Generating Datasets with Pretrained Language Models: News Classification

Katie Baldwin, Nada Elfazary, Jason Yuan

Introduction

- Machine learning as a discipline relies on access to data for the machines to learn from. However, this data can be difficult to come by.
- Schick et. al. present a solution to this issue which involves using large language models (LLMs) to generate training datasets, given instructions as prompts. [2] Our work attempts to apply their method of instruction-based dataset generation to a new task – news headline classification.

Background & Motivation

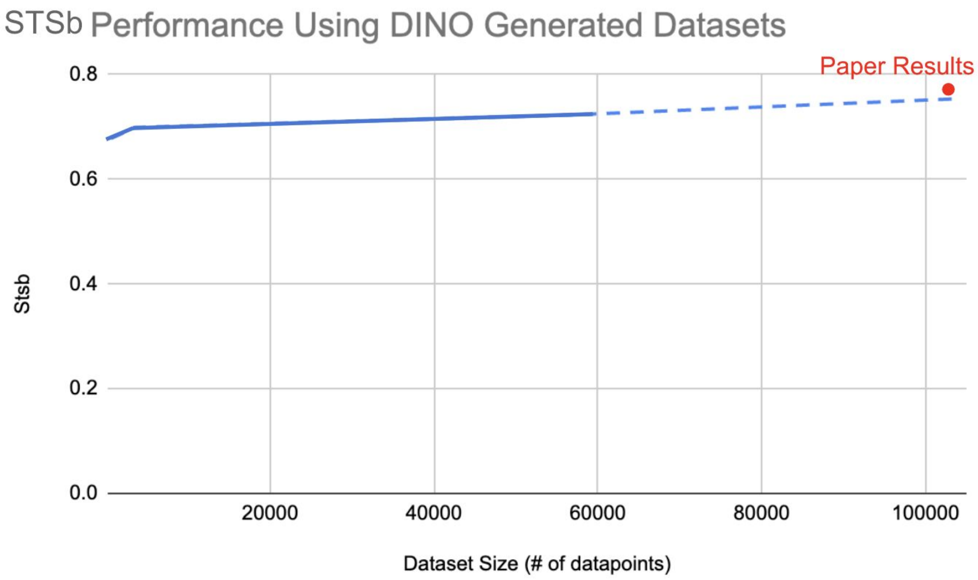
- High-quality datasets for use in training of machine learning models can be expensive and practically difficult to acquire. Models also tend to perform better when trained on supervised data than on unsupervised data, but acquiring supervised datasets is especially expensive.
- However, state-of-the-art language generation models are fairly competent at producing realistic language. Schick and Schütze harness this for generating datasets, creating a method they call Dataset from Instructions (DINO). [2] They generate a textual similarity dataset using LLMs and evaluate it on the STS (semantic textual similarity) task introduced by Agirre et al. [1]

Methods: Reproducing DINO and STS

- STS task: Given two sentences, rate on a 0-1 scale how similar they are.
- DINO Algorithm: An *unsupervised* approach to STS, by generating a dataset with GPT
    a. Essentially prompt GPT with “Generate two sentences that are {the same|similar|different}”
    b. Instead of sampling the highest probability next token, modify to penalize tokens that have high probability under a different label (Equation 1)
    c. Evaluate downstream performance of a BERT model trained on generated dataset
- We reproduced the DINO results, using the instructions they wrote, but on a smaller scale for computational reasons.
- The dotted line represents the projected performance of a model of the same size as theirs; the solid line connects the results we reproduced.

Equation 1

$$\delta_y = p_y - \max_{y' \in \mathrm{CL}(y)} p_{y'}$$



Methods: DINO on a new task

News Classification Task: Given a news headline, classify it under the topic it discusses (e.g. is it about world news or science and technology or…)

1. We use the DINO approach to prompt GPT2-large to generate news headlines about topics corresponding to the labels of the Hugging Face *ag_news* dataset.
    a. We generate two datasets using LLMs - one with “naive” prompts, and one with human-designed prompts (based on empirically seeing what prompts generated the most realistic headlines)
2. We restrict each dataset to size 6.2K to make the comparison fair.
3. We fine-tune a DistilBERT sequence classification model on these generated datasets, as well as the original dataset. We train for 2 epochs, and evaluate every ⅓ epochs.
4. We measure the quality of the training dataset by evaluating the fine-tuned DistilBERT model on the benchmark validation dataset (from Hugging Face, a held out set of size 7K).

Naive prompts:

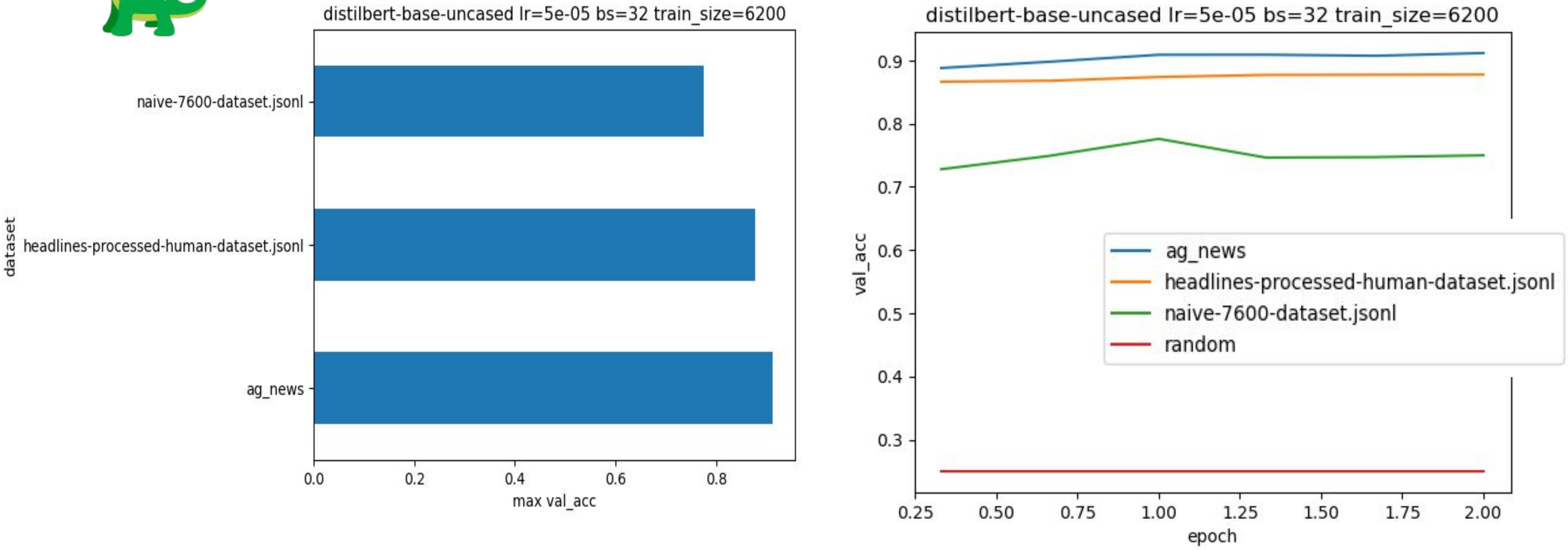
→ Write a **world** news headline.
→ Write a **sports** news headline.
→ Write a **business** news headline.
→ Write a **science** news headline.

Human-crafted prompts:

→ Write a headline about **international incidents**.
→ Write a headline about **international politics**.
→ Write an **Olympic sports** news headline.
→ Write a **national sports** news headline.
→ Write a news headline about **financial institutions**.
→ Write a news headline about **corporations**.
→ Write a **science** news headline.
→ Write a **technology** news headline.

Results

Accuracy Curves for DistilBERT Model



- Both datasets produce fine-tuned results that are decent, demonstrating the potential of this technique.
- The human dataset produces results that are very close to training on the actual dataset.
  - But the performance gap increases as the number of epochs grows
- The naive dataset does significantly worse, and does not consistently get better with more training.

Discussion

An overfitting issue

- We observe that the model overfits on both our datasets but not the real dataset.
  - Evidenced by the loss curves
- We hypothesize that this is due to lack of diverse outputs in generated data
  - Evidenced by Dataset Examples #1
  - Despite all datasets being of the same number of examples, the generated dataset is functionally “less” in size than the real dataset.

Prompting matters

- We observe that improving the quality of the prompts significantly improves the quality of generated data.
- With small changes in our prompting, the generated dataset is much better qualitatively (see Dataset Examples #2) and quantitatively (see the accuracy results).
- But our generated dataset still underperforms in generating enough length and details.
  - Evidenced by Dataset Examples #2

Dataset Examples #1

Naive

- "A new method for detecting cancer cells from a single DNA sequence"
- "A new method for detecting cancer cells using a single drop of urine"
- "A new method for detecting cancer cells using an optical technique"
- "A new method for detecting cancer from DNA in a living organism"

Dataset Examples #2

Naive

- "I just bought some stock."

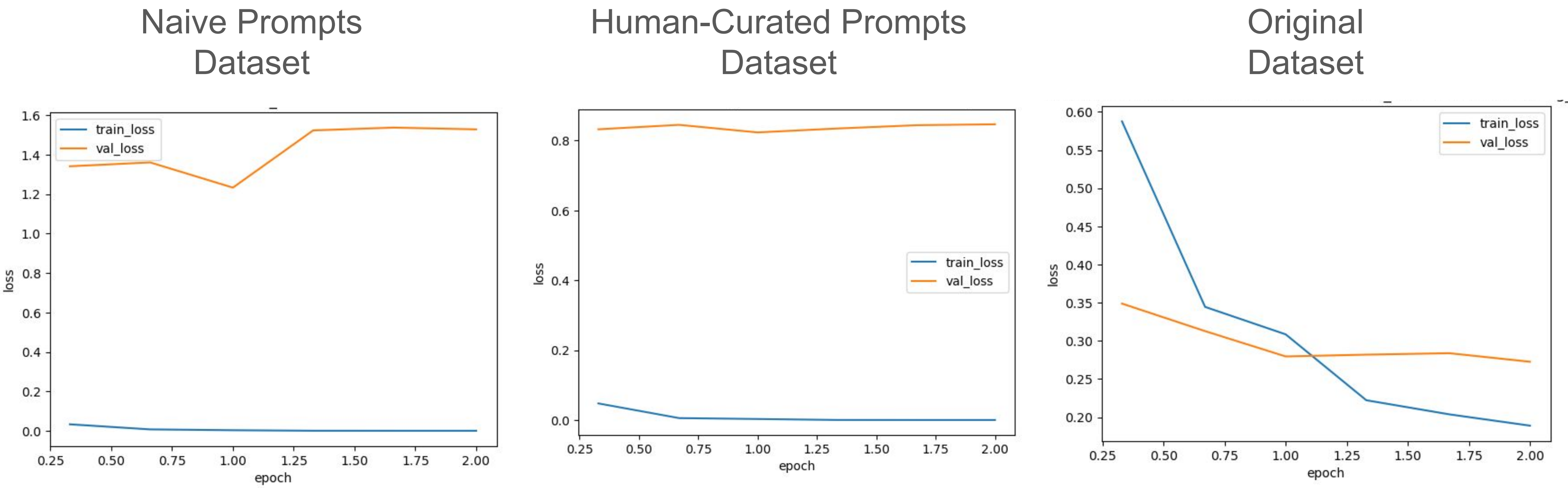
Human

- "FED-Backed Bank Feds Shut Down $2B of Assets"

Actual

- Fed President Calls Oil Troublesome Record oil prices are creating headwinds for the U.S. economy but do not place its recovery at risk, Dallas Federal Reserve Bank President Robert McTeer said on Monday.

Loss Curves for Distilbert Model



Naive Prompts Dataset

Human-Curated Prompts Dataset

Original Dataset

Conclusion

- The model trained on the LLM-generated dataset given processed prompts achieves a performance close to the model trained on the ag_news dataset
- The diversity of dataset examples reduces overfitting
- Human-engineered LLM prompts improve performance model accuracy.

References

[1] Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, page 385–393, USA. Association for Computational Linguistics.

[2] Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'21), Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 6943–6951. https://doi.org/10.18653/v1/2021.emnlp-main.555