Large language models for aspect-based sentiment analysis

AUTHORS Paul Simmering, Dr. Paavo Huoviala

8 January

GPT-4 equals the performance of the top specialized ABSA (Aspect Based Sentiment Analysis) models that have been developed until this year, while learning from just six examples. GPT-3.5, when fine-tuned, achieves the highest accuracy currently known.

Artificial Intelligence

6 min read

Share

Large language models (LLMs), like the GPT-3.5 and GPT-4 models that power ChatGPT, offer unprecedented text processing capabilities. They can fulfil a wide range of roles, including those of models designed for specific tasks. Aspect-based sentiment analysis (ABSA) is one such task. ABSA is best explained with an example. When given the review: "The food was great, but the service was slow", the ABSA result is "food: positive, service: negative". It goes beyond classic sentiment analysis by identifying the positive, negative and neutral aspects of a product or service.

ABSA yields valuable insights for market research, so we tested the performance of the new GPT models on the task. Are they better than specialised models?

Model performance is scored by comparing the model's answers to hand-made benchmark datasets of inputs and outputs. We used the SemEval 2014 dataset (Pontiki et al. 2014), which is a commonly used benchmark for ABSA and consists of 7331 laptop and restaurant reviews. Specifically, we tested the joint aspect extraction and sentiment classification task.

The experiment: Zero-shot, Few-shot or Fine-tuned

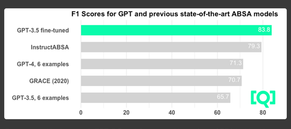
Unlike classic language models, GPT-3.5 and GPT-4 can accomplish tasks without specific training on a large dataset. All that's required is detailed instruction called a prompt. The prompt can just be a description of the task you are expecting the model to perform, or it can additionally contain example texts and the correct answers to them. In AI research, the former situation is called a zero-shot setting, whereas the latter is called a few-shot setting if the number of examples is low. It is also possible to provide the GPT model with a significant number of training examples (here, 5759 example text-correct answer pairs) in order to fine-tune the model to the task. In general, the more examples you give to the model, the better it will learn the task at hand and the better it will perform. On the other hand, performing a task well with zero or few examples is considered a more impressive feat.

We compared the models under three settings:

1. Zero-shot GPT-3.5 and GPT-4
2. Few-shot GPT-3.5 and GPT-4 using hand-picked examples
3. Fine-tuned GPT-3.5 using the training dataset of 5,759 examples

Each of these was evaluated on the test set of 1572 examples. We also tested the impact of variations of the prompt. The performance is judged using the F1 score as a metric. F1 scores can have values from 0 to 100 and are a composite of precision (the percentage of extracted aspects that are correct) and recall (the percentage of aspects that were extracted).

Results: Finetuning the GPT-3.5 model for the task leads to a world record benchmark score in ABSA



Given a good task prompt, GPT-3.5 and 4 managed to perform ABSA but had low accuracy without examples. However, equipped with 6 examples, GPT-4 performed on par with the best specialist models from 2018–2020. These specialist models were fine-tuned on the whole training dataset with 5759 examples.

When GPT-3.5 is fine-tuned in the same way, the results were remarkably good. It reached a world record 83.76 F1 score, outperforming the 2023 specialist model InstructABSA by 4.46 percentage points.

Prompt engineering: GPT-4 instructs itself well

The best prompt proved to be one written by GPT-4 itself. We provided it with the detailed labelling instructions that human analysts used for building the SemEval 2014 dataset and asked it to write a summary of them. This summary prompt worked just as well as the detailed instructions and better than all other prompt variants. The fine-tuned GPT-3.5 didn't need a prompt because it learned the rules implicitly during training.

Economics: Large models are expensive to operate

The performance and ease of use of LLMs come at a cost. LLMs are behemoths compared to specialised neural networks, sporting up to 1000 times more trainable parameters. This is reflected in their operating costs: labelling 1000 ABSA examples costs $15.02 with GPT-4 using 6 examples and less than $0.05 with Instruct ABSA. Running the fine-tuned GPT-3.5 costs $0.36 per 1000 examples. At small scales, these costs don't add up to much, but for pipelines that have millions of reviews, the cost-efficiency of the model becomes critical.

Fine-tuning GPT models is an attractive option, providing greater accuracy than GPT-4 at a lower price. However, it requires a high-quality dataset of examples. For well-researched tasks like ABSA, the HuggingFace hub offers a plethora of open-source datasets one can train on, but when dealing with languages other than English or when domain-specific examples are required, custom labelling may be needed.

Takeaways: Large language models provide previously unseen flexibility and performance highs

Prompt engineering and few-shot learning can bring GPT-4 to an acceptable performance level in a specific task like ABSA. This enables its use even in settings without any training data, though at comparatively high operating costs. Freed from the need for training data, the labelling schema can be changed at will, and projects can get off the ground faster.

On the other hand, if training data is available, world-class performance can be achieved using fine-tuning of the GPT-3.5 model at a lower operating cost than GPT-4.

While we only tested OpenAI's GPT models in the study, we expect that similar results can be achieved with open-source LLMs like Llama 2 and Mistral.

The original paper is available on arXiv.

Artificial Intelligence

Share

Paul Simmering
Data Scientist at Q Agentur für Forschung GmbH

Dr. Paavo Huoviala
Data Scientist at Q Agentur für Forschung GmbH

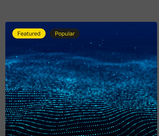
RELATED

23 May 2023 7 min read
by David Smith, Adam Riley

How to survive AI... the skills we need to stay relevant:...

6 November 2023 2 min read
by Ilya Hull

Decoding colour science in market research



15 August 19 min read
by Crispin Beale, Simon Chadwick, Finn Raben, Mike Stevens

Synthetic Data – Get on Board but do it wisely!

ESOMAR NETWORK

ESOMAR Corporate
ESOMAR Directory
Resource Library by ANA
Trust in Research
Foundation
Job Board