Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 165-177 recibido 31-03-2023 revisado 07-05-2023 aceptado 10-05-2023

# Entailment-based Task Transfer for Catalan Text Classification in Small Data Regimes

## *Transferencia de Tareas basada en Implicación Textual para la Clasificación de Textos en Catalán en Escenarios de Pocos Datos*

**Irene Baucells de la Peña,**[1,2] **Blanca Calvo Figueras,**[2]
**Marta Villegas,**[2] **Oier Lopez de Lacalle**[1]
[1]HiTZ Center - University of the Basque Country UPV/EHU
[2]Barcelona Supercomputing Centre
{irene.baucells, blanca.calvo, marta.villegas}@bsc.es, oier.lopezdelacalle@ehu.eus

**Abstract:** This study investigates the application of a state-of-the-art zero-shot and few-shot natural language processing (NLP) technique for text classification tasks in Catalan, a moderately under-resourced language. The approach involves reformulating the downstream task as textual entailment, which is then solved by an entailment model. However, unlike English, where entailment models can be trained on huge Natural Language Inference (NLI) datasets, the lack of such large resources in Catalan poses a challenge. In this context, we comparatively explore training on monolingual and (larger) multilingual resources, and identify the strengths and weaknesses of monolingual and multilingual individual components of entailment models: pre-trained language model and NLI training dataset. Furthermore, we propose and implement a simple task transfer strategy using open Wikipedia resources that demonstrates significant performance improvements, providing a practical and effective alternative for languages with limited or no NLI datasets.

**Keywords:** Entailment, Few-Shot, Multilingual Models, Text Classification.

**Resumen:** El presente trabajo investiga una reciente técnica de aprendizaje zero-shot y few-shot, en que la tarea objetivo se reformula como un problema de implicación textual y se resuelve mediante un modelo de implicación (un modelo de lenguaje entrenado con un corpus de implicación o NLI (Natural Language Inference)), para abordar tareas de clasificación textual en catalán, una lengua con recursos limitados que dispone de un corpus de NLI de tamaño moderado. Comparamos su aplicación con los recursos en esta lengua frente a los multilingües, de tamaño muy superior. Así mismo, identificamos las ventajas y limitaciones de ambas aproximaciones y el impacto del tamaño y la lengua del modelo de lenguaje y corpus de NLI. Finalmente, implementamos una estrategia de transferencia de aprendizaje, empleando datos extraídos de Wikipedia, que consigue mejoras significativas y demuestra ser una opción interesante para lenguas que disponen de un corpus de NLI reducido o carecen de él.

**Palabras clave:** Implicación, Few-shot, Recursos multilingües, Clasificación textual.

## 1 Introduction

Over the past years, the prevailing paradigm in natural language processing (NLP) has been to pre-train a language model (LM) through task-agnostic, self-supervised training and fine-tune it on annotated data from the target task, typically around a thousand examples. However, as the demand for NLP applications in industry continues to grow, the need to address new domains, tasks, and languages, where annotated data is often scarce or non-existent, becomes increasingly critical. In the quest for systems that learn from a small number of examples (few-shot) or even without specific data (zero-shot), several proposals have emerged in recent years, generally aiming at exploiting the knowledge already contained in pre-trained LMs.

One such proposal is the entailment-based approach, where the target task is reformulated as a natural language inference (NLI) or textual entailment (TE) task and passed as input to an entailment model, whose output is then mapped to that of the target task.

ISSN 1135-5948 DOI 10.26342/2023-71-13

The method has been mainly studied for NLP classification tasks, obtaining promising results (Yin, Hay, and Roth, 2019; Wang et al., 2021). Its main advantages include providing a common framework for unifying different NLP tasks, and the ability to leverage large, general-purpose NLI datasets to train the entailment model used for inference.

However, the usefulness of the entailment-based approach in data-poor scenarios has been demonstrated primarily for tasks in English, where huge NLI datasets and powerful models are readily available, raising the question of the approach's dependence on these large resources —which seems paradoxical given its intended use in data-scarce scenarios. Our research aims to investigate the **feasibility and potential improvements of the entailment-based approach for languages with fewer resources**. Specifically, we focus on Catalan, a medium-resource language for which a limited NLI dataset is available, and we investigate a multi-class text classification task (TC), due to its similarity to other classification tasks already studied within the entailment-based framework.

In addition to investigating the capabilities and limitations of monolingual resources for entailment-based TC, we experiment with pre-trained LMs and NLI datasets to build partially and fully multilingual entailment models and address the following questions: Are multilingual and larger resources more effective than monolingual and fewer resources? What is the individual contribution of monolingual and multilingual pre-trained LMs and NLI datasets? At this point, our research intersects with the ongoing debate surrounding the use of monolingual compared to multilingual resources. Notably, Armengol-Estapé et al. (2021) and Agerri et al. (2020) have examined the performance of monolingual Catalan and Basque LMs, respectively, against state-of-the-art multilingual models on several NLP tasks and have concluded the superiority of language-specific models within the pre-training and fine-tuning paradigm. Our work builds on this line of research by investigating the comparison of monolingual and multilingual resources in the context of the entailment-based approach. Finally, our research looks at task transfer learning, where we reuse data from a similar task transformed into NLI, to seek potential improvements.

In summary, our main contributions are:

- Applying the entailment-based approach in zero- and few-shot settings to address a TC task in Catalan.
- Providing valuable insight into the advantages and limitations of using monolingual and multilingual resources, which can be useful for guiding future efforts in resource creation.
- Implementing a simple task transfer strategy that significantly improves the zero-shot capabilities of entailment models for TC using monolingual resources.[1]

This article is organized as follows. In Section 2, we contextualize the entailment-based approach within the broader landscape of zero- and few-shot methods, and identify the key existing works that have utilized it for classification tasks. Section 3 presents the methodology followed for the experiments, including our research objectives. Finally, in Section 5, we summarize the main insights from our study and outline possible directions for future research.

## *2 Related Work*

### 2.1 Zero- and Few-shot Learning

Zero-shot (ZS) and few-shot (FS) learning surge as powerful solutions to address real-world scenarios where the limited availability of annotated data renders the standard fine-tuning approach inadequate to attain satisfactory performance levels (Schick and Schütze, 2021a).[2] Several methods have been developed for ZS and FS learning. One such method is Parameter-Efficient Fine-Tuning (PEFT) for FS (Liu et al., 2022), which fine-tunes only a subset of the model parameters using limited training examples. SetFit (Tunstall et al., 2022), a sentence transformer-based method, has recently

[1] We have made available our task transfer entailment model and the Catalan Wikipedia-based TC dataset (CaWikiTC) created as part of these experiments. They can be accessed at `https://huggingface.co/projecte-aina/roberta-base-ca-v2-cawikitc` and `https://huggingface.co/datasets/projecte-aina/CaWikiTC`, respectively. The code used is also publicly available at `https://github.com/ibaucells/entailment_based_catalan_tc`.

[2] Beyond practical objectives, ZS and FS is the search for models with true generalization abilities that can learn new tasks in a manner that mirrors human learning, relying on small explanations or just a few examples.

emerged as a leading technique for classification tasks. Other methods are based on data augmentation (Xie et al., 2019), intermediate task learning such as STILTs (Phang, Févry, and Bowman, 2018), and improvements over standard fine-tuning to handle a small number of training examples (Lee, Cho, and Kang, 2019; Zhang et al., 2020).

However, some of the main ZS and FS techniques for NLP are based on eliciting the kwnowledge of pre-trained LMs through a reformulation of the final task. Those that rely on prompting the LM with a target task that is reformulated to be similar to the LM's pre-training goals are often referred to as prompt-based approaches, and have emerged powerfully, even as a new paradigm —pretrain, prompt, predict— able to replace the current pre-training and fine-tuning approach (Liu et al., 2021). One possibility is to prompt a generative LM with a task description and demonstrations, where GPT-3 (Brown et al., 2020), a 175B-parameter model, represented a breakthrough proof of the powerful ZS and FS capabilities of huge pre-trained LMs by achieving near-SOTA performances on various NLP tasks. To leverage Masked LMs, the target task is converted into a cloze-question problem. One such prominent technique for FS is PET[3] (Schick and Schütze, 2021a; Schick and Schütze, 2021b), of which later improvements have been proposed (Tam et al., 2021; Mahabadi et al., 2022), and LM-BFF[4] (Gao, Fisch, and Chen, 2021).

An alternative research direction investigates methods that reformulate the final task into a different non-language modeling NLP task serving as a bridge. In particular, the Natural Language Inference (NLI) task, a two-input classification task that requires deciding whether or not the meaning of a sentence (premise) entails a second one (hypothesis), has been proposed as a common, task-agnostic formulation for solving various NLP tasks (Yin, Hay, and Roth, 2019; Wang et al., 2021). This approach, referred to as entailment-based approach, has demonstrated its usefulness across diverse classification and, more recently, information extraction tasks (Sainz et al., 2021; Sainz et al., 2022b; Sainz et al., 2022a).

## 2.2 Entailment-based Approach for Classification Tasks

Yin, Hay, and Roth (2019) identify the unique challenges of dealing with TC tasks across different domains (news, reviews, etc.), aspects (topic, emotion, etc.) and label space characteristics, without target task data. Furthermore, they establish a benchmark for comparing different systems and propose an entailment-based approach to address these challenges.[5] Using entailment models trained on NLI datasets in English, their method overperforms the (scarce) existing baselines for ZS, such as Explicit Semantic Analysis (Chang et al., 2008) and Word2Vec (Mikolov et al., 2013), in the proposed benchmark. Other works that adopt the approach in ZS include those by Sainz and Rigau (2021) for domain labeling and Obamuyide and Vlachos (2018) for relation classification. More recently, in the context of ZS, a parallel work (Pàmies et al., 2023) investigates entailment-based TC on the scientific domain through a task transfer approach comparable to ours, consisting of training a model with in-domain data from another task reformulated as entailment. They demonstrate significant improvements over standard entailment-based TC. Our task transfer experiments further support the advantages of this approach, both in ZS and FS, in the context of an under-resourced language and general domain TC.

Wang et al. (2021) propose the entailment-based framework in few-shot settings to address any classification task. Their best method, EFL, involves training an entailment model on a large-scale general NLI dataset (MNLI) and fine-tuning it on the few training examples reformulated as NLI. With 8 examples per class, EFL outperforms the standard fine-tuning and the other few-shot techniques considered (majority, LM-BFF, and STILTS) in 15 NLP tasks, with an average 8.2 % improvements over them.

While the approach has shown impressive results across tasks, its limitations have al-

[3] In PET, the few training examples are reformulated as cloze-phrases using various patterns, and each is used to train a separate LM; the resulting models are ensembled together in order to annotate unlabeled data with soft labels, which are finally used to fine-tune a standard classification model.

[4] LM-BFF uses automatically generated prompts and task demonstrations.

[5] They state two significant advantages of the approach over the standard (supervised) classification formulation: it removes the need to specify the number of output classes and uses the label names for the task (instead of converting them to indexes).

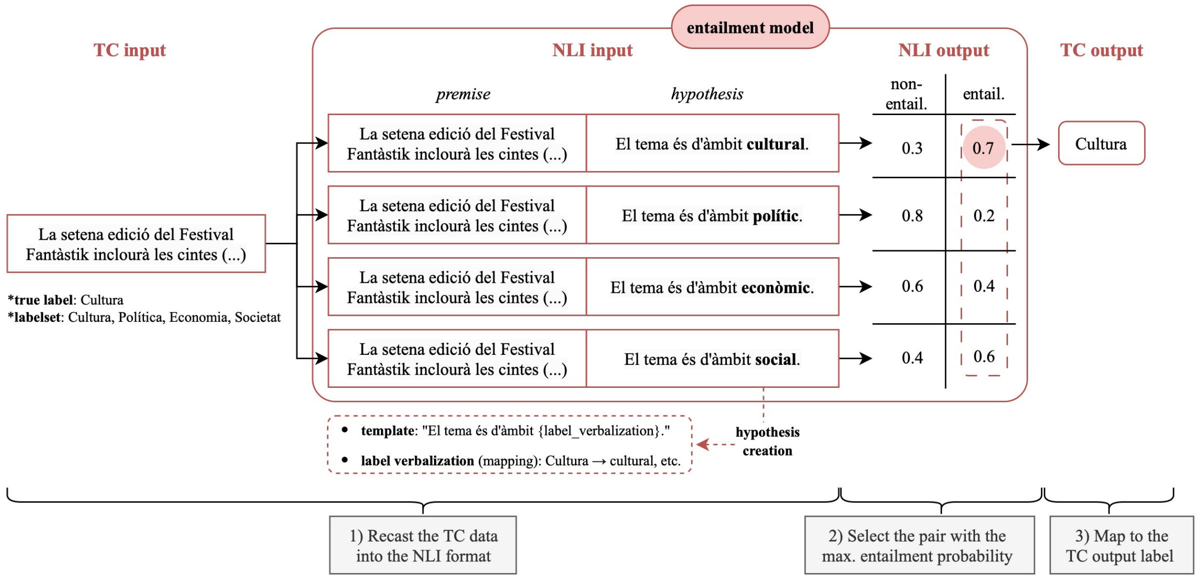

TC input
entailment model
NLI input
NLI output
TC output
premise
hypothesis
non-entail.
entail.
La setena edició del Festival Fantàstik inclourà les cintes (...)
*true label: Cultura
*labelset: Cultura, Política, Economia, Societat
El tema és d'àmbit cultural.
El tema és d'àmbit polític.
El tema és d'àmbit econòmic.
El tema és d'àmbit social.
0.3
0.7
0.8
0.2
0.6
0.4
0.4
0.6
Cultura
template: "El tema és d'àmbit {label_verbalization}."
label verbalization (mapping): Cultura → cultural, etc.
hypothesis creation
1) Recast the TC data into the NLI format
2) Select the pair with the max. entailment probability
3) Map to the TC output label

Figure 1: Steps involved in the entailment-based text classification (TC) at inference time.

so been brought to light by some studies. In particular, Ma et al. (2021) have questioned the value of the NLI training, showing that a BERT model (without any fine-tuning) can achieve similar or better results by reusing the Next Sentence Prediction objective. Besides, they suggest that limited NLI datasets may negatively affect performance, as also highlighted in our comparison of monolingual and multilingual resources. Despite these challenges, the approach remains a prominent option for zero- and few-shot TC, with many still unexplored fronts and significant potential for improvement.

# *3 Methodology*

## 3.1 Research Objectives

The experimentation is divided into two branches, each with its own research objective(s):

(a) **Monolingual vs. multilingual**.

1. Evaluate the capabilities of the entailment-based approach for TC in zero- and few-shot using Catalan monolingual resources.
2. Compare the performance of (smaller) monolingual to (larger) multilingual resources for the entailment approach in the Catalan setting and determine the impact of the larger size and language specificity of the pre-trained model and NLI dataset.
3. Explore the robustness of monolingual vs. multilingual models in zero-shot scenarios when faced with changes in the premise length and hypothesis template.

(b) **Task transfer**. Seek performance improvements of the approach through task transfer learning.

## 3.2 Target Task

We focused on the TeCla dataset[6] as the target multi-class text classification (TC) task in Catalan. TeCla is a collection of 113,376 news articles labeled based on a hierarchical class structure, with each article assigned a coarse-grained class from among four possible classes and a fine-grained class from among 53 possible classes. To leverage the inherent two-level difficulty, we treated both categorizations as separate tasks. Given the highly imbalanced class distribution of the dataset, we employed weighted F1 as the main metric for evaluating the performance of our models.

## 3.3 Approach

Our application of the entailment-based approach aligns with previous works (Wang et al., 2021; Yin, Hay, and Roth, 2019) and can be summarized into the three steps illustrated in Figure 1 with an example from the TeCla dataset. Firstly, the TC data is converted into the NLI format as follows: each TC example generates a number of premise-hypothesis pairs equal to the number of la-

[6] Available at `https://huggingface.co/datasets/projecte-aina/tecla`.

| Entailment model | Pre-trained LM | NLI dataset |
|---|---|---|
| RoBERTa-ca-Teca | RoBERTa-base-ca-v2<br>125M params.<br>34.9GB train. data<br>*monolingual (ca)* | Teca<br>21,163 NLI pairs<br>*monolingual (ca)* |
| XLMR-Teca | XLM-RoBERTa-base<br>270M params.<br>2.5TB train data<br>*multilingual* | |
| XLMR-SMAX | | SNLI, MNLI, ANLI, XNLI<br>15 languages (mainly en)<br>1284k NLI pairs<br>*multilingual* |

Table 1: Entailment models used in the zero- and few-shot experiments with their respective pre-trained LM and NLI dataset(s).

bels in the task, all using the same premise (i.e. the text from the TC task), but different hypotheses. Each hypothesis consists of a sentence indicating that the text belongs to one of the possible labels, and has been created using two elements: a template, with a fill-in gap for the label (for example, "This text is about {label}."), and a label verbalization, which is the mapping from the label to a word or description to be replaced in the template. In the second step, the entailment model receives this NLI data as input and returns the probabilities for entailment and non-entailment (or entailment, neutral, and contradiction, if the NLI training data makes this three-way distinction). The NLI pair with the highest entailment probability is subsequently selected, and, in the last step, the label verbalization used to form its hypothesis will be mapped to the original label to obtain the final prediction.

### 3.4 Experimental design

In the **monolingual vs. multilingual** experimental branch (a), RoBERTa-ca-Teca, a Catalan monolingual entailment model, is evaluated against two multilingual variants: a partially multilingual model (XLMR-Teca), trained on the Catalan NLI dataset, and a fully multilingual model (XLMR-SMAX), trained on various English and multilingual NLI datasets.[7] Table 1 summarizes the key specifications of each model. Note the considerably bigger size of the multilingual pre-trained model and NLI datasets.

For the template experimentation in the zero-shot scenario, we developed 17 templates in Catalan divided into two sets, which can be found in Appendix A: the first, with twelve of them, are slight linguistic variations of a commonly used template in literature (the English translation is "This text is about {label}.") and use the lowercased label as the verbalization; the second, with the remaining five, are designed to allow for a verbalization consisting of the adjectivized label (e.g., "culture" becomes "cultural"). Additionally, given that NLI datasets typically use one-sentence premises —in contrast to texts, which result from the conversion of TC data to NLI—, the zero-shot experiments were conducted with two different premise-shortening setups: using the entire text as the premise vs. using only the first sentence (corresponding to the title of the article in the original data). Both experimentations, on templates and premise shortening, were performed on the TeCla development set, and the best setting for each model was chosen for the test evaluation.

In the few-shot scenario, each model was fine-tuned on a small amount of data from the target task reformulated as NLI.[8] For the reformulation, we used the best-performing templates from zero-shot and created all possible non-entailment pairs per each TC example. Four data regimes are explored: 1-1, 8-4, 16-8, and 32-16, where the first and second digits refer to the number of training and development examples per class, respectively. For training, the hyperparameters are kept fixed at a learning rate of 3e-5, a batch size of 16, and a maximum of 10 epochs, and the development set is used to select the best checkpoint according to the highest weighted F1 score in the classification task. The two premise-shortening setups from zero-shot are again tested in the small development partition, and only the best one is evaluated on the test set. In Appendix B, we provide details on additional experiments we conducted to investigate the impact of the ratio of non-entailment pairs generated and the checkpoint selection strategy, which support

[7] RoBERTa-ca-Teca and XLMR-SMAX were already available at `https://huggingface.co/projecte-aina/roberta-base-ca-v2-cased-te` and `https://huggingface.co/symanto/XLM-RoBERTa-base-snli-mnli-anli-xnli`, respectively. XLMR-Teca, on the other hand, was fine-tuned for this work with a learning rate of 1e-5, 10 maximum epochs, and a batch size of 16. It achieved 79 % accuracy in the NLI Teca dataset, while RoBERTa-ca-Teca reached 83 % accuracy.

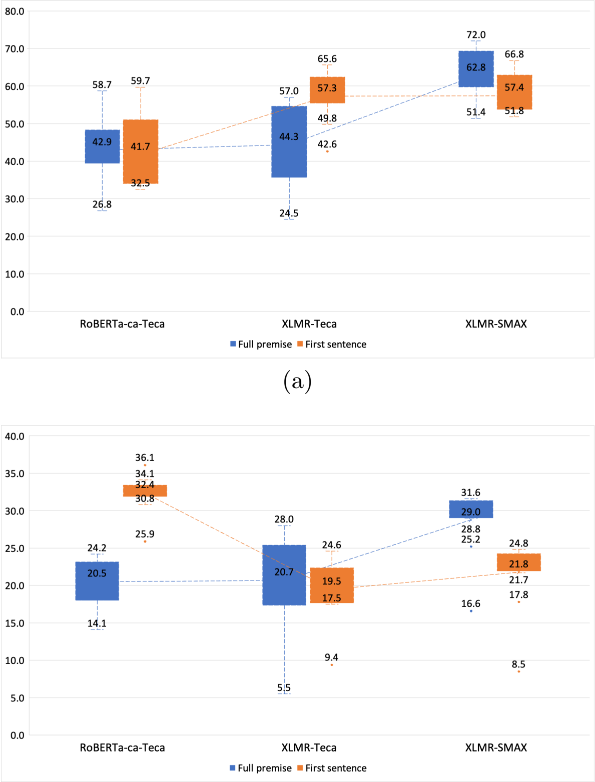
[8] The reformulation stage is the same as shown in Figure 1, but with the addition of the correct label for the entailment task (i.e., "entailment" if the label in the template is the correct category of the text, and "non-entailment" otherwise) and the optionality of creating all possible non-entailment pairs.

the effectiveness of our choices.

In the **task transfer** branch (b), we experiment with a task transfer strategy to build an improved monolingual entailment model for TC that consists of three simple steps: 1) obtaining data from a related task (i.e., another TC task), 2) converting it to NLI data, and 3) using it to fine-tune an entailment model. To this end, we scraped 21,002 article summaries from the Catalan Wikipedia and their corresponding labels, yielding a total of 67 exclusive classes, to be used as the task transfer source.[9] To convert this TC dataset, which we called CaWikiTC, into NLI, we used the template with which the monolingual model achieved the highest performances in zero-shot[10] and generated one non-entailment pair per each entailment pair.[11] Two entailment models were developed using this NLI data: RoBERTa-ca-CaWikiTC, by directly fine-tuning the monolingual LM on it, and RoBERTa-ca-Teca-CaWikiTC, by further fine-tuning our model trained on the Catalan NLI dataset, Teca. Moreover, to gain deeper insights into the importance of these strategies in the few-shot setting, we also trained the monolingual LM using the available target task data from each data regime without prior training on Teca or CaWikiTC. Throughout these experiments, we used the same training configuration presented earlier for the few-shot experiments.

**Baselines.** The following baselines are used in the zero-shot (ZS) and few-shot (FS) experiments:

- **Majority.** The most common label from the full training set.
- **Random.** A random uniform classifier.
- **Prompt-based approach (ZS).** The text to classify is input to the monolingual LM concatenated with a template (the same templates used for the entailment models are evaluated in development, and the best is chosen for testing), where the label verbalization is replaced by a mask token. The output space is restricted to the possible output classes.
- **Supervised models (FS).** A standard fine-tuning of RoBERTa-base-ca-v2 and XLM-RoBERTa-base using the available training and development data from the target task.
- **SetFit (FS).** SetFit models involve converting the available data into contrastive pairs to fine-tune a sentence transformer, encoding the original text with it, and using the resulting sentence embeddings to train a classification head.[12]



(a)

(b)

Figure 2: Zero-shot performances over the TeCla development set using the full text vs. the first sentence as premise; (a) refers to the coarse-grained task across 17 templates (template sets 1 and 2) and (b) to the fine-grained task across 13 templates (template set 1).

[9] To build the dataset, we extracted all the texts belonging to the specified categories (manually selected from the possible categories of a similar thematic hierarchy level in the Catalan Wikipedia) and removed the texts associated with more than one category.

[10] The template is "Aquest article tracta sobre {label}.", meaning "This article is about {label}.".

[11] We made this choice to limit the size of the NLI training data and prevent an increase in computational cost during posterior fine-tuning.

[12] For its implementation, since we did not find any sentence transformer (ST) in Catalan, we used a multilingual ST, paraphrase-multilingual-mpnet-base-v2, available at `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`, which has been trained on parallel data from over 50 languages, including Catalan. We used the default training configuration options from the official li-

| Task | Model | | ZS | 1-1 | 8-4 | 16-8 | 32-16 |
|---|---|---|---|---|---|---|---|
| coarse-grained task | entailment models | RoBERTa-ca-Teca | 59.7 | 56.9 ± 8.8 | 79.2 ± 3.3 | 82.4 ± 2.9 | **89.2 ± 0.6** |
| | | XLMR-Teca | 63.9 | **63.1 ± 2.0** | **81.5 ± 2.1** | **86.7 ± 1.3** | 86.7 ± 2.8 |
| | | XLMR-SMAX | **71.1** | 56.5 ± 8.3 | 79.7 ± 7.4 | **86.7 ± 1.0** | 87.7 ± 2.1 |
| | baselines | prompt-based | 52.4 | - | | | |
| | | majority | 23.5 | - | | | |
| | | random | 25.5 | - | | | |
| | | supervised-RoBERTa-base-ca-v2 | - | 28.5 ± 4.4 | 63.0 ± 9.8 | 74.7 ± 8.1 | 83.5 ± 2.9 |
| | | supervised-XLM-RoBERTa-base | - | 8.9 ± 22.1 | 40.6 ± 24 | 65.4 ± 20.9 | 87.8 ± 1.6 |
| fine-grained task | entailment models | RoBERTa-ca-Teca | **36.3** | **48.5 ± 4.2** | **60.2 ± 1.4** | **62.4 ± 1.3** | **63.2 ± 1.3** |
| | | XLMR-Teca | 27.0 | 41.2 ± 4.1 | 51.3 ± 0.7 | 56.7 ± 1.8 | 60.7 ± 0.5 |
| | | XLMR-SMAX | 31.4 | 40.0 ± 3.1 | 53.8 ± 2.2 | 57.0 ± 0.8 | 60.0 ± 0.9 |
| | baselines | prompt-based | 22.8 | - | | | |
| | | majority | 2.2 | - | | | |
| | | random | 2.3 | - | | | |
| | | supervised-RoBERTa-base-ca-v2 | - | - | 50.0 ± 5.1 | 54.4 ± 4.1 | 61.8 ± 2.4 |
| | | supervised-XLM-RoBERTa-base | - | - | 44.0 ± 2.7 | 51.6 ± 1.9 | 61.3 ± 0.6 |

Table 2: Monolingual vs. multilingual entailment models' performance (weighted F1) on the coarse-grained (4 classes) and fine-grained (53 classes) tasks of the TeCla test set in the zero-shot (ZS) scenario and four different few-shot regimes (where x-y denotes x examples/class for training and y for development, etc.). The results in few-shot are the mean and standard deviation across three training and development samples within each data regime. A hyphen indicates that the model has been unable to learn in the given setting.

# *4 Results and Discussion*

## 4.1 Monolingual vs. Multilingual

### 4.1.1 Zero-shot experiments

The box plots in Figure 2 summarize the results of the zero-shot experiments with the shortening of the premise and templates on the TeCla development set. When comparing the full premise to the first sentence setups, XLMR-SMAX is the only model that consistently performs better when using the full premise setup, while the others demonstrate unstable behavior across the coarse- and fine-grained tasks. This result is consistent with another finding that emerges from a closer look at XLMR-Teca: in the full premise setting, the two models that share the monolingual NLI training dataset obtain very similar mean scores, which are always far below XLMR-SMAX; in parallel, in the first sentence setting, the two models that share the pre-trained LM (XLMR) yield quite similar mean scores. One possible explanation is that the monolingual NLI dataset limits the model's ability to leverage longer textual premises, due to the significant shift in data distribution between the NLI training data and the test data, i.e. the monolingual NLI dataset contains only one-sentence premises, while some of the NLI datasets in XLMR-SMAX[13] contain much longer premises. This also suggests that the role of the NLI dataset is only to enable the knowledge from the pre-trained LM, which plays the most important role.

The template variations also led to large fluctuations in model performance without any discernible patterns, which aligns with other studies (Sainz and Rigau, 2021; Ma et al., 2021). Not only does each entailment model show a preference for different templates, but the premise setup also significantly affects this preference on the same model. Furthermore, even small, semantically irrelevant differences in templates lead to drastic changes in results. In a comparison between models, XLMR-Teca was found to be the most inconsistent model across templates, with the largest standard deviations and the lowest overall performance in both tasks; in addition, further exploration revealed that this model's worst F1 results occur when using adjectival labels, while other models tend to achieve their best results with them.

### 4.1.2 Few-shot experiments

Zero- and few-shot results of the monolingual, partially, and fully multilingual entailment models, as well as the baselines, are presented in Table 2. Notably, the model performance varies significantly between the coarse- and fine-grained TC tasks. In ZS, XLMR-SMAX performs best in the coarse-grained task, followed by XLMR-Teca and then RoBERTa-ca-Teca with a margin of 7.2 and 11.4 points, respectively. In the fine-

brary at `https://github.com/huggingface/SetFit`: batch size of 16, 1 epoch, cosine-similarity loss, and 20 iterations to generate sentence pairs.

[13] MNLI and XNLI cover multiple text genres and provide premises of varying lengths.

| **Task** | **Model** | | **ZS** | **1-1** | **8-4** | **16-8** | **32-16** |
|---|---|---|---|---|---|---|---|
| coarse-grained task | entailment models | RoBERTa-base-ca-v2 | - | 36.1 ± 9.5 | 45.0 ± 10.5 | 65.9 ± 4.4 | 78.1 ± 8.2 |
| | | RoBERTa-ca-Teca | 59.7 | 56.9 ± 8.8 | 79.2 ± 3.3 | 82.4 ± 2.9 | 89.2 ± 0.6 |
| | | RoBERTa-ca-CaWikiTC | **75.0** | **74.8 ± 0.6** | **80.9 ± 5.8** | **87.7 ± 0.9** | **89.6 ± 0.1** |
| | | RoBERTa-ca-Teca-CaWikiTC | 66.4 | 66.5 ± 0.9 | 79.9 ± 4.8 | 86.5 ± 1.1 | 88.2 ± 1.6 |
| | SetFit | | - | 47.7 ± 6.8 | 79 ± 6.2 | 84.7 ± 3.2 | 87.0 ± 1,7 |
| fine-grained task | entailment models | RoBERTa-base-ca-v2 | - | 13.8 ± 6.7 | 60.8 ± 2.8 | 62.1 ± 1.3 | 63.6 ± 2.1 |
| | | RoBERTa-ca-Teca | 36.3 | 48.5 ± 4.2 | 60.2 ± 1.4 | **62.4 ± 1.3** | 63.2 ± 1.3 |
| | | RoBERTa-ca-CaWikiTC | 49.1 | 51 ± 2.9 | **60.9 ± 0.3** | 60.9 ± 0.4 | 64.2 ± 0.9 |
| | | RoBERTa-ca-Teca-CaWikiTC | **49.8** | **53.4 ± 2.4** | 59.7 ± 1.2 | 61.5 ± 1.1 | **64.3 ± 0.4** |
| | SetFit | | - | 22.0 ± 3.6 | 50.3 ± 0.6 | 53.3 ± 1.2 | 56.7 ± 1.5 |

Table 3: Task transfer experiments' results (weighted F1) on the TeCla test set for the coarse-grained (4 classes) and fine-grained (53 classes) tasks. The two task transfer strategies, RoBERTa-ca-CaWikiTC and RoBERTa-ca-Teca-CaWikiTC, can be compared with the standard monolingual model from the previous section, RoBERTa-ca-Teca, and with the monolingual LM directly trained on the available few-shot regime.

grained task, RoBERTa-ca-Teca outperforms the other models, followed by XLMR-SMAX at 4.9 points and XLMR-Teca at 9.3 points behind. The lower performance of XLMR-Teca compared to XLMR-SMAX suggests that a larger (multilingual) NLI dataset consistently leads to better performances. However, the usefulness of monolingual and multilingual LM varies based on the task characteristics, with the multilingual LM performing better on the coarse-grained task and the monolingual LM on the fine-grained task. We hypothesize that exposure to a large amount of text, even if not in the target language, is critical for acquiring the general-domain inference skills needed in the coarse-grained task, whereas the fine-grained task may require more specific language-related knowledge that is not present in multilingual data. Compared to the baselines, the three entailment models significantly outperform them in both TC tasks. The best entailment model translates into a 35.7 % and 59.2 % of relative improvement over the best baseline in the coarse- and fine-grained tasks, respectively, and even the worst entailment model achieves a relative improvement of 13.9 % and 18.4 %.

In the FS scenario[14], the fine-grained results follow the tendencies from ZS: the monolingual entailment model outperforms the XLMR models in all data ratios, starting in the 1-1 scenario from an absolute difference of 8.5 and 7.3 with respect to XLMR-Teca and XLMR-SMAX, respectively, and eventually reaching a difference of 3.2 and 2.5 when the data ratio is 32-16. In the coarse-grained task, however, XLMR-SMAX loses its dominance, probably because of the language shift between the pre-training NLI dataset and the target task data: in the 1-1 and 8-4 data regimes, XLMR-Teca stands out significantly with the highest results; in the 16-8 stage, XLMR-SMAX catches up to XLMR-Teca; finally, RoBERTa-ca-Teca is 2.5 and 1.5 points ahead of the XLMR models in the 32-16 scenario, demonstrating the greatest ability to improve as more data is provided when compared to the others, which show little or no improvement from the last stage. It is also worth noting that the 1-1 regime shows a significant drop in performance with respect to ZS in the coarse-grained task. This is likely due to the extremely limited training data available, which is insufficient for effective generalization. Conversely, significant improvements are observed in the fine-grained task, where the larger number of output classes results in more training data which is further increased in its conversion to NLI.

Compared to the supervised baselines, the three entailment models outperform them up to the 32-16 data regime, where the worst entailment model is slightly outperformed. However, the best performance is consistently achieved with one entailment model, with the greatest improvements in the scarcest data scenarios: the absolute improvement gradually decreases from 34.6 to 1.4 in the coarse-grained task, and from 10.2 (8-4 set-

[14] The three models performed better when trained on the full premise setup (compared to the first sentence) in the development data of each few-shot regime, with the exception of RoBERTa-ca-Teca model in the 16-8 regime. Consequently, only in this particular case was this setup chosen for testing. In addition, during development, the XLMR models showed an inability to learn with the predefined hyperparameters in the fine-grained task. This was successfully handled by switching to a smaller learning rate, 1e-5, and leaving the other hyperparameters unchanged.

ting) to 1.4. Overall, the supervised models demonstrate a weaker learning capacity in low data regimes and a more unstable training[15], but show an accelerated progression as more data becomes available.

## 4.2 Task Transfer Experiments

Task transfer results, presented in Table 3 against a few-shot SOTA method (SetFit), demonstrate that the two task transfer strategies significantly improve the monolingual entailment model's performance in the ZS scenario and when few target task data is available. In ZS, the best of the two strategies achieves an impressive absolute improvement of 15.3 and 13.5 over RoBERTa-ca-Teca, trained on the Catalan NLI dataset. However, as more target task data becomes available, the improvements become less prominent, even with respect to the model without training on an NLI dataset before fine-tuning with target task data. Nevertheless, this previous NLI training remains essential in the most extreme data-scarce scenarios, as already observed in Wang et al. (2021). When a sufficient amount of training data is available, the distinctions between the models begin to blur: in the coarse-grained task, noticeable improvements can still be achieved in the 32-16 data regime, while, in the fine-grained task, the four models begin to perform very similarly already in the 8-4 data regime. Regarding the two task transfer strategies considered, their effectiveness was found to depend on the task characteristics. Specifically, training on CaWikiTC recast as NLI was more effective for the coarse-grained task, while a combination of Teca and CaWikiTC was generally preferable for the fine-grained task.

Compared to SetFit, RoBERTa-ca-Teca already exhibits better performance, but task transfer leads to further improvements, especially in the 1-1 regime, where absolute gains can reach up to 27.1 and 31.4 in the coarse-grained and fine-grained tasks, respectively. In the coarse-grained task, however, the improvement decreases significantly from the 8-4 regime onward, reaching up to 3 points of gain, while it remains about 9 points behind in all data regimes in the fine-grained task.

## 5 *Conclusions and Future Work*

The entailment-based approach proved to be an effective technique for tackling TC tasks in a medium-resource language, Catalan, providing significant improvements over our baselines (including supervised models and SetFit) in both zero-shot and few-shot scenarios, especially in the scarcest data regimes. By comparing the utility of monolingual and multilingual resources in the approach, we concluded that the size of the NLI dataset is a key factor in zero-shot: a larger NLI dataset not only improves the model's inference capabilities but also reduces the potential for introducing limiting bias —for example, in terms of the model's ability to understand premises of different lengths. The larger size of the LM is also advantageous, but the language-specific knowledge of the monolingual LM proved more valuable for certain TC tasks requiring nuanced categorization.

In the scarcest few-shot data scenarios, the advantage of using a model trained on larger multilingual NLI datasets disappears and even translates into worse performance, most likely due to the language shift. Besides, as more target task data becomes available, the weight of the NLI dataset decreases and the performance of entailment models using the same pre-trained LM tends to converge. In this context, again, the monolingual model performed better when the task required a more fine-grained, language-related categorization, but generally worse than the multilingual options otherwise. However, we observed that the monolingual entailment model had the greatest ability to improve as more target task data became available, whereas the progress of the multilingual models stalled earlier. Finally, we presented a task transfer learning setup where a different TC dataset was created from a Wikipedia crawl, converted to NLI, and used to train an entailment model. This strategy proved highly effective, yielding significant performance gains over all models and offering an attractive option for languages with limited or without NLI datasets.

The experiments also highlighted some limitations of the entailment-based method that future work might aim to address, such as the strong reliance on handwritten templates and verbalizations, particularly in zero-shot. A potential solution could be to incorporate advances from prompt-based learning, such as automatic retrieval or genera-

[15] In the one-shot data regime (1-1) of the fine-grained task, they were unable to learn, neither with a learning rate of 3e-5 nor 1e-5.

tion in natural language or continuous embedding space. Furthermore, because of the potential bias introduced by small and homogeneous NLI datasets in zero-shot, techniques for enriching them (for instance, with a varying premise length), may be especially helpful for less-resourced languages. Finally, the potential of entailment-based task transfer should be corroborated in the context of other under-resourced languages and explored in relation to different classification tasks, such as sentiment analysis, which may require more abstract inference skills. This could include identifying resources to reformulate as NLI that provide broader inference capabilities to enhance applicability across tasks.

## *Acknowledgements*

This work was funded by the Generalitat de Catalunya (Projecte AINA), the Basque Government (excellence research group IT1570-22) and by the DeepKnowledge (PID2021-127777OB-C21) project funded by MCIN/AEI/10.13039/501100011033.

## *References*

Agerri, R., I. S. Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, and E. Agirre. 2020. Give your text representation models some love: the case for basque.

Armengol-Estapé, J., C. P. Carrino, C. Rodriguez-Penagos, O. de Gibert Bonet, C. Armentano-Oller, A. Gonzalez-Agirre, M. Melero, and M. Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, 7.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language models are few-shot learners. 5.

Chang, M.-W., L. Ratinov, D. Roth, and V. Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, AAAI'08, page 830–835. AAAI Press.

Gao, T., A. Fisch, and D. Chen. 2021. Making pre-trained language models better few-shot learners. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3816–3830.

Lee, C., K. Cho, and W. Kang. 2019. Mixout: Effective regularization to finetune large-scale pretrained language models. 9.

Liu, H., D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, and C. Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. 5.

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Ma, T., J. G. Yao, C. Y. Lin, and T. Zhao. 2021. Issues with entailment-based zero-shot text classification. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2:786–796.

Mahabadi, R. K., L. Zettlemoyer, J. Henderson, M. Saeidi, L. Mathias, V. Stoyanov, M. Yazdani, and M. Ai. 2022. Perfect: Prompt-free and efficient few-shot learning with language models.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings.*

Obamuyide, A. and A. Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium, November. Association for Computational Linguistics.

Pàmies, M., J. Llop, F. Multari, N. Duran-Silva, C. Parra-Rojas, A. Gonzalez-Agirre, F. A. Massucci, and M. Villegas. 2023. A weakly supervised textual entailment approach to zero-shot text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 286–296, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Phang, J., T. Févry, and S. R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks.

Sainz, O., O. L. de Lacalle, G. Labaka, A. Barrena, and E. Agirre. 2021. Label verbalization and entailment for effective zero- and few-shot relation extraction.

Sainz, O., I. Gonzalez-Dios, O. Lopez de Lacalle, B. Min, and E. Agirre. 2022a. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States, July. Association for Computational Linguistics.

Sainz, O., H. Qiu, O. L. de Lacalle, E. Agirre, and B. Min. 2022b. Zs4ie: A toolkit for zero-shot information extraction with simple verbalizations. *NAACL 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pages 27–38.

Sainz, O. and G. Rigau. 2021. Ask2transformers: Zero-shot domain labelling with pre-trained language models. *GWC 2021 - Proceedings of the 11th Global Wordnet Conference*, pages 44–52, 1.

Schick, T. and H. Schütze. 2021a. Exploiting cloze questions for few shot text classification and natural language inference. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 255–269.

Schick, T. and H. Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 2339–2352.

Tam, D., R. R. Menon, M. Bansal, S. Srivastava, and C. Raffel. 2021. Improving and simplifying pattern exploiting training. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 4980–4991.

Tunstall, L., N. Reimers, U. E. S. Jo, L. Bates, D. Korat, M. Wasserblat, O. Pereg, and H. Face. 2022. Efficient few-shot learning without prompts.

Wang, S., H. Fang, M. Khabsa, H. Mao, and H. Ma. 2021. Entailment as few-shot learner. 4.

Xie, Q., Z. Dai, E. Hovy, M. T. Luong, and Q. V. Le. 2019. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 2020-December, 4.

Yin, W., J. Hay, and D. Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *EMNLP-IJCNLP 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3914–3923, 8.

Zhang, T., F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi. 2020. Revisiting few-sample BERT fine-tuning.

## A Appendix 1: Templates and label verbalizations

The two sets of templates used in the zero-shot experiments are listed in Table 4 and 5.

## B Appendix 2: Entailment model's checkpoint selection and negative hypotheses generation strategies

In the few-shot learning experiments conducted, two specific configuration decisions were consistently applied. Firstly, for the training of each entailment model with the available training data, the checkpoint that achieved the highest F1 score in the target task (text

| | | | |
|---|---|---|---|
| **Templates** | 1 | Aquest text tracta sobre {label}. | *original* |
| | 2 | Aquest text va sobre {label}. | *verb change* |
| | 3 | Aquest text és sobre {label}. | |
| | 4 | Aquest text tracta de {label}. | *preposition change* |
| | 5 | El text tracta sobre {label}. | *article change* |
| | 6 | Aquest exemple tracta sobre {label}. | *noun change* |
| | 7 | Aquest article tracta sobre {label}. | |
| | 8 | Això tracta sobre {label}. | *noun phrase change* |
| | 9 | ø Tracta sobre {label}. | |
| | 10 | Aquest text tracta sobre {label} | *punctuation change* |
| | 11 | {label} | *only label* |
| | 12 | Pregunta: El text tracta sobre {label}? Resposta: Sí. | *QA form* |
| **Label verbalization** | Original label names, all lowercased except for 3 label names corresponding to proper nouns in the fine-grained task: "Unió Europea", "Parlament", "Govern". | | |

Table 4: First set of templates with their corresponding label verbalization, applicable to coarse-grained and fine-grained tasks.

| | | |
|---|---|---|
| **Templates** | 13 | És un tema {label}. |
| | 14 | Aquest text tracta un tema {label}. |
| | 15 | El tema és de caire {label}. |
| | 16 | El tema és d'àmbit {label}. |
| | 17 | L'article és de caire {label}. |
| **Label verbalization** | Label names converted into their adjective form and lowercased, as in the following mapping (original: verbalization):<br>- Cultura: cultural<br>- Política: polític<br>- Economia: econòmic<br>- Societat: social | |

Table 5: Second set of templates with their corresponding label verbalization, only applicable to coarse-grained categories.

| | | *coarse-grained task* | | | *fine-grained task* | | |
|---|---|---|---|---|---|---|---|
| **ratio of negative hypotheses** | **ckp. selection strategy** | **8-4** | **16-8** | **32-16** | **8-4** | **16-8** | **32-16** |
| 1 negative hip. per positive hip. | best ckp. according to the NLI task | **79.4 ± 4.0** | 82.5 ± 1.4 | 88.1 ± 1.2 | 41.9 ± 8.4 | 42.2 ± 18 | 38.9 ± 17.6 |
| | best ckp. according to the CLS task | 78.9 ± 3.7 | **83.8 ± 1.9** | 87.6 ± 2.2 | 46.2 ± 0.4 | 53.7 ± 4.4 | 57.9 ± 2.7 |
| all possible negative hip. per positive hip. | best ckp. according to the CLS task | 79.2 ± 3.3 | 82.4 ± 2.9 | **89.2 ± 0.6** | **60.2 ± 1.4** | **62.4 ± 1.3** | **63.2 ± 1.3** |

Table 6: Test set results for the coarse- and fine-grained tasks obtained with RoBERTa-ca-Teca in three few-shot setups (8-4, 16-8, 32-16) using three different decisions with respect to the ratio of negative hypotheses created for training and to the checkpoint selection strategy.

classification) on the development set was selected, rather than using the results from the NLI task. Secondly, during the generation of the NLI training data, for each entailment hypothesis (generated using the correct label), all possible negative hypotheses (one for each of the remaining labels) were also generated. To investigate the impact of these decisions, additional experiments were conducted using the RoBERTa-ca-Teca model as the base entailment model: in the 8-4, 16-8, and 32-16 few-shot setups, we converted the available data to the entailment format by creating one non-entailment hypothesis for each entailment one, and we kept the best checkpoint based on both the classification and the NLI task. These results were then compared to those obtained from the initial experimental setup.

The results of the experiments in the coarse- and fine-grained tasks are presented in Table 6. In the coarse-grained task, there is minimal fluctuation in the results across experiments within each training data regime, and the best-performing model among the three configurations changes at each step. In contrast, in the fine-grained task, the results significantly improve when the best checkpoint is selected based on the classification task performance (by 4.3, 11.5, and 19.0 points compared to the best checkpoint selected according to the NLI task performance). This impact becomes more pronounced as more data is available, and the model becomes increasingly unstable (as indicated by the high standard deviations obtained).

Furthermore, when the model is trained using all possible non-entailment hypotheses, which in this case implies 53 hypotheses for each example, the results further improve by an astounding 14.0 in the 8-4 setup, by 8.7 in the 16-8 setup, and by 5.3 points in the 32-16 setup. Overall, the results suggest that both choices become particularly important when the number of categories is large. In such cases, the results on the target task are more reliable than the NLI task, and the model appears to be able to benefit from the augmented number of examples in the dataset.