Are Large Language Models Better than Small Language Models?: A kind approach

David Cabestany and **Clara Adsuar**
dcabestany001@ikasle.ehu.eus
cadsuar001@ikasle.ehu.eus

Abstract

In this work we made a comparison of the results obtained in zero and few-shot BERT model with the results obtained in same characteristics DistilBERT pre-trained with Masked Language Modeling, both making a sentimental classification with the same dataset modified with prompting in each instance. The expectation is that BERT shows a better behaviour than DistilBERT, but we aim to know how much better is the evaluation score and make a brief approach to both behaviours.

Code and Data allocated together in GitHub for an easier access.

1 Introduction

Language Models (LMs) have shown an exponential growth in size and popularity during the last decades. This size growth has been measured by the number of parameters and the size of the training data used. The popularity impact of LMs in Deep Learning and NLP field created a competition that in the beginning was about achieving better performances, but it seemed to resume on the idea "the more quantity, the more quality".

Bender et al. (2021) exposed some of the risks of having large LMs. Some learning patterns apprehended by LMs are highly influenced by biases and harmful attitudes, and this leads to the problem of collecting uncurated data. The lack of attention in gathering training data affects in a bidirectional way to the model, which has been encoded by hegemonic views, and marginalized population, which happen to be discrimination targets. Furthermore, the boost of environmental and financial cost could be managed through size control mechanism for LMs.

The creation of diverse and smaller versions of large Language Models opened the door to a progressive production of "greener" Language Models (in terms of parameter count). Schick and Schütze (2020b) illustrated how ALBERT (as an underlying LM) combined with PET/iPET, in comparison to GPT-3 outperformed the latter having three times fewer parameters (ALBERT and PET/iPET have 223M, meanwhile GPT-3 has 175B). Pattern-exploiting training (PET) was proposed by Schick and Schütze (2020a) as "an alternative to priming [...] which combines the idea of reformulating tasks as cloze questions with regular gradient-based finetuning". Even smaller models could show success as Sanh et al. (2019) shown in their experiment. Their proposal was concentrated in "using a method to pre-train a smaller general-purpose language representation model – DistilBERT – which can be fine-tuned with good-performances on a wide range of tasks", and it revealed favourable results.

The models proved to reach similar performances, and also happened to be lighter and faster at inference time, and lower in terms of computational training cost. Consequently, DistilBERT was a real training success as "a general-purpose Language Model with distillation knowledge, and also analysed various component with ablation study" (Sanh et al., 2019). Ablation study is a technique to test or observe the performance of an AI system when one or more components are removed and or damaged.

2 Masked Language Modeling

Masked Language Modeling (MLM) pre-training is broadly exploited in Natural Language Processing (Yamaguchi et al., 2021; Devlin et al., 2018; Lan et al., 2019; Liu et al., 2019; Wang et al., 2019). This self-supervised pre-training method is mostly used for learning text representations. The result of MLM pre-training is that the model predicts random samples of specific tokens which has been supplanted by a [MASK] placeholder. This replacement occurs along the whole vocabulary batch in a multi-class setting.

MLM has a wide variety of extensions, for in-

stance masking a contiguous segment (instead of independent token treatment) (Song et al., 2019; Sun et al., 2020; Joshi et al., 2020), masking attention weights (Yang et al., 2019), addressing a binary classification task (Clark et al., 2020), predicting if two sentences are in correct order or reversed (Lan et al., 2019). In this work, MLM is applied to predict the label of the text representation taken by our database. This prediction could only have two values, either positive or negative, which is displayed in the manually added prompt: "This review is [MASK]". Prompt-based learning (Liu et al., 2021) is based on language models and it is used to perform predictions tasks by modelling the probability of text.

The application of this learning is by modifying the input text by using a template (a string of characters) with one or more empty slots. In the sample displayed in Figure 1, the model language fills the missing information according to the probabilistic calculations and outputs a final string with all the wanted information.

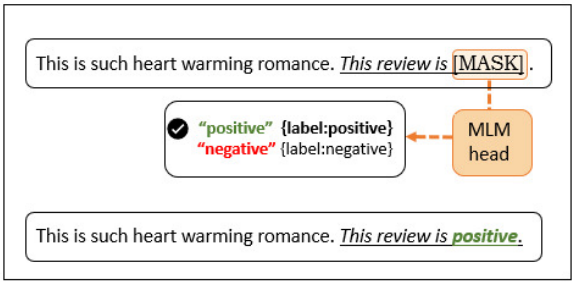

Figure 1: Prompt-based fine-tuning example of a correct label classification

3 BERT

Since 2017, Transformers have attracted many researchers and companies due to their great modelling of long-range dependencies performance (Vaswani et al., 2017). BERT is a transformer-based model for language representation, indeed its name stands for Bidirectional Encoder Representations from Transformers (Devlin et al., 2018).

One of the most outstanding features about BERT is taking unlabelled text for pre-training bidirectional representations, considering both left and right context in each one of the layers. The fine-tuning of this model can be done adding one layer which establishes the chosen NLP task such us Natural Language Inference, instance classification, or Question-Answering. Furthermore, BERT challenged the unidirectionality of previous models by the implementation of masked language model pre-training objective. In this way, BERT's model architecture is described as a multi-layer bidirectional Transformer encoder, and for this work we have selected $\text{BERT}_{\text{Base}}$ model.

The type chosen for developing this project consists in 12 layers of transformers block with a hidden size of 768 and 12 self-attention heads. Moreover, $\text{BERT}_{\text{Base}}$ includes 110M of trainable parameters.

4 DistilBERT

In less than a decade we have seen the spreading trending in NLP to creating bigger pre-trained language models. On the other hand, it has created simultaneously a tendency to corroborate if small pre-trained language models are also suitable for the success of natural language processing tasks. In this section, we introduce the technique of knowledge distillation (Buciluǎ et al., 2006; Hinton et al., 2015) which compresses the behaviour of a large model and transmits it to a compact or small model. DistilBERT – also considered the student, being BERT the teacher – was created through the application of this technique to BERT, and the distilled version shown similar performances on many downstream tasks (Sanh et al., 2019). In addition, DistilBERT kept the flexibility of its teacher larger model (Rogers et al., 2020), but with the advantage of running on the edge (for instance, on mobile devices).

BERT and DistilBERT share the same general architecture, but the token-type embeddings and the pooler were eliminated, and also reduced by half the number of layers. As a consequence, DistilBERT's architecture is composed by 6 hidden layers of transformers with a hidden size of 768 and 12 self-attention heads for each attention layer. The number of parameters in the version we have chosen of DistilBERT is 67M.

5 Learning Methods

Nowadays, our current models achieve remarkable results in tackling tasks with big amounts of labelled data. However, we cannot always find these labelled datasets – they require resources like time, human effort and financial support – and when models are not fed with these large labelled datasets, the performance success could descend. Therefore, in this section we are going to introduce the few

and zero-shot learning.

The basis is training a model with substantial amount of data samples on various categories. Then, in testing time, "the model is provided with novel categories" (Bendre et al., 2020), usually the total number of categories is until five and each categories contains few data samples.

The ultimate goal is to generalize or learn the learning process, and simultaneously identifying the best hyperparameters and model weights for the model to adapt and avoid overfitting. This technique is called transfer learning and its objective is obtaining knowledge or experience from a set of tasks and relocate it to a task in the similar domain (Pan and Yang, 2010). Hence, when we apply transfer learning, the classification layers are trained for the new tasks, but previous layers keep the previous weights (obtained by pre-training with other tasks) (Yosinski et al., 2014).

Few and zero-shot learning are three techniques which core is equal: the creation of a deep learning model that learns how to generalize with a low (or null) quantity of examples with iterative training based on prior knowledge or experience (Bendre et al., 2020).

The difference among these learning techniques aforementioned is the different number of examples we are given to a Language Model for learning a new classification task. Consequently, in few-shot learning the model is given two to five samples per each class; whereas, in zero-shot the number of samples is null.

6 Data

The dataset chosen for sentiment analysis in text classification is Amazon product data created by Julian McAuley (He and McAuley, 2016). This dataset contains product reviews and metadata from Amazon, the total number of the reviews collection ascends to 142 million divided into 24 categories. The data collected in this dataset is huge; therefore, we chose one category: Movies and TV. Regarding the size of the dataset, we chose the 5-core subset which have been reduced to extract the users and products that have at least 5 reviews. The absolute number of reviews in this fold is 1,697,533; at first, the percentage of extracted data from the original dataset was 11.7% (200,000 reviews). After some experiments, we decided to reduce this number by its 5%, being the final number of reviews selected: 10,000.

	overall	verified	reviewTime	reviewerID	asin
0	3.0	True	04 6, 2017	A2P2EJ6SG8PAK8	6302256615
1	4.0	True	01 24, 2016	A1064205QOBO19	6302260973
2	5.0	True	11 16, 2015	A5E3H5F6DAJ5F	6302260973
3	5.0	True	08 19, 2015	A34EY7MQR6QFRA	6302260973
4	4.0	True	02 2, 2009	A3CUFZAL76NM5P	6302168465
...	...	...	...	...	...
199995	5.0	False	06 10, 2010	A2UHIQNHUD3XFR	6304432402
199996	5.0	False	03 1, 2010	A25ZZMC5VB6MC8	6304432402
199997	1.0	False	08 1, 2009	A9XKE4OE48BNK	6304432402
199998	5.0	False	06 15, 2009	A178LLVAZ2B97E	6304432402
199999	5.0	True	09 10, 2008	A6G390ZQB8IF0	6304432402

\\

style	reviewerName	reviewText
{'Format:': ' Blu-ray'}	Craig Barker	So great one liners in this movie, but no...
{'Format:': ' DVD'}	Amazon Customer	Very satisfied with the service, VERY dis...
{'Format:': ' Amazon Video'}	Daniel J. Gainsboro	Great movie! Fun for everyone. Well acted...
{'Format:': ' DVD'}	A Time Traveler	"The Player" seems to be considered the g...
{'Format:': ' Blu-ray'}	Torrlisand67	real sharp beautiful detail. great extras...
...	...	...
{'Format:': ' DVD'}	Jes	Here we see the Daleks come looking for t...
{'Format:': ' DVD'}	Jacob	In Doctor Who episdoe entitled "Destiny o...
{'Format:': ' DVD'}	Byron	Another horrible installment from the hor...
{'Format:': ' DVD'}	Jorge Dejesus Sanchez	Great First adventure of the very cute La...
{'Format:': ' DVD'}	James V. Blecha	what can I say, the more Doctor Who that ...

\\

summary	unixReviewTime	vote	image
Hmmm.	1491436800	NaN	NaN
Good service, bad movie	1453593600	NaN	NaN
Mind expanding movie	1447632000	NaN	NaN
Probably the Best-Disguised Film About th...	1439942400	4	NaN
classic sci-fi 1950's	1233532800	NaN	NaN
...	...	...	...
Key to the 'New' Dr. Who	1276128000	NaN	NaN
The Doctor travels takes him to a meeting...	1267401600	NaN	NaN
Bad story from a bad season,	1249084800	5	NaN
Great First Lala Ward (Romana II) advent...	1245024000	NaN	NaN
awsome	1221004800	NaN	NaN

Figure 2: Amazon Movies and TV Shows Dataset

The structure of the product reviews in Movies and TV (shown in Figure 2) is the division of the data in 10 sections: overall (rating of the product), verified (whether the account is verified or not), reviewTime (publication time of the review), reviewerID (ID or nickname of the reviewer), asin (ID of the product), style (in this subset, the format of the movie or TV-series), reviewerName (name of the reviewer), reviewText (the actual information that we analyse, the opinion of the reviewer

about the product), summary (summary of the review) and unixReviewTime (publication time of the review, unix time).

For the classification, the only section that we take into account is the reviewText, which is the one we analyse in order to extract the sentimental data of the review. In the evaluation period, the subcategory overall has been simplified, if the rating is between 4 and 5, the label is "positive"; whereas, if the rating is from 1 till 3, the review will be labelled as "negative". Therefore, overall and reviewText are the categories we use in the evaluation process.

7 Results

In Table 1 and 2, we can see the confusion matrices for few-shot classification for DistilBERT and $\text{BERT}_{\text{Base}}$. In Table 1, we observe that 3998 samples have been classified as positive when the real label was negative, meanwhile DistilBERT in the case of negatives samples only 68 samples have been classified as positive. On the other hand, in Table 2 the number of samples classified by $\text{BERT}_{\text{Base}}$ as negative when the correct label is positive is lower in comparison to DistilBERT, 12 samples. However, the wrong classification of real negative samples is higher than in DistilBERT, with 4737 samples.

negative	4552	68
positive	3998	321
	negative	positive

Table 1: Few-shot DistilBERT Confusion Matrix

negative	4906	12
positive	4737	81
	negative	positive

Table 2: Few-shot $\text{BERT}_{\text{Base}}$ Confusion Matrix

In Table 3 and 4, we find the results of the performance for the few-shot classification by DistilBERT and $\text{BERT}_{\text{Base}}$ respectively. These results are distributed by three metrics: precision, recall and F1, and we measured them by micro and macro average. As we can see, $\text{BERT}_{\text{Base}}$ shows a little bit higher results than DistilBERT, whereas DistilBERT demonstrates a fairly good behaviour even when the number of parameters is lower than the one in $\text{BERT}_{\text{Base}}$.

Precision	Recall	F1 Macro
0.452	0.324	0.263
Precision	Recall	F1 Micro
0.487	0.487	0.487

Table 3: Metrics Few-shot DistilBERT

Precision	Recall	F1 Macro
0.459	0.332	0.233
Precision	Recall	F1 Micro
0.498	0.498	0.498

Table 4: Metrics Few-shot $\text{BERT}_{\text{Base}}$

negative	4846	72
positive	4462	356
	negative	positive

Table 5: Zero-shot DistilBERT Confusion Matrix

The confusion matrices in Table 5 and 6 correspond to the performance for zero-shot classification in DistilBERT and $\text{BERT}_{\text{Base}}$ respectively. In DistilBERT's confusion matrix (Table 5), the wrong classification of real negative samples (classificated as positive) is almost as high as the right classification of negative samples. Notwithstanding, the wrong classification of real positives is 72 samples, and in the case of $\text{BERT}_{\text{Base}}$ this case is even lower (12 samples).

negative	4906	12
positive	4739	79
	negative	positive

Table 6: Zero-shot $\text{BERT}_{\text{Base}}$ Confusion Matrix

In Table 6, the confusion matrix for zero-shot classification $\text{BERT}_{\text{Base}}$ shows also this tendency of wrong classification of real negative as positives getting even a higher number than in his distilled version (4739 samples has been labelled as positive when were negative).

Regarding the metrics, in Table 7 and 8, we should highlight the good behaviour of DistilBERT in comparison to $\text{BERT}_{\text{Base}}$.

The precision achieved by zero-shot DistilBERT is higher than itself in few-shot and, also, than $\text{BERT}_{\text{Base}}$ in the same conditions.

The lowest score we can perceive in this table is the F1 Macro in case of $\text{BERT}_{\text{Base}}$, with a value of 0.233 points, which was the same result that we got in few-shot classification. Indeed, if we compare

the metrics scores from few-shot and zero-shot in the case of $\text{BERT}_{\text{Base}}$, we can detect that almost all the results are equivalent.

Precision	Recall	F1 Macro
0.450	0.346	0.269
Precision	Recall	F1 Micro
0.520	0.520	0.520

Table 7: Metrics Zero-shot DistilBERT

Precision	Recall	F1 Macro
0.458	0.332	0.233
Precision	Recall	F1 Micro
0.498	0.498	0.498

Table 8: Metrics Zero-shot $\text{BERT}_{\text{Base}}$

8 Conclusions

DistilBERT shows a really good performance in low-resources environments. As it has been demonstrated in the experiment, DistilBERT needs less time of training due to the usage of fewer parameters. Therefore, when the dataset is small the scores for DistilBERT are slightly higher in comparison to the ones for $\text{BERT}_{\text{Base}}$. However, the results are far from accurate and favourable. The best score achieved in precision was 0.52, and as we perceived in the confusion matrix, the classification cannot be considered successful. Regarding the differences of using few-shot and zero-shot, $\text{BERT}_{\text{Base}}$ displays a better performance in few-shot classification whereas DistilBERT in zero-shot classification. As a consequence, in order to generate a good classification model for sentiment analysis in text classification we need to take more samples. For future experiments, we advise to determine which is the limit or the boundary line for DistilBERT to perform good enough without using huge datasets but getting satisfactory metric scores.

Acknowledgements

This project has been developed as a final assignment for the subject Deep Learning in the Master's programm HAP-LAP at the University of the Basque Country (EHU/UPV). This project has been supervised by Ander Barrena.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Nihar Bendre, Hugo Terashima Marín, and Peyman Najafirad. 2020. Learning from few samples: A survey.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Pre-training transformers as energy-based cloze models. *arXiv preprint arXiv:2012.08561*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ruining He and Julian McAuley. 2016. Ups and downs. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *CoRR*, abs/2002.12327.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Timo Schick and Hinrich Schütze. 2020a. Exploiting cloze questions for few-shot text classification and natural language inference. *CoRR*, abs/2001.07676.

Timo Schick and Hinrich Schütze. 2020b. It's not just size that matters: Small language models are also few-shot learners. *CoRR*, abs/2009.07118.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. *CoRR*, abs/1905.02450.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

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.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*.

Atsuki Yamaguchi, George Chrysostomou, Katerina Margatina, and Nikolaos Aletras. 2021. Frustratingly simple pretraining alternatives to masked language modeling. *arXiv preprint arXiv:2109.01819*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?