A weakly supervised textual entailment approach to zero-shot text classification

Marc Pàmies[1]* **Joan Llop**[1]*
Francesco Multari[2] Nicolau Duran-Silva[2] César Parra-Rojas[2]
Aitor Gonzalez-Agirre[1] Francesco Alessandro Massucci[2] Marta Villegas[1]

[1]Barcelona Supercomputing Center [2]SIRIS Academic

Abstract

Zero-shot text classification is a widely studied task that deals with a lack of annotated data. The most common approach is to reformulate it as a textual entailment problem, enabling classification into unseen classes. This work explores an effective approach that trains on a weakly supervised dataset generated from traditional classification data. We empirically study the relation between the performance of the entailment task, which is used as a proxy, and the target zero-shot text classification task. Our findings reveal that there is no linear correlation between both tasks, to the extent that it can be detrimental to lengthen the fine-tuning process even when the model is still learning, and propose a straightforward method to stop training on time. As a proof of concept, we introduce a domain-specific zero-shot text classifier that was trained on Microsoft Academic Graph data. The model, called SCIroShot, achieves state-of-the-art performance in the scientific domain and competitive results in other areas. Both the model and evaluation benchmark are publicly available on HuggingFace[1] and GitHub[2].

1 Introduction

Ever since the first BERT (Devlin et al., 2019) and GPT (Radford et al., 2018) models were introduced to the world, the Transformer (Vaswani et al., 2017) has become the dominant architecture in the Natural Language Processing (NLP) field. As a consequence, in the years that followed, the pretrain-then-finetune paradigm (Howard and Ruder, 2018) has been widely adopted to progressively push the state-of-the-art in a wide variety of downstream tasks and languages (Nozza et al., 2020).

Even though the current training regime is far from being environmentally friendly due to the computational cost of pre-training (Patterson et al., 2021; Strubell et al., 2019), transfer learning removes the need of having to train a new model from scratch for each application. However, the fine-tuning of models for every single task is expensive both in terms of time and money as it is always preceded by a labor-intensive data labelling process (Wang et al., 2021a). In order to overcome this issue, as well as the fact that real world data can be scarce in many scenarios, the field has started to shift towards techniques that require smaller amounts of labelled examples or even none at all (Wang et al., 2021b; Schick and Schütze, 2021; Radford et al., 2019; Brown et al., 2020).

In particular, in the scientific domain, the growing amount of publications in an ever-increasing number of fields makes the classification task very challenging for neural language models (Larsen and Von Ins, 2010; Bornmann et al., 2021). The impossibility to predict the emergence of new fields of study and the high cost associated with the creation of new datasets (which often requires domain experts) generate a need for systems that are capable of adapting to new situations. Furthermore, the complexity and technicality of scientific language makes general-domain models perform poorly in comparison to domain-specific ones (Lee et al., 2020; Cohan et al., 2020).

This work addresses these problems by training an entailment-based zero-shot classifier for scientific text. Instead of using a general domain dataset such as the popular XNLI (Conneau et al., 2018) or MNLI (Williams et al., 2018), a textual entailment dataset of scientific documents was built from scratch in a weakly supervised manner. By training a vanilla model on the entailment task, it is then able to classify documents into unseen classes with a high degree of success. As a by-product of this, the study of the relation between the training and target tasks led to intriguing questions about the strengths and limitations of the entailment approach to Zero-Shot Text Classification (ZSTC).

*Equal contribution.
[1]https://huggingface.co/BSC-LT/sciroshot
[2]https://github.com/bsc-langtech/sciroshot

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 286–296

2 Related Work

Zero-Shot Learning is a widely studied problem in Machine Learning that consists in completing a task for which no training examples were provided. Although the term became initially popular in computer vision (Lampert et al., 2009; Xian et al., 2018), it soon made the leap into NLP with an early paper (Chang et al., 2008) that presented a classifier capable of interpreting the Explicit Semantic Analysis (ESA) representations of documents and labels from a semantic point of view, using Wikipedia as a source of world knowledge.

This highlighted the importance of capturing the labels' semantics in their representations, unlike in conventional text classification where labels are mapped to meaningless indices. This became easier to capture with the arrival of word embeddings, and can be enhanced by simply replacing their name with a short description (Song and Roth, 2014).

In subsequent research, zero-shot tasks were occasionally tackled from another problem's perspective (Levy et al., 2017; Obamuyide and Vlachos, 2018), which implies training a model on a task for which annotated data is available and performing inference on a different one. Recognizing Textual Entailment (RTE) (Dagan et al., 2005) is arguably the most versatile task because of its generality, which is why it is commonly used to model other downstream tasks (Wang et al., 2021b).

The most popular approach at the time of writing is to reformulate ZSTC as a textual entailment problem as proposed by Yin et al. (2019), aiming to imitate the way humans would address this Natural Language Understanding task. The underlying idea is that a model trained on the entailment task should be able to perform classification of unseen classes by computing the entailment score between the input text, which acts as a premise, and candidate labels conveniently converted into hypotheses.

Later research showed that the Next Sentence Prediction (NSP) objective for sentence pair classification can also be used as a strong baseline for ZSTC, since competitive results where obtained using raw BERT models that were not fine-tuned on any Natural Language Inference (NLI) data (Ma et al., 2021). Even though entailment-based zero-shot text classifiers have shown to have certain limitations like a high instability or an excessive reliance on spurious lexical patterns, the current literature offers no better alternatives for a problem that is still far from being mastered by machines.

3 Creation of weakly supervised data

This section describes the methodology followed to transform a text classification dataset into entailment data that could potentially be used to fine-tune a domain-specific (or not) zero-shot text classifier. This approach takes advantage of the fact that there are plenty of publicly accessible labelled examples for classification while there is not so much available for entailment tasks, most likely due to the difficulty of producing this type of data.

In this work, a weakly supervised NLI dataset was constructed using Microsoft Academic Graph[3] (MAG) data as a starting point. To do so, all labels were converted to natural language sentences that would serve as hypotheses. Therefore, the generated training examples consist of pairs of sequences (premise and hypothesis) that are delimited by EOS tokens, where the first part of the text contains the abstract section from a scientific publication (premise) and the second part is an artificially generated sentence that somehow embeds the class label of the scientific text (hypothesis). Table 1 shows this idea in a simplified form.

Input Sequence	Label
<s>Text1</s></s>This example is X</s>	1
<s>Text1</s></s>This example is Y</s>	0
<s>Text2</s></s>This example is Z</s>	1

Table 1: Format of the training examples. The label is 1 if the premise entails the hypothesis and 0 otherwise.

Once the classification data has been turned into the entailment format, any model can be fine-tuned by predicting whether the premise of each input sequence entails the corresponding hypothesis or not. Note however that this is rather an adaptation of the RTE task, since the second sentence does not really "entail" the first in positive examples. It is basically like performing a text classification task where labels are not converted into numeric indices, ensuring that their semantic content is preserved.

Figure 1 illustrates how a ZSTC model would operate both during fine-tuning (left) and at inference time (right), providing a full picture of the methodology followed in this work. We basically adopt the approach introduced by Yin et al. (2019) but going one step further by fine-tuning on a weakly supervised domain-specific dataset instead of using already existing general-domain NLI data.

[3] https://academic.microsoft.com/

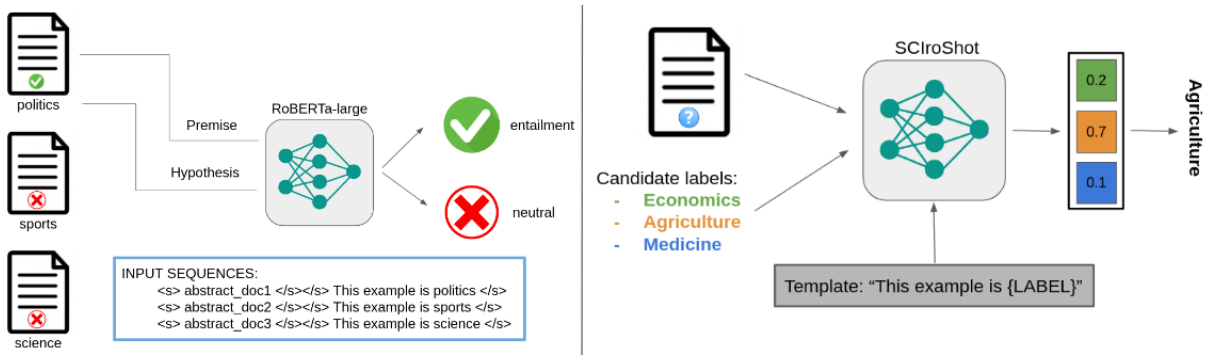

Figure 1: Overview of the entailment approach to ZSTC. On the left hand side it can be seen how the model is fine-tuned on a two-class textual entailment task, by providing the input text as a premise and the label name embedded in a natural language sentence that represents the hypothesis. On the right, the trained model classifies documents into unseen classes by computing the entailment score between the input text and each candidate label.

4 Data

4.1 Training Dataset

As mentioned in Section 3, our training dataset builds on top of scientific-domain annotated data from Microsoft Academic Graph (Sinha et al., 2015). This database consists of a heterogeneous graph with billions of records from both scientific publications and patents, in addition to metadata information such as the authors, institutions, journals, conferences and their citation relationships. The documents are organized in a hierarchical structure composed of hundreds of thousands of scientific concepts, creating a six-level hierarchy with a subsumption-based model (Shen et al., 2018), although the two top-most levels are manually curated to guarantee accuracy. As an example, the 0-level field of study (FoS) in the MAG taxonomy covers the following 19 scientific concepts: {*Art, Biology, Business, Chemistry, Computer Science, Economics, Engineering, Environmental Science, Geography, Geology, History, Materials Science, Mathematics, Medicine, Philosophy, Physics, Political Science, Psychology, Sociology*}.

Dataset	Labels	Examples
train	240	2,104,493
dev_{seen}	240	233,833
$\text{dev}_{\text{unseen}}$	52	5,200
Total	292	2,343,526

Table 2: Number of labels and examples in the train and development sets of our dataset. Note that the 240 labels from the train and dev_{seen} set are the same, while the 52 labels from the $\text{dev}_{\text{unseen}}$ set were purposely kept aside for *label-fully-unseen* setups.

Due to the descriptive broadness of the 0-level MAG taxonomy, we have created our training corpus focusing exclusively the 1-level MAG taxonomy, which is composed of 292 FoS classes. The higher granularity of this level provides more descriptive information in the form of narrower scientific concepts, such as "*Computational biology*", "*Transport engineering*" or "*Civil engineering*".

Using the relationship between scientific texts and their matching concepts in the 1-level MAG taxonomy we are able to generate the premise-hypothesis pairs corresponding to the entailment label. Conversely, we generate the pairs for the neutral label by removing the actual relationship between the texts and their scientific concepts and creating a virtual relationship with those to which they are not matched (see Table 3).

Input Sequence	Label
One plus one is two.	Maths
Cancer is a disease.	Health
One plus one is two. This text is Maths.	entails
One plus one is two. This text is Health.	neutral
Cancer is a disease. This text is Maths.	neutral
Cancer is a disease. This text is Health.	entails

Table 3: Toy example of the initial classification dataset (top) and the adapted entailment dataset (bottom).

For each of the 292 classes, a random sample of scientific articles with a publication year between 2000 and 2021 was extracted with their respective titles and abstracts in English. We have collected a maximum of 5k and 10k positive and neutral textual entailment samples, respectively, for each of the possible 1-level FoS classes. In total, SCIroshot has been fine-tuned using 919k documents with a

total of 465M words. In order to perform experiments in *label-fully-unseen* setups, the examples associated to 52 labels were kept aside. For computational reasons, the size of the development set was reduced considerably to the point that it became a fully balanced set of 100 examples per class. The number of labels and examples of each set are summarized in Table 2.

4.2 Evaluation Datasets

We evaluate the performance of our models on a collection of disciplinary-labeled textual datasets. For the in-domain evaluation, we gathered cross-disciplinary and domain-specific datasets of scientific publications. For the out-of-domain case, we use the three datasets from the benchmark provided by Yin et al. (2019), which study 3 aspects of ZSTC: topic categorization (Yahoo! Answers), emotion detection (UnifyEmotion) and situation frame detection (Situation Typing). Table 4 provides an overview of the number of examples and labels for each dataset.

Dataset	Labels	Examples
arXiv	11	3,838
SciDocs-MeSH	11	16,433
SciDocs-MAG	19	17,501
Konstanz	24	10,000
Elsevier	26	14,738
PubMed	109	5,000
Yahoo! Answers	10	60,000
UnifyEmotion	10	15,689
Situation Typing	12	3,311

Table 4: Statistics of each dataset from the scientific-domain (top) and general-domain (bottom) benchmarks.

4.2.1 Scientific-domain datasets

arXiv (He et al., 2019). 11-label dataset of papers from the arXiv repository. The labels are a set of sub-categories within the branches of Computer Science and Mathematics. While the original dataset contains the full publication texts, we only gathered titles and abstracts from the 3,838 publications for which a DOI was available in the API[4].

SciDocs-MeSH (Cohan et al., 2020). Over 16k papers from the medical domain. Each paper is assigned one of 11 high-level disease classes derived from the MeSH vocabulary (Lipscomb, 2000).

SciDocs-MAG (Cohan et al., 2020). More than 17k cross-disciplinary publications labelled using the 0-level MAG taxonomy (Wang et al., 2020). None of the labels are included in our training data.

Konstanz[5]. 10k journal articles produced by researchers at the University of Konstanz, extracted from the Konstanz Online Publication System (KOPS). Publications from this open-access repository are manually labelled by the research staff with a category taken from the DDC taxonomy (Dewey, 1876), which unfolds into more than 30 classes describing different scientific domains. We only consider English journal articles labelled within a reduced set of 24 categories.

Elsevier (Kershaw and Koeling, 2020). Cross-disciplinary corpus of 14.7k open access articles from Elsevier's journals. The document labels are given by their ASJC Subject Classification scheme, which links publication venues (and, transitively, each single publication) to 27 scientific subject domains. We removed all publications labelled with more than 1 subject area. Publications annotated with the non-informative "*Multidisciplinary*" label were removed as well.

PubMed[6]. We collected 5k publications labelled with a manually-selected subset of 109 MeSH terms within the *Disciplines and Occupations* and *Technology, Industry, and Agriculture* branches of the MeSH taxonomy. The chosen categories are general-domain and well-known concepts out of specific medical terminology, as most MeSH terms.

4.2.2 Out-of-domain datasets

Yahoo! Answers (Zhang et al., 2015). Topic categorisation dataset with questions and their corresponding best answer in *Yahoo! Answers*. We only use the test set, which consists of 60k examples that belong to exactly one of the 10 largest main categories in the website.

UnifyEmotion (Oberländer and Klinger, 2018). Emotion detection dataset with texts from a variety of sources (tweets, emotional events, tales, and artificial sentences) classified into 9 emotions and "none" when no emotion fits the case. We use the modified version from Yin et al. (2019), which removes all multi-label instances.

Situation Typing (Mayhew et al., 2019). Multi-label event-type classification dataset of 11 classes, designed for low-resource situation detection.

[4] https://arxiv.org/help/api/

[5] https://kops.uni-konstanz.de/

[6] https://pubmed.ncbi.nlm.nih.gov/

5 Experiments

5.1 Design choice

The newly-created scientific dataset from Section 4.1 was used to fine-tune a 355M parameters RoBERTa (Liu et al., 2020) and a 400M parameters BART (Lewis et al., 2020) models, in an attempt to determine which architecture (i.e. encoder or encoder-decoder) is best suited for the task at hand. As already noted in section 4.1, 52 labels from the training data were kept apart so that they could be used as a development set of fully-unseen classes. For a given input text, the entailment score with each candidate label has to be computed by the model. The final prediction will be the highest scoring class in a single-label classification setup, or the N classes above a certain threshold in a multi-label scenario. Table 5 shows the accuracy score of the last checkpoints evaluated in RTE's dev_{seen} set as well as the ZSTC $\text{dev}_{\text{unseen}}$ set.

Model	RTE	ZSTC
RoBERTa-large	98.00	48.78
BART-large	98.30	45.69

Table 5: Accuracy scores of our two models in the Recognizing Textual Entailment (RTE) and Zero-Shot Text Classification (ZSTC) tasks.

Even though both models achieved a similar accuracy in the entailment task, we could not help noticing that the best performing model on RTE (even if only by a small margin) was doing worse on ZSTC by three full points. This raised a concern as to whether our fine-tuning task, which at the end is no more than an adaptation of the real RTE, was positively correlated with the target task of ZSTC or not.

5.2 Correlation between the RTE-ZSTC tasks

In order to verify the correlation between both tasks, we conducted an exhaustive evaluation of all checkpoints using the 52-labels $\text{dev}_{\text{unseen}}$ set.

As it can be observed in Figure 2, after a certain point the performance in the ZSTC task begins to gradually worsen while on RTE it is still getting better at a slow but steady pace. This means that somehow, as the training progresses, the model forgets the meaning of the unseen labels, with the exception of the initial checkpoints where an exponential growth is experienced simultaneously in both tasks. We can observe a peak ZSTC performance when the model has an evaluation RTE score

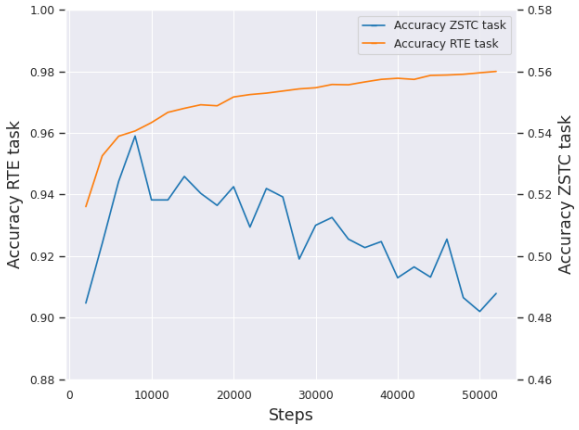

Figure 2: Accuracy scores obtained by the RoBERTa-large checkpoints. Each y-axis represents a different range of accuracies for better visualization.

of roughly 0.96. Despite the high variability, it is clear that from that point onwards the zero-shot capacities of the model decrease. To mitigate this effect we propose an early stopping technique.

5.3 Early stopping

We concluded that, at training time, the validation of the model should be done on the target task rather than the training task. This way the training process can be interrupted as soon as the model stops improving on ZSTC, something that we expect to happen at an earlier stage. We propose to evaluate each checkpoint on the subset of 52 unseen classes described in Section 4.1 and use early stopping with a patience of 10.

Model	RTE	ZSTC
$\text{RoBERTa-large}_{\text{last}}$	**98.00**	48.78
$\text{RoBERTa-large}_{\text{selected}}$	96.07	**53.90**
$\text{BART-large}_{\text{last}}$	**98.30**	45.69
$\text{BART-large}_{\text{selected}}$	96.59	**52.76**

Table 6: Accuracies in the RTE and ZSTC tasks. The *last* subscript indicates that it was the last checkpoint stored during fine-tuning, while the *selected* subscipt refers to the checkpoint selected with early stopping.

In Table 6 it is clear that the early stopping technique improves the results in the ZSTC evaluation task, since both architectures obtain a substantial boost when early stopping is applied. Overall, the RoBERTa-large model performs better than BART-large, what we hypothesize that might be caused by a loss of generality from models that achieve higher scores in the RTE task.

Model	arXiv	SciDocs-MesH	SciDocs-MAG	Konstanz	Elsevier	PubMed
fb/bart-large-mnli	33.28	**66.18**	51.77	54.62	28.41	**31.59**
bart-large-rte$_{\text{selected}}$	36.71	56.57	63.98	63.57	48.48	27.46
bart-large-rte$_{\text{last}}$	28.58	44.21	57.75	62.25	42.51	21.72
SCIroShot$_{\text{selected}}$	**42.22**	59.34	**69.86**	**66.07**	**54.42**	27.93
SCIroShot$_{\text{last}}$	35.44	52.27	65.27	60.74	50.92	22.85

Table 7: Label wise weighted F1 score of different models in our scientific benchmark. For simplicity, the hypothesis template was set to *"This example is {}."* in all cases.

6 Results

This section assesses the performance of our zero-shot classifiers, which were trained on a weakly supervised entailment dataset of scientific text. In an effort to obtain a more complete picture, we perform both an in-domain and out-of-domain study.

6.1 Scientific domain

For an in-domain evaluation, the in-house generated scientific benchmark from Section 4.2.1 was used. We compare our SCIroShot with the strong baseline set by Facebook's *bart-large-mnli* model[7]. This NLI-based model is the most downloaded zero-shot classifier in the HuggingFace Hub (over 1M monthly downloads) and the one used by default in the *zero-shot-classification* pipeline from their transformers library (Wolf et al., 2020). For an apples-to-apples comparison, we also consider a BART-large model that was trained following the same methodology employed for SciroShot, as it has the same architecture as the Facebook model.

The results presented in Table 7 prove the importance of domain-specific training, something that has been repeatedly seen in traditional text classification. The models trained on MAG data obtain the best results in four out of six datasets, by large margins in all cases, and interestingly enough the Facebook model only wins in the two medical datasets: PubMed and SciDocs-MeSH. The numbers also support our theory that too much training can be detrimental, as the models selected with early stopping score higher than their "last" counterpart in all cases.

6.2 General domain

For the out-of-domain study, we include ourselves in the benchmark proposed by Yin et al. (2019). This amounts to a total of three datasets that cover a variety of topics, with classes ranging from news article topics to human emotions.

(Yin et al., 2019)	Topic	this text is about {}
	Emotion	this text expresses {}
	Situation	The people there need {}
(Ma et al., 2021)	Topic	It is related with {} .
	Emotion	This person feels {} .
	Situation	The people there need {} .
SCIroShot (ours)	Topic	This example is {}
	Emotion	
	Situation	

Table 8: Hypothesis templates used for each dataset. The {braces} indicate the label name location.

The results reported in Table 9 show that our SCIroShot model is competitive in other domains as well. Actually, it is quite impressive that it was able to outperform the rest in two tasks and obtain the second highest score in the third one.

It is important to note that the Situation dataset has the added difficulty of being multi-label, meaning that a piece of text can be linked to an arbitrary number of labels. In a single-label setting, softmax is applied over all the labels logits (so that they sum up to one) and the highest scoring class is chosen as the final prediction. On the other hand, when performing multi-label classification, the softmax function is applied to each label separately (so they are independent variables that do not add up to one) and all labels with a probability above a certain threshold are selected. We did not tune such threshold, so the model's predictions included all labels with a score higher than 0.5.

7 Analysis

7.1 RTE vs ZSTC in the scientific benchmark

This section analyses the relation between the training task and the ZSTC task to assess the effectiveness of the early stopping technique presented in Section 5.3. We evaluate every checkpoint in our ZSTC Scientific Benchmark in the same way that several models were evaluated in Section 5.2.

[7] https://huggingface.co/facebook/bart-large-mnli

Model	Topic	Emotion	Situation
RTE (Yin et al., 2019)	43.8	12.6	**37.2**
FEVER (Yin et al., 2019)	40.1	24.7	21.0
MNLI (Yin et al., 2019)	37.9	22.3	15.4
NSP (Ma et al., 2021)	50.6	16.5	25.8
NSP (Reverse) (Ma et al., 2021)	53.1	16.1	19.9
$\text{SCIroShot}_{\text{last}}$	51.51	22.63	23.70
$\text{SCIroShot}_{\text{selected}}$	**59.08**	**24.94**	27.42

Table 9: Results obtained in the general benchmark. Accuracy is used for Topic classification and label-wise weighted F1 for the rest. For simplicity, the hypothesis template was set to "This example is ." in all cases.

Figures 3 and 4 corroborate our hypothesis that the performance in the ZSTC task does not necessarily improve hand-by-hand with the performance on RTE. Taking a closer look to the plots we can appreciate how the ZSTC accuracy drops after the model reaches an accuracy of 96% on RTE. The checkpoints selected using the early stopping technique based on the ZSTC task achieve 96.07% RTE accuracy in the case of RoBERTa-large and 96.60% for BART-large. We argue that, although the selected checkpoints are not the best possible option for all the datasets in the benchmark, our technique has stopped training before the overall ZSTC performance diminishes, and it has done so using a subset of unseen labels from the training dataset. This implies obvious savings in time and computation that would otherwise have been wasted for no good reason.

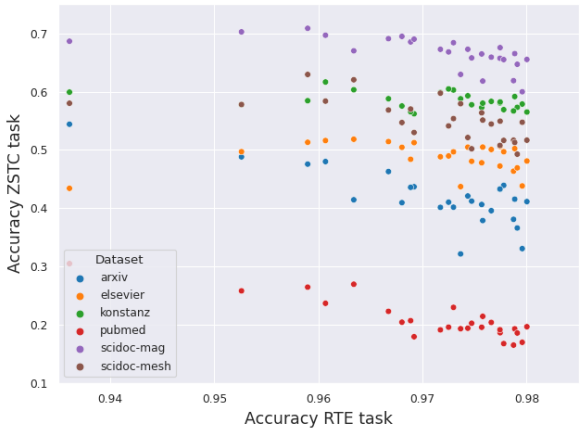

Figure 3: Evaluation of all RoBERTa-large checkpoints in the training (RTE) and testing (ZSTC) tasks.

7.2 Robustness to hypothesis templates

This section is an attempt to measure the importance of the hypothesis template and its impact on the final performance of a zero-shot model. With

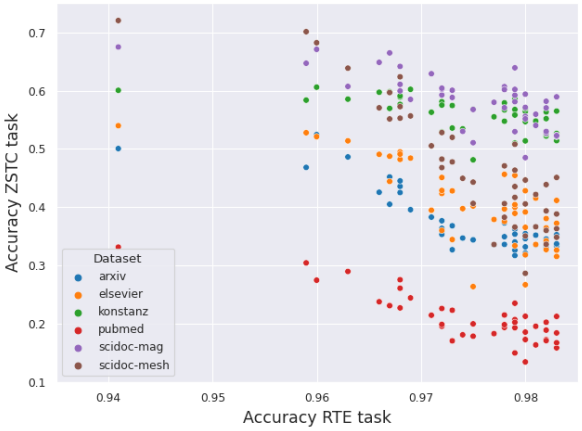

Figure 4: Evaluation of all BART-large checkpoints in the training (RTE) and testing (ZSTC) tasks.

this goal in mind, we evaluate our SCIroShot and Facebook's *bart-large-mnli* with two hypothesis templates that are virtually the same: "*This example is {LABEL}*" and "*This example is {LABEL}.*". Note that their semantic content is exactly the same, being the only difference that the first template does not contain a punctuation mark at the end.

Figure 5 shows that SCIroShot is quite robust against changes in the hypothesis template. On the other hand, as it can be seen in Figure 6, *bart-large-mnli* can experience severe performance drops caused by an apparently insignificant change in the template. We hypothesize that this might happen because the model was trained with high-quality NLI data where dots were always present at the end of the hypothesis, and thus it is not used to the absence of these type of anchor tokens. It can also be inferred that our training task is quite robust to different hypothesis templates.

We would like to point out that the high sensitivity of the Facebook model was accidentally detected during our experiments. We noticed that using our default template the model obtained sur-

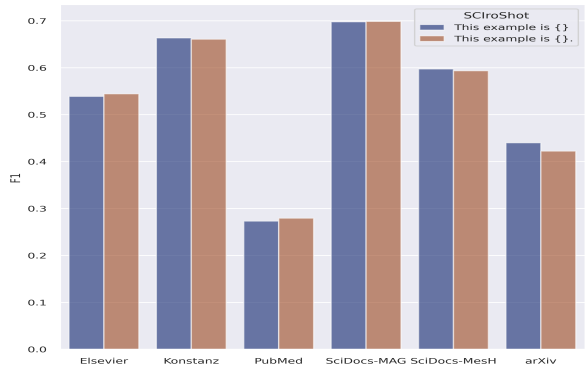

Figure 5: SCIroShot performances when using slightly different hypothesis templates.

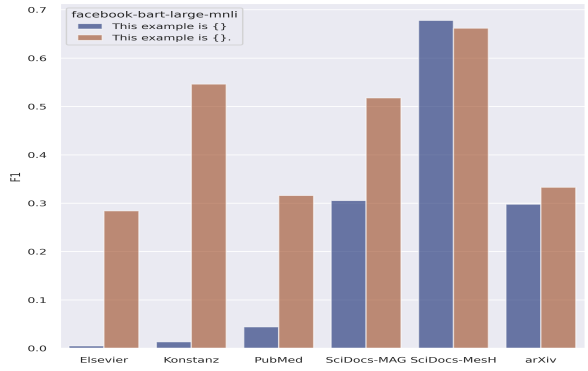

Figure 6: Facebook's bart-large-mnli performances when using slightly different hypothesis templates.

prisingly low results in several tests, something highly unexpected considering its great capabilities. In pursue of intelectual honesty, we decided to report the results obtained with the template that best suited the model's interests. The exact same template was given to our models so that there was one less thing to take into account when comparing performances, even though it might not be the one providing the best results.

Conclusion

The benefits offered by a zero-shot text classifier are four-pronged: (i) it removes the need for time-consuming annotation processes carried out by domain experts, (ii) reduces the computational cost of having to fine-tune a model for each application, (iii) allows the classification of documents in real-world scenarios where there is a scarcity of data, and (iv) is able to handle new classes that might not even have existed at the time of training.

This work proposes the usage of readily available classification datasets for effortless generation of entailment data, which can be used to train ZSTC models. By not using conventional RTE datasets of generic nature, the resulting model exhibits superior performance in the domain for which it has been trained. We show that this is the case for the scientific domain, but the idea could certainly be extrapolated to other fields (e.g. a model trained on news articles should excel at topic classification of this kind of documents). As a proof of concept, we present a scientific-domain zero-shot text classifier that achieves state-of-the-art performance in the scientific domain and competitive results in other areas.

Furthermore, our experiments and analysis suggest that entailment-based classifiers are no panacea: they are very sensitive to the input sequences and do not exhibit the linear correlation that one would expect between the performance on the training (RTE) and testing (ZSTC) tasks. We have empirically proven that the model can become worse at ZSTC as it improves in RTE, which is counter-intuitive and goes against the idea of entailment being a unified method to model other downstream tasks. Our analysis also show that our technique does not suffer from the instability observed in models trained with conventional RTE datasets, which can occasionally experience severe performance drops with minor changes such as removing a punctuation mark from the hypothesis template.

Future Work

In future work, we will further investigate the correlation between different tasks as this could only be the tip of the iceberg. It might also be interesting to increase the difficulty of the fine-tuning task by working with thousands of fine-grained labels from deeper levels of the MAG taxonomy, aiming to delay the point at which the model performance starts worsening in the ZSTC task.

Our findings also motivate an interesting research direction that we would like to explore: employing novel prompt tuning techniques to find the ideal hypothesis template. Having seen that zero-shot text classifiers experience dramatic performance drops caused by apparently insignificant modifications in the hypothesis text, it is clear to us that there is a need to find a suitable text in an automatic manner. We consider this to be a major pitfall of zero-shot models, and thus we are willing to study the feasibility of applying prompt tuning to this particular case as future work.

Limitations

Our main limitations had to do with time and computational constraints. Given the low efficiency of entailment-based classifiers, which need to execute a forward pass per class, the time required to traverse certain datasets was simply too high. Specially when the number of candidate labels is large, because every sequence-label pair has to be fed through the model to compute the logits of all possible combinations. This low ability to scale is certainly a drawback with respect to traditional classifiers, and the reason that forced us to discard datasets with over 500 labels as well as a few experiments that we intend to leave for future work.

Ethics Statement

We believe that this work meets the ACL Code of Ethics as it provides an already trained zero-shot text classifier that can be used in an endless number of situations that would otherwise require a task-specific fine-tuning. Moreover, one of the main findings of this work is that entailment-based text classifiers should not be over-trained as it negatively affects their final performance. This should encourage fellow NLP practitioners to shorten the training time of their ZSTCs thus minimizing their carbon footprint, which is in line with the idea of moving towards more sustainable language models.

Acknowledgements

This work has received funding from two projects within the European Union's Horizon 2020 research and innovation programme.

The work carried out by Barcelona Supercomputing Center was funded by the IntelComp project, under grant agreement number 101004870.

The work carried out by SIRIS Academic was partially funded by the INODE project, under grant agreement number 863410.

SIRIS Academic also acknowledges the support provided by the PRACE initiative for awarding access to the MareNostrum supercomputer.

References

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pages 830–835.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1*, pages 4171–4186. Association for Computational Linguistics.

Melvil Dewey. 1876. *A classification and subject index, for cataloguing and arranging the books and pamphlets of a library*. Brick row book shop, Incorporated.

Jun He, Liqun Wang, Liu Liu, Jiao Feng, and Hao Wu. 2019. Long document classification from local word glimpses via recurrent attention learning. *IEEE Access*, 7:40707–40718.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Daniel Kershaw and Rob Koeling. 2020. Elsevier oa cc-by corpus.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE.

Peder Larsen and Markus Von Ins. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. 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. Association for Computational Linguistics.

Carolyn E. Lipscomb. 2000. Medical subject headings. *Bulletin of Medical Library Association*, 88 3:265–6.

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.

Tingting Ma, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. Issues with entailment-based zero-shot text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 786–796.

Stephen Mayhew, Tatiana Tsygankova, Francesca Marini, Zihan Wang, Jane Lee, Xiaodong Yu, Xingyu Fu, Weijia Shi, Zian Zhao, Wenpeng Yin, Karthikeyan K, Jamaal Hay, Michael Shur, Jennifer Sheffield, and Dan Roth. 2019. University of pennsylvania lorehlt 2018 submission. Technical report, University of pennsylvania lorehlt 2019 submission.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78.

Laura Ana Maria Oberländer 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.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguía, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269. Association for Computational Linguistics.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*, pages 87–92, Melbourne, Australia. Association for Computational Linguistics.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021a. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021b. Entailment as few-shot learner.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. 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 *Conference on Empirical Methods in Natural Language Processing*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Fine-tuning hyperparameters

Hyper-parameter	Value
Learning Rate	8e-6
Learning Rate Decay	Linear
Weight Decay	0.0
Warmup Steps	0
Batch Size	256
Max. Training Epochs	10
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Gradient Clipping	1.0

Table 10: Hyper-parameter values.

B Computing infrastructure

The fine-tuning of each model took around 2 days on 16 HPC nodes[8] equipped with an AMD EPYC 7742 (@ 2.250GHz) processor with 128 threads and 2 AMD MI50 GPUs each.

[8] https://www.bsc.es/innovation-and-services/technical-information-cte-amd

C Proportion of *Entailment* and *Neutral* samples in the training data

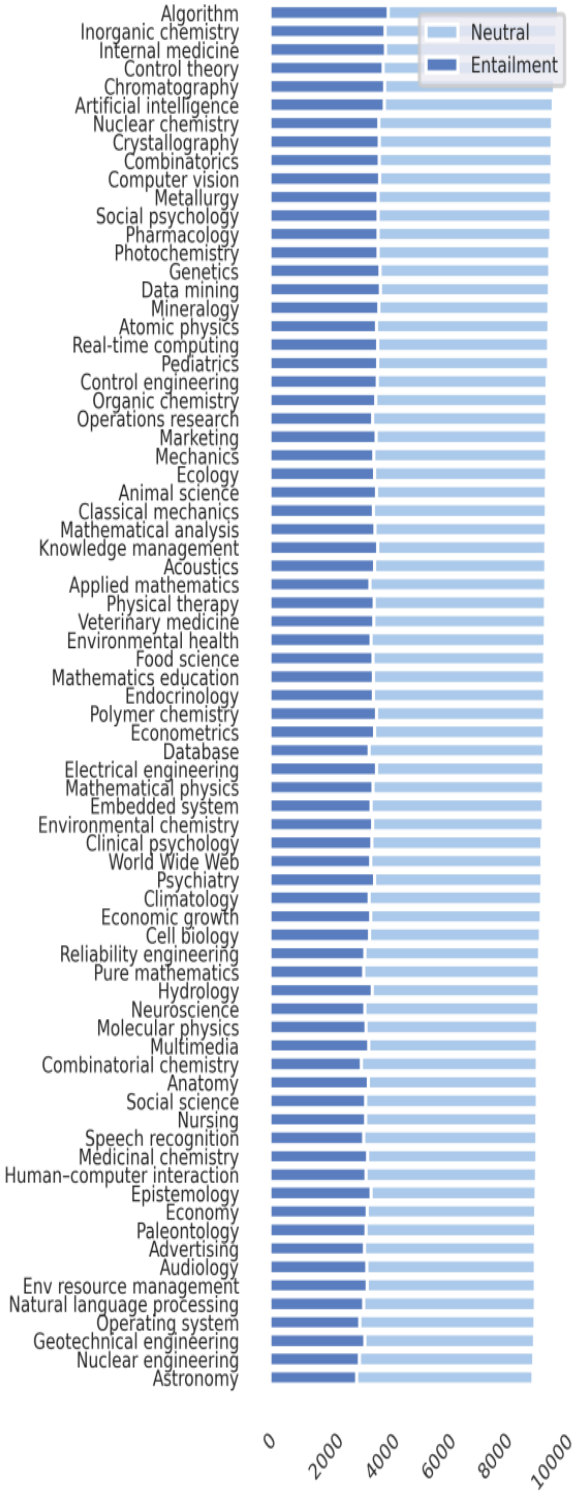

Figure 7: Proportion of entailment and neutral samples in the training data. For space limitations it is not possible to display all 292 labels, so the bar plot has been purposely limited to the top 75 classes.