Digital Socrates: Classifying Action Statements as (Un)Ethical

Harper Lyon
Tulane University / New Orleans, LA
hlyon@tulane.edu

1 Problem Overview

The idea of ethical machines is attractive, but challenging to implement in practice. One approach is to borrow techniques from sentiment analysis (Hendrycks et al., 2023) to classify everyday descriptions of actions as ethical or not, creating ideally a system which can give human-like responses to human questions about decisions. This problem contains many of the same inherent challenges as sentiment analysis, as subtle differences in language use can shift the meaning of documents massively - consider the difference between ”I will not break my promise even for money” and ”I will not break my promise except for even more money” - but also in that it seemingly requires more modeling of meaning than simple sentiment analysis, since ethical judgements are inherently tied to actions in a way that movie reviews or teaching evaluations are not.

The ”hook” for my project is the creation of a webpage where a digital facsimile of Socrates, the famously wise and annoying Athenian, tells you whether or not what you are planning to do is ethical, but for that I will need a classification model capable of accurately making that decision.

2 Data

My primary data source is the ETHICS (Hendrycks et al., 2023) dataset, which is a dataset combining five semi-distinct text classification tasks related to various areas of ethical reasoning. For my purposes, only two subsections of the dataset are useful, and these are as follows:

Table 1: Dataset Divisions

	Rows	**Classes**
Commonsense	21.8k	0 : 11.49k, 1 : 10.26k
Justice	26.5k	0 : 12.31k, 1 : 14.23k

These two subsets are similar, but not exactly the same. The commonsense dataset consists of two types of document class pairs, the first being straightforward action descriptions like (”I left her bleeding on the snowy hillside.”, 1) and the other being posts sourced from reddit.com/r/AITAH, an internet forum where users post stories from their lives and receive crowdsourced moral judgements. We can question of the wisdom of relying on these judgements, but they are a convinient corpus for this project.

The justice dataset consists of statements of the for I (deserve/did) *action* because *reason*, for example (”I deserve to get a nice haircut from my barber because I paid him to make my hair look nice.”, 1). Some work is required to reconcile these datasets to a single task - mainly switching the labels to match - I see no reason that they cannot be accomplished by the same model or at least by similar models trained in similar ways.

I am also in the early stages of finding additional data sources to augment my training data, more information on those potential sources (Lourie et al., 2021)(Ziems et al., 2022) in Related Work.

3 Methods

My goal is to evaluate three distinct approaches to this task at different ”weights”, in particular

1. Logistic Regression
2. Finetuned BERT
3. Llama

Logistic regression serves as a baseline model, since it will ultimately work through simple word inclusion, but it should give us a good idea, more or less, of how difficult the task is. Otherwise I don't expect it to be all that interesting or play a large part in my final application.

Much of the existing work in this area focuses on BERT based models (Lourie et al., 2021)(Hendrycks et al., 2023), so these are a natural place to begin the real work. There are many BERT varieties, so I expect to try several options (currently I have only worked with distilbert). However, the best BERT based performance so far outside of the commonsense data subset is moderate at best, with ALBERT(Lan et al., 2020) achieving 59.9% on the justice task and 85.5% on commonsense. I don't expect to beat those numbers given my limited resources and time, but it's clear that especially on justice BERT is not up to the task of top notch performance.

What I really want to explore is using an LLM like Llama to generate classifications since this is an area left open by the originators of the ETHICS dataset. I plan to use Llama-2-7b since I am limited by computing power, and that is the most moderate of the new models available. I have already been given access by Meta to the weights for this model, so I will be beginning work here soon.

Evaluation will be straightforward, I am interested in achieving high accuracy and not much else. If I can match the median results of the original paper I'll be happy with my outcomes. I'm also interested in examining the loss variance

4 Preliminary Results

So far I have (on the commonsense subset only) implemented the Logistic Regression model and made a first clumsy pass at finetuning a BERT variant called DistilBERT (Sanh et al., 2020), which I chose for no other reason than that is quick to train as a proof of concept and seems to be the ”standard” BERT variant used for classification absent other considerations. I've achieved extremely poor performance with both models, hitting exactly chance on Logistic Regression (recall that we have a 50/50 class split) for Logistic Regression 1 and actually performing worse than chance with DistilBERT 2. I'm not surprised by the results of logistic regression, but I am disappointed at how poorly DistilBERT performed - there is definitely an overfitting or analogous issue since the model acheived 85% accuracy on the validation split.

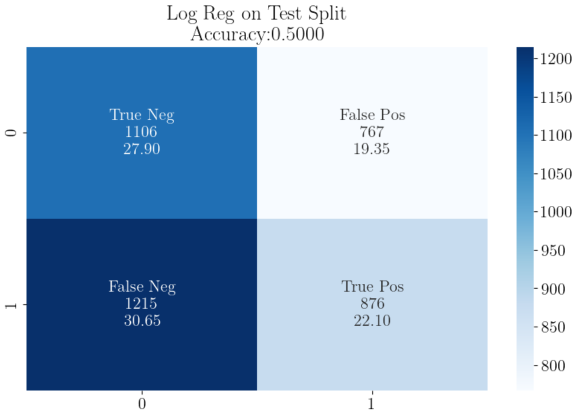

Figure 1: Logistic Regression Confusion Matrix

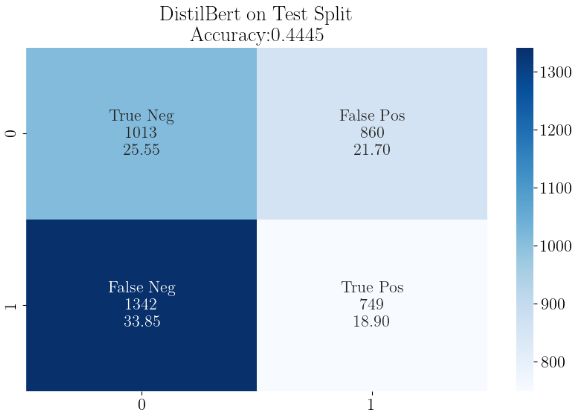

Figure 2: DistilBERT Confusion Matrix

5 Related Work

5.1 BERT Related Work

- **Title:** Aligning AI With Shared Human Values
  Citation: (Hendrycks et al., 2023)
  Description: This is the paper which originated the ETHICS dataset and inspired a lot of this project - to a large extent all I'm trying to do is extend their work slightly. They focus almost entirely on BERT based approaches, especially RoBERTa (see below), so I hope that by taking advantage of improvements in BERTology (as it where) and by introducing LLMs I can make some meaningful improvements to their work.

- **Title:** Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life Anecdotes
  Citation: (Lourie et al., 2021)
  Description: This is a potential source of additional data, which is especially important since I seem to be having finetuning problems. These ”Community Judgements” are also sourced from reddit.com/r/AMITAH, so they should dovetail well with the ETHICS dataset.

- **Title:** RoBERTa: A Robustly Optimized BERT Pretraining Approach
  Citation: (Liu et al., 2019)
  Description: This paper describes many of the finetuning techniques used by Hendrycks et al. to achieve their best performance on ETHICS, so I'm hoping that I can use RoBERTa, and their finetuning techniques more broadly, to improve my abysmall current DistilBERT accuracy.

5.2 LLM Related Work

- **Title:** The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems
  Citation: (Ziems et al., 2022)
  Description: I'll admit, this paper is more for me than for this project specifically. They propose a Seq-Seq moral reasoning process/prompt engineering technique which injects essentially attacks on the LLM's moral reasoning in the form of sentences like ”You should not judge people negatively based on race” to test and improve output. I'm still not sure if any output beyond classification from Llama is going to play a role in my final product, but if so I do think there are some valuable insights here.

- **Title:** Language Models are Few-Shot Learners
  Citation: (Brown et al., 2020)
  Description: This paper describes a variety of techniques for adapting LLM's Seq-Seq functioning for other tasks including text classification. This is more or less the paper on doing so, so it should be helpful as I move into the third stage of the project.

6 Timeline

I plan to hold to the following rough timeline of tasks:

Task	**Estimated Completion**
Install Llama-2-7b, begin fine tuning	3/29
Tune Additional BERT models	4/4
Determine best model for web integration	4/11
Implement ”Digital Socrates” page	4/16
Complete presentation	4/20
Complete report	4/28

Table 2: Timeline

With some expected slop and slippage here and there.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2023. Aligning ai with shared human values.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Caleb Ziems, Jane A. Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems.