche SNCF
{gael.guibon,matthieu.labeau,chloe.clavel}@telecom-paris.fr
{ext.gael.guibon,helene.flamein,luce.lefeuvre}@sncf.fr

Abstract

In this paper, we place ourselves in a classification scenario in which the target classes and data type are not accessible during training. We use a meta-learning approach to determine whether or not meta-trained information from common social network data with fine-grained emotion labels can achieve competitive performance on messages labeled with different emotion categories. We leverage few-shot learning to match with the classification scenario and consider metric learning based meta-learning by setting up Prototypical Networks with a Transformer encoder, trained in an episodic fashion. This approach proves to be effective for capturing meta-information from a source emotional tag set to predict previously unseen emotional tags. Even though shifting the data type triggers an expected performance drop, our meta-learning approach achieves decent results when compared to the fully supervised one.

1 Introduction

Training a model for a classification task without having access to the target data nor the precise tag set is becoming a common problem in Natural Language Processing (NLP). This is especially true for NLP tasks applied to company data, highly specialized, and which is most of the time raw data. Annotating these data requires to set up a lengthy and costly annotation process, and annotators must have specific skills. It also raises some data privacy issues. Our study is conducted in this context. It deals with private messages, that shall be annotated with emotions as labels. This task is highly difficult because of the subjective and ambiguous nature of the emotions, and because of the nature of the data. We tackle this problem in an emotion classification task from short texts. We assume that meta-learning can serve for emotion classification in different text structures along with a different tag set.

Predicting and classifying emotions in text is a widely spread research topic, going from polarity-based labels (Strapparava and Mihalcea, 2007; Thelwall et al., 2012; Yadollahi et al., 2017) to more complex representations of emotion (Alm et al., 2005; Bollen et al., 2009; Yu et al., 2015; Zhang et al., 2018a; Zhu et al., 2019; Zhong et al., 2019; Park et al., 2019). In this paper, we place ourselves in a situation where we have no access to target data or models of target classes. Therefore, we want to learn information from related data sets to predict labels on our target data, even though label sets differ. Thus, we apply meta-learning using a few-shot learning approach to predict emotions in messages from daily conversations (Li et al., 2017) based on meta-information inferred from social media informal texts, *i.e.* Reddit comments (Demszky et al., 2020a).

With this setup, our goal is to investigate if combining few-shot learning and meta-learning can yield competitive performance on data of a different kind from those on which the model was trained. Indeed, recent work already showed meta-learning is useful when shifting to different topics on a classification task with the Amazon data set (Bao et al., 2020) or different entity relations on the dedicated Few-Rel data set (Han et al., 2018; Gao et al., 2019a). In this paper, we take another step forward by leveraging meta-learning when shifting not only emotional tag sets but also data sources, involving different topics, lexicons and phrasal structures. For instance, the "surprise" emotion is set for "*Wow you found the answer, wish you were on top, will link to you in my post*" in GoEmotions (Demszky et al., 2020a) and for "*Are you from south?*" in DailyDialog (Li et al., 2017), varying both the lexicon used (post related vocabulary for GoEmotions) and the sentence structure (cleaner syntactic structures in DailyDialog).

Our contribution relies on the implementation of a two-level meta-learning distinguishing data

Proceedings of the 1st Workshop on Meta Learning and Its Applications to Natural Language Processing (MetaNLP 2021), pages 76–89
Bangkok, Thailand (online), August 5, 2021.

by their label set and data source at the same time. We also try to quantify the impact of switching data sources in this framework. After summarizing the related work (Section 2), we present the data sets and labels (Section 3) that we consider in our methodology and experiments (Section 4). We then present the results (Section 5) before discussing some key points (Section 6) and conclude (Section 7).

The data preparation code and files, and the implementations are available in a public repository: https://github.com/gguibon/metalearning-emotion-datasource.

2 Related Work

Emotion classification approaches (Alm et al., 2005; Strapparava and Mihalcea, 2007; Bollen et al., 2009; Thelwall et al., 2012; Yu et al., 2015; Yadollahi et al., 2017; Zhang et al., 2018a; Zhu et al., 2019; Zhong et al., 2019; Park et al., 2019) usually benefit from using as many examples as possible when training the classifier. However, it is not always possible to obtain large data sets for a specific task: we need to learn from a few examples by applying specific strategies. Few-shot learning (Lake, 2015; Vinyals et al., 2016; Ravi and Larochelle, 2016) is an approach dedicated to learn from a few examples per class and thus to create efficient models on a specific task.

Meta-Learning. While they can be used for different purposes, few-shot learning frameworks are often used for meta-learning (Schmidhuber, 1987), defined as "learning to learn". Like few-shot learning, meta-learning considers tasks for training but with the aim of being effective at a new task in the testing stage (Yin, 2020). To do so, meta-learning can focus on different aspects such as learning a meta-optimizer (various gradient descent schemes, reinforcement learning, *etc.*), a meta-representation (embedding by metric learning, hyper parameters, *etc.*), or a meta-objective (few-shot, multi-task, *etc.*), three aspects respectively represented as "How", "What" and "Why" (Hospedales et al., 2020). Both few-shot learning and meta-learning approaches have mainly been developed in computer vision using different optimization schemes. The main meta-learning approaches use an episodic setting (Ravi and Larochelle, 2016) which consists in training on multiple random tasks with only a few examples per class. Then, each task is an episode made of a number of *shots* (examples per class), a *support set* (set of examples to train from), a *query set* (set of examples to predict and compute a loss), and a number of *ways* (classes).

Optimization-based. Optimization-based meta learning is an approach represented mainly by the Model Agnostic Meta Learning (MAML) (Finn et al., 2017a) which learns parameters meta-initialization and meta-regularization. It possesses multiple variations, such as First-Order MAML (Finn et al., 2017b), which reduces computation; Reptile (Nichol et al., 2018), which considers all training tasks and requires target tasks to be close to training tasks; and Minibatch Proximal Updates (Zhou et al., 2019), which learns a prior hypothesis shared across tasks. Another recent approach focuses on learning a dedicated loss (Bechtle et al., 2021).

Metric learning. Meta-representation and meta-objective aspects of meta-learning are often used together. In this work, regarding the meta-representation aspect, we focus on approaches aiming to learn a distance function, usually named metric-learning. Among these approaches, Siamese Networks (Koch et al., 2015) do not take tasks into account and only focus on learning the overall metric to measure a distance between the examples. Matching Networks (Vinyals et al., 2016) use the support set examples to calculate a cosine distance directly. Prototypical Networks (Snell et al., 2017), for their part, consider class representations from the support set and use an euclidean distance instead of the cosine one. Lastly, Relation Networks (Sung et al., 2018) consider the metric as a deep neural network instead of an euclidean distance, using multiple convolution blocks and the last sigmoid layer to compute relation scores. When applied to image data sets, a recent work showed Prototypical Networks (Snell et al., 2017) possess better efficiency with the lowest amount of training examples (Al-Shedivat et al., 2021) which leads us to use this approach due to our data configuration.

Meta-learning and NLP. Other approaches have recently made use of several optimization schemes (Bernacchia, 2021; Al-Shedivat et al., 2021) and have been adapted to NLP tasks (Bao et al., 2020) especially on Few-Rel dataset, a NLP corpus dedicated to few-shot learning for relation classification (Gao et al., 2019b; Han et al., 2018; Sun et al., 2019). For text classification, meta-

learning through few-shot learning has been used on Amazon Review Sentiment (ARSC) dataset (Yu et al., 2018; Geng et al., 2019; Bao et al., 2020; Bansal et al., 2020) by training sentiment classifiers while varying the 23 topics. We draw on their work on Amazon topics to better tackle another type of labels, emotions, while further adapting Prototypical Networks on texts by considering attention in the process.

Meta-learning and Emotions. Recent studies on acoustic set up a generalized mixed model for emotion classification from music data (Lin et al., 2020), or even meta-learning for speech emotion recognition whether it is monolingual (Fujioka et al., 2020) or multilingual (Naman and Mancini, 2021). On the other hand, on textual data one used distribution learning (Zhang et al., 2018b) through sentence embedding decomposition and K-Nearest Neighbors (Zhao and Ma, 2019) while others studied emotion ambiguity by meta-learning a BiLSTM (Huang et al., 2015) with attention in the scope of 4 labels (Fujioka et al., 2019).

Considering both our use-case scenario and the aforementioned recent meta-learning efficiency comparison (Al-Shedivat et al., 2021), we focus on using Prototypical Networks for this work, while varying the encoders to better adapt Prototypical Networks to textual data in a few-shot and meta-learning setting. Thus, we contribute by using metric learning based meta learning while considering emotion classes as tasks for NLP. Moreover, as far as we know, this work is the first one on meta-learning considering a two-level meta-learning by transferring knowledge to new tasks, despite the use of new data sources at the same time.

3 Datasets and Tag Sets

We consider two different English data sets to stay in line with our will to use a source data set on which the meta-model will be trained and a target data set on which we will evaluate the transferring capabilities of our model.

GoEmotions (Demszky et al., 2020a) is the data set we use to train and tune hyper-parameters. It is a corpus made of 58,000 curated Reddit comments labeled with 27 emotion categories. We split it into 3 tag sets (*EmoTagSets*) for meta-training afterwards which detail later on. GoEmotions (Demszky et al., 2020a) also comes with predefined train/val/test splits by ratio, ensuring the presence of all labels in each split. We use them to apply the fully supervised learning.

DailyDialog (Li et al., 2017) corresponds to the target data to be labeled using the meta-trained model. This corpus is initially structured as 13,118 human-written daily conversations, going through multiple topics; but for the purpose of our study, we only use it as individual utterances. We chose this corpus because of its propinquity with our case study: messages from conversational context are usually private and unlabeled. We retrieve utterances from the official test set with their associated emotion label, because studying the conversational context exceeds the scope of this paper. We only focus on utterances, language structure differences, and different emotion tag sets for meta-learning. This leads to a total of 1,419 utterances for 6 emotion labels (*EmoTagSet3*). As for GoEmotions, DailyDialog comes with official train/val/test splits that we use for comparison purposes while using supervised or meta learning approaches.

Tag Sets. To apply meta-learning on emotion labels we consider 3 different tag sets named *EmoTagSets*. As previously said, we made these tag sets considering the different labels from each data set: let Z_G represent the set of GoEmotions' labels and Z_D the set of DailyDialog's labels, we consider the intersection $Z_D \cap Z_G$ as the target labels named *EmoTagSet3*. These target labels are the labels we want to hide from both training and validation phases to only use them during the test phase. The purpose of using the intersection is to enable results comparison on both data sets. The complement of the resulting intersection is then used to create *EmoTagSet1* and *EmoTagSet2*, while taking into account class balance and polarity distribution to ensure each *EmoTagSet1* and *2* possesses a variety of classes. The resulting tag sets and their dedicated usage are visible in Table 1. Table 1 also shows the mapping between the 6 target emotion classes of *EmoTagSet3* and their possible correspondences in regard to other labels. This mapping comes directly from GoEmotions' mapping[1].

[1] https://github.com/google-research/google-research/blob/master/goemotions/data/ekman_mapping.json

EmoTagSet3 *(DailyDialog tags)* *For test and supervised* ↓	**EmoTagSet1** *For training* ↓	**EmoTagSet2** *For validation* ↓
anger →	annoyance	disapproval
disgust →	/	/
fear →	nervousness	/
joy →	amusement, approval, excitement, love, pride, admiration	gratitude, optimism, relief, desire, caring
sadness →	remorse	disappointment, embarrassment, grief
surprise →	realization, confusion	curiosity

Table 1: Tag set mapping to the 6 basic emotions of EmoTagSet3. All these labels are present in GoEmotions while only the EmoTagSet3 is present in DailyDialog. EmoTagSet1 and 2 are mapped to EmoTagSet3 following the GoEmotions' official mapping (Demszky et al., 2020b).

4 Methodology and Experimental Protocol

First, the objective is to retrieve label-level meta-information using Reddit comments (GoEmotions) and the different label sets (*EmoTagSets*). Then, we seek to transfer the meta-information to daily conversation-extracted utterances (DailyDialog), hence varying in data structure and vocabulary.

Meta-training. The first step consists of an emotion-based meta-learning on GoEmotions' training and validation sets in order to learn meta-information that we evaluate on DailyDialog's test set later on. Figure 1 shows this approach. We want to meta-train a classifier from few examples by using few-shot learning with 5 examples per class from GoEmotions' train set, our classes being the different emotion labels. We adopt the Prototypical Networks (Snell et al., 2017) in an episode training strategy to apply few-shot learning to the meta-learning process. For each episode, Prototypical Networks apply metric-learning to few-shot classification by computing a prototype $\mathbf{c}_k$ for each class k (*way*) with a reduced number of examples from the support set S_k (*shots*). Each class prototype being equal to the average of support examples from each class as follows:

$$\mathbf{c}_k \leftarrow \frac{1}{N_C} \sum_{(\mathbf{x}_i, y_i) \in S_k} f_\phi(\mathbf{x}_i)$$

where f_ϕ corresponds to the encoder. We then minimize the euclidean distance between prototypes and elements from the query set Q_k to label them and compare the resulting assignments $d\left(f_\phi(\mathbf{x}), \mathbf{c}_k\right))$ where $\mathbf{x}$ represents an element from the query set. This follows the standard Prototypical Networks with the following loss

$$\frac{1}{N_C N_Q}\left[d\left(f_\phi(\mathbf{x}), \mathbf{c}_k\right)) + \log \sum_{k'} \exp\left(-d\left(f_\phi(\mathbf{x}), \mathbf{c}_{k'}\right)\right)\right]$$

One key element of the Prototypical Networks is the encoder f_ϕ, which will define the embedding space where the class prototypes are computed. Moreover, it is in fact the encoder which is meta-learned during the training phase. In our experiments, we use various encoders to represent a message as one vector: the average of the word embeddings (AVG), convolutional neural networks for sequence representation (CNN) (Kim, 2014) or a Transformer encoder layer (Vaswani et al., 2017) (Tr.). We define our episodic composition by setting $N_c = 6$, $N_s = 5$ and $N_q = 30$ making it a 5-shot 6-way 30-query learning task where N_c is constrained by the number of test classes: indeed, down the line, the model will be tested on the 6 basic emotions from the DailyDialog tag set. This setting renders obsolete the notion of an unbalanced data set.

Episodic composition for training and validating are the same. We meta-train for a maximum of 1,000 epochs, one epoch being 100 random episodes from training classes (*EmoTagSet1*). We set early stopping to a patience of 20 epochs without best accuracy improvement. Validation is also done using 100 random episodes but from validation classes (*EmoTagSet2*). For testing, however, we test using 1,000 random episodes from test classes (*EmoTagSet3*), in which the query set (N_q) is randomly chosen from the test split in a 6-way 5-query fashion. This means 5 elements to classify in one of the 6 target emotions. Figure 1 shows a global view of our meta-learning strategy, from meta-training to evaluation.

Experimental protocol details are as follows. For each data set, we follow previous studies (Bao et al., 2020) and use pre-trained fastText (Joulin et al., 2017) embeddings as our starter word representation. We also compare the different approaches by using a fine-tuned pre-trained BERT (Devlin et al., 2019) as encoder, provided by Hugging Face Transformers (Wolf et al., 2019) (*bert-base-uncased*), and by using the ridge regressor with attention gen-

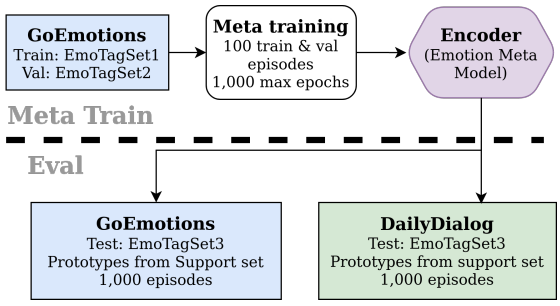

Figure 1: Global view of the meta-learning strategy. While testing on DailyDialog, only utterances from the official test set are considered. EmoTagSet1 ∪ EmoTagSet2 ∪ EmoTagSet3 = ∅.

erator representing distributional signatures (Bao et al., 2020).

Supervised Learning for comparison. We first apply supervised learning by using only DailyDialog's training, validation, and test sets (official splits by ratio) in order to enable later comparison with the meta-learning approach. We use the supervised results as reference scores illustrating what can be achieved in ideal conditions. Ideal conditions also means this does not follow our previously defined scenario. Indeed, a classic supervised learning approach learns using the same labels during training, validation and testing phases, which differ from our scenario. In these supervised results we only used the 6 emotions from *EmoTagSet3* by filtering GoEmotions' elements. Moreover, the encoder and classifier are not distinct as we simply add a linear layer followed by a softmax and use a negative log likelihood loss to compute cross entropy over the different predictions.

The objective here is to enable comparison between our approach and a direct naive supervised one. By naive, we mean that no transfer learning method is used; rather, it only consists in training a fully supervised model on GoEmotions or DailyDialog training and validation sets and applying it on DailyDialog or GoEmotions test sets. Table 2 shows the results of this naive fully supervised approach along with the meta-learning one. However, even with the advantage of using the target labels during training, this fully supervised approach yields lesser scores than our meta-learning approach. This confirms that meta-learning is a viable solution for our use-case scenario which adapts itself to unknown target labels while allowing faster training due to the episodic composition approach (*i.e.* smaller number of batches).

Hyper-parameters tuning. In this paper, we consider the case in which we want to train an emotion classifier while having no access to the target data information. However, to ensure a fair comparison, we use the hyper-parameters obtained through a limited grid-search in our baseline supervised setup. This makes the whole experiment less dependent on specific parameters, leading to a better evaluation process despite not representing a 'real' application case. Hyper parameters are as follows.

The Prototypical Networks' hidden size is set to $[300, 300]$ which is equal to the base embedding size (300 from pre-trained FastText on Wiki News[2]), global dropout is set to 0.1. The CNN encoder consists in three filter sizes of 3, 4 and 5 and is the same architecture as Kim's CNN (Kim, 2014) except for the number of filters which we set to 5000. For the Transformer encoder, we set the learning rate at $1e-4$, the dropout at 0.2, the number of heads at 2 and the positional encoding dropout to 0.1. The embedding and hidden sizes follow the same size as the input embedding with $d = 300$. We considered using multiple Transformer encoder layers but sticking to only 1 layer gave the most optimal results and efficiency.

During supervised learning, we consider an encoder learning rate of $1e-3$ except for the Transformer layer where a learning rate of $1e-4$ gave better results. However, for meta-learning phases we follow optimization methods from recent literature by searching the best learning rate, positive or negative, in a window close to zero and finally set it to $1e-5$ (Bernacchia, 2021). Hence, the learning rate is the only parameter that we do not directly copy from the supervised learning phase's hyper parameters.

Evaluation Metrics. We evaluate the performance of the models by following previous work on few-shot learning (Snell et al., 2017; Sung et al., 2018; Bao et al., 2020) and using few-shot classification accuracy. We go further in the evaluation by adding a weighted F1 score and the Matthews Correlation Coefficient (MCC) (Cramir, 1946; Baldi et al., 2000) as suggested by recent studies in biology (Chicco and Jurman, 2020), but in its multiclass version (Gorodkin, 2004) to better suit our task. Reported scores are the mean values of each

[2] https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip

metrics on all testing episodes with their associated variance ±.

5 Results

Table 2 shows two main different result sets: the ones obtained using supervised learning, and those obtained using meta-learning.

Supervised Learning Results. Results presented in Table 2 come from using the official splits from DailyDialog. As explained in Section 4, we tuned hyper-parameters for each classifier and encoder using this supervised learning phase. Using the Transformer (Vaswani et al., 2017) as classifier requires carefully setting up hyper parameters to converge, especially if the data set size is relatively small. This is the case in this study, and we believe it to be the main reason for the Transformer classifier to perform below the CNN classifier in this fully supervised setting.

Supervised results (top section of Table 2) can be divided into two sub-parts: the supervised learning trained using GoEmotions' training and validation sets then applied on either GoEmotions' test set or DailyDialog's test set, and the results using only DailyDialog's splits. These results serve as a good indication of performance goals for the later meta learning phase. We can see that the naive strategy to use a model trained on GoEmotions to predict DailyDialog's test set yields poor results with up to 34.58% F1-score even though it only considers the same 6 labels (*EmoTagSet3*) during training, validation and test to befit a standard supervised approach.

Meta Learning Quantitative Results. The bottom section of Table 2 shows two sets of results: the meta-training phase on GoEmotions (Demszky et al., 2020a) using splits by emotion labels (the *EmoTagSets* from Table 1) and evaluation of these models on the DailyDialog official test set. As expected, meta-learning yields results lesser than the supervised learning when the datasets come from the same source, but highly better ones when the dataset is from a different source. Indeed, the meta-learning process trains on data from different sources, with different tag sets, sentence lengths and conversational contexts. Results show that the more similar the linguistic structure of the train and target data are, the easier the work of meta-learning is, yielding better performance. Indeed, results of meta-learning obtained on GoEmotions are better

Supervised Learning

Supervised Learning trained on GoEmotions tested on GoEmotions (val set – 6 filtered classes)

Enc	Clf	Acc	±	F1	±	MCC	±
AVG	MLP	72.67	00.8	0.7254	00.8	67.23	00.9
CNN	MLP	76.37	00.7	0.7617	00.7	71.74	00.8
Tr.	MLP	**98.94**	00.7	**98.94**	00.6	**98.73**	00.8

Eval models trained on GoEmotions on DailyDialog (6 classes)

AVG	MLP	32.93	13.6	31.07	13.1	19.14	15.7
CNN	MLP	34.71	13.9	32.18	13.4	21.28	15.8
Tr.	MLP	**39.88**	18.5	**34.58**	18.2	**27.42**	23.2

Supervised Learning on DailyDialog Splits (6 classes)

Enc	Clf	Acc	±	F1	±	MCC	±
AVG	MLP	49.73	18.9	42.06	19.2	42.32	23.7
CNN	MLP	**62.57**	18.7	**54.89**	20.6	**59.12**	22.0
Tr.	MLP	55.35	21.11	48.52	21.4	49.24	26.1

Meta-Learning

Meta-Learning using GoEmotions 6 way 5 shot 30 query

Enc.	Clf	Acc	±	F1	±	MCC	±
AVG	Proto	25.20	03.5	23.92	03.6	10.61	04.4
CNN	Proto	31.35	04.5	29.82	04.6	17.95	05.5
BERT	Proto	39.82	04.9	39.11	05.1	28.11	05.9
Dist.	RR	31.92	04.9	31.1	05.1	18.81	06.0
Tr.	Proto	**93.02**	04.6	**91.64**	06.1	**92.08**	05.2

Eval Meta-Learned Models on DailyDialog's test set (1,000 episodes)

AVG	Proto	23.95	06.9	22.52	07.0	09.11	08.6
CNN	Proto	17.61	07.5	15.36	07.2	01.23	09.5
BERT	Proto	42.59	09.7	41.50	09.7	31.80	11.9
Dist.	RR	25.78	08.1	24.38	07.8	11.28	10.0
Tr.	Proto	**61.77**	20.8	**58.55**	24.1	**58.82**	22.4

Fine-tuning meta-learned models on GoEmotions test set (1 epoch of 10 episodes) Eval on DailyDialog's test set (1,000 episodes)

Enc.	Clf	Acc	±	F1	±	MCC	±
AVG	Proto	20.82	06.9	19.23	07.1	05.07	08.5
CNN	Proto	20.34	05.7	18.91	05.4	04.73	07.6
Tr.	Proto	28.59	09.9	21.13	10.6	17.22	13.1

Table 2: Top section: Supervised learning on utterances (official DailyDialog splits). Bottom section: meta learning trained by splitting classes from GoEmotions (train on 11, validate on 10, test on 6). The trained meta model is then applied on DailyDialog's test set. Evaluated using accuracy (Acc), F1-score (F1) and multi-class Matthews Correlation Coefficient (MCC). ± represents the variance over test episodes.

than the ones obtained on Daily Dialog. Contrary to what can be observed in supervised learning results, the Transformer, here associated with Prototypical Networks for meta-training, significantly outperforms other encoders. Even though, using the fine-tuned BERT as encoder yields a slightly better F1-score than recent models such as ridge regressor with distributional signature in our use-case scenario but, more importantly, BERT results show less variance ($\pm$) than our best model. However, our data being not segmented at the sentence level and possessing excessive variable numbers of tokens, BERT cannot be used to its full extent. This confirms prior conclusions from related work (Bao et al., 2020). We believe the poor results yielded by using the CNN (Kim, 2014) as encoder demonstrate the need of attention in the training process to better capture usable meta-information. These results using a Transformer layer (Tr.), BERT or attention generator with ridge regressor (RR) as encoders would confirm previous studies making the same observation (Sun et al., 2019).

If we compare our approaches, using attention based algorithms, to the architecture using distributional signatures with Ridge Regressor presented by Bao *et al.* (Bao et al., 2020), we can see we constantly outperform it on the evaluation metrics used. Moreover, fine-tuning the models trained on GoEmotions using GoEmotions' test set for 10 additional episodes did not improve the final scores. We believe this is due to the fine-tuning starting to change the model's parameters but, by doing so, changing the previously learned meta information.

Meta Learning Qualitative Results. Our best model manages to obtain good results based on quantitative evaluation even if those scores decrease a lot when applied on data from another source and phrasal structure, as shown in Section 1. Table 3 presents one mistake example for each emotion label in the test set. These examples show the most common mistake for each emotion. For instance, the **True** label "joy" is most commonly mistaken with "surprise" (the predicted – **Pred** – label) by the model; "sadness" is most commonly mistaken with "surprise", and so on. These two datasets coming from different platforms, further analysis is needed to dive into the different topics tackled in these messages, which may be one of the main obstacles to obtaining higher performance. We discuss it in the next section (Section 6). The message structure relates to the type of conversations: GoEmotions (*i.e.* Reddit) seems to have a higher number of general comments about a third object/topic/person, while DailyDialog seems to be made of personal discussions between people that are close to each other.

Text	True	Pred
Oh, yes, I would!	joy	surprise
Yelling doesn't do any good.	sadness	surprise
Yes. Then I noticed he was on the sidewalk behind me. He was following me.	anger	disgust
What's wrong with you? You look pale.	fear	surprise
This is all too fast. He's my best friend, and now he's gone.	disgust	surprise
What? What kind of drugs was he using?	surprise	anger

Table 3: Some mistakes made by our best meta-model (Table 2) meta-trained on GoEmotions and applied on DailyDialog. Each line is one example from the most frequent label confusion (*eq.* "joy" mistaken for "surprise" by the model).

6 Discussions

How do meta-trained models manage to perform on previously unseen tags? Prototypical Networks use the support set to compute a prototype for each class (*i.e. way*), hence new prototypes are computed for each episode. This means the trained encoder does not rely on predicting classes, but gathers representative information that will determine the position of the elements in the embedding space. Because it is the *relative* proximity that serves to assign a query element to a specific prototype, having a different tag set that will be embedded "far away" should not hinder how well the model can classify data.

Emotion Label Ambiguity. The 21 emotions from GoEmotions that we use for training and validation are fine-grained but could have overlaps ("annoyance" and "embarrassment" for instance); this is why a mapping to the same 6 emotions as the *EmoTagSet3* is provided with the data set (Table 1). Considering how well the meta-learning works on the emotion label part (see GoEmotions results in Table 2), achieving 91.64% in F1 score, labels' ambiguity and the different granularity seem to be handled well. Moreover, it should be noted that the labels were obtained differently for the two data sets: in isolation for GoEmotions and consider-

ing the conversation context for DailyDialog. This makes the task even more difficult.

Meta-learning through Different Data Sources. We want here to investigate whether the difficulty of this meta-learning task comes from varying tag sets or data sources. We fine-tune the models meta-trained on GoEmotions in order to slightly adapt the encoder to the target tag set (*EmoTagSet3*) by leveraging meta-information related to emotion labels. The training tag set is now the same as DailyDialog. The fine-tuning consists of 1 epoch of 10 more episodes instead of a maximum of 1,000 epochs made of 100 episodes during training. Results are reported at the bottom of Table 2. This fine-tuning produced worse results compared to simply meta-training and applying on a different target tag set. This leads to the hypothesis that the different linguistic structures from the two data sources (social network and daily communications) are the main sources of errors in this setup.

To confirm this, we look further in the data sources' specifics of GoEmotions (User Generated Content) and DailyDialog (an idealized version of dyadic daily conversations) by using machine learning based exploration. We study the most frequent nouns that are specific to each corpus. We use SpaCy[3] in order to obtain the Universal Part-of-Speech (UPOS) tags (Nivre and *al.*, 2019) along with the lemmas for both corpora. Then, we retrieve the sets of nouns for each corpus and compute the symmetric difference between both sets in order to see the differences in language level. GoEmotions being User Generated Content (UGC) from Reddit, its top 5 most frequent exclusive nouns are "lol", "f**k" (censored), "op", "reddit", and "omg". On the other hand, the top 5 most frequent exclusive nouns in DailyDialog are "reservation", "madam", "doesn" (tagging error), "taxi", and "courses". It shows a first indication both of language register and lexical field differences[4]. To further confirm the language structure differences, we retrieved the UPOS tags frequencies for both corpora. GoEmotions' top 3 UPOS are "NOUN", "VERB", and "PUNCT" while DailyDialog's top 3 are "PUNCT", "PRON", "VERB". This indicates DailyDialog's language follows a well formed structure with punctuation and pronouns while GoEmotions' language structure is more direct with mainly nouns and verbs[5]. All these data sources' specifics can provide explanation for the lower performance of our system on DailyDialog. The data sources' differences lead to prototypes differences during the two testing phases. Table 4 shows that the average euclidean distance between query elements $\mathbf{x}$ and class prototypes $\mathbf{c}_{k'}$ from the same class $-d\left(f_\phi(\mathbf{x}), \mathbf{c}_{k'}\right)$ is greater when tested on GoEmotions than on DailyDialog.

happiness	sadness	anger	fear	disgust	surprise
-9.30	-9.65	-8.80	-9.23	-9.32	-8.71
-7.91	-8.12	-8.15	-8.18	-8.09	-8.11

Table 4: Average euclidean (l2) distance from queries to predicted emotions using our best model (Tr.+Proto), on GoEmotions (go) and DialyDialog (dd).

Varying Pre-Trained Language Models. To confirm our preliminary results on pre-trained language models on this task, we further explore fine-tuning several of them. Results are visible in Table 5. In addition to BERT, we fine-tune XLNet (Yang et al., 2019) (*xlnet-base-cased*) and RoBERTa (Liu et al., 2019) (*roberta-base*) from the Transformers library (Wolf et al., 2019) along with their distilled variants. Results show fine-tuning BERT is better than other pre-trained language models on this task. This confirms our initial results on Table 2 of our model being better at retaining meta-information while only considering static pre-trained embeddings from FastText (Joulin et al., 2017).

Enc.	Acc	±	F1	±	MCC	±
DistilBERT	23.24	±04.0	22.98	±04.1	08.11	±04.8
XLNET	25.80	±04.2	25.85	±04.1	11.06	±04.8
roBERTa	25.58	±04.1	25.17	±04.0	10.76	±05.0
distilroBERTa	27.38	±04.5	26.83	±04.4	12.86	±05.3
BERT	**42.59**	09.7	**41.50**	09.7	**31.80**	11.9

Table 5: Results on DailyDialog's test set using multiple pre-trained language models for meta learning following the same scenario as Table 2's bottom section: meta trained on GoEmotions and meta test on DailyDialog. These language models are fine-tuned during meta-training.

Using Empathetic Dialogues as Training Source. We consider the same meta learning scenario using a different data set to train the meta-models. We choose utterances from the Empathetic Dialogues (Rashkin et al., 2019) full

[3] https://spacy.io/

[4] For more details, see the tables 8 and 9 in appendix.

[5] For more details, see figures 3 and 4 in the appendix.

data set while considering the dialogues label (i.e. the "context" column) as the label for each utterance. To apply meta learning on emotion labels, we select labels based on balancing polarity and numbers of occurrences, leading us to consider the following sets: 13 labels for training (*caring, confident, content, excited, faithful, embarrassed, annoyed, devastated, furious, lonely, terrified, sentimental, prepared*), 13 different labels for validation (*grateful, hopeful, impressed, trusting, proud, embarrassed, annoyed, devastated, furious, lonely, terrified, sentimental, prepared*) and 6 test emotions, keeping the set from DailyDialog (*joyful, sad, angry, afraid, disgusted, surprised*). Results for this meta learning experiment using Empathetic Dialogues are shown in Table 6.

Meta-Learning using ED
6 way 5 shot 30 query

Enc.	Clf.	Acc	±	F1	±	MCC	±
AVG	Proto	27.43	±04.2	25.95	±04.3	13.16	±05.2
Dist.	RR	31.73	±04.7	31.11	±05.1	18.51	±05.8
Tr.	Proto	97.80	±03.4	97.54	±04.1	97.49	±03.8
Eval Meta-Learned Models on DailyDialog's test set (1,000 episodes)							
AVG	Proto	18.07	±03.0	16.58	±03.1	02.21	±03.8
Dist.	RR	26.29	±08.1	24.90	±08.1	11.86	±10.0
Tr.	Proto	**66.24**	±18.2	**66.09**	±18.0	**60.43**	±21.9

Table 6: Meta learning trained on Empathetic Dialogues (ED) before applying the model on DailyDialog's test set.

Empathetic Dialogues is a merge of multiple data sets, with DailyDialog among them. Hence, evaluating the meta model learnt using Empathetic Dialogues on DailyDialog's test set does not allow for fair comparison with our previous model. Indeed, we obtain here significantly better results on DailyDialog's test set. However, results show similar trends between evaluation sets and types of models as our main meta learning scenario (Table 2), which confirms our overall conclusions on this task.

7 Conclusion

In this paper, we are interested in a classification scenario where we only possess a certain kind of training data, with no guarantee that the testing data will be of the same type nor use the same labels. We choose our training data from common social media sources (Reddit) with fine-grained emotion labels. We address this problem using meta-learning and few-shot learning, to evaluate our model on conversation utterances with a simpler emotion tag set.

We consider metric learning based meta-learning by setting up Prototypical Networks with a Transformer encoder, trained in an episodic fashion. We obtained encouraging results when comparing our meta-model with a supervised baseline. In this use-case scenario with a two-level meta-learning, our best meta-model outperforms both other encoder strategies and the baseline in terms of meta-learning for NLP. Moreover, our approach works well for learning emotion-related meta-information but still struggles while varying data types.

For future work, we wish to investigate if this meta-learning approach could integrate the conversational context for classifying the utterances of the target dialog data. We also plan on applying this approach to another language than English.

Acknowledgements

We want to thank the anonymous reviewers for their insights and useful suggestions. This allowed us to better put forward our contributions by specifying additional comparisons.

References

Maruan Al-Shedivat, Liam Li, Eric Xing, and Ameet Talwalkar. 2021. On data efficiency of meta-learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1369–1377. PMLR.

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, pages 579–586, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

P. Baldi, Søren Brunak, Y. Chauvin, and Henrik Nielsen. 2000. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16(5):412–424.

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534, Online. Association for Computational Linguistics.

Yujia Bao, Menghua Wu, Shiyu Chang, and Regina Barzilay. 2020. Few-shot text classification with dis-

tributional signatures. In *International Conference on Learning Representations*.

Sarah Bechtle, Artem Molchanov, Yevgen Chebotar, Edward Grefenstette, Ludovic Righetti, Gaurav Sukhatme, and Franziska Meier. 2021. Meta-learning via learned loss.

Alberto Bernacchia. 2021. Meta-learning with negative learning rates.

Johan Bollen, Alberto Pepe, and Huina Mao. 2009. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv:0911.1583 [cs]*. ArXiv: 0911.1583.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6.

Harald Cramir. 1946. Mathematical methods of statistics. *Princeton U. Press, Princeton*, 500.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020a. 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.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020b. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017a. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia. PMLR.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017b. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR.

Takuya Fujioka, Dario Bertero, Takeshi Homma, and Kenji Nagamatsu. 2019. Addressing ambiguity of emotion labels through meta-learning. *CoRR*, abs/1911.02216.

Takuya Fujioka, Takeshi Homma, and Kenji Nagamatsu. 2020. Meta-learning for speech emotion recognition considering ambiguity of emotion labels. *Proc. Interspeech 2020*, pages 2332–2336.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019a. Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6407–6414.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019b. FewRel 2.0: Towards More Challenging Few-Shot Relation Classification. *arXiv:1910.07124 [cs]*. ArXiv: 1910.07124.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction Networks for Few-Shot Text Classification. *arXiv:1902.10482 [cs]*. ArXiv: 1902.10482.

Jan Gorodkin. 2004. Comparing two k-category assignments by a k-category correlation coefficient. *Computational biology and chemistry*, 28(5-6):367–374.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. *arXiv:1810.10147 [cs, stat]*. ArXiv: 1810.10147.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2020. Meta-Learning in Neural Networks: A Survey. *arXiv:2004.05439 [cs, stat]*. ArXiv: 2004.05439.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image Recognition. *ICML*, page 8.

Brenden Lake. 2015. LakeEtAl2015Science-startOfFewShot.pdf. *Sciences Mag*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

Sung-Chiang Lin, Chih-Jou Chen, and Tsung-Ju Lee. 2020. A multi-label classification with hybrid label-based meta-learning method in internet of things. *IEEE Access*, 8:42261–42269.

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.

Anugunj Naman and Liliana Mancini. 2021. Fixed-maml for few shot classification in multilingual speech emotion recognition.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Joakim Nivre and *al.* 2019. Universal dependencies 2.4. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Sungjoon Park, Jiseon Kim, Jaeyeol Jeon, Heeyoung Park, and Alice Oh. 2019. Toward dimensional emotion detection from categorical emotion annotations. *arXiv preprint arXiv:1911.02499*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning.

Jürgen Schmidhuber. 1987. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. Ph.D. thesis, Technische Universität München.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical Attention Prototypical Networks for Few-Shot Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485, Hong Kong, China. Association for Computational Linguistics.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Wenpeng Yin. 2020. Meta-learning for Few-shot Natural Language Processing: A Survey. *arXiv:2007.09604 [cs]*. ArXiv: 2007.09604.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xue-jie Zhang. 2015. Predicting Valence-Arousal Ratings of Words Using a Weighted Graph Method. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 788–793, Beijing, China. Association for Computational Linguistics.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse Few-Shot Text Classification with Multiple Metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215, New Orleans,

Louisiana. Association for Computational Linguistics.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018a. Text Emotion Distribution Learning via Multi-Task Convolutional Neural Network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 4595–4601, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence Organization.

Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018b. Text emotion distribution learning via multi-task convolutional neural network. In *IJCAI*, pages 4595–4601.

Zhenjie Zhao and Xiaojuan Ma. 2019. Text emotion distribution learning from small sample: A meta-learning approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3957–3967, Hong Kong, China. Association for Computational Linguistics.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. *arXiv:1909.10681 [cs]*. ArXiv: 1909.10681.

Pan Zhou, Xiaotong Yuan, Huan Xu, Shuicheng Yan, and Jiashi Feng. 2019. Efficient meta learning via minibatch proximal update. *Advances in Neural Information Processing Systems*, 32:1534–1544.

Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial Attention Modeling for Multi-dimensional Emotion Regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 471–480, Florence, Italy. Association for Computational Linguistics.

A Open Source Code

The anonymous code is available to reviewers in supplementary materials. A link to the Public Github repository containing the code to run experiments along with data will be added to the article. The code base has been implemented in Python using, among others, PyTorch and Hugging Face Transformers (Wolf et al., 2019) for BERT. All training runs were made using an Nvidia V100 Tensor Core GPU[6].

B Hyper Parameters

Prototypical networks hidden size is set to $[300, 300]$ which is equal to the base embedding size (300 from pre-trained FastText on Wiki News[7]), global dropout is set to 0.1.
CNN hyper parameters:

- cnn filter sizes: 3, 4, 5
- number of filters: 5000
- learning rate: 0.001

Transformer hyper parameters:

- learning rate: 0.0001
- transformer dropout: 0.2
- embedding size: 300 (from FastText)
- attention heads: 2
- hidden size: 300
- transformer encoder layers: 1
- position encoding dropout: 0.1

Please note that these hyper parameters are the one inferred from the supervised learning. During meta-learning we only change the learning rate and set it to $1e-5$ as explained in Section 4 of the paper.

[6] https://www.nvidia.com/en-us/data-center/v100/

[7] https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M.vec.zip

C Training Additional Information

Models trained for 72 epochs using average embeddings as encoder, 42 epochs using Transformer encoder, and 35 epochs using CNN as encoder. Depending on the run, our best meta-model (Transformers with Prototypical Networks using a learning rate of 1e-5) converges between the 87th epoch and the 165th epoch. The total training time does not exceed one hour.

D Additional Results Information

Figure 2 shows the confusion matrix for our best meta-model trained on GoEmotions and applied on DailyDialog (the row obtaining 58.55% F1 score in Table 2).

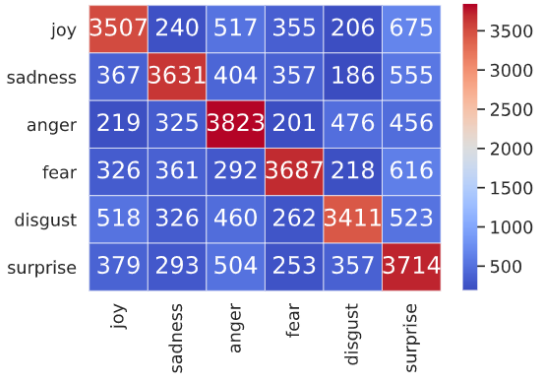

Figure 2: Confusion matrix for our Tr.+Proto meta-learning trained on GoEmotions and tested on DailyDialog. This is the 1,000 test episodes' outputs merged together. Rows represent reference labels while columns represent predicted labels.

To ensure the relative stability of our best model, we did 3 meta-learning runs using our Transformer encoder in Prototypical Networks using a learning rate a 1e-5. The results of these runs (including the one reported in Table 2) are visible in Table 7.

E Data Comparison & Information

In Section 6 we discussed data sources differences. Here you can see more in-depth information. On the other hand, Tables 8 and 9 shows side by side the top ten most frequent tokens for the predicted NOUN UPOS. Figures 3 and 4 show the predicted part-of-speech distribution for each corpus.

Runs (trained and applied on GoEmotions)				
Encoder	Classifier	Accuracy	F1-score	MCC
Transformer	Proto	0.9302 ± 0.0463	0.9164 ± 0.0607	0.9208 ± 0.0515
Transformer	Proto	0.9183 ± 0.0423	0.9016 ± 0.0572	0.9075 ± 0.0468
Transformer	Proto	0.9301 ± 0.0464	0.9163 ± 0.0608	0.9207 ± 0.0516
Runs (same model applied on DailyDialog)				
Encoder	Classifier	Accuracy	F1-score	MCC
Transformer	Proto	0.6177 ± 0.2078	0.5855 ± 0.2408	0.5882 ± 0.2241
Transformer	Proto	0.6573 ± 0.2016	0.6256 ± 0.2354	0.6248 ± 0.2179
Transformer	Proto	0.6253 ± 0.2093	0.5929 ± 0.2442	0.5937 ± 0.2258

Table 7: Additional runs of our best model to ensure results' stability.

token	count
lol	576
f**k	248
op	204
reddit	147
omg	145
lmao	143
' '	133
congrats	115
*	110
meme	106

Table 8: Top 10 frequent nouns (SpaCy) exclusive to GoEmotions

token	count
reservation	267
madam	143
doesn	142
taxi	127
courses	102
shipment	79
noon	50
aren	49
aisle	47
exhibition	45

Table 9: Top 10 frequent nouns (SpaCy) exclusive to DailyDialog

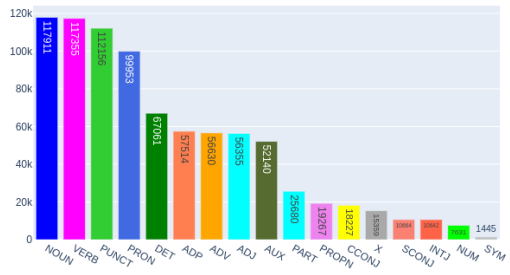

Figure 3: GoEmotions POS distribution (POS tagged using SpaCy)

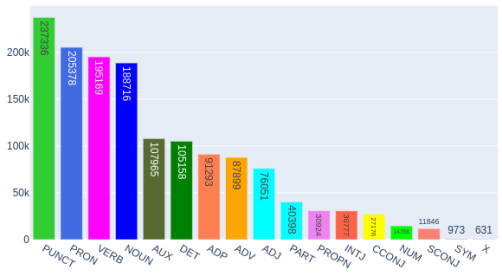

Figure 4: DailyDialog POS distribution (POS tagged using Spacy)