Datasaur

Datasaur Predictive Labeling: Achieve High-Performing AI Models with 60% Less Data Labeling Effort

June 2024 | Whitepaper

01 Introduction

In the field of machine learning, the effectiveness of models is closely linked to the quality and quantity of the training data. At Datasaur, a common question at the start of a project is: "How much data is needed?" This question highlights the critical balance between having enough data and achieving good model performance, recognizing that data preparation can be both costly and strategically important.

Recent advances aim to reduce the amount of data needed to achieve desired performance levels. Techniques such as data augmentation, transfer learning, and meta-learning are increasingly used to enhance model robustness and lower data requirements. Pretrained models, particularly Transformers (Vaswani et al., 2017), benefit from extensive pretraining on large datasets, enabling effective knowledge transfer to specific tasks.

02 Understanding Few-Shot Learning

Few-shot learning is a specialized area within machine learning that concentrates on training models to achieve accurate predictions using only a few labeled examples (Parnami & Lee, 2022). Unlike traditional machine learning approaches that typically require extensive labeled data to generalize effectively and make precise predictions, few-shot learning addresses scenarios where obtaining abundant labeled data is impractical or costly.

The fundamental challenge of few-shot learning lies in developing models capable of effective learning from a limited number of labeled examples and subsequently generalizing this learning to unseen data. This setup closely resembles human learning processes, where individuals can often recognize new concepts or objects with just a few instances or even a single example.

Here are the core elements of Few-Shot Learning:

1. **Limited Labeled Training Set (X, Y)**: In few-shot learning, the training dataset consists of only a few labeled examples, represented as (X, Y), where X denotes the input data and Y denotes the corresponding labels. Typically, these labeled examples constitute a small subset of the overall available dataset.
2. **Unseen Testing Data (X)****: The primary objective of few-shot learning is to train models capable of accurately predicting labels for new, unseen testing data denoted as X. These unseen examples are distinct from the initial labeled training set and serve to assess the model's ability to generalize effectively.

3. **Generalization and Adaptation**: Few-shot learning models strive to generalize well beyond the limited labeled examples used during training. They need to adapt quickly to new tasks or classes with only a few labeled examples, making accurate predictions for unseen data points. These models are designed to capture underlying data patterns and characteristics to ensure reliable predictive performance.

03 Few Shot Text Classification

Few-shot text classification accurately classifies text data using only a few labeled examples. Unlike traditional supervised learning approaches that depend on abundant labeled data, few-shot text classification leverages pretrained language models (PLMs) like BERT or GPT. These models, pretrained on large text corpora, quickly adapt to new tasks with limited labeled examples. Techniques like parameter-efficient fine-tuning (PEFT) (Liu et al., 2022) and pattern exploiting training (PET) (Schick & Schütze, 2020). are employed to enhance performance in limited data scenarios .

04 Sentence Transformer Fine-tuning

Although prompting is no longer a significant issue, the size and cost of LLMs still pose challenges. To address these concerns, innovative approaches are needed to optimize model efficiency and performance while minimizing resource requirements. Sentence Transformer Fine-tuning or SetFit (Tunstall et al., 2022) is a new approach to modelling that uses pretrained models along with a technique called contrastive learning. It leverages pretrained language models to achieve high accuracy even with minimal labeled data, making it a powerful tool for various natural language processing tasks.

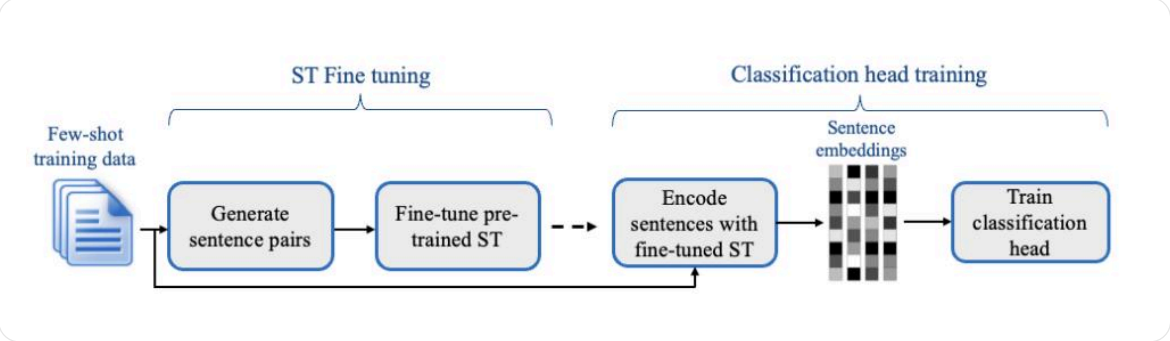

Fig. 1: SetFits's fine-tuning and training block diagram.

SetFit Features:

- **No prompts or verbalisers**: Current techniques for few-shot fine-tuning require handcrafted prompts or verbalisers to convert examples into a format that's suitable for the underlying language model. SetFit dispenses with prompts altogether by generating rich embeddings directly from a few labeled text examples.
- **Fast to train**: SetFit doesn't require large-scale models like T0 or GPT-3 to achieve high accuracy. As a result, it is typically an order of magnitude (or more) faster to train and run inference with.
- **Multilingual support**: SetFit can be used with any Sentence Transformer on the Hub, which means you can classify text in multiple languages by simply fine-tuning a multilingual checkpoint.

05 Datasaur Predictive Labeling

Datasaur Predictive Labeling feature leveraging the power of SetFit which can significantly enhance the efficiency and effectiveness of data labeling processes. SetFit, as a few-shot learning technique utilizing pretrained models and contrastive learning, offers several key advantages that align well with the goals of Predictive Labeling.

SetFit eliminates the need for prompts or verbalisers typically used in few-shot fine-tuning approaches. Instead, it directly generates rich embeddings from a few labeled text examples. This feature is particularly advantageous for Predictive Labeling in Datasaur, as it streamlines the labeling process by efficiently learning from existing labeled data to predict future labels without the need for additional manual intervention or prompt crafting.

Integrating SetFit within Datasaur Predictive Labeling framework harnesses the benefits of efficient few-shot learning, enabling smarter and quicker AI predictions based on past labeling efforts. By leveraging SetFit's capabilities, Datasaur enhances its predictive labeling functionality, making data annotation more accessible, efficient, and accurate for users across various industries and use cases.

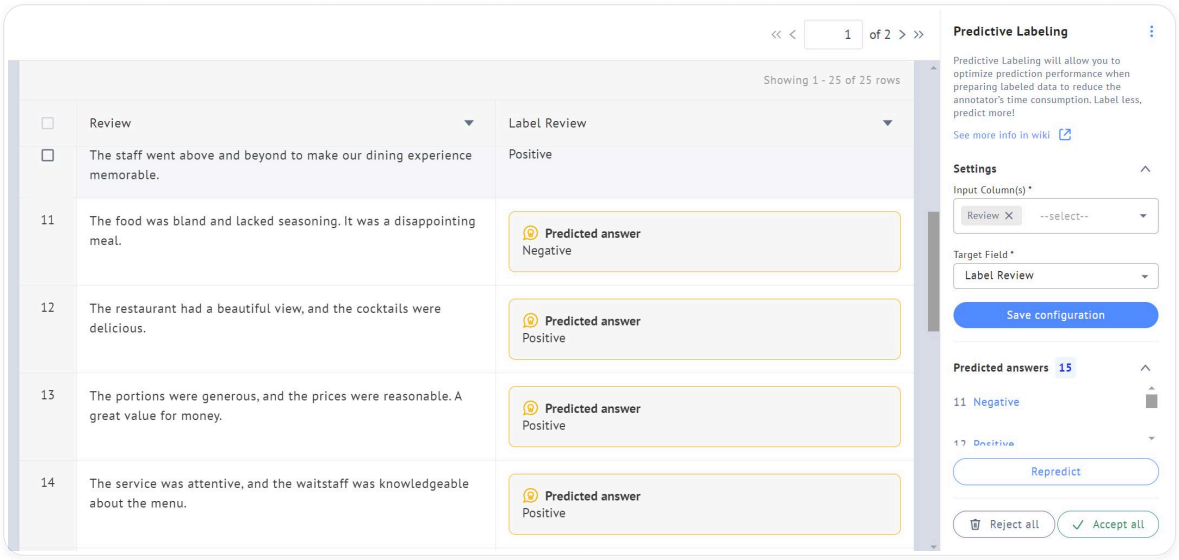

Fig. 2: Predictive Labeling in Action.

06 Experiments

To demonstrate the effectiveness of Predictive Labeling powered by SetFit, we conducted experiments using various datasets representing different text classification tasks. Specifically, we trained and evaluated SetFit on the datasets listed in Table 1. Our objective was to leverage SetFit's pretrained models and contrastive learning approach to showcase its ability to accurately and efficiently predict labels across datasets with varying numbers of classes.

Dataset	Num class	URL
yelp_polarity	2	https://huggingface.co/datasets/yelp_polarity
ag_news	4	https://huggingface.co/datasets/ag_news
emotion	6	https://huggingface.co/datasets/dair-ai/emoti...
20_newsgroups	20	https://huggingface.co/datasets/SetFit/20_n...
banking77	77	https://huggingface.co/datasets/banking77

Table 1: Dataset used for Experimentation.

These experiments demonstrate how Predictive Labeling, enhanced by SetFit, adapts to various classification tasks across different datasets, providing efficient and accurate labeling solutions for diverse industry needs. We will compare the SetFit with the Transformer's model.

07 Results and Analysis

The experiment results illustrate the superior performance of SetFit compared to the Transformers model across various datasets and shot settings

Dataset	Model	1-shot	2-shot	4-shot	8-shot	12-shot	16-shot	20-shot
yelp_polarity	Transformers	51,8	53,27	53,1	57,7	64,77	67,9	74,3
	Setfit	64,9	75,77	86,1	91,87	93,1	93,33	93,07
ag_news	Transformers	30,37	31,6	50,5	74,97	79,7	80,67	81,37
	Setfit	55,3	66,1	75,53	80,77	83,33	82,43	83,5
emotion	Transformers	24,6	22,03	23,2	31	37	38,87	48,13
	Setfit	18,53	20,97	33,7	43,97	52,6	55,47	57,93
20_newsgroups	Transformers	10,6	21,37	41,13	54,07	59,4	60,63	61,67
	Setfit	27,1	47,9	57,6	59,6	62,1	65	63,6
banking77	Transformers	15,47	33,3	60,5	76,33	83,03	85,93	88,1
	Setfit	47,2	60,2	75,9	82,1	81,4	84,5	86,4

Table 2: Detailed Metric Performance.

To summarize the experiment results and performance comparison between Transformers and SetFit models across various datasets and shot settings:

Performance Across Different Datasets:

- **yelp_polarity**: SetFit consistently outperformed Transformers across all shot settings, achieving up to 93.33% accuracy with 16-shot.
- **ag_news**: SetFit demonstrated superior performance, especially noticeable in higher shot settings, reaching 83.5% accuracy with 20-shot.
- **emotion**: Despite the dataset complexity, SetFit showed competitive results, with 57.93% accuracy at 20-shot.
- **20_newsgroups and banking77**: SetFit consistently surpassed Transformers in accuracy across varying shot settings, showcasing its adaptability and effectiveness.

Effect of Shot Settings:

In general, both models benefited from increased shot settings, with SetFit often achieving higher accuracy levels compared to Transformers, especially noticeable in fewer shot scenarios.

Additionally, we evaluated the average metric performance of both models:

Model	1-shot	2-shot	4-shot	8-shot	12-shot	16-shot	20-shot
SetFit	26.57	32.31	45.69	58.81	64.78	66.8	70.71
Transformers	43.91	54.61	65.13	70.62	74.17	75.67	76.44

Table 3: Average Metric Performance.

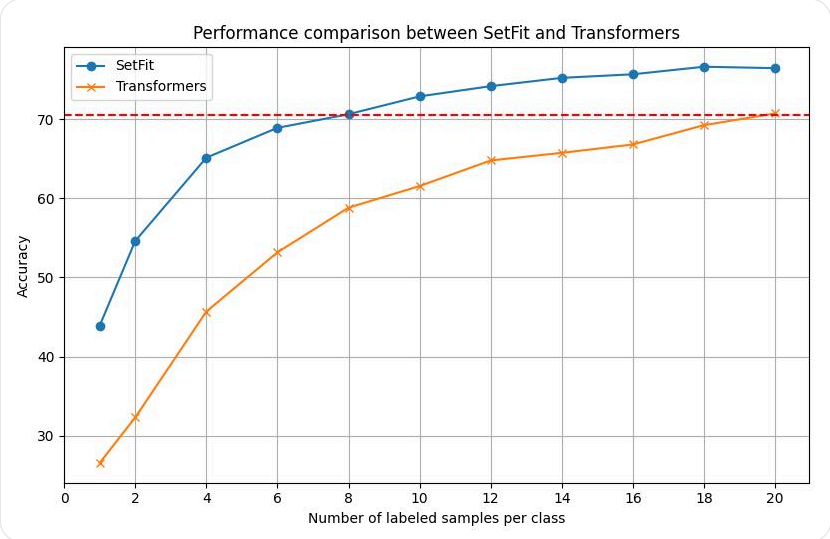

Fig. 3: Performance comparison between SetFit and Transformers models across varying numbers of labeled samples per class.

The results strongly suggest that the SetFit model achieves higher accuracy with fewer labeled samples compared to the Transformers model. Specifically, SetFit reaches a performance level approaching 70% accuracy with approximately 8 labeled samples per class, while the Transformers model requires nearly twice as many samples to approach the same level of performance with 20 labeled samples. This demonstrates the efficiency of SetFit in leveraging labeled data to achieve comparable or superior model accuracy with significantly less labeling effort.

The ratio of these minimum data requirements (20 / 8) equates to 2.5, indicating that Transformers need 2.5 times more data than SetFit. To determine the reduction in data labeling effort, we use the inverse of this ratio: 1 / 2.5, which equals 0.4. Converting this to a percentage gives us the value of (1 - 0.4) x 100% = 60%.

Therefore, utilizing Predictive Labeling allows for achieving high-performing AI models with 60% less data labeling effort compared to the approach using Transformers, showcasing a substantial efficiency gain in AI model development and deployment.

Training time comparison

We also measured the training time performance between the SetFit and Transformers models, and the results are summarized below:

Model	1-shot	2-shot	4-shot	8-shot	12-shot	16-shot	20-shot
SetFit	26.57	32.31	45.69	58.81	64.78	66.8	70.71
Transformers	43.91	54.61	65.13	70.62	74.17	75.67	76.44

Table 4: Training Time across numbers of labeled samples per class.

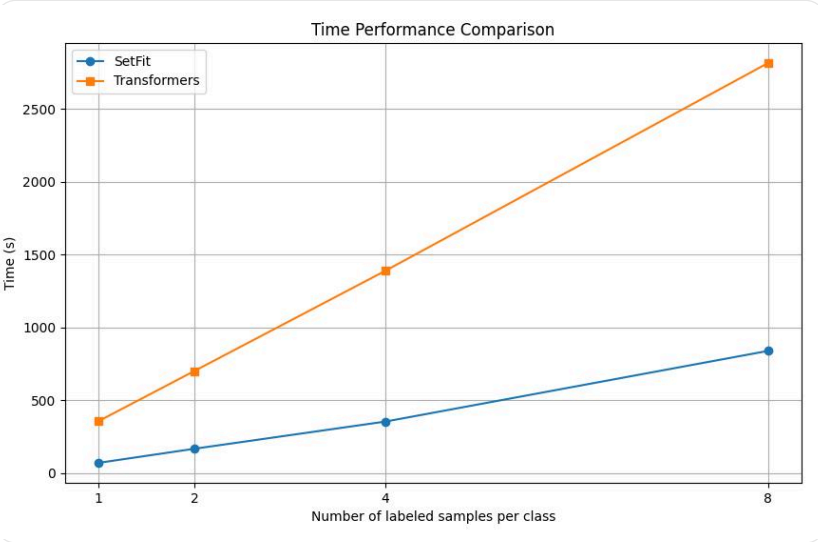

Fig. 4: Training time performance comparison.

Observations and Insights

- **Accuracy Comparison**: Across various datasets and label sizes, SetFit consistently outperforms distilbert-base-uncased in terms of accuracy, especially as the number of labeled samples increases.
- **Efficiency in Training**: SetFit demonstrates significantly lower training times compared to distilbert-base-uncased, showcasing its efficiency in leveraging smaller labeled datasets.

These experimental results underscore the effectiveness of predictive labeling techniques, particularly when coupled with optimized model architectures like SetFit, in enhancing the efficiency and performance of machine learning workflows.

08 Conclusion

SetFit's utilization of smaller pretrained models and contrastive learning allows for remarkable accuracy and efficiency with limited labeled data. This capability reduces reliance on large labeled datasets and extensive training, making advanced machine learning more accessible and scalable.

The adoption of few-shot learning techniques in predictive labeling marks a significant shift towards more efficient and cost-effective machine learning operations. These innovations promise to streamline data-centric workflows and expand the potential of AI applications, ensuring organizations can leverage AI more effectively.

By using Datasaur Predictive Labeling powered by SetFit, organizations can achieve high-performing AI models with 60% less data labeling effort, showcasing a substantial efficiency gain in AI model development and deployment.

09 References

- Parnami, A., & Lee, M. (2022). Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. https://arxiv.org/abs/2203.04291
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. https://arxiv.org/abs/1706.03762
- Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient Few-Shot Learning Without Prompts. https://arxiv.org/abs/2209.11055
- Schick, T., & Schütze, H. (2020, January 21). Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference. arXiv.org. https://arxiv.org/abs/2001.07676
- Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. (2022, May 11). Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. arXiv.org. https://arxiv.org/abs/2205.05638

About Datasaur

Datasaur is a private LLM provider and data labeling platform designed for companies to build their AI ecosystem with ease and efficiency. It assists organizations and universities in setting up custom LLMs and annotating data more efficiently and accurately through automation, quality control, and human-in-the-loop workflows. For more information, visit www.datasaur.ai.

Schedule a demo