9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Text classification through iterative few-shot learning with generative LLMs

Keywords: natural language processing, supervised learning, large language models

Extended abstract

Modern supervised learning techniques have enabled rapid advancement in applications of text data. While much attention is rightly focused on the sophistication of these models, the utility of supervised learning hinges on the quality of manually labeled text samples for training and validation. This stage poses non-trivial human challenges: limited attention, fatigue, and changes in perception of the conceptual categories can cause inconsistencies. This process, supposedly generating "gold standard" data for model training and evaluation, is often carried out by MTurkers or undergraduates. This typical classification workflow poses issues beyond inconsistent data quality: demographically similar annotators may produce correlated, unobserved errors and manual labeling requires extensive resources in cost and time, which can slow the research process and exacerbate researcher inequality. This project proposes an alternative workflow. We investigate the capacity of generative Large Language Models (LLMs) to augment–and, in some cases, replace–human annotators. Our proposed workflow involves an iterative human-in-the-loop feedback system. By harnessing LLMs for classification with iterative feedback, our approach can produce accurate, rapid, reproducible classification at a fraction of the cost and time of traditional approaches. Here, we conduct two pilot studies using OpenAI's newly-released GPT-3.5-TURBO API, which is the same model used in ChatGPT. Our initial results show that our approach achieves similar performance metrics to undergraduate coders as well as state-of-the-art techniques for text classification. Overall, labeling 2,000 text samples ten times each (20,000 total samples) across our pilot experiments cost under 10 dollars and took fewer than two hours in total.

Process. Rather than fine-tuning an LLM for a specific classification task, we instead provide a general purpose generative LLM with task-specific instructions (i.e., an objective and a codebook) and data to code. We build queries in Python, prompting the LLM with instructions and text samples to classify, then concatenate the output labels. The workflow includes five stages: first, creation of classification instructions (i.e., a codebook); second, expert annotation of a small subset of samples to create a "gold standard" for validation; third, LLM classification of provided text samples along the dimensions defined by the provided codebook; fourth, refinement of the instructions based on LLM performance on the validation data–a process we refer to as a human-in-the-loop feedback system; fifth, repeated classification of the desired corpus.

In the fifth stage, we introduce a level of diversity across LLM classifications by utilizing the LLM's "temperature." Each text sample is classified repeatedly across iterations, then we assign the modal answer as the predicted label. With multiple LLM labels for each text sample, we can approximate a degree of "confidence" in each LLM classification. Labels assigned with lower confidence may warrant extra attention (see Figure 1).

Pilot study 1: Sentiment detection Pilot study 1 features a common classification task: sentiment labeling. We employ the IMDB movie reviews dataset, in which the text samples are user-submitted movie reviews and the labels are a binary measure of sentiment (positive or negative). We use the LLM to label a random sample of 1,000 reviews. As shown in Table 1, the results are remarkably accurate. These benchmarks would rank among the top performing models on the HuggingFace leaderboard, although our performance is compared to a random sample of the training set and not the held-out validation set.

Pilot study 2: Domain-specific classification Pilot study 2 replicates a subset of the analyses in Card et al. (2022). In Card et al. (2022), undergraduate coders manually reviewed text samples of political speech for references to immigration. Using this manually labeled data, the researchers then fine-tune a RoBERTa classifier to predict labels in the complete corpus. To replicate this annotation process using our generative LLM approach, we drafted a brief codebook that describes the conceptual category of interest (i.e., references to immigration). To create a "gold standard," we randomly sample 100 text segments labeled by the Card et al.'s undergraduate annotators and had 3 experts (doctoral students in relevant fields) carefully classify the segments. This "gold standard" allows us to compare the performance of the LLM against that of the undergraduate annotators (see Table 2). Prior to performing any codebook refinement, we find that the LLM's performance produces comparable results to the RoBERTa classifer in Card et al. (2022) (see Table 1). It is possible that our human-in-the-loop feedback system to improve the codebook would further enhance the classification performance as well.

Dataset	Model	Accuracy	Recall	Precision	F1
IMDB	Generative LLM	.93	.92	.93	.93
	DistilBERT	.93	N/A	N/A	.93
Card et al. (2022)	Generative LLM	.81	.80	.81	.81
	RoBERTa	.87	.89	.85	.87

Table 1: Generative LLM performance with 1,000 samples labeled 10 times.

Source of predicted labels	Accuracy	Recall	Precision	F1
Undergraduate coders	0.89	0.89	0.91	0.90
Generative LLM	0.88	0.89	0.89	0.89

Table 2: Domain-specific classification performance compared with "gold standard" dataset

References

Card, D, Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R., & Jurafsky, D. (2022). Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences, 119*(31).

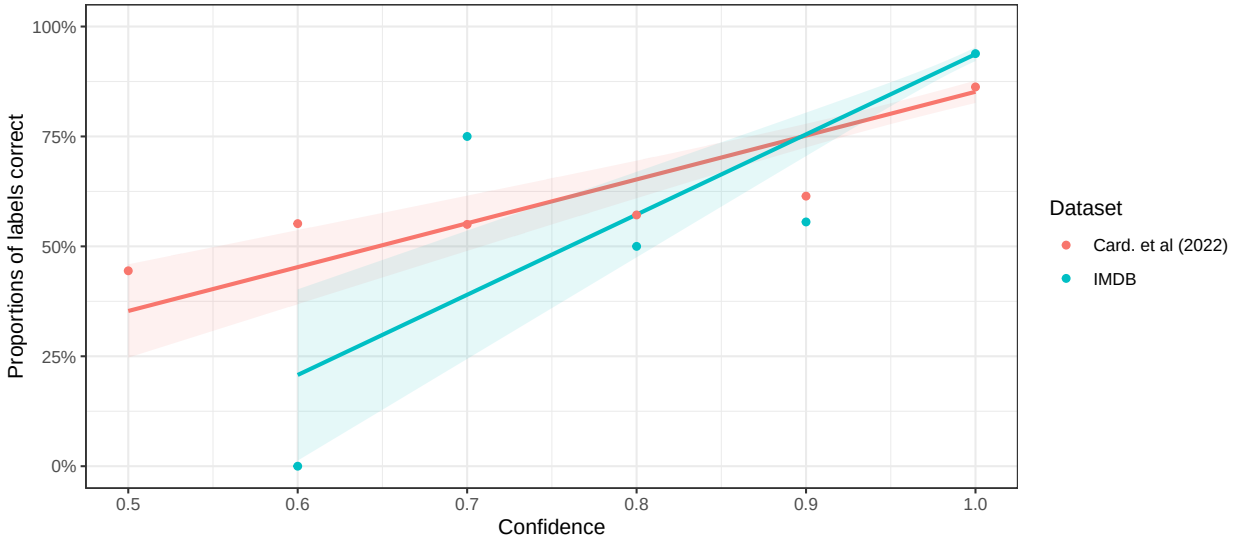

Figure 1: Classification accuracy increases as label confidence increases in both pilot studies.