Minimum Viable Tuning for Intent and Emotion Classification with DistilBERT

Kieffer Thomas
UC Berkeley School of Information
kieffert@berkeley.edu

Abstract

Inferring the underlying intents and emotions in small snippets of text is central to many applications of machine learning. These classifications can provide a high-level motivational map for search queries, give a deeper understanding of posts on social media, and are crucial for developing accurate question-answering and dialog systems. Focusing on contexts where resources may be limited in terms of the scale of data, the length of text to be classified, and compute resources available, utilizing the DistilBERT model to fine-tune a classifier on small amounts of data is found to be effective at building accurate text classifiers with relatively few data points.

1 Introduction

Intent detection plays a critical role in many real-world applications for machine learning. For instance, when choosing what ad to show when serving advertisements on a search engine, being able to differentiate the nuanced motivations behind similar queries like ”best deal on new hiking boots” (shop for deals), ”buy new hiking boots” (high intent to purchase), and ”best hiking boots for high altitudes” (research) would enable an advertiser to tailor specific messaging that matches a high-level contextual understanding of what that short snippet of text truly *means* for the searcher. More directly, there are often scenarios such as dialog systems on websites or automated phone systems that rely on specifically answering questions from a user, and correctly classifying the type of question being posed allows for efficient retrieval of answers and relevant information.

Similary, emotion detection from text can play an important role in developing better automated solutions. In *SemEval-2019 Task 3: EmoContext, Contextual Emotion Detection in Text* (Chatterjee et al., 2019), the researchers note that in addition to deep informational knowledge, if digital assistants like Siri and Alexa can possess a degree of emotional intelligence they can achieve more human-like interaction with users, and highlight the challenge of inferring emotion strictly from text without additional information such as facial expressions. On a social media site like Twitter, the ability to accurately gauge a user's emotion from a short tweet can enable more efficient moderation, content recommendation algorithms, or even identify when users are struggling with mental health crises (Hasan et al., 2019).

However, the ability to create a robust text classifier for intents and emotions will often be difficult due to resource constraints. This could be the case when adapting to a new environment, like working to tailor advertisements for a small company where little data exists for queries of their niche products, attempting to build an automated question-answering system from the ground up, or when a user or community has little textual information to build a statistical model from.

This analysis looks at low-resource data scenarios, and using DistilBERT – a pared down version of the larger BERT model that is ”40% smaller, 60% faster, [and] retains 97% of the language understanding capabilities” (Sanh et al., 2019) – finds classification models fine-tuned using just 2000 data points are generally sufficient to produce models that perform within 5 percentage points of the best accuracy scores of larger BERT models and best performers overall.

2 Background

The highly adaptable pre-trained language model BERT has proven very effective in low-resource

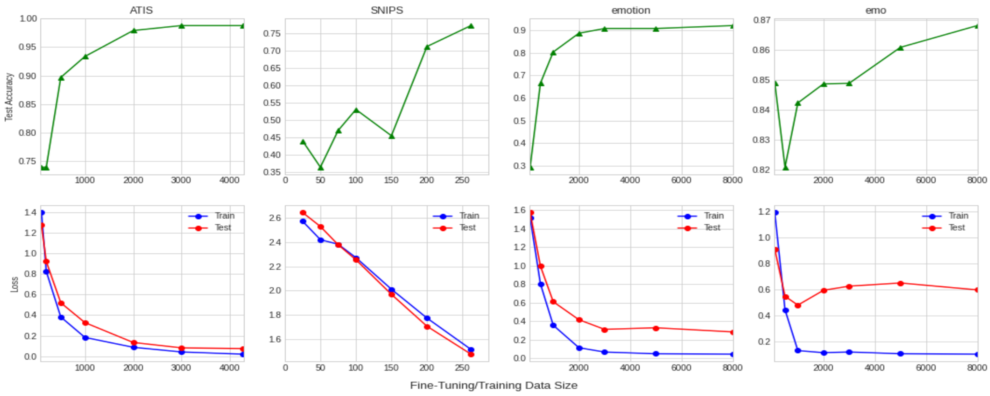

Figure 1: Classification accuracy and train/test cross-entropy loss across training data sizes used for fine-tuning.

environments, leveraging fine-tuning of the base model on a wide range of NLP tasks to achieve state of the art performance, often with small datasets used for fine-tuning (Devlin et al., 2018). DistilBERT is a distilled version of the larger BERT transformer model developed by HuggingFace that has a smaller number of model parameters, can be fine-tuned in less time, and still achieves similar performance as BERT (Sanh et al., 2019).

In low-resource settings, Grießhaber et al. have managed to fine-tune BERT models for downstream tasks effectively using fewer than 1000 data points using BERT and employing active learning techniques (Grießhaber et al., 2020). Similarly, employing transfer learning with very small datasets has proven effective in the task of sentiment classification (Gupta et al., 2018).

Intent detection, often paired with a slot-filling task for question answering systems, has seen nearly perfect 99% accuracy achieved on the relatively small Airline Travel Information Systems (ATIS) dataset (Wang et al., 2018), and researchers have also used BERT to achieve 99% accuracy classifying intents in the SNIPS dataset (Wu et al., 2020).

Emotion detection from text, with wide-ranging applications in human-computer interactions and content understanding, has garnered a lot of attention in recent years (Sailunaz et al., 2018). Most notably, the ACL's recent 2019 SemEval posed the challenge of classifying emotion from a string of text exchanges (Chatterjee et al., 2019), and strides have been made in properly assessing emotions behind posts on social media sites like Twitter (Silveira et al., 2014).

The aim of this analysis is to build on this work and illustrate that a decent model for intent and emotion classification can be created quickly and with little context for model fine-tuning.

3 Methods

Sequence classification models were created by fine-tuning the DistilBERT model (Sanh et al., 2019), available in the PyTorch Transformers library on HuggingFace, on varying sizes of training data. For comparison, the largest training size was also used for sequence classification fine-tuning and evaluation using the BERT$_{\text{base}}$ model (Devlin et al., 2018), also available on HuggingFace. All models were tuned and evaluated on Google Colab using a GPU (see Appendix for relative pre-training times on DistilBERT models).

3.1 Data

Fine-tuning and performance is evaluated for four datasets. The first two are the benchmark datasets for intent detection; the emotion datasets have been used in recent papers and ML competitions and provide excellent examples for classification from short texts akin to the intent detection datasets:

ATIS: Airline Travel Information Systems. Consists of transcripts of requests for flight information and the intent of the request. n=4274 training examples, 512 test. 17 intent classes.

SNIPS: Benchmark dataset created at snips.ai. Consists of text of requests to automated assistants.

Comparison of Classification Accuracy					
Dataset	**Baseline (MCC)**	**DistilBERT (up to 2K)**	**DistilBERT (up to 8K)**	**BERT$_{\text{base}}$ (up to 8K)**	**SOTA**
ATIS	74.2	**97.5**	98.8	98.8	99.0
SNIPS	19.8	**77.3**	77.3	62.1	99.0
emotion	33.5	**88.6**	92.0	92.2	94.1
emo	49.6	**84.5**	86.8	87.0	89.7

Table 1: Fine-tuning DistilBERT on 2000 training examples achieves performance close to BERT$_{\text{base}}$ and state of the art approaches. In table, ”up to 2K” and ”up to 8K” indicate model was fine-tuned on 2000 and 8000 training examples if data available; it dataset is not large enough, full training data was used to fine-tune and results reported on test set. Baseline is most common classifier (MCC). State of the art (SOTA) results for ATIS were achieved by (Wu et al., 2020), SNIPS by (Wang et al., 2018), emotion by (Silveira et al., 2014), emo by (Chatterjee et al., 2019).

Split to n=262 training examples, 66 test. 10 intent classes.

emotion: Dataset collected by Elvis Saravia et al. (Saravia et al., 2018). Consists of English language twitter messages. n=16000 training examples, 2000 test. 6 emotion classes.

emo: Dataset created by Chatterjee et al. (Chatterjee et al., 2019). Each example has three ”turns” of a text exchange, the third of which is classified with an emotion. n=30160 training examples, 5509 test. 4 emotion classes.

3.2 Learning Parameters

In all models the learning rate parameter is set slightly above the default to 2e-5, a rate that should allow for somewhat aggressive learning while avoiding ”catastrophic forgetting”, where pre-trained knowledge in the underlying models gets erased in tuning for the new task (Sun et al., 2019). The number of training epochs in all models is set to 4, as at least 3 is recommended for BERT models, but up to 5 can be recommended for classification tasks.

3.3 Class Imbalance

Most of the datasets had some degree of class imbalance, but research has shown BERT models deal well with class imbalances if the training data and test data are sufficiently similar (Madabushi et al., 2020). No augmentation or sampling techniques are used to balance classes, and strong performance of these models seems to indicate sufficient similarity in test and train data across the 4 datasets analyzed.

3.4 Evaluation

Accuracy on test data classification is used as the primary evaluation metric for this analysis. This is the standard metric seen for the intent detection benchmarks, and in all cases here is an appropriate gauge for overall performance on multi-class classification tasks (weighted F1 scores are also shown in Appendix). The baseline chosen (shown in Table 1) is the most common classifier, a low threshold to outperform to show that these models are indeed gaining performance with minimal fine-tuning.

4 Results

DistilBERT proves very effective at tuning for classification tasks with little data. As highlighted in Figure 1, performance on the test data rapidly increases, and by roughly 2000 training examples, is nearing both the larger BERT$_{\text{base}}$ performance on data four times that size, as well as overall best benchmark performances on these datasets. For the ATIS, emotion, and emo datasets, accuracy for a DistilBERT classification model fine-tuned on just 2000 examples is within about 5 points of the accuracy for top performing classifiers that utilize more complex models and fine-tune on all available data. For comparison, a DistilBERT model fine-tuned on 2000 examples will train in about 1/7 the time it takes to fine-tune a BERT$_{\text{base}}$ model on 8000 examples. If top-notch performance isn’t mandatory and learning quickly is the goal, or if faced with constraints in available data and compute resources, one should expect solid performance on similar tasks utilizing DistilBERT.

Of note, the available SNIPS benchmark

dataset only provided 328 total examples. The trends seen in performance do indicate rapid learning, however with such small amounts of data for fine-tuning it is clear that improvements are not steady, though one might expect similar characteristics to the ATIS and emotion accuracy plots with a larger volume of data.

The results in Figure 1 also highlight an interesting trend in the cross-entropy loss for the emo dataset. While the training loss approaches zero quickly with the addition of training examples to the fine-tuning data, the test loss actually slightly *increases*. It is possible the model is overfitting, however as the training size continues to increase we do see accuracy improving and a slight decline in the training loss. Performance of the model is still near that of the best classifiers created in the SemEval-2019 Emotion Detection task.

5 Conclusion

Utilizing DistilBERT for intent and emotion detection in low-resource settings shows strong results, and corroborates the HuggingFace researchers' claims that the distilled version retains nearly all of the larger BERT model's underlying language understanding while achieving meaningful gains in efficiency for fine-tuning and evaluation. Further, strong performance can be achieved with a relatively small (around n=2000) sample size used for fine-tuning.

Future work that may prove beneficial for these classification tasks in low-resource settings would include leveraging even more efficient variations of the initial BERT model like ALBERT, data augmentation when training examples are very limited, and exploring if model performance can be maintained while reducing training time with techniques like freezing model layers.

References

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.

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*.

Daniel Grießhaber, Johannes Maucher, and Ngoc Thang Vu. 2020. Fine-tuning bert for low-resource natural language understanding via active learning. *arXiv preprint arXiv:2012.02462*.

Rahul Gupta, Saurabh Sahu, Carol Espy-Wilson, and Shrikanth Narayanan. 2018. Semi-supervised and transfer learning approaches for low resource sentiment classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5109–5113. IEEE.

Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. 2019. Automatic emotion detection in text streams by analyzing twitter data. *International Journal of Data Science and Analytics*, 7(1):35–51.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2020. Cost-sensitive bert for generalisable sentence classification with imbalanced data. *arXiv preprint arXiv:2003.11563*.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):1–26.

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*.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697.

Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. 2014. A gold standard dependency corpus for english. In *LREC*, pages 2897–2904. Citeseer.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Yu Wang, Yilin Shen, and Hongxia Jin. 2018. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. *arXiv preprint arXiv:1812.10235*.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. *arXiv preprint arXiv:2010.02693*.

Appendix

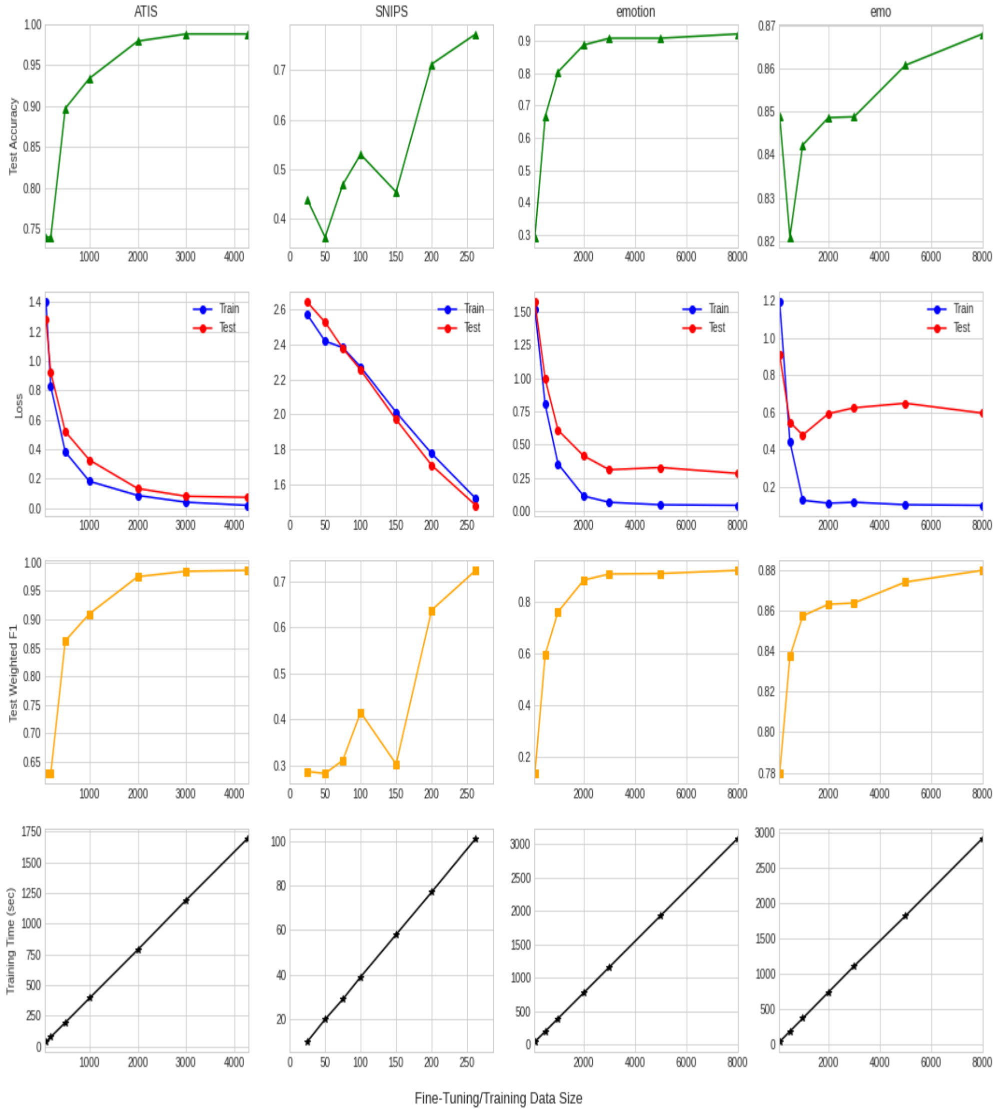

Figure 2: Summary of full results across training sizes: classification accuracy, train/test cross-entropy loss, weighted F1 score on test data, and training time for DistilBERT