DistilBERT Sentiment Analysis on IMDb Movie Reviews vs Amazon Fine Foods Reviews

Horia-Gabriel Radu
k55592hr

Alexandru-Liviu Bratosin
m97519ab

Abstract

In this paper, we compare the performance of a DistilBERT model on two different sentiment analysis datasets: movie reviews from the IMDb Movie Reviews dataset and food reviews from the Amazon Fine Food Reviews. For a fair comparison, we reduced the Amazon Fine Foods dataset such the sizes of the training splits are the same as the ones of the smaller IMDb Movie Reviews dataset, and ensured that the sentiment distribution is uniform. We trained DistilBERT on each of these two datasets, with and without pre-training, and cross-tested them on the other dataset. We expected to see one of the models with no pre-training outperform its counterpart trained on a different dataset on the out-of-domain test sets due to the domain adaptation effects. However, we observed the opposite: both models performed well on the in-domain test set, though they both suffered a performance drop of about 10% on the out-of-domain test set. We then fine-tuned pre-trained DistilBERT models on these two datasets to find if this could potentially help the model in better adapting to the out-of-domain test set. We observed that fine-tuning helps both models regain about 10% of the performance lost on the out-of-domain test set.

1 Introduction

Sentiment analysis is one of the most popular natural language processing tasks, with applications in fields such as marketing, politics and economics. Sentiment analysis is a text classification task, where the objective is to predict the sentiment of a given text (Pang et al., 2002). Sentiment can be expressed in terms of positive/negative, high/low, etc.

There are many techniques that have been proposed for sentiment analysis. Traditional techniques such as handcrafted features and rule-based methods are still in use, but the majority of the state-of-the-art techniques are deep learning based methods.

In recent years, transfer learning has gained significant popularity. In general, transfer learning is the process of learning to solve a task by reusing the knowledge gained while solving a different but related task. Transfer learning can be used to leverage existing models to achieve better performance on a new task.

As is the case with all natural language processing applications, the amount of available data is a limiting factor. For instance, there may not be enough data to train a deep learning model from scratch. In such cases, transfer learning can be used to overcome this issue by using pre-trained models. Pre-trained models are models that have been trained on a very large dataset (for instance, a general domain dataset such as Wikipedia). These models are then used as the starting point for fine-tuning on a new, smaller dataset.

In recent years, there has been significant progress in the field of transfer learning for natural language processing tasks. For example, the task of text classification has been tackled successfully using transfer learning. Some of the approaches that have been proposed are based on the use of pre-trained models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 1907), and DistilBERT (Sanh et al., 2019). In most of these models, the pre-training task has been the same, namely language modelling. The pre-trained models are fine-tuned for different tasks, such as text classification, summarization and question answering.

The objective of this paper is to study how a sentiment analysis model adapts to new domains, from movies to foods and vice versa. We compare the performance of four different DistilBERT models, one trained from scratch on IMDb, one fine-tuned on IMDb Movie Reviews, and two more of the same but on Amazon Fine Foods Reviews. We then cross-test the models: the models trained on the

IMDb Movie Reviews are tested on the Amazon Fine Food Reviews and vice-versa. We are testing how the models perform on the in-domain test set compared to how they perform on the out-of-domain test set. We then fine-tune the pre-trained models on each of the two datasets to observe if fine-tuning helps the models better adapt to the out-of-domain test set.

We expected to observe that fine-tuning does help the models regain the performance drop on the out-of-domain test set.

2 Related Work

Transfer learning is a technique that allows knowledge to be transferred from one task to another. This means that a sentiment analysis model trained on one dataset can be used to improve the performance of a sentiment analysis model on a different dataset.

In recent years, transfer learning has been used for sentiment analysis. Some of the approaches that have been proposed are based on the use of pre-trained models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 1907), DistilBERT (Sanh et al., 2019), and other Transformers (Wolf et al., 2019).

The BERT model was proposed by Devlin et al. (Devlin et al., 2018). The BERT model is a transformer-based model trained on a large corpus of text. The BERT model has been fine-tuned on many tasks, such as text classification, summarization, and question answering. The model has been shown to outperform existing methods such as LSTMs, MLPs and CNNs on the tasks stated above, including sentiment analysis.

The DistilBERT model was proposed by Sanh et al. (Sanh et al., 2019). The DistilBERT model is a distilled version of the BERT model. Distilling a model is a process of training a smaller model to mimic the behavior of a larger model.

The model is trained on the same task as the larger model. Distilling a model is usually coupled with an objective of preserving the performance of the larger model. Sanh et al. (Sanh et al., 2019) reported that the DistilBERT model achieves 97% of the performance of the larger BERT model on GLUE benchmarks (Wang et al., 2018).

The currect state-of-the-art performance on the IMDb dataset is 97.4% (Thongtan and Phienthrakul, 2019), a bit larger than what we achieved with our DistilBERT fine-tuned model. But, for the Amazon Fine Foods dataset, we were not able to find a benchmark.

3 Dataset

3.1 The IMDb Movie Reviews Dataset (I)

The IMDb Movie Reviews dataset (Maas et al., 2011) consists of 50,000 movie reviews. The reviews have been labelled as positive or negative. The training set contains 22,500 reviews, the validation set contains 2,500 reviews, and the test set contains 25,000 reviews. The dataset is primarily used for sentiment analysis.

3.2 The Amazon Fine Foods Reviews Dataset (A)

The Amazon Fine Foods Reviews dataset (McAuley and Leskovec, 2013) is a review dataset from Amazon. The dataset consists of 568,454 reviews.

The reviews labelled for helpfulness and score. The helpfulness is a ratio and is defined as the number of reviews that designated the review as helpful over the total number of reviews that designated that review as helpful or not helpful. The score is an integer between 1 and 5 which defines the rating each review writer gave to the certain product. We have taken reviews of 1 and 2 as negative reviews, and reviews of 3, 4 and 5 as positive reviews, to compare to the IMDb model's performance.

We split the Amazon Fine Foods Reviews dataset into three parts having the same sizes as the IMDb training, validation, and testing splits. We converted the scores to positive and negative labels by considering reviews with a score greater or equal to 3 as positive. We also ensured that the sentiment distribution is uniform for all three parts. This means that the total number of positive and negative reviews is the same for all three splits. Finally, we made sure that there was no data leakage among the training splits through thorough uniqueness assertions.

The training split is used to train the model. The validation split is used to tune the hyperparameters of the model and also serves as a rough estimate of the model's performance on unseen data. The test split is used to evaluate the model. The test split is a completely unseen set of data.

Split	I+	I−	A+	A−
Train	11250	11250	11250	11250
Val	1250	1250	1250	1250
Test	12500	12500	12500	12500
Total	50000		50000	

Table 1: IMDb Movie Reviews and Amazon Fine Foods dataset splits. $+$ and $-$ refer to positive and negative reviews, respectively.

The DistilBERT tokenizer is used for preprocessing the reviews in our data splits. It is a pre-trained tokenizer that is already available in the Huggingface library. For this, the texts are all lowercased and tokenized using the WordPiece subword segmentation algorithm (Wu et al., 2016), with a vocabulary size of 30000.

4 Training

4.1 Model and Hyperparameters

The base DistilBERT model along with pre-trained weights have also been taken from the Huggingface Transformers library (Wolf et al., 2020). The pre-trained model along with the tokenizer were pre-trained on BookCorpus, a dataset consisting of 11,038 unpublished books, and English Wikipedia.

To train the model, we implemented a PyTorch-Lightning wrapper which automatically saves training logs and checkpoints. We used the AdamW optimizer to minimize the loss function. The hyperparameters used for all of the models that we have trained are listed in Table 2.

Name	Value
epochs	10
batch size	256
max seq. length	256
seq. padding	max length
seq. truncation	yes
learning rate	5e-5
schedule step size	8
schedule decay gamma	0.1
weight decay	1e-4
layers	6
attention heads	12
encoder and pooler dim.	768
feed-forward dim.	3072
dropout	0.1
attention dropout	0.1

Table 2: Hyperparameters used throughout training.

After training the models on the two datasets with and without pre-training, we expect to obtain similar performance on both the in-domain splits and out-of-domain splits. Otherwise, if they differ by a significant margin, it would mean that one dataset captures a wider spectrum of emotions in language.

5 Results and Discussions

All models performed equally well in both the in-domain and out-of-domain test sets. With that said, it is interesting to note that the domain shift of the out-of-domain test sets result in an accuracy drop of ≈10%. This shows that the model is not very robust to domain shifts. However, fine-tuning the models on the out-of-domain test set helps the model regain the performance lost on the out-of-domain test set. This is likely due to the fact that fine-tuning the model on the out-of-domain test set allows the model to adapt to the out-of-domain test set. Nonetheless, the results are still quite surprising, as this means that the model does not perform very well on out-of-domain test sets. One potential reason for this could be the different vocabularies used in the two domains. As the model is trained on the in-domain data, it is likely that the model learn the in-domain vocabulary better.

	Training dataset			
Testing dataset	I	†I	A	†A
I	85%	90%	73%	85%
A	73%	86%	87%	92%

Table 3: Accuracy of our models on the test splits, with and without pre-training (†).

When the model is tested on the out-of-domain data, it is likely that the model encounters words that it is not familiar with, which could potentially impact the model's performance. For example, if the in-domain data is movie reviews and the out-of-domain data is food reviews, the model might encounter words such as “food” and “restaurant” in the out-of-domain data that it is not familiar with. This could potentially impact the model's performance, as the model would not be able to correctly identify the sentiment of the review.

Another potential reason for the discrepancy between in-domain and out-of-domain accuracy values could be the different sentence structures used in the two domains. For example, if the in-domain data is movie reviews and the out-of-domain data is food reviews, the model might encounter sentence structures such as “I went to the restaurant for dinner” in the out-of-domain data that it is not familiar with.

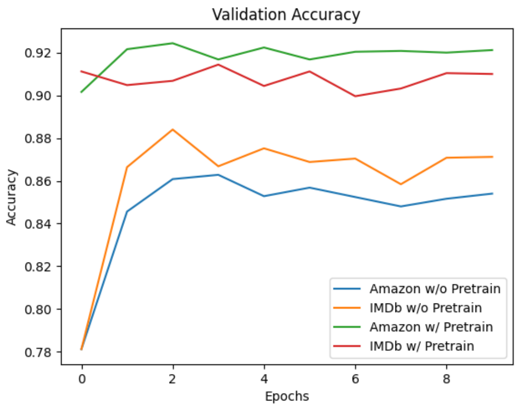

Figure 1: Accuracy of our 4 trained models on the validation split during training.

We observed that fine-tuning the pre-trained DistilBERT model helps in better adapting to the out-of-domain test set. Through the experiments, we noticed that pre-training the model on a different domain helps with the adaptation to the out-of-domain dataset. In light of these findings, we propose to pre-train DistilBERT on multiple datasets, and fine-tune it on the desired task. This can potentially help the model better adapt to a new domain, and improve its performance on out-of-domain test sets.

6 Conclusion

Compared to our initial expectations, of one of the model outperforming another, no dataset resulted in a noticeably better performing model on the out-of-domain test sets.

Furthermore, the fine-tuned models were, as expected, much better than the models trained from scratch, and had a smaller drop in accuracy on the out-of-domain test sets. This proves that in sentiment analysis, generalization of models training data improves performance on any type of domain that the sentiment is to be extracted from.

We recommend to always fine-tune the DistilBERT model on the target domain, to help it adapt better. However, if that is not possible, we recommend pre-training the model on multiple domains, as that will help improve performance and adaptation to the target domain.

References

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*.

Yinhan Liu, Myle Ott, Naman Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 1907. Roberta: A robustly optimized bert pretraining approach. arxiv 2019. *arXiv preprint arXiv:1907.11692*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Tan Thongtan and Tanasanee Phienthrakul. 2019. Sentiment classification using document embeddings trained with cosine similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 407–414.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim

Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.