Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora

Flor Miriam Plaza-del-Arco[1,2], **María-Teresa Martín-Valdivia**[1], **Roman Klinger**[2]
[1]SINAI, Computer Science Department, CEATIC, Universidad de Jaén, Spain
[2]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
{fmplaza, maite}@ujaen.es, klinger@ims.uni-stuttgart.de

Abstract

Within textual emotion classification, the set of relevant labels depends on the domain and application scenario and might not be known at the time of model development. This conflicts with the classical paradigm of supervised learning in which the labels need to be predefined. A solution to obtain a model with a flexible set of labels is to use the paradigm of zero-shot learning as a natural language inference task, which in addition adds the advantage of not needing any labeled training data. This raises the question how to prompt a natural language inference model for zero-shot learning emotion classification. Options for prompt formulations include the emotion name *anger* alone or the statement "This text expresses *anger*". With this paper, we analyze how sensitive a natural language inference-based zero-shot-learning classifier is to such changes to the prompt under consideration of the corpus: How carefully does the prompt need to be selected? We perform experiments on an established set of emotion datasets presenting different language registers according to different sources (tweets, events, blogs) with three natural language inference models and show that indeed the choice of a particular prompt formulation needs to fit to the corpus. We show that this challenge can be tackled with combinations of multiple prompts. Such ensemble is more robust across corpora than individual prompts and shows nearly the same performance as the individual best prompt for a particular corpus.

1 Introduction

To enable communication about emotions, there exists a set of various emotion names, for instance those labeled as *basic emotions*, by Ekman (1992) or Plutchik (2001) (*anger*, *fear*, *joy*, *sadness*, *disgust*, *surprise*, *trust*, *anticipation*). While such psychological models influence natural language processing and emotion categorization approaches, the choice of emotion concepts is context-dependent. For instance, Scherer and Wallbott (1997) and Troiano et al. (2019) opted to use *guilt* and *shame* as self-directed emotions in addition to Ekman's basic emotions, to analyze self-reports of events. For the context of the perception of art it is more appropriate to consider *aesthetic emotions* (Menninghaus et al., 2019; Haider et al., 2020), like *beauty*, *sublime*, *inspiration*, *nostalgia*, and *melancholia*.

This leads to a potential gap between concepts in emotion-related training data and the application domain, purely because the label set is not compatible. One solution is to resort to so-called dimensional models, in which emotion names are located in vector spaces of affect (valence, arousal, Preoţiuc-Pietro et al., 2016; Buechel and Hahn, 2017) or cognitive appraisal (e.g., regarding *responsibility*, *certainty*, *pleasantness*, *control*, *attention* with respect to a stimulus event, Hofmann et al., 2020; Troiano et al., 2023). In these vector spaces, classes can be assigned to predicted points with a nearest-neighbor approach, even if these classes have not been seen during training. This approach, however, has the disadvantage of the so-called hubness problem (Lazaridou et al., 2015), namely that the distance between predictions and concepts that have been seen during training tends to be smaller than to novel concepts. We acknowledge ongoing research to tackle this problem (Park et al., 2021;

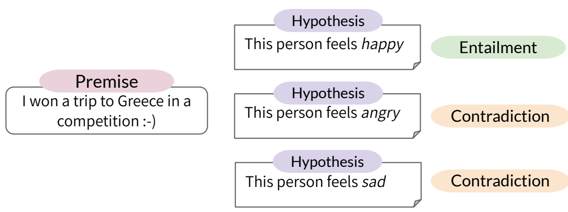

Figure 1: An example of the application of NLI to ZSL emotion classification. Given the premise "I won a trip to Greece in a competition", three hypotheses represent the emotions (*joy*, *anger*, *sadness*). The representation of *joy* is entailed and therefore predicted.

Buechel et al., 2021).

We take a different, more direct, route to obtaining classifiers for discrete emotion categories which are not known at system development time, namely zero-shot learning (ZSL). One instantiation of such ZSL systems is via natural language inference models (ZSL-NLI), in which the inference task needs to perform reasoning (Yin et al., 2019). Consequently, the idea of implementing ZSL-NLI models is not by exemplification and optimizing a classifier, but developing appropriate natural language class name representations which we refer to as *prompts*. We see an example for the application of an NLI model to ZSL emotion classification in Figure 1 – the NLI model needs to decide if the hypothesis (a prompt which represents the class label) entails the premise (which corresponds to the instance to be classified). This paradigm raises the question (which we answer in this paper) of how to formulate the emotion prompt and how much the design choice of the prompt needs to fit the dataset.

Manually developing intuitive templates based on human data introspection may be the most natural method to produce prompts. In this paper, we provide manually created templates to probe emotion classification in an NLI-ZSL setup and we analyze whether prompts are language-register dependent according to various corpora (tweets, event descriptions, blog posts). To accomplish this aim, we perform experiments on an established set of emotion datasets with three NLI models and we show that (1) prompts are indeed corpus-specific and that the differences follow the same pattern across different pretrained NLI models, (2) that an ensemble of multiple prompts behaves more robustly across corpora, and (3) the representation of the emotion concept as part of the textual prompt is an important element, benefiting from representations with synonyms and related concepts, instead of just the emotion name. Our code is publicly available at https://github.com/fmplaza/zsl_nli_emotion_prompts.

2 Related Work

2.1 Emotion Classification

Emotion analysis has become a major area of research in NLP which comprises a variety of tasks, including emotion stimulus or cause detection (Li et al., 2021; Doan Dang et al., 2021) and emotion intensity prediction (Mohammad and Bravo-Marquez, 2017; Köper et al., 2017). The task of emotion classification received most attention in recent years (Bostan and Klinger, 2018; Mohammad et al., 2018a; Plaza-del-Arco et al., 2020, i.a.).

Emotion classification aims at mapping textual units to an emotion category. The categories often rely on psychological theories such as those proposed by Ekman (1992) (*anger*, *fear*, *sadness*, *joy*, *disgust*, *surprise*), or the dimensional model of Plutchik (2001) (adding *trust* and *anticipation*). However, neither are all these basic emotions relevant in all domains, nor are they sufficient. For instance, in the education field, D'mello and Graesser (2007) found *boredom*, *confusion*, *flow*, *frustration*, and *delight* to be more relevant than *fear* or *disgust*. Sreeja and Mahalaksmi (2015) reveal that emotions such as *love*, *hate*, and *courage* are necessary to model the emotional perception of poetry. Bostan et al. (2020) identify *annoyance*, *guilt*, *pessimism*, or *optimism* to be important to analyze news headlines.

A strategy to avoid specification of discrete categories is the use of dimensional spaces that consider valence, arousal, and dominance (VAD, Russell and Mehrabian, 1977). Smith and Ellsworth (1985) claim that this model does not represent important difference between emotions and propose an alternative dimensional model based on cognitive appraisal, which has recently been used for text analysis (Hofmann et al., 2020; Stranisci et al., 2022; Troiano et al., 2023). Independent of the classification or regression approach, nearly all recently proposed systems rely on transfer learning from general language representations. We refer the reader to recent shared task surveys for a more comprehensive overview (Mohammad et al., 2018b; Tafreshi et al., 2021; Plaza-del Arco et al., 2021).

2.2 Zero-shot Learning

Zero-shot learning (ZSL) aims at performing predictions without having seen labeled training examples specific for the concrete task. Zero-shot methods typically work by associating seen and unseen classes using auxiliary information, which encodes observable distinguishing properties of instances (Xian et al., 2019). In NLP, the term is used predominantly either to refer to cross-lingual transfer to languages that have not been seen at training time (change of the language), or to predict classes that have not been seen (change of the labels, Wang et al., 2019). Our work falls in the second category.

Various approaches exist to perform zero-shot text classification. One approach represents labels in an embedding space (Socher et al., 2013; Sappadla et al., 2016; Rios and Kavuluru, 2018, i.a.). A model is trained to predict the respective embedding vectors for categorical labels. At test time, embeddings of novel labels need to be known and will be assigned if the distance between the predicted embedding and the label embedding is small. This method suffers from the hubness problem, that is, when the semantic label embeddings are close to each other, the projection of labels to the semantic space forms hubs (Radovanovic; et al., 2010).

Another approach is to use transformer language models to classify if a label embedding is compatible with an instance embedding (Brown et al., 2020). To this end, no labeled examples are provided at training phase but an instruction in natural language is given to the model to interpret the label class (the *prompt*). An instance of this approach is *Task-Aware Representations* (TARS, Halder et al., 2020) who separate the instance text and the class label by the special separator token `[SEP]` in BERT (Devlin et al., 2019).

An alternative is to treat ZSL as textual entailment. Following this approach, Yin et al. (2019) propose a sequence-pair classifier that takes two sentences as input (a premise and a hypothesis) and decides whether they entail or contradict each other. They study various formulations of the labels as hypotheses and evaluate the method in various NLP tasks including topic detection, situation detection, and emotion classification. In their evaluation, emotion classification turns out to be most challenging. Another study that conducted prompt engineering in NLI models proposes probabilistic ZSL ensembles for emotion classification (Basile et al., 2021). The authors experiment with the same prompts as Yin et al. (2019) and aggregate the predictions of multiple NLI models using Multi-Annotator Competence Estimation (MACE), a method developed for modelling crowdsourced annotations.

Our work on ZSL for emotion classification differs from previous approaches as follows. We analyze whether prompts are corpus-specific and propose an ensemble of multiple prompts to achieve a classifier which is more robust across corpora (in contrast to an ensemble of multiple NLI models in the work by Basile et al. (2021)). Further, we analyze if the introduction of more knowledge about the emotion in the prompt through emotion synonyms and related concepts helps its interpretation in the NLI models.

3 Methods

In this section, we explain how we apply NLI for ZSL emotion classification and propose a collection of prompts to contextualize and represent the emotion concept in different corpora. In addition, we propose a prompt ensemble which is more robust across corpora.

3.1 Natural Language Inference for Zero-shot Emotion Classification

The NLI task is commonly defined as a sentence-pair classification in which two sentences are given: a *premise* s_1 and a *hypothesis* s_2. The task is to learn a function $f_{\mathrm{NLI}}(s_1, s_2) \rightarrow \{E, C, N\}$, in which E expresses the entailment of s_1 and s_2, C denotes a contradiction and N is a neutral output.

We treat ZSL emotion classification as a textual entailment problem, but represent each label under consideration with multiple prompts, in contrast to Yin et al. (2019). Given a sentence to be classified x (*premise*) and an emotion e, we have a function $g(e)$ that generates a set of prompts (*hypothesis*) out of the class $e \in E$ (with E being the set of emotions under consideration). Under the assumption of an NLI model m, which calculates the entailment probability $p_m(\gamma, x)$ for some emotion representation $\gamma \in g(e)$, we assign the average entailment probability across all emotion representations as

$$\bar{p}^g_m(e, x) = \frac{1}{|g(e)|} \sum_{\gamma \in g(e)} p_m(\gamma, x)$$

for a particular prompt generation method g. The classification decision

$$\hat{e}^g_x = \arg\max_{e \in E} \bar{p}^g_m(e, x)$$

returns the emotion corresponding to the maximum entailment probability.

3.2 Emotion Prompts

In the context of emotion analysis, two important questions arise when formulating a prompt: (*i*) How to contextualize the emotion name, and (*ii*) How to represent the emotion concept?

ID	Prompt	Example
Emo-Name	*emotion name*	*joy*
Expr-Emo	This text expresses *emotion name*	This text expresses *joy*
Feels-Emo	This person feels *emotion name*	This person feels *joyful*
WN-Def	This person expresses *WordNet def.*	This person expresses *a feeling of great pleasure and happiness*
Emo-S	*emotion synonym*	*happy*
Expr-S	This text expresses e*motion syn.*	This text expresses *happiness*
Feels-S	This person feels *emotion syn.*	This person feels *happy*
EmoLex	*emotion lexicon*	*party*

Table 1: Emotion prompts. Words in *italics* represent placeholders for the emotion concept representation.

3.2.1 Prompt Generation

We generate a set of prompts with the function $g(e) = c + r(e)$, in which c represents what we call the *context* and $r(e)$ represents a set of emotion representations.[1] As c, we use either an empty string ϵ, the text "*This text expresses*", "*This person feels*", or "*This person expresses*", motivated by our choice of the language register presented in the datasets used in our experiments (see § 4).

3.2.2 Prompts for Zero-Shot Emotion Classification

Each prompt in this paper consists of context and the emotion representation. There are three prompts which have in common the emotion name representation, namely Emo-Name, Expr-Emo, and Feels-Emo. Variations of these prompts are Emo-S, Expr-S, and Feels-S, where the emotion name representation is replaced by multiple emotion synonyms and EmoLex where the emotion name is replaced by entries from an emotion word lexicon. In detail, we use the following prompts (Table 1 shows examples):

Emo-Name. $c = \epsilon$ and $r(e) = \{e\}$.

Expr-Emo. c = "This text expresses", $r(e) = \{e\}$.

Feels-Emo. c = "This person feels", $r(e) = \{e\}$.

WN-Def. c = "This person expresses" and $r(e) = \{\text{WN-Def}(e)\}$, where WN-Def$(e)$ is the WordNet definition for e (Miller, 1995).

Emo-S. We aim to see whether incorporating additional information using a set of abstract emotion-related names leads to a better model. Hence, we set $r(e)$ to return a set of emotion synonyms for e. Table 3 shows the emotion synonyms considered for each emotion.[2]

Expr-S. We set $r(e)$ to be the same as in Emo-S, but additionally set c = "This text expresses". Therefore, $g(e)$ returns all combinations of this string with each synonym.

Feels-S. This prompt is the same as Expr-S with the difference that we set c = "This person feels".

EmoLex. This prompt is different from the previous ones, which consisted of small sets of context/emotion representation combinations. Here, $c = \epsilon$, but for the emotion representation we use a large popular lexicon, namely Emolex (Mohammad and Turney, 2013) to assign all entries associated with e in this lexicon. This generates prompts which contain abstract emotion synonyms as well as concrete objects (like *gift* for *joy*).

3.3 Ensemble of prompts

In practical applications, the choice of a particular prompt could not be performed manually by some user. Under the assumption that the choice of prompts is indeed corpus-specific, we combine multiple prompt sets in an ensemble.

The ensemble model takes as input a text x and a set of prompt-generating models G with $\bar{p}^g_M(e,x)$ ($g \in G$). The ensemble decision is then

$$\hat{e}(x,m) = \arg\max_{e\in E} \frac{1}{|G|} \sum_{g\in G} \bar{p}^g_m(e,x)\,.$$

4 Experiments

We aim at answering the following research questions: **(RQ1)** Do NLI models behave the same across prompts? **(RQ2)** Should we use synonyms for the emotion representation? **(RQ3)** Is an ensemble of multiple prompts more robust across corpora? **(RQ4)** Are synonyms sufficient? Would it be even more useful to use more diverse representations of emotions?

[1] In principle, c could also be a set. $g(e)$ would then need to use a cross-product instead of the element-wise concatenation $+$, which we use in our experiments.

[2] Each synonym is grammatically adapted for the context of the prompts Expr-S and Feels-S.

Dataset	Labels	Size	Source	Avail.
TEC	Ekman	21,051	tweets	D-RO
BLOGS	Ekman + *no emotion*	5,205	blogs	R
ISEAR	Ekman − *Su* + *G* + *Sh*	15,302	events	GPLv3

Table 2: Datasets used in our experiments (Su: surprise, G: guilt, Sh: shame) [D-RO] available to download, research only, [R] Available upon request, [GPLv3] GNU Public License version 3.

4.1 Experimental Setting

4.1.1 Datasets

We compare our methods on three English corpora, to gain an understanding of the role of the respective corpus. TEC (Mohammad, 2012) contains 21,051 tweets weakly labeled according to hashtags corresponding to the six Ekman emotions (Ekman, 1992): *#anger*, *#disgust*, *#fear*, *#happy*, *#sadness*, and *#surprise*. ISEAR (Scherer and Wallbott, 1997) includes 7,665 English self-reports of events that triggered one of the emotions (*joy*, *fear*, *anger*, *sadness*, *disgust*, *shame*, and *guilt*). BLOGS (Aman and Szpakowicz, 2007) consists of 5,205 sentences from 173 blogs compiled from the Web using a list of emotion-related seed words. It is human-annotated according to Ekman's set of basic emotions and an additional *no emotion* category. TEC and ISEAR are publicly available for research purposes and BLOGS is available upon request. All datasets are anonymized by the authors.

These corpora differ in various parameters (see Table 2): the annotation scheme (variations of Ekman's model), the corpus source (tweets, events, blogs), the annotation procedure (hashtag, crowdsourcing, self-reporting), and the size. Note that the annotation procedure that the ZSL method needs to reconstruct varies in complexity.

4.1.2 NLI Models and Baseline

We compare our ZSL models with an empirical upper bound, namely a RoBERTa model fine-tuned with supervised training (Liu et al., 2020) on each emotion dataset described in § 3.2.2. We fine-tune RoBERTa for three epochs, the batch size is set to 32 and the learning rate to $2 \cdot 10^{-5}$. No hyperparameter search has been applied. We perform 10-fold cross-validation and report the results on the whole data set (as we do with the NLI models).

For our ZSL experiments, we explore three state-of-the-art pretrained NLI models publicly available within the Hugging Face Transformers Python library (Wolf et al., 2020), and fine-tuned on the MultiNLI dataset (Williams et al., 2018). Concretely, we choose RoBERTa, BART and DeBERTa as they cover different architectures and represent competitive approaches across a set of NLP tasks.

Emotion	**Emo-S**
anger	anger, annoyance, rage, outrage, fury, irritation
fear	fear, horror, anxiety, terror, dread, scare
joy	joy, achievement, pleasure, awesome, happy, blessed
sadness	sadness, unhappy, grief, sorrow, loneliness, depression
disgust	disgust, loathing, bitter, ugly, repugnance, revulsion
surprise	surprise, astonishment, amazement, impression, perplexity, shock
guilt	guilt, culpability, blameworthy, responsibility, misconduct, regret
shame	shame, humiliate, embarrassment, disgrace, dishonor, discredit

Table 3: Emotion synonyms per emotion category considered in Emo-S prompt (details in the Appendix).

RoBERTa. The Robustly Optimized BERT Pretraining Approach (Liu et al., 2020) is a modified version of BERT which includes some changes such as the removal of the next-sentence prediction task, the replacement of the WordPiece tokenization with a variation of the byte-pair encoding, and the replacement of the static masking (the same input masks are fed to the model on each epoch) with dynamic masking (the masking is generated every time the sequence is fed to the model). For the NLI task, we use the *roberta-large-mnli* model from Hugging Face which contains over 355M of parameters.

BART. The Bidirectional and Auto-Regressive Transformer (Lewis et al., 2020) is a model that combines the bidirectional encoder with an autoregressive decoder into one sequence-to-sequence model. We use the *facebook/bart-large-mnli* model from Hugging Face with over 407M parameters.

DeBERTa. The Decoding-enhanced BERT with Disentangled Attention model (He et al., 2021) improves BERT and RoBERTa using two techniques, namely disentangled attention and an enhanced mask decoder. We use *microsoft/deberta-xlarge-mnli* from Hugging Face, which contains over 750M of parameters.

All experiments are performed on a node equipped

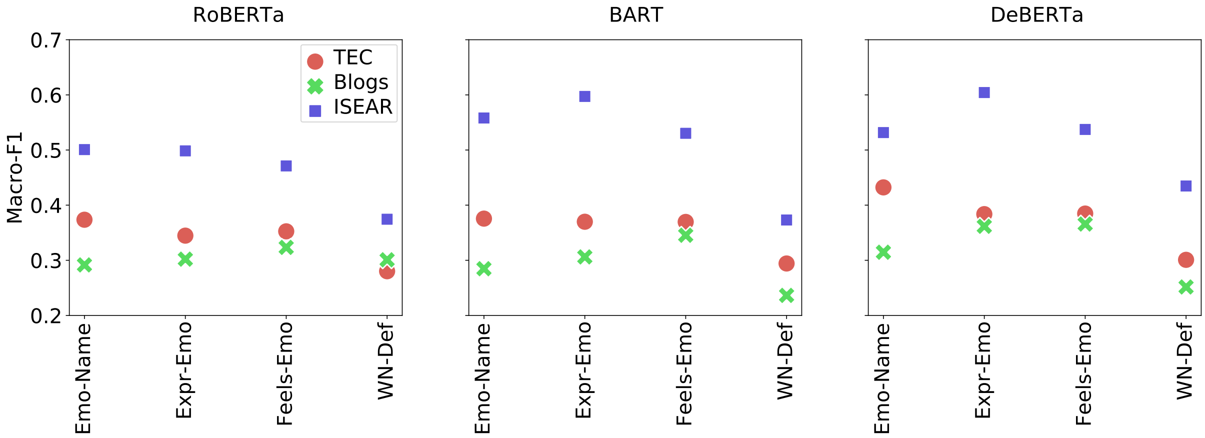

Figure 2: Results of Experiment 1. Comparison of prompts across NLI models and emotion datasets.

with two Intel Xeon Silver 4208 CPU at 2.10GHz, 192GB RAM, as main processors, and six GPUs NVIDIA GeForce RTX 2080Ti (with 11GB each).

4.2 Results

In order to answer the research questions formulated in this study, we conduct different ZSL-NLI emotion classification experiments.

4.2.1 Experiment 1: Are NLI models behaving the same across prompts?

With the first experiment, we aim at observing if different NLI models behave robustly across emotion datasets and prompts. We use each model described in § 4.1.2 with each emotion representation that is not a set of multiple prompts, but only consists of a single prompt, namely Emo-Name, Expr-Emo, Feels-Emo and WN-Def. We evaluate each model using all datasets (§ 4.1.1).

Figure 2 (and Table 6 in the Appendix) show the results. Each plot shows the performance of one NLI model on the three emotion datasets using the four prompts. We see that the performances follow the same patterns across NLI models and emotion datasets. Emo-Name is the best performing prompt for TEC, Expr-Emo for ISEAR and Feels-Emo for BLOGS. The lowest performance is achieved with WN-Def. The most successful NLI model across the prompts is DeBERTa followed by BART and RoBERTa.

Therefore, NLI models do behave robustly across prompts. Particularly low performance can be observed with WN-Def. This finding is in line with previous research (Yin et al., 2019): These definitions may be suboptimal choices, for instance, *sadness* is represented via "This person expresses emotions experienced when not in a state of well-being". This is ambiguous since not being in a state of well-being may also be associated with other negative emotions such as *anger* or *fear*. Interestingly, the best-performing emotion representation on TEC is Emo-Name, which resembles the annotation procedure of just using an emotion-related hashtag for labeling. Similarly, Expr-Emo shows the best performance for the self-reports of ISEAR ("This text expresses") and Feels-Emo on BLOGS ("This person feels"). These subtle differences in the prompt formulations indicate that there are particular factors in the dataset that influence the interpretation of the prompt, for instance, the annotation procedure, the data selection or the language register employed in the corpus, and therefore, they affect the interpretation of the emotion by the NLI-ZSL classifier.

4.2.2 Experiment 2: Should we use synonyms for emotion representation?

In this experiment, we aim at observing whether the incorporation of synonyms in the prompt helps the emotion interpretation. Instead of considering only the emotion name, we use six close emotion synonyms (see Emo-S, Expr-S, Feels-S in Table 7 in the Appendix).[3] This leads to six prompts for each emotion. For simplicity, we now only consider DeBERTa, which showed best performances in the previous experiment.

Figure 3 (and Table 6 in the Appendix) shows the results of each context with just the emotion

[3] We assume that larger numbers might show better performance in general, but this set of six synonyms focuses on close, unambiguous synonyms which undoubtedly represent the emotion in most contexts. We evaluate the impact of larger sets with the EmoLex approach.

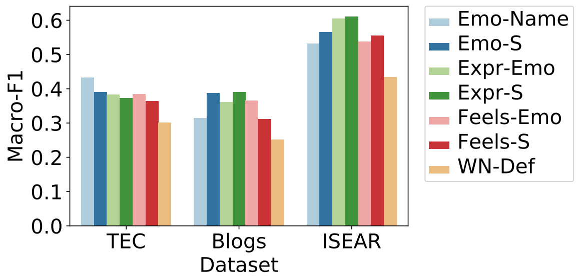

Figure 3: Results of Experiment 2. Comparison of prompts including synonym emotion representations across three emotion datasets (TEC, BLOGS and ISEAR) using the DeBERTa model.

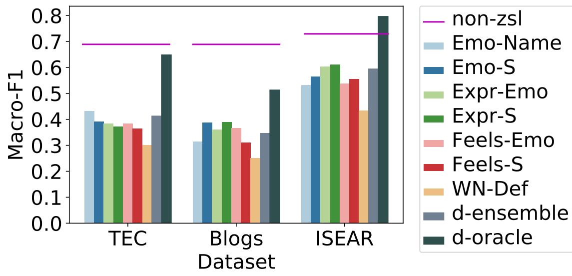

Figure 4: Results of Experiment 3. Comparison of the prompt individual models and the proposed ensemble models along with the non-zsl experiments.

name and with the synonyms in comparison. In general, synonym use leads to an improvement, with some notable exceptions. For TEC, the single use of the emotion (Emo-Name) works better than using synonyms (Emo-S). This might stem from a similarity of the prompt with the annotation procedure, in which single hashtags were used for labeling. Another exception is Feels-Emo/Feels-S in BLOGS. Therefore, to answer RQ2 we conclude that both context and emotion concept representation are corpus-dependent and in some cases synonyms support the emotion classification.

4.2.3 Experiment 3: Is an ensemble of multiple prompts more robust across corpora?

The previous experiments demonstrate the challenge of engineering an emotion prompt that fits different corpora which stem from various sources. To cope with this challenge, we analyze if the combination of sets of prompt-generation methods in an ensemble improves the generalizability. We use the ensemble method described in § 3.3 that combines the predictions given by the set of model prompts described in § 3.2.2 with the DeBERTa model (d-ensemble). In addition to this realistic ensemble model, we want to understand which performance could be achieved with an ideal (oracle) ensemble (which we refer to as d-oracle), which always picks the correct decision by an ensemble component, if one is available. This serves as an upper bound and analyzes the complementarity of the individual models.

Figure 4 shows the performance for the individual models discussed before, which participate in both the realistic and the oracle ensemble (individual results in Table 6 in the Appendix, ensemble results also in Table 5). In addition, we see both ensemble methods and (as a horizontal line) the supervised learning upper bound. We observe that the realistic ensemble (d-ensemble), which is based on averaging the individual probabilistic outputs of the individual models, shows a performance nearly en par with the individual best model: For TEC, we have an F_1 =.41 in comparison to the individual best F_1 =.43, for BLOGS, we have F_1 =.35 in comparison to F_1 =.39, and for ISEAR, we achieve F_1 =.59 in comparison to F_1 =.61 – but without the necessity to pick the prompt-generating approach beforehand or on some hold-out data.

We further see that the oracle ensemble performs better than all other models – this shows the variance between the models and suggests a reason for their corpus-dependency, but also shows the potential for other ensemble models. This oracle also approaches (or is even slightly higher than) the supervised upper-bound. All of our current (non-oracle) ZSL learning methods clearly underperform supervised learning, but to various degrees. The oracle performance suggests that sets of prompts, combined with a good ensembling method, might exist that outperform supervised learning in emotion classification.

We conclude that an ensemble model is indeed more robust across emotion datasets with different language registers and prompts, with a performance nearly en par with the best corpus-specific prompt. This raises the question what differences and commonalities instances have in which models perform the same or differently. To this end, we show examples in Table 4, in which *all* individual models did output the correct label. As we can see, these instances contain explicit words related to the emotion conveyed. For instance, "lost" for *sadness*, "love" for *joy*, "angry" for *anger*, "nervous" for *fear*, "ashamed" for *shame*, and "felt bad" for *guilt*. Therefore, prompt-NLI models succeed

Emotion	Text
anger	The sports fishermen who catch gulls instead of fish with their hooks. It is often a mistake but it makes me angry. (ISEAR)
disgust	my sister got this purse, It smell like straight up KITTY LITTER. (TEC)
fear	Oh well its nothing too too bad but its making me nervous. (BLOGS)
guilt	While at primary school, I did not let a friend ring a bell although he would have liked to do it. Afterwards I felt bad. (ISEAR)
joy	When I get a hug from someone I love. (ISEAR)
sadness	When I lost the person who meant the most to me. (ISEAR)
surprise	Snow in October! (BLOGS)
shame	We got into a fight with some chaps in front of our family house. The value of the property destroyed was approximately 15 000 FIM. I felt ashamed when my parents came to know about this. (ISEAR)

Table 4: Instances where all the prompt models agree with the emotion prediction.

in interpreting emotions that are clearly expressed in the text, but vary performance-wise when the emotion is implicitly communicated.

4.2.4 Experiment 4: Are synonyms sufficient? Would it be even more useful to use more representations of emotions?

In Experiment 2 we found that the use of synonyms is beneficial in some cases (ISEAR and BLOGS). This raises the question if more terms that represent the emotion would lead to an even better performance. We evaluate this setup with the EmoLex model introduced above, in which each emotion concept is represented with a set of prompts, where each prompt is a concept from an emotion lexicon. Notably, in this prompt-generating methods, emotions are not only represented by abstract emotion names or synonyms, but in addition with (sometimes concrete) concepts, like "gift" or "tears".

Table 5 shows the performance of the DeBERTa model using the Emolex concepts (d-emolex), next to the ensemble results. The additional concepts which cover a wide range of topics associated with the respective emotions particularly help in the BLOGS corpus, which is the one resource that has been manually annotated in a traditional manner. This manual annotation process might include complex inference by the annotators to infer an emotion category, instead of only using single words

	TEC			BLOGS			ISEAR		
Model	P	R	F_1	P	R	F_1	P	R	F_1
d-ensemble	.42	.44	.41	.40	.65	.35	.67	.62	.59
d-oracle	.63	.69	.65	.51	.80	.51	.82	.80	.80
d-emolex	.37	.36	.33	.52	.48	.48	.47	.42	.40
non-zsl	.69	.69	.69	.72	.71	.69	.73	.73	.73

Table 5: Results of Experiments 3 and 4. We report macro-average precision (P), macro-average recall (R), and macro-average F_1 (F_1) for each model. d-emolex: DeBERTa using EmoLex prompt, d-ensemble: ensemble model of prompts using DeBERTa, d-oracle: oracle ensemble model using DeBERTa), non-zsl: Supervised RoBERTa model fine-tune on the three emotion datasets.

to trigger an event description (ISEAR) or using words as hashtags (TEC). Lexicons can therefore aid in the injection of background knowledge in the prompt. However, this comes at the cost of considerably longer runtimes, because the NLI models is queried for every entry in the lexicon.

5 Conclusion and Future Work

We presented an analysis of various prompts for NLI-based ZSL emotion classification. The prompts that we chose were motivated by the various particularities of the corpora: single emotions for TEC (tweets), "The person feels/The text expresses" for BLOGS (blogs), and ISEAR (events). In addition, we represented the emotions with emotion names, synonyms, definitions, or with the help of lexicons. Our experiments across these data sets showed that, to obtain a superior performance, the prompt needs to fit well to the corpus – we did not find one single prompt that works well across different corpora. To avoid the requirement for manually selecting a prompt, we therefore devised an ensemble model that combines multiple sets of prompts. This model is more robust and is nearly on par with the best individual prompt. In addition, we found that representing the emotion concept more diversely with synonyms or lexicons is beneficial, but again corpus-specific.

Our work raises a set of future research questions. We have seen that the oracle ensemble showed a good performance, illustrating that the various prompts provide complementary information. This motivates future research regarding other combination schemes, including learning a combination based on end-to-end fine-tuned NLI models.

We have further seen that including more concepts with the help of a dictionary helps in one corpus, but not across corpora; however, synonyms constantly help. This raises the question about the right trade-off between many, but potentially inappropriate, noisy concepts and hand-selected, high-quality concepts. A desideratum is an automatic subselection procedure, which removes concepts that might decrease performance and only keeps concepts that are "compatible" to the current language register and annotation method. Ideally, this procedure would not make use of annotated data, because that would limit the advantages of ZSL.

The main limitation of our current work is that we manually designed the prompts under consideration, based on the corpora we used for evaluation. This is a bottleneck in model development, which should either be supported by a more guided approach which supports humans in developing prompts, or by an automatic model that is able to automatically generate prompts based on the language register and concept representation in the dataset.

Acknowledgements

We thank Enrica Troiano and Laura Oberländer for discussions on the topic of emotion analysis.

Roman Klinger's work is supported by the German Research Council (DFG, project number KL 2869/1-2). Flor Miriam Plaza-del-Arco and María-Teresa Martín Valdivia have been partially supported by the LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and ERDF A way of making Europe, and a grant from the Ministry of Science, Innovation and Universities of the Spanish Government (FPI-PRE2019-089310).

References

Saima Aman and Stan Szpakowicz. 2007. Identifying Expressions of Emotion in Text. In *Text, Speech and Dialogue, 10th International Conference, TSD 2007, Pilsen, Czech Republic, September 3-7, 2007, Proceedings*, volume 4629 of *Lecture Notes in Computer Science*, pages 196–205. Springer.

Angelo Basile, Guillermo Pérez-Torró, and Marc Franco-Salvador. 2021. "Probabilistic Ensembles of Zero- and Few-Shot Learning Models for Emotion Classification". In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 128–137, Held Online. 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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sven Buechel and Udo Hahn. 2017. "EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis". In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Sven Buechel, Luise Modersohn, and Udo Hahn. 2021. "Towards Label-Agnostic Emotion Embeddings". In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9231–9249, Online and Punta Cana, Dominican Republic. 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.

Sidney D'mello and Arthur Graesser. 2007. Mind and body: Dialogue and posture for affect detection in learning environments. In *Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, page 161–168, NLD. IOS Press.

Bao Minh Doan Dang, Laura Oberländer, and Roman Klinger. 2021. Emotion Stimulus Detection in German News Headlines. In *Proceedings of the 17th*

Conference on Natural Language Processing (KONVENS 2021), Düsseldorf, Germany. German Society for Computational Linguistics & Language Technology.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. "PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry". In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1652–1663, Marseille, France. European Language Resources Association.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. "Task-Aware Representation of Sentences for Generic Text Classification". In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *International Conference on Learning Representations*.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. "Appraisal Theories for Emotion Classification in Text". In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. "IMS at EmoInt-2017: Emotion Intensity Prediction with Affective Norms, Automatically Extended Resources and Deep Learning". In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57, Copenhagen, Denmark. Association for Computational Linguistics.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015. "Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning". In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China. 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.

Xiangju Li, Wei Gao, Shi Feng, Yifei Zhang, and Daling Wang. 2021. "Boundary Detection with BERT for Span-level Emotion Cause Analysis". In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 676–682, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Winfried Menninghaus, Valentin Wagner, Eugen Wassiliwizky, Ines Schindler, Julian Hanich, Thomas Jacobsen, and Stefan Koelsch. 2019. What are aesthetic emotions? *Psychological Review*, 126(2):171–195.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

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.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion Intensities in Tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018a. 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 Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018b. 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 and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. 2021. Dimensional Emotion Detection from Categorical Emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Salud M. Jiménez Zafra, Arturo Montejo Ráez, M. Dolores Molina González, Luis Alfonso Ureña López, and María Teresa

Martín Valdivia. 2021. Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67(3):273–294.

Flor Miriam Plaza-del-Arco, Carlo Strapparava, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France. European Language Resources Association.

Robert Plutchik. 2001. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. "Modelling Valence and Arousal in Facebook posts". In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

Milos Radovanovic;, Alexandros Nanopoulos, and Mirjana Ivanovic. 2010. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research*, 11(86):2487–2531.

Anthony Rios and Ramakanth Kavuluru. 2018. "Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces". In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142, Brussels, Belgium. Association for Computational Linguistics.

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

Prateek Veeranna Sappadla, Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. Using Semantic Similarity for Multi-Label Zero-Shot Classification of Text Documents. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN-16)*, Bruges, Belgium. d-side publications.

Klaus R. Scherer and Harald G. Wallbott. 1997. The ISEAR Questionnaire and Codebook. Geneva Emotion Research Group.

Craig A. Smith and Phoebe C. Ellsworth. 1985. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813.

Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. 2013. Zero-Shot Learning through Cross-Modal Transfer. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 935–943, Red Hook, NY, USA. Curran Associates Inc.

P.S. Sreeja and G.S. Mahalaksmi. 2015. Applying vector space model for poetic emotion recognition. *Advances in Natural and Applied Sciences*, 9(6 SE):486–491.

Marco Antonio Stranisci, Simona Frenda, Eleonora Ceccaldi, Valerio Basile, Rossana Damiano, and Viviana Patti. 2022. Appreddit: a corpus of reddit posts annotated for appraisal. In *Proceedings of the Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France. European Language Resources Association.

Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. "WASSA 2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories". In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104, Online. Association for Computational Linguistics.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1).

Enrica Troiano, Sebastian Padó, and Roman Klinger. 2019. "Crowdsourcing and Validating Event-focused Emotion Corpora for German and English". In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.*, 10(2).

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, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. "Transformers: State-of-the-Art Natural Language Processing". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online. Association for Computational Linguistics.

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. 2019. Zero-Shot Learning—A

Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265.

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 Experiment Results

Dataset	Model	Emo-Name P	Emo-Name R	Emo-Name F_1	Expr-Emo P	Expr-Emo R	Expr-Emo F_1	Feels-Emo P	Feels-Emo R	Feels-Emo F_1	WN-Def P	WN-Def R	WN-Def F_1
TEC	r	.39	.42	.37	.36	.38	.34	.39	.40	.35	.37	.31	.28
	b	.39	.42	.38	.38	.42	.37	.40	.41	.37	.32	.32	.29
	d	.42	.47	.43	.41	.42	.38	.42	.42	.38	.44	.33	.30
	d-synonyms	.42	.42	.39	.39	.40	.37	.39	.40	.36	—	—	—
BLOGS	r	.32	.62	.29	.36	.60	.30	.41	.59	.32	.47	.47	.30
	b	.33	.58	.28	.35	.62	.31	.47	.56	.35	.35	.40	.24
	d	.35	.64	.31	.41	.65	.36	.49	.58	.37	.38	.48	.25
	d-synonyms	.41	.62	.39	.42	.63	.39	.36	.60	.31	—	—	—
ISEAR	r	.58	.50	.50	.53	.50	.50	.55	.47	.47	.50	.37	.37
	b	.62	.56	.56	.64	.60	.60	.68	.53	.53	.57	.40	.37
	d	.63	.56	.53	.66	.62	.60	.68	.54	.54	.54	.45	.43
	d-synonyms	.64	.57	.57	.64	.62	.61	.63	.58	.55	—	—	—

Table 6: Results from the set of prompts across emotion datasets (TEC, BLOGS and ISEAR) and NLI models. We report macro-average precision (P), macro-average recall (R), and macro-average F_1 (F_1) for each model. (r: RoBERTa, b: BART, d: DeBERTa, d-synonyms: DeBERTa using as prompts synonyms. In cases where no experiments have been performed, we use '—'. Figures 2 and 3 in the paper depict these experiments.

B List of Emotion Representations as Part of Prompts

Emotion	**Emo-S**	**Expr-S**	**Feels-S**	**WN-Def**
Context	ϵ	"This text expresses..."	"This person feels..."	"This person expresses..."
anger	anger, annoyance, rage, outrage, fury, irritation	anger, annoyance, rage, outrage, fury, irritation	anger, annoyed, rage, outraged, furious, irritated	a strong feeling of annoyance, displeasure, or hostility
fear	fear, horror, anxiety, terror, dread, scare	fear, horror, anxiety, terror, dread, scare	fear, horror, anxiety, terrified, dread, scared	an unpleasant emotion caused by the belief that someone or something is dangerous, likely to cause pain , or a threat
joy	joy, achievement, pleasure, awesome, happy, blessed	joy, an achievement, pleasure, the awesome, happiness, the blessing	joyful, accomplished, pleasure, awesome, happy, blessed	a feeling of great pleasure and happiness
sadness	sadness, unhappy, grief, sorrow, loneliness, depression	sadness, unhappiness, grief, sorrow, loneliness, depression	sadness, unhappy, grieved, sorrow, lonely, depression	emotions experienced when not in a state of well-being
disgust	disgust, loathing, bitter, ugly, repugnance, revulsion	disgust, loathing, bitterness, ugliness, repugnance, revulsion	disgusted, loathing, bitter, ugly, repugnance, revulsion	a feeling of revulsion or strong disapproval aroused by something unpleasant or offensive
surprise	surprise, astonishment, amazement, impression, perplexity, shock	surprise, astonishment, amazement, impression, perplexity, shock	surprised, astonishment, amazement, impressed, perplexed, shocked	a feeling of mild astonishment or shock caused by something unexpected
guilt	guilt, culpability, blameworthy, responsibility, misconduct, regret	guilt, culpability, responsibility, blameworthy, misconduct, regret	guilty, culpable, responsible, blame, misconduct, regretful	a feeling of having done wrong or failed in an obligation
shame	shame, humiliate, embarrassment, disgrace, dishonor, discredit	shame, humiliation, embarrassment, disgrace, dishonor, discredit	shameful, humiliated, embarrassed, disgraced, dishonored, discredit	a painful feeling of humiliation or distress caused by the consciousness of wrong or foolish behavior

Table 7: Emotion representation in prompts Emo-S, Expr-S, Feels-S, and WN-Def.