

# Probabilistic Ensembles of Zero- and Few-Shot Learning Models for Emotion Classification

**Angelo Basile**  
Symanto Research  
Nurenberg, Germany  
angelo.basile@symanto.com

**Guillermo Pérez-Torró**  
Symanto Research  
Valencia, Spain  
guillermo.perez@symanto.com

**Marc Franco-Salvador**  
Symanto Research  
Valencia, Spain  
marc.franco@symanto.com

## Abstract

Emotion Classification is the task of automatically associating a text with a human emotion. State-of-the-art models are usually learned using annotated corpora or rely on hand-crafted affective lexicons. We present an emotion classification model that does not require a large annotated corpus to be competitive. We experiment with pretrained language models in both a zero-shot and few-shot configuration. We build several of such models and consider them as biased, noisy annotators, whose individual performance is poor. We aggregate the predictions of these models using a Bayesian method originally developed for modelling crowdsourced annotations. Next, we show that the resulting system performs better than the strongest individual model. Finally, we show that when trained on few labelled data, our systems outperform fully-supervised models.

## 1 Introduction

A large part of Natural Language Processing (NLP) research is focused on building technology to automatically extract information from large collections of texts. However, text contains often not just mere factual information, but also opinions, attitudes and emotions. Applications of emotion-aware NLP models range from established tasks such as analysis of product reviews (Blitzer et al., 2007) and development of “emotional” chatbots (Chatterjee et al., 2019) to less obvious tasks such as the analysis of developer experience on Stack Overflow (Novielli et al., 2018), author profiling (Rangel and Rosso, 2016) and the prediction of mental health disorders (Uban et al., 2021).

In this work, we focus on the fine-grained emotion classification task (Strapparava and

Mihalcea, 2007): given a document, an emotion label must be provided. For example, the sentence *The angry wolf ate the happy boy* could be associated with *fear* or *sadness* if the emotion is being modelled from the reader’s perspective; alternatively, it could be associated with *anger* or *joy* considering the text’s perspective. In the emotion classification literature, the targeted emotion perspective is rarely made explicit (Bostan et al., 2020), which — in addition to the subjectivity involved in the annotation task — makes it difficult to obtain large amounts of high quality data (Bobicev and Sokolova, 2017; Troiano et al., 2021). Furthermore, with few exceptions (Mohammad et al., 2018; Lamprinidis et al., 2021), most of the research has been conducted on English corpora: most of the other languages can be considered low-resourced with respect to affective corpora.

In this work, we aim to minimize the amount of annotated data needed to obtain competitive performance in the task of emotion classification. Several emotion theories exist, which differ on the emotion inventory and representation type (categorical vs. continuous): in this work we focus on the categorical paradigm, using the mix of different inventories available from the Unify Emotion dataset (Bostan and Klinger, 2018).

We describe the use of pretrained language models (PLMs) for emotion classification in both *few-shot* and *zero-shot* scenarios (Section 4). Few-shot models are supposed to solve a task using only few annotated instances; zero-shot models are supposed to use none. These models have been shown to perform well in different tasks (Yin et al., 2019; Schick and Schütze, 2021b; Wang et al., 2021). However, in a real unsupervised scenario (i.e., without

an evaluation split), their performance is by definition unknown. To mitigate the risks involved in deploying such models, we experiment with a probabilistic ensemble that combines the individual – and potentially biased and noisy – outputs of those models. We use the Multi-Annotator Competence Estimation model (Hovy et al., 2013), a Bayesian method designed to deal with noisy crowdsourced annotations (Section 5). Experimental results (Section 6) show that our ensemble performs better than the strongest individual model. In addition, we show that just fine tuning with few labeled data, our system outperforms fully-supervised models.

## 2 Related Work

Attempts to minimize the amount of hand-labelled data required to train emotion-aware NLP models have mostly focused on using distant supervision (Go, 2009) to collect large amounts of silver labels for training models in a supervised fashion: emoji (Felbo et al., 2017), emoji description (Eisner et al., 2016), and hashtags (Mohammad, 2012) have been shown to be good proxies for emotion classification.

The idea of using label templates for unsupervised classification can be traced back at least to Hearst patterns (Hearst, 1992). Within the neural paradigm, Cloze (Taylor, 1953) label templates are used by Schick and Schütze (2021a), who obtain strong results on few-shot classification. Yin et al. (2019) and Wang et al. (2021) use templates to generate synthetic data for framing text-classification as entailment. An alternative neural approach to unsupervised classification embeds both the input sequence of text and the set of possible label names in the same semantic space and selects the one which maximizes a defined similarity metric (Gabrilovich and Markovitch, 2007). Pushp and Srivastava (2017) concatenates both input and label embeddings and classify their relatedness. A recent survey of template-based or “prompt-based” learning can be found in Liu et al. (2021). Perhaps the closest work to ours is Yin et al. (2019), who evaluate zero-shot text classification on the Unify Emotion dataset (Bostan and Klinger, 2018).<sup>1</sup> Our main

<sup>1</sup>We don’t compare our results directly to Yin et al. (2019): even though both datasets stem from the Unify

contribution with respect to Yin et al. (2019) is the suggestion of a principled way for *a*) aggregating multiple predictions without having any access to a model’s performance and *b*) inferring the most probable emotion label given multiple models.

An overview of Bayesian models of annotation can be found in Paun et al. (2018). The idea of using a generative model to infer a latent label from multiple signals has recently been presented in a unified framework by Ratner et al. (2016). An alternative to Bayesian models for aggregating predictions can be found in Poerner and Schütze (2019), who apply Generalized Canonical Correlation Analysis to build an ensemble of unsupervised BERT models for Duplicate Question Detection in a low-resource scenario.

## 3 Data

	train	validation	test
anger	5147	1714	1717
anticipation	191	64	63
confusion	77	26	26
disgust	2701	900	899
fear	9592	3196	3199
guilt	656	218	219
joy	22338	7448	7446
love	2292	764	764
noemo	62692	20897	20898
sadness	9185	3061	3061
shame	658	219	219
surprise	5392	1795	1798
trust	485	162	161

Table 1: Overview of the dataset.

For all our experiments we use a section of the Unify Emotion dataset (Bostan and Klinger, 2018) which aggregates several annotated corpora in a common format. Specifically, we use the following datasets: Grounded-Emotions (Liu et al., 2007), CrowdFlower (Crowdfower), DailyDialog (Li et al., 2017), TEC (Mohammad, 2012), Electoral-Tweets (Mohammad et al., 2015), ISEAR (Scherer and Wallbott., 2017), Emotion-Stimulus (Ghazi et al., 2015), Tales-Emotion (Alm et al., 2005; Alm and Sproat, 2005; Alm, 2008), and EmoInt (Mohammad et al., 2017). We aggregate all the corpora and then sample from each class 60% of the data for the train split, and 20% for the

Emotion dataset, the actual instances and label set used are different.

development and the test split, respectively. The annotation quality, domain, annotation procedure (manual vs. semi-automatic), and topic differ among the various datasets. We refer to (Bostan and Klinger, 2018) for additional details about the specific datasets.

Table 1 highlights the label distribution in the dataset. As it is the case for most available emotion corpora, the labels are heavily unbalanced.

#### 4 Entailment as Zero-Shot and Few-Shot Learning

Given two sentences, a premise and a hypothesis, they can be related by an *entailment*, *contradiction* or *neutral* relation. The task of Natural Language Inference (NLI) (Dagan et al., 2005) aims at predicting such relations. Recently, the creation of large NLI datasets (Bowman et al., 2015; Williams et al., 2018; Thorne et al., 2018; Conneau et al., 2018) has allowed deep learning methods to achieve state-of-the-art performance on the NLI task, outperforming logic-based approaches. The high performance of BERT-like models (Devlin et al., 2019) on NLI tasks can be exploited to successfully tackle general classification tasks by recasting them as entailment problems: pretrained language models can be finetuned on NLI datasets and these finetuned models can be then re-purposed to attack different problems (Wang et al., 2021). For modelling the emotion classification problem, we follow Yin et al. (2019) and given a text to classify (the hypothesis), we build pseudo-sentences to serve as premises, one for each target label. For instance, the input sentence “John said he loved the pizza” can be classified as JOY, if an NLI models predicts that it entails the artificial sentence “This person expressed a feeling of *pleasure*”. We can substitute *pleasure* with other emotion-expressing words and map them to specific target labels (e.g., *pleasure* to JOY, *sad* to SADNESS, etc.) to build a system for zero-shot, multi-label emotion classification.

Following Yin et al. (2019), in this work we explore two options to formulate the hypotheses: based on the label’s name and on the label’s WordNet (Fellbaum, 1998) definition. We show the details about our hypotheses for emotion classification in Table 2.

We experiment with six different pretrained NLI models that differ in terms of the underlying pretrained language model (BART (Lewis et al., 2020), RoBERTa (Liu et al., 2019) and XLM-RoBERTa (Conneau et al., 2020)) and NLI dataset used for training (Multigenre NLI (MNLI) (Williams et al., 2018), Adversarial NLI (ANLI) (Nie et al., 2020) and XNLI (Conneau et al., 2018)). In Table 3.B of Section 6 we include the details about our models.<sup>2</sup> The zero-shot setup motivates the usage of a variety of pretrained NLI models: given that in a true zero-shot scenario no development dataset is available, assessing how different NLI training data and pretrained language models impact the performance is of crucial importance.

We conduct the few-shot learning experiments by fine-tuning a pretrained entailment model. To build the training data, we fill the templates used in the zero-shot setup with the gold labels and training follows the standard sentence pair classification task used to train the original entailment model.

#### 5 A Probabilistic Ensemble

In a true zero-shot classification scenario, no development set is available and therefore a method for estimating the performance of the model on the specific input data is required. In this work, we propose to use several different models and to infer the best possible answer using a probabilistic model. Such a model has two advantages over a simple majority voting strategy: first, it has been shown to outperform majority voting (Snow et al., 2008); second, it provides a confidence value for each instance and estimates the models’ accuracy.

To aggregate the predictions from the different unsupervised models, we use the Multi-Annotator Competence Estimation (MACE) model (Hovy et al., 2013). This model has been originally developed to analyse crowdsourced annotations for both identifying unreliable annotators and retrieving the true labels. Figure 1 shows the plate diagram of the model and we refer to the original publication for further details. Algorithm 1 describes the generative process.

We generalize the notion of annotator to also

<sup>2</sup>We downloaded the models from <https://huggingface.co>.

Label	Label-based hypothesis	WordNet-based hypothesis
anger	(...) feels angry	(...) expresses a strong feeling of annoyance, displeasure, or hostility
anticipation	(...) has a feeling of anticipation	(...) is anticipating something, expecting or predicting something about to happen
confusion	(...) is feeling confused	(...) is feeling disoriented and can not think clearly or focus to do something
disgust	(...) feels disgusted	(...) expresses a feeling of revulsion or strong disapproval aroused by something unpleasant or offensive
fear	(...) is afraid of something	(...) expresses an unpleasant emotion caused by the belief that someone or something is dangerous, likely to cause pain, or a threat
guilt	(...) feels guilty	(...) expresses a feeling of having done wrong or failed in an obligation
joy	(...) feels joyful	(...) expresses a feeling of great pleasure and happiness
love	(...) loves that	(...) expresses a great interest and pleasure in something
noemo	(...) does not feel any emotion	(...) is insensitive, showing unfeeling and unresponsive behaviour, with a lack of emotion about the situation
sadness	(...) feels sad	(...) expresses emotions experienced when not in a state of well-being
shame	(...) feels shameful	(...) expresses a painful feeling of humiliation or distress caused by the consciousness of wrong or foolish behavior
surprise	(...) feels surprised	(...) expresses a feeling of mild astonishment or shock caused by something unexpected
trust	(...) feels trusty about this	(...) has a strong belief in the reliability, truth, or ability of someone or something

Table 2: Formulation of label as hypotheses for entailment. All our hypotheses start with *This person* (...).

---

**Algorithm 1:** Generative process of the MACE model.

---

```

for item  $i \in I$  do
  draw  $G_i \sim \text{Uniform};$ 
  for annotator  $n \in N$  do
    draw  $B_{i,n} \sim \text{Bernoulli}(1 - \theta_j);$ 
    if  $B_{i,n} == 0$  then
       $y_{i,n} = G_i;$ 
    else
       $y_{i,n} \sim \text{Multinomial}(\xi_j)$ 

```

---

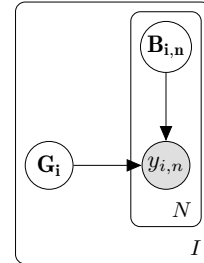


Figure 1: The MACE model. Given  $I$  instances and  $N$  annotators, the observed label  $y_{i,n}$  is dependent on the gold label  $G_i$  and  $B_{i,n}$ , which models the behaviour of annotator  $n$  on instance  $i$ . The model parameters  $\theta$  and  $\xi$  are left out.

include model annotations. The latent variable  $B$  of the model has been originally introduced to model the behaviour of crowdworkers as spammer or not spammer, while in our setup,  $B$  represents the “fitness” of an unsupervised model. As it is usually hard to know if an annotator is spamming, similarly, in a true zero-shot classification scenario (i.e., without a validation set), it is not possible to know if a model is fit to the task. We use a custom Python implementation of the model.<sup>3</sup>

## 6 Evaluation

We evaluate the unsupervised models in a zero-shot configuration against two supervised baselines. We then experiment with adding increasing amounts of supervision to both the baselines and the entailment model in a few-shot setting. We conduct all the evaluation using

<sup>3</sup>The original Java implementation can be found at <https://github.com/dirkhov/MACE>

the standard classification metrics: precision, recall, and macro-averaged f1-score.

**Baselines** As upper-baselines for our experiments, we train two supervised models on all the available training data: we train both a neural network based on RoBERTa-base (Liu et al., 2019) and a linear SVM using character  $n$ -grams as representations (henceforth referred to as Char-SVM).<sup>4</sup>

**Entailment Models** We assemble 12 different unsupervised zero-shot classification models finetuned on the NLI task. The models differ in terms of the pretrained language model, label template and NLI finetuning data. Yin et al. (2019) explore two different evaluation scenarios: *label-partially-unseen* and *label-*

<sup>4</sup>We use a custom implementation of a RoBERTa and use the pretrained model from <https://huggingface.co/models>. For the SVM model, we use the LinearSVM implementation contained in scikit-learn (Pedregosa et al., 2011).
















	model	precision	recall	f1-score	lang	NLI dataset	Hypothesis
A	Char-SVM	47.13	36.12	39.77	English	-	-
	RoBERTa-base	<b>48.88</b>	<b>39.60</b>	<b>41.29</b>	English	-	-
B	bart-large	26.43	28.50	17.91	English	MNLI	L
	bart-large	16.11	21.74	12.76	English	MNLI	W
	distilbart-12-1 	26.61	<b>30.40*</b>	<b>20.12</b>	English	MNLI	L
	distilbart-12-1 	24.15	19.40	13.11	English	MNLI	W
	distilbart-12-9 	25.96	<b>30.48*</b>	18.91	English	MNLI	L
	distilbart-12-9 	22.33	20.73	12.39	English	MNLI	W
	roberta-large	20.93	25.99	14.16	English	MNLI	L
	roberta-large	20.71	23.95	11.20	English	MNLI	W
	xlm-roberta-large	23.50	18.46	10.62	Multilingual	XNLI-ANLI	L
	xlm-roberta-large	16.63	17.93	8.01	Multilingual	XNLI-ANLI	W
	xlm-roberta-large	<b>27.74*</b>	21.05	10.85	Multilingual	XNLI	L
xlm-roberta-large	19.77	17.22	9.05	Multilingual	XNLI	W	
C	majority	<b>25.44</b>	<b>29.26</b>	16.43	-	-	-
	MACE	22.53	28.57	<b>20.96*</b>	-	-	-
D	distilbart-fs-8 	26.41 $\pm$ 0.02	32.75 $\pm$ 0.02	19.54 $\pm$ 0.04	English	MNLI	L
	distilbart-fs-16 	27.75 $\pm$ 0.03	36.55 $\pm$ 0.01	22.88 $\pm$ 0.01	English	MNLI	L
	distilbart-fs-32 	25.91 $\pm$ 0.01	42.17 $\pm$ 0.00	23.81 $\pm$ 0.01	English	MNLI	L
	distilbart-fs-64 	27.57 $\pm$ 0.02	46.72 $\pm$ 0.01	26.93 $\pm$ 0.02	English	MNLI	L
	distilbart-fs-128 	29.35 $\pm$ 0.01	51.64 $\pm$ 0.01	31.91 $\pm$ 0.01	English	MNLI	L
	distilbart-fs-256 	32.15 $\pm$ 0.01	55.25 $\pm$ 0.01	36.13 $\pm$ 0.01	English	MNLI	L
	distilbart-fs-512 	34.64 $\pm$ 0.01	57.86 $\pm$ 0.00	39.34 $\pm$ 0.01	English	MNLI	L
	distilbart-all 	46.97 $\pm$ 0.00	50.43 $\pm$ 0.01	48.27 $\pm$ 0.00	English	MNLI	L

Table 3: Overview of the evaluation results. Scores are macro-averaged. L: embedded label name for hypothesis representation; W: WordNet definition for hypothesis representation. : distilled model. Rows in A: fully supervised. Rows in B: zero-shot. Rows in C: aggregations. Rows in D: few-shot learning, each row denotes the number of training instances; the results are averaged over three runs and the standard deviation is shown in subscript. Statistically significant results according to a  $\chi^2$  test, per sub-table, are highlighted in bold. Significant results between Zero-shot and the aggregations (B and C sub-tables), are highlighted with \*.

*fully-unseen*. In the partially-unseen setup, a model is trained on a subset of the label set and then evaluated on the full dataset; in the fully-unseen setup, no labelled data is shown to the model. In this work we do not take into account the *label-partially-unseen* because, as stated by Yin et al. (2019), this is a restrictive definition of the zero-shot paradigm, unlike the *label-fully-unseen* scenario. For the few-shot evaluation, we only train the best-performing model (distilbart-mnli-12-1). Information about the used hyperparameters can be consulted in A

**Probabilistic Ensemble** We train a MACE model using Variational-Bayes on the predictions of the 12 zero-shot entailment models. We train the model for 100 iterations and 50 restarts; we use the default values (0.5) for the  $\alpha$  and  $\beta$  parameters.

**Results** As reported in Table 3, emotion classification is a challenging task even for

fully-supervised models trained on large annotated datasets. Interestingly, RoBERTa-base, a large neural model, outperforms the Char-SVM model only by few points. The battery of zero-shot models shows a large variation in terms of performance, ranging from 8.01% f1-score for XLM-RoBERTa-large with WordNet-based hypothesis to 20.12% for distilbart with label names. The performance of that distilled entailment model is remarkable, considering that bart-large uses twice the number of parameters of its distilled version. On average, name-based templates outperform WordNet-based ones. Aggregating the predictions of the zero-shot models using MACE leads to a much higher f1-score when compared to majority voting. The MACE-based ensemble outperforms the strongest zero-shot model in terms of f1-score by a small but statistically significant margin (+0.84% f1-score). However, in a true zero-shot scenario, where no evaluation

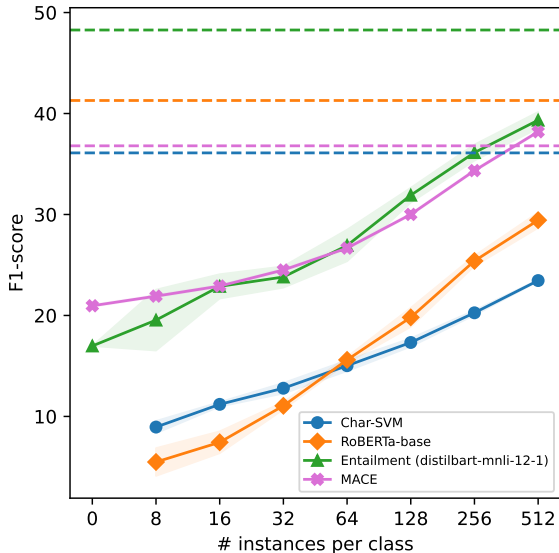


Figure 2: Supervised training with different amounts of labelled data. The results are averaged over three runs. The filled area around lines represents the standard deviation among runs. Horizontal dashed lines: upper-bound computed using all the available training data.

set is available, simply discarding the predictions from weak, unfit models and selecting the best ones available, is crucial for deploying zero-shot models in a production environment. Our results show that for emotion classification, the model-based aggregation can not only automatically select the best available model, but also improve its performance.

When few annotated instances are available, our results show that entailment models perform notably better than supervised models: Figure 2 shows that a finetuned entailment model outperforms by a large margin not only a linear baseline model using shallow features but also a strong neural LM-based model. The RoBERTa-base model is outperformed by the entailment model by a large margin (+6.98% f1-score) when trained on the full dataset. This highlights a key advantage of few-shot learning for under-resourced scenarios.

Given the diverse nature of the data that compose the Unify Emotion dataset, we evaluate four different models on the individual datasets contained in Unify Emotion: Table 4 highlights the results. As shown already in Bostan and Klinger (2018), some datasets are easier to model than others. CrowdFlower and DailyDialog are relatively noisy datasets and

	RoBERTa-base	ZS	FS	All
CrowdFlower	17.04	13.16	13.87	19.85
DailyDialog	18.90	9.57	13.80	34.52
Electoral-Tweets	28.04	13.88	16.10	30.19
EmoInt	34.42	15.21	15.41	41.06
Emotion-Stimulus	69.43	28.78	32.24	79.69
Grounded-Emotions	11.97	1.93	2.52	13.36
ISEAR	33.61	27.98	25.96	46.26
Tales-Emotion	24.76	15.26	17.97	31.01
TEC	25.29	15.48	14.41	28.49

Table 4: Overview of model performance (macro-averaged f1-score) across datasets. RoBERTa-base is a fully supervised baseline. ZS: true zero-shot. FS: DistilBart finetuned using 8 instances. All: DistilBart finetuned using all the available data.

all the models struggle on them. Datasets containing noisy text written in non-canonical language (e.g., Electoral-Tweets, Grounded-Emotions), seem to challenge the zero-shot models more than corpora like ISEAR and Tales-Emotion which contains more standard text. ISEAR’s annotation format is particularly close to the pseudo-sentences that we used (i.e., “This person feels [...]”), which can explain the high performance achieved by the zero-shot model.

## 7 Conclusions

In this work we presented an emotion classification model that does not require large annotated data to be competitive on the Unify Emotion dataset. We experimented with pretrained language models in both the zero-shot and few-shot settings. We aggregated the predictions of these models using MACE, a Bayesian method developed for modelling noisy, crowdsourced annotations. Experimental results showed that the resulting system performs better than the strongest individual zero-shot model. When evaluated on a diverse dataset, our zero- and few-shot models behave in a comparable way to fully-supervised models, without requiring the same amount of annotated data. Noisy text seems to challenge the NLI models trained on canonical text, while zero-shot models perform well when the annotation scheme matches the pseudo-sentences used for building the synthetic data: this suggests that different domains might need different templates that take into account elements like vocabulary or stylistic variation. Finally, we showed that when the MACE and the few-shot systems are trained

on few labelled data, they outperform fully-supervised models.

In future works we will further explore how to apply zero and few-shot learning for text classification tasks, and how to better aggregate the outputs of different models in a unsupervised manner.

## References

- 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*, pages 579–586. Association for Computational Linguistics.
- Cecilia Ovesdotter Alm and Richard Sproat. 2005. Perceptions of emotions in expressive storytelling. In *Ninth European Conference on Speech Communication and Technology*.
- Ebba Cecilia Ovesdotter Alm. 2008. *Affect in\* text and speech*. University of Illinois at Urbana-Champaign.
- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. [Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification](#). In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.
- Victoria Bobicev and Marina Sokolova. 2017. [Inter-annotator agreement in sentiment analysis: Machine learning perspective](#). In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102, Varna, Bulgaria. INCOMA Ltd.
- Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. [GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.
- Laura-Ana-Maria Bostan and Roman Klinger. 2018. [An analysis of annotated corpora for emotion classification in text](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. [SemEval-2019 task 3: EmoContext contextual emotion detection in text](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [Xnli: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Crowdfunder. [The emotion in text](#), published by crowdfunder.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. [The pascal recognising textual entailment challenge](#). In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment, MLCW’05*, page 177–190, Berlin, Heidelberg. Springer-Verlag.
- 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.
- Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. [emoji2vec: Learning emoji representations from their description](#). In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54, Austin, TX, USA. Association for Computational Linguistics.

- Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. [Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.
- Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI’07*, page 1606–1611, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Association for Computational Linguistics (ACL)*.
- Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.
- A. Go. 2009. [Sentiment classification using distant supervision](#).
- Marti A. Hearst. 1992. [Automatic acquisition of hyponyms from large text corpora](#). In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.
- Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. [Learning whom to trust with MACE](#). In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.
- Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. 2021. [Universal joy a data set and results for classifying emotions across languages](#). In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- 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 Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#).
- V. Liu, C. Banea, and R. Mihalcea. 2007. Grounded emotions. In *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, Texas.
- 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*.
- Saif Mohammad. 2012. [#emotional tweets](#). In *\*SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.
- Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. [SemEval-2018 task 1: Affect in tweets](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.
- Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.
- Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural



- language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Nicole Novielli, Fabio Calefato, and Filippo Lanubile. 2018. [A gold standard for emotion annotation in stack overflow](#). In *Proceedings of the 15th International Conference on Mining Software Repositories, MSR '18*, page 14–17, New York, NY, USA. Association for Computing Machinery.
- Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. [Comparing Bayesian models of annotation](#). *Transactions of the Association for Computational Linguistics*, 6:571–585.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Nina Poerner and Hinrich Schütze. 2019. [Multi-view domain adapted sentence embeddings for low-resource unsupervised duplicate question detection](#). 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 1630–1641, Hong Kong, China. Association for Computational Linguistics.
- Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. [Train once, test anywhere: Zero-shot learning for text classification](#).
- Francisco Rangel and Paolo Rosso. 2016. [On the impact of emotions on author profiling](#). *Inf. Process. Manage.*, 52(1):73–92.
- Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. [Data programming: Creating large training sets, quickly](#). In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Klaus R. Scherer and Harald G. Wallbott. 2017. [International survey on emotion antecedents and reactions \(isear\)\(1990\)](#).
- Timo Schick and Hinrich Schütze. 2021a. [Exploiting cloze-questions for few-shot text classification and natural language inference](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.
- Timo Schick and Hinrich Schütze. 2021b. [It’s not just size that matters: Small language models are also few-shot learners](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.
- Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Ng. 2008. [Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.
- 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, Prague, Czech Republic. Association for Computational Linguistics.
- W. L. Taylor. 1953. “cloze procedure”: A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, 30:415 – 433.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). 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 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021. [Emotion ratings: How intensity, annotation confidence and agreements are entangled](#). In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49, Online. Association for Computational Linguistics.
- Ana Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. [Understanding patterns of anorexia manifestations in social media data with deep learning](#). In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 224–236, Online. Association for Computational Linguistics.
- Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. [Entailment as few-shot learner](#).
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). 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 1112–1122. Association for Computational Linguistics.

randomly sampled training sets from the whole Unified dataset.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. [Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

## A Appendix

### A.1 Hyperparameters

Training the Entailment model in a few-shot scenario was carried out using a single NVIDIA GeForce GTX 1080 Ti, which allows to allocate up to 11GB of RAM. This has constraint some of the hyperparameters we have chosen to try.

Aside from that, we have also followed the recommendations shared by [Wang et al. \(2021\)](#) because our experiments with few-shot learning were very similar to theirs.

**Batch size and Maximum Length** A batch size of 8 samples was used. The maximum length was adapted to the maximum number of tokens seen when encoding the whole Unified dataset with the corresponding tokenizer of the chosen model (distilbart-mnli-12-1). When computing this, it was also taken into account the fact that, for the entailment approach, the label description or hypothesis is encoded as an additional input to the model. The final selected value was 286 tokens.

**Learning Rate** Typical learning rate values recommended for fine tuning an Adam optimizer are: 5e-5, 3e-5, 2e-5 ([Devlin et al., 2019](#)). Following the implementation of [Wang et al. \(2021\)](#), we used a constant and smaller value of 1e-5.

**Epochs** As a practical consideration we decided to train just 1 epoch because we observed that training on more steps reduced the overall performance.

**Number of trials** In order to avoid instability among reported results, mainly caused by the small number of samples used in few-shot experiments ([Wang et al., 2021](#); [Gao et al., 2021](#)), the metrics measuring the performance of the model are averaged among 3 different runs that are trained over its corresponding