(https://www.algoanalytics.com/)

Blog 6 min read

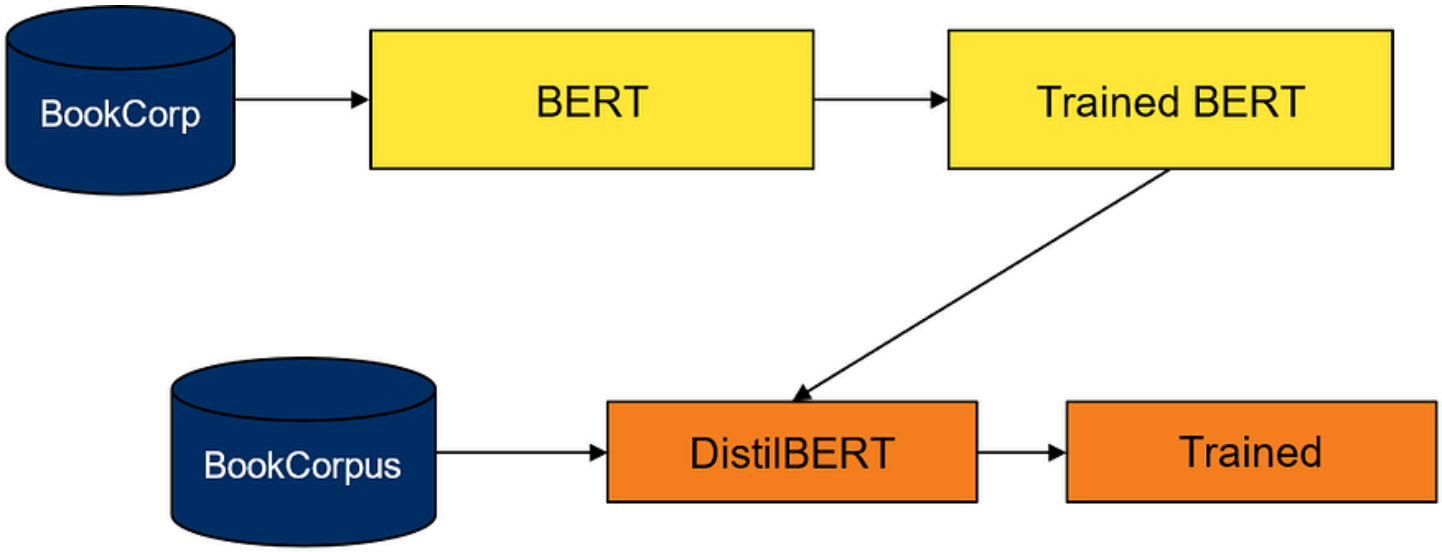
A Comparative Analysis: Fine-tuning Multilabel Classification Models Using DistilBERT vs. GPT-3

By Team Algo

Reading Time: 6 minutes

by Etisha Gurav

NLP is one the most rapidly evolving fields in Machine Learning and with the introduction of many paid and open source Large Language Model, it has become very easy to downstream various tasks in various domains involving text such as text generation, question answering and text classification. In this blog, we would be focusing on multilabel text classification. It is a machine learning technique that assigns predefined categories to texts and categorizes them in groups. To achieve this, pretrained models such as DistilBERT were fine-tuned on a custom dataset but with the advent of advanced models such as GPT-3, there is an opportunity to explore a new approach for fine-tuning multilabel classification models. This blog post aims to compare the traditional approach of fine-tuning using DistilBERT with the new approach of fine-tuning using GPT-3.



Overview of Traditional Approach: Fine-Tuning with DistilBERT

DistilBERT is smaller, cheaper and a faster version of BERT where the model is compressed by training on the same dataset as of BERT using the student-teacher learning approach (Knowledge Distillation). DistilBERT tries to mimic the output distribution of BERT by training with cross-entropy over the soft-targets. DistilBERT has the same architecture as BERT but with minor differences such as token type embeddings and pooler are removed and there is a 2x reduction in number of layers.

It was pretrained with three objectives:

- Distillation loss: the model was trained to return the same probabilities as the BERT base model.
- Masked language modeling (MLM): this is part of the original training loss of the BERT base model. When taking a sentence, the model randomly masks 15% of the words in the input, then runs the entire masked sentence through the model and has to predict the masked words.
- Cosine embedding loss: the model was also trained to generate hidden states as close as possible as the BERT base model.

This way, the model learns the same inner representation of the English language than its teacher model, while being faster for inference or downstream tasks.

To use DistilBERT for text classification, the model has to be finetuned on a custom dataset first. To achieve this, all the text has to be tokenized and each word be encoded into their input ids and attention masks(so that model knows which word is more important than the others). Once this is done, encoded inputs are then passed to the pretrained DistilBertSequenceClassification Model with a defined optimizer and loss function. Now this model is ready and can be used on the test data.

DistilBERT is an efficient model which can be used in resource-constrained environments and requires better results. It retains 97% of the performance with fewer parameters(66 M parameters in comparison to BERT's 304M parameters).

Introduction to GPT-3: A New Paradigm

GPT-3 is an autoregressive, only decoder language model, which uses an attention mechanism, created by OpenAI. It is trained on 175 Billion parameters and 45TB of data, making it one of the largest models to be available. There are many models within the GPT-3 family catering to various types of use cases. Some of them are shown in the table given below.

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}
GPT-3 Small	125M	12	768	12	64
GPT-3 Medium	350M	24	1024	16	64
GPT-3 Large	760M	24	1536	16	96
GPT-3 XL	1.3B	24	2048	24	128
GPT-3 2.7B	2.7B	32	2560	32	80
GPT-3 6.7B	6.7B	32	4096	32	128
GPT-3 13B	13.0B	40	5140	40	128
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128

Family of GPT-3 models with various parameters

All these models can be accessed by an API key provided by OpenAI.

Like DistilBERT, GPT-3 can also be down-streamed for various tasks. This is done by giving an example of the input in the form of a simple prompt and output in the form of completion of the prompt. For a classification task, the prompt could be the data to be classified and the output, the class that is assigned. An example is given below:

	prompt	completion
0	Tata Power net debt has reduced further by Rs2...	Debt \n
1	The company reduced its gross debt to ₹47,424 ...	Debt \n
2	"Its net debt as of March 31, 2023 stood at Rs...	Debt \n

Example of Prompt and Completion for Multilabel Classification

Since GPT-3 models take text as input, there is no need for an extra step of tokenizing the text as well as the output(in this case, class) which makes it very flexible for any kind of textual data. This greatly reduces the data preparation steps required for training a model. Once this step is done, the data can be passed through the desired GPT-3 model. Another benefit of this model is that it can handle very large text sequences due to its large context window, which is helpful in classification where the entire textual data would be required to understand which class is to be assigned. This greatly enhances the performances of the model.

Comparative Analysis: DistilBERT vs. GPT-3

For the comparative study of the performance of these two models, we have considered 13 labels with 100 texts under each label for fine tuning of the models.

For DistilBert, We had used the pretrained model 'distilbert-base-uncased' available from HuggingFace and created a Tokenizer as well as for Sequence Classification. However for GPT-3, the training data was created with prompt -> completion format and stored in jsonl file which could be passed into the GPT-3 models using the OpenAI API.

To evaluate the performances of these models, we have used the classification metrics such as precision, recall and accuracy. Precision would tell us how the model predicted the outcome correctly with respect to all its predictions whereas Recall would tell us

how the model predicted the outcome with respect to all positive data points present in the training dataset. These metrics would tell us about the accuracy of the model.

Other parameters to consider for the comparison are the number of epochs and learning rate(in case of GPT-3, a learning rate multiplier is used). The following are the classification metrics results for DistilBert and GPT-3

Class	Precision	Recall	F1-Score	Support
Debt	1.00	1.00	1.00	20
Dividend	0.96	1.00	0.98	51
Employment	0.90	0.86	0.88	21
Financial Results	0.81	0.96	0.88	27
Investment & Funding	0.95	1.00	0.97	39
Litigation	1.00	0.86	0.93	22
Macroeconomics	0.93	0.70	0.80	20
Merger & Acquisition	0.95	0.91	0.93	23
Partnership & JointVenture	0.78	0.78	0.78	27
Rating_and_Recommendation	0.96	1.00	0.98	27
Share Repurchase	0.96	1.00	0.98	27
Products And Services	1.00	0.96	0.98	25
No	1.00	1.00	1.00	25
Accuracy			0.94	354
Macro Avg	0.94	0.93	0.93	354
Weighted Avg	0.94	0.94	0.94	354

Classification metrics for GPT-3 model(ada-model)

	precision	recall	f1-score	support
Debt	0.89	0.94	0.92	18
Dividend	1.00	0.98	0.99	53
Employment	0.91	0.95	0.93	22
Financial Results	0.88	0.94	0.91	16
Investment & Funding	0.95	0.95	0.95	38
Litigation	0.95	0.95	0.95	19
Macroeconomics	1.00	0.89	0.94	19
Merger & Acquisition	0.96	0.86	0.91	28
No	0.81	0.81	0.81	27
Partnership & JointVenture	1.00	1.00	1.00	26
Products And Services	1.00	1.00	1.00	23
Rating_and_Recommendation	0.94	1.00	0.97	31
Share Repurchase	0.97	1.00	0.98	30
accuracy			0.95	350
macro avg	0.94	0.94	0.94	350
weighted avg	0.95	0.95	0.95	350

Classification for DistilBERT Model

As seen above, Both the models, GPT-3 and DistilBERT, have a good accuracy scores of 94% and 95%, with the latter performing a little better. However, due to the advantage of minimal data preparation, GPT-3 would be preferred over DistilBERT.

Real-World Use Cases and Applications

Even after having similar accuracy scores on the same tasks, GPT-3 models and DistilBERT could be used for different cases and applications. Below are some of the real-world uses cases for each of the model:

DistilBERT

1. Sentiment analysis: DistilBERT models can be used to classify text as positive, negative, or neutral. This can be used to understand the sentiment of social media posts, customer reviews, and other text data.

2. Topic classification: DistilBERT models can be used to classify text into different topics. This can be used to organize documents, filter news articles, and recommend content to users.

3. Entity recognition: DistilBERT models can be used to identify entities in text, such as people, organizations, and locations. This can be used to extract information from documents, improve search results, and power chatbots.

GPT-3

1. Customer support: GPT-3 models can be used to classify customer support tickets by topic, sentiment, and other factors. This can help customer support agents to quickly and accurately identify the issue and provide the best possible resolution.

2. Product recommendation: GPT-3 models can be used to recommend products to users based on their past purchase history, interests, and other factors. This can help users to discover new products that they are likely to be interested in.

3. Fraud detection: GPT-3 models can be used to classify transactions as fraudulent or legitimate. This can help businesses to protect themselves from financial losses.

Conclusion

GPT-3 models and DistilBERT tend to have similar performances with respect to multilabel classification, however it is also important to consider factors such as the resources available for fine-tuning and cost. DistilBERT, due to its smaller size, can be used when there is resource constraint. Moreover DistilBERT works well when the classification is done over a dataset pertaining to a single domain and can also be easily fine tuned. It is also an open-source model which is available commercially, which when compared to GPT-3 models, which is not open-source, compares significantly higher. However, GPT-3 models show their strength when the dataset is not pertaining to a single domain and the outcome depends on varied sentiment and longer token length. Fine-tuning of models has become easier with the introduction of models like GPT-3, causing other open-source models to be at par with the recently introduced models which in turn will bring a lot of enhancements in the field of Natural Language Processing.

References

https://www.analyticsvidhya.com/blog/2022/11/introduction-to-distilbert-in-student-model/
(https://www.analyticsvidhya.com/blog/2022/11/introduction-to-distilbert-in-student-model/)https://www.kaggle.com/code/pritishmishra/text-classification-with-distilbert-

92-accuracy
(https://www.kaggle.com/code/pritishmishra/text-classification-with-distilbert-92-accuracy)https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/
(https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/)https://en.wikipedia.org/wiki/GPT-3 (https://en.wikipedia.org/wiki/GPT-3)

https://swatimeena989.medium.com/distilbert-text-classification-using-keras-c1201d3a3d9d (https://swatimeena989.medium.com/distilbert-text-classification-using-keras-c1201d3a3d9d)

Recent Posts

Transforming Customer Service with Conversational AI: Trends and Innovations (https://blog.algoanalytics.com/2024/10/25/transforming-customer-service-with-conversational-ai-trends-and-innovations/)

October 25, 2024

Siamese Networks: AI's Dynamic Duo for Smarter Similarity Learning! (https://blog.algoanalytics.com/2024/09/18/siamese-networks-ais-dynamic-duo-for-smarter-similarity-learning/)

September 18, 2024

Implementing a basic Retrieval-Augmented Generation (RAG) Pipeline – Part II (https://blog.algoanalytics.com/2024/08/30/implementing-a-basic-retrieval-augmented-generation-rag-pipeline-part-ii/)

August 30, 2024

Implementing a basic Retrieval-Augmented Generation (RAG) Pipeline – Part I (https://blog.algoanalytics.com/2024/08/30/implementing-a-basic-retrieval-augmented-generation-rag-pipeline-part-i/)

August 30, 2024

(https://www.algoanalytics.com/)

(htt ps:// ww (htt ps:/ /w

Follow Us