

Philosophize That: Using Document Embeddings to Predict Philosophical Theory

Lucas Dionisopoulos

University of California San Diego

ldionisopoulos@ucsd.edu

Abstract

Deep learning has unlocked significant capabilities in the field of natural language processing, particularly with the combination of scale and the transformer architecture. One of the domains that has benefited from this progress is that of embedding models, which have seen performance improvements on down-stream tasks.

This paper leverages a popular open-source text embedding model to analyze documents from several schools of philosophical thought. We are able to generate meaningful embeddings for philosophical texts across a range of writing styles, allowing for dimensionality reduction visualization and classification of unseen text into philosophical categories¹.

Additionally, by supplementing our training data with synthetic data, we are able to improve classification performance on the evaluation task.

1 Introduction

Built upon prior work on document embeddings (Raghavan and Wong, 1986) and neural language modeling (Bengio et al., 2000), neural document embeddings have taken advantage of deep learning to attain downstream capabilities (Le and Mikolov, 2014). Notably, training embedding models through the objective of language modeling has paved the way for techniques such as ELMo (Peters et al., 2018), BERT (Devlin, 2018), and their variants, which have further expanded the capabilities of embedding models.

Specifically, BERT and BERT-variants leverage the transformer encoder architecture (Vaswani,

2017) with a masked language modeling objective to generate semantically-rich embeddings at the word and text-level (e.g., sentence-level, document-level).

Combining the efficacy of modern embedding models with distillation techniques (Sanh, 2019), small yet powerful embedding models can run efficiently on consumer-grade hardware. Specifically, MiniLM (Wang et al., 2020) and open-sourced derivatives offer highly-performative, lightweight text embedding models that can be effectively applied on consumer-grade hardware.

Leveraging these capabilities, this work:

- Effectively embeds a variety of philosophical texts from select schools of thought using open-source text embedding models.
- Trains a classification model that categorizes text into philosophical schools of thought.
- Improves classification on a custom evaluation task through supplementation with synthetically generated data.

2 Related Work

Application of Document-Level Embeddings:

Document embeddings have been used for a variety of downstream tasks, notably passage retrieval (Karpukhin et al., 2020) within retrieval-augmented generation (RAG) settings (Borgeaud et al., 2022). Additionally, recent work has used document-level embeddings for de-duplication of image datasets (Abbas et al., 2023). Further, document-level embeddings have been effective in domain-specific settings such as analyzing research papers – notably SPECTER2 (Singh et al., 2022), which is effective for downstream tasks of classification, regression, search, and proximity tasks.

¹Full repo available at https://github.com/lucasdino/philosophical_oracle, Colab for playing around at <https://tinyurl.com/philosophizethat>

NLP in Philosophy: The application of NLP techniques in the broader field of philosophy remains nascent. Some recent work explores leveraging large language models (LLMs) as pedagogical tools for philosophy (Smithson and Zweber, 2024). Another paper used expert-labeling of a set of philosophers across 8 philosophical dimensions. From this, correlation and dimensionality reduction techniques were applied (Fabbri et al., 2010).

Thus, applying NLP techniques to the field of philosophy offers an understudied direction for research.

3 Datasets

Similar to (Fabbri et al., 2010), this analysis identifies a set of foundational philosophies to use for analysis. The goal is to identify a set of ideologies that represents a spanning basis of philosophical views to allow for holistic classification.

The final set of philosophies chosen includes Absurdism, Buddhism, Epicureanism, Existentialism, Rationalism, Stoicism, and Transcendentalism. These were chosen as they each offer some unique dimension of belief – though we acknowledge that there exists overlapping themes across these schools of thought.

For the data, several different sources were leveraged:

- **Source Texts:** Seminal works from philosophers of these schools of thought, sourced from Project Gutenberg and other publicly accessible data sources (Project Gutenberg, nd).
- **Wikipedia:** Relevant Wikipedia sections on each specific school of thought (Wikipedia contributors, nd).
- **Philosophize This Transcripts:** Transcripts from the Philosophize This Podcast related to each school of thought (West, nd).
- **Synthetic Data:** Philosophical summaries from GPT-4o (Achiam et al., 2023) and automated creation of 'philosophy quizzes' from Llama 3.2 (3B) (Dubey et al., 2024).
- **Evaluation Data:** Comments manually scraped from Subreddits and other forums related to each philosophy.

Detail is provided in Appendix A related to the philosophers and source texts used. For all sources, pre-processing included both manual and scripted text processing. Some common cleaning tasks encountered included removal and standardization of formatting characters, links and citations, and other auxiliary metadata within files.

3.1 Data loading

When loading in our data, we utilize Langchain's `RecursiveCharacterTextSplitter` to split our document into smaller chunks that are then passed into the `SentenceTransformer all-MiniLM-L6-v2` model for embedding. The key hyperparameter experimented with was the chunk size – 2,500 characters was found to perform the best.

Additionally, when loading in data we further apply a data balancing function to account for discrepancies in the sizes of our various datasets. There is a hyperparameter that can tune the maximum multiple of samples from one dataset compared to the smallest, and ultimately we used 1.0 for this value.

3.2 Data diversity

Initial tests solely using source texts displayed poor generalization. This was likely due to the model overfitting on writing style – for example, Transcendentalist texts are much closer to modern writing style than Stoic texts. To counteract this, there is a necessity for data diversity – ideally composed of diverse writing styles across history.

To address this, three primary methods were used to capture philosophically dense yet stylistically similar sources for each of our schools of thought: Wikipedia, Philosophize This transcripts, and synthetic data. All three offer the benefit of modern text in different styles, which allowed us to improve performance on our evaluation task.

More detail on synthetic data generation is provided in Appendix B. It is worth noting that synthetic data generation was prone to low-quality samples. Recent work has found synthetic data generation performs better on tasks with low subjectivity (Li et al., 2023) – unfortunately, this task does require a high degree of subjectivity. Regardless, the inclusion of the synthetic data improved evaluation performance, highlighting the value of diverse data sources.

3.3 Evaluation data

Initially, we used a held out subset from our training data as evaluation data. However, this quickly yielded accuracy in-line with training accuracy (+90%). We determined that the most effective evaluation should measure performance of a downstream task of interest: Using our classification model to label modern texts.

To generate this dataset, we searched through several subreddits to manually scrape comments of modern practitioners of each philosophy. Ideally, this will allow us to gauge how well our classifier understands the semantic meaning of each philosophy – not just the stylistic writing pattern. Additionally, this directly measures a desired downstream capability in classifying modern philosophical text (e.g., journal entries, essays, etc.). See Appendix C for detail on the evaluation samples.

4 Approach

Our approach begins by loading in our training data, chunking it, embedding it, and conducting data balancing.

To visualize the quality of our dataset embeddings, we conducted principal component analysis (PCA) on the mean embedding computed across each training sample. Appendix D has a 3-D visualization of these mean embeddings (all human samples and GPT summaries). Figure 1 shows a 2-D visualization of the mean embedding for each training data sample after passing it through the first fully connected layer in our trained classifier model.

To train our classifier model, we leveraged a simple 2-layer fully-connected feed forward neural network with hidden dimensionality of 64, ReLU activations, and dropout of 0.35. Our objective function uses cross-entropy loss and we train for 10 epochs.

5 Classification Results

Table 1 displays our classification results compared against an array of dataset splits. It is worth noting that training accuracy was often in excess of 90%.

- **Split 1:** Using purely synthetic data outperforms human summaries and source texts, though performance against human summaries (e.g., Wikipedia, Philosophize This)

Table 1: Evaluation Performance vs. Training Split

Split	Data Used	Loss	Accuracy
Split 1	source texts	1.857 ± 0.058	0.258 ± 0.036
	human summaries	1.583 ± 0.060	0.419 ± 0.065
	synthetic data	1.515 ± 0.044	0.431 ± 0.044
Split 2	human summaries	1.583 ± 0.060	0.419 ± 0.065
	+ source texts	1.719 ± 0.124	0.442 ± 0.037
Split 3	human summaries	1.583 ± 0.060	0.419 ± 0.065
	+ GPT summaries	1.469 ± 0.060	0.482 ± 0.054
	+ synthetic game	1.289 ± 0.074	0.508 ± 0.043
	+ source texts	1.910 ± 0.098	0.408 ± 0.029

Note: Data presented as mean with 1 standard deviation. Statistics computed over 20 training runs.

is within a single standard deviation.

- **Split 2:** The benefit of adding source texts to human summaries is mixed; it improves accuracy but comes with increased loss.
- **Split 3:** Our best performance occurs when we train using human summaries and both of the synthetic datasets – avoiding the source texts in total.

However, simply comparing results on the evaluation set conveys that training on the source text is not beneficial. One the other hand, a core goal of this project was to build a classifier that could infer an accurate distribution of philosophical beliefs when provided unseen text.

For this, we shall use an essay written by the author about his reasons for leaving investment banking to take a gap year, study machine learning, and start graduate school (Dionisopoulos, 2024). When running the classification model against this essay using the optimized parameters, the performance was poor and over-attributed to a specific school of thought. However, when training using all data, the performance was interesting and insightful. In other words, the evaluation set is likely bad – see next section.

6 Error Analysis

There were several issues encountered and addressed during this project, and there are several others that could be further improved upon.

Issues addressed during the project:

- **Overfitting to Style:** The first training of the classifier overfit to style – essentially classifying any modern text as that of Absurdism, Existentialism, or Transcendentalism.

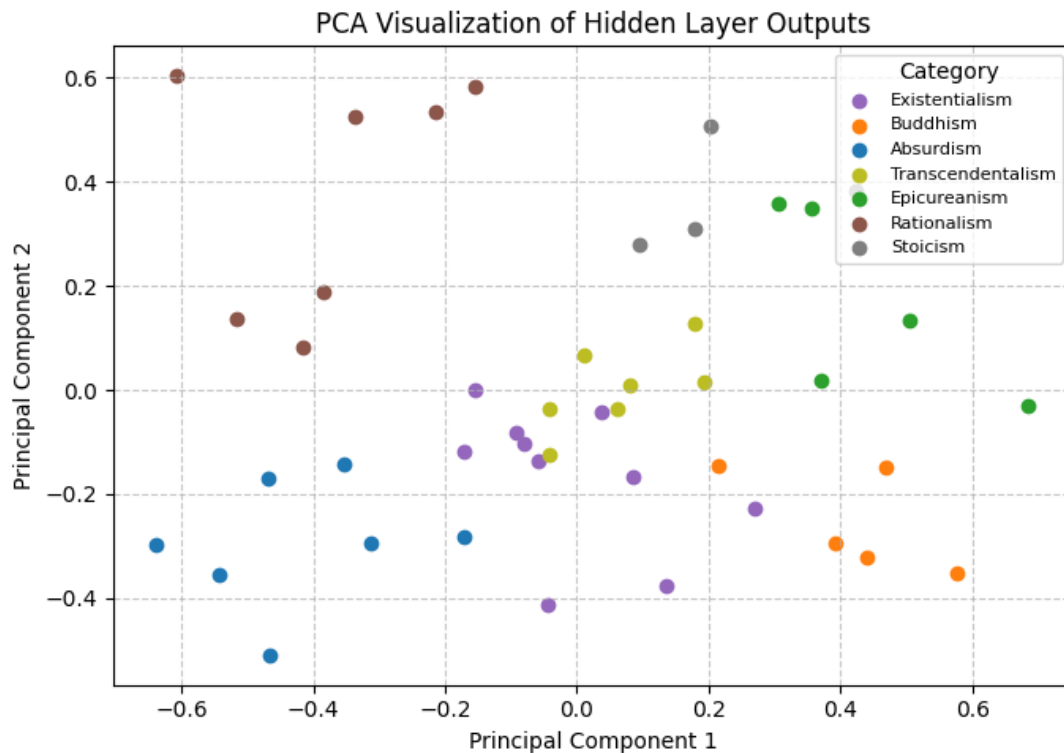


Figure 1: 2-D visualization of hidden-layer values on training data.

To counteract this, several diverse datasets were sourced – providing reasonable success.

- **Using Llama to Clean Data:** Initially, the hope was to use a locally-ran version of Llama to clean up datasets with the goal to also use this to summarize each source document into different styles – that way, we could create diverse datasets using only the source data. However, the quantized versions of Llama used were very fickle and commonly returned poor results. This approach was ultimately replaced by the other forms of synthetic data generation.
- **Determining Set of Philosophies:** Initially the set included philosophies such as Confucianism, Taoism, and Effective Altruism. However, these yielded unsatisfying results; Effective Altruism was over-represented in modern text inference, likely because the nuance that truly differentiates Effective Altruism from 'being a good person' wasn't being captured by the model. For Confucianism and Taoism, they failed to be differentiated from others, likely due to their broad wisdom and less stringently defined code of ethics.

The primary items that would heavily benefit

from improvement are train and evaluation data. The goal was to get philosophically dense but stylistically diverse data samples for training. This is difficult and aggregating the correct sources is time-consuming and requires expertise.

Regarding evaluation, effective benchmarks would likely require expert scoring over texts spanning various styles and time periods, providing a distribution of philosophies for each passage. Specifically a distribution is likely necessary as many times it is not absolute. Two identified issues with our evaluation set are that, first, evaluation samples are significantly shorted in length than the training samples – single sentences or paragraphs vs. full chunks (2,500 characters). Second, sourcing modern interpretations of philosophies such as Rationalism and Transcendentalism are very difficult; while arguably still quite relevant, they don't have the same defined active communities.

7 Conclusion

This project was incredibly interesting to work on and has created more questions than it has answered.

Results on personal writing were surprisingly accurate and insightful – *though like a horoscope, this may be fueled by confirmation bias.*

That said, using natural language techniques to analyze philosophical texts presents an interesting approach. For example, you could use dimensionality analysis to generate a spanning set of underlying ideologies – similar to the work on Philosometrics (Fabbri et al., 2010). You could apply a classifier on historical data – for example war journals – to track how personal philosophies were impacted by war. For a modern consumer application, you could amplify a journaling app: For the novice, you could help categorize their beliefs or offer reading suggestions; for the expert, you could use it to track how their beliefs have changed over time. Lastly, you could couple this with LLMs to conduct large-scale studies on philosophical beliefs across entire populations.

"The unexamined life is not worth living"
- Socrates

References

- Abbas, A., Tirumala, K., Simig, D., Ganguli, S., and Morcos, A. S. (2023). Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Alteschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, 13.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lepiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.
- Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dionisopoulos, L. (2024). Why pivot? <https://lucasdino.notion.site/Why-Pivot-7736665e93394c3886f78950ca66f7d2?pvs=4>.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Fabbri, R., Oliveira Jr, O. N., and Costa, L. d. F. (2010). Philosometrics. *arXiv preprint arXiv:1010.1880*.
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.
- Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.
- Li, Z., Zhu, H., Lu, Z., and Yin, M. (2023). Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849*.
- Ollama (n.d.). Ollama. <https://ollama.com/>.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations.
- Project Gutenberg (n.d.). Project gutenberg. <https://www.gutenberg.org>.
- Raghavan, V. V. and Wong, S. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for information Science*, 37(5):279–287.
- Sanh, V. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Singh, A., D’Arcy, M., Cohan, A., Downey, D., and Feldman, S. (2022). Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.
- Smithson, R. and Zweber, A. (2024). Reviving the philosophical dialogue with large language models.
- Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.
- West, S. (n.d.). Philosophize this! <https://www.philosophizethis.org/>.
- Wikipedia contributors (n.d.). Wikipedia. <https://www.wikipedia.org>.

A Philosophers and Works Referenced

Below are the philosophers chosen for each school of thought and the source texts used:

- **Absurdism:** Albert Camus (*The Myth of Sisyphus*, *The Stranger*).
- **Buddhism:** Ernest Bowden and Sir Edwin Arnold (*The Essence of Buddha*) Mark Siderits (*Buddha*, *Stanford Encyclopedia of Philosophy*).
- **Epicureanism:** Epicurus (*Letter to Menoceus*, other fragments of text).
- **Existentialism:** Simone de Beauvoir (*Ethics of Ambiguity*), Søren Kierkegaard (*The Present Age*), Friedrich Nietzsche (*Thus Spoke Zarathustra*).
- **Rationalism:** Gottfried Leibniz (*Theodicy*), Baruch Spinoza (*Ethics*).
- **Stoicism:** Marcus Aurelius (*Meditations*).
- **Transcendentalism:** Henry David Thoreau (*Walden*), Ralph Waldo Emerson (*Self-Reliance*).

B Synthetic Data Generation

Synthetic data generation took two forms:

GPT Summaries

Summaries were generated from prompting GPT-4 with the following (example provided for Stoicism):

You are philosopherGPT, a philosophy professor who is sage in the ways of summarizing entire schools of philosophical thought into their core themes.

Please provide a summary of the Stoic school of thought. Please do not be concise - elaborate as much as is necessary.

The resulting generations were long-form summaries breaking down the various dimensions of each school of thought. This was conducted for specific schools of thought as well as for specific philosophers (e.g., Kierkegaard, Beauvoir, etc.).

Philosophy Quizzes

In order to generate stylistically diverse yet philosophically dense data, we ran a quantized version of Llama 3.2 (3B) (Dubey et al., 2024) locally using Ollama (Ollama, nd). The prompt was of the following form:

Let's play a game. You are a philosopher of {philosophy}. However, you are going to pretend you are a {style} who deeply believes in these ideas. I am a philosophy student - and you are trying to see if I can guess what your true philosophy is. Do not tell me your philosophy. Do not tell me names of famous philosophers in this category. The philosophy expert should be able to tell who you are, but it should be difficult for me. You may do this through a story, a sermon, or a diary entry. Do not give me a riddle. Once I say go, you will be fully in character - don't break character! Go!

where the philosophies were each of our philosophies and style referred to character types such as 'Roman poet', 'Victorian writer', 'beatnik writer', and 'modern technologist'. The goal was to generate philosophically dense yet stylistically diverse samples. A sample from this dataset is the following:

A young seeker of wisdom, eager to learn from one who has walked the path of contemplation for many years. Very well, let us sit beneath the shade of this ancient tree and engage in a conversation that may reveal the depths of your understanding.

Come, tell me, have you ever stopped to consider the fleeting nature of desire? How often do we find ourselves consumed by the things that promise to bring us happiness, only to discover that they are ultimately unsatisfying? I recall a fellow traveler who, like many before him, was blinded by the allure of sensual pleasure. He sought it in all the wrong places— in fleeting moments of joy, in the company of others, even in the pursuit of material wealth. But true happiness eluded him, leaving only a hollow sense of discontent.

And yet, I see that you are not one to be easily swayed by the whims of desire. You possess a quiet confidence, a stillness that speaks to a deeper understanding of the nature of reality. Tell me, how do you think this fellow traveler might have been freed from his grasp on such fleeting pleasures?

The answer to the above is 'Buddhism' as it focuses on the insatiability of desire. 177 samples like the above were generated and we manually pruned instances where the model clearly missed the instructions. There were varying degrees of quality in responses – highlighting the need for further study of synthetic data generation in future work.

C Evaluation Data

Below are a couple examples of evaluation data:

If you mean that once you accept nothing matters you can experience everything else with the freedom of creating your own meaning and living in the moment, then yes, this absolutely resonates with me. (**Absurdism**)

It is so focused on personal advancement and obtaining more things, a better job, or a better house. Many still choose to reject this construct— anyone who gets out in nature, a hiker or rock climber perhaps. The surfer, however, has a special itch to get back out to the ocean. Surfing has provided a rebirth of this living. It forces men and women to follow their inner self, reject society's systems and chase waves. (**Transcendentalism**)

We are often forced to choose between pleasure now/pain later, or the opposite, pain now, but pleasure later. We can try to fully enjoy the pleasure of the present, "because tomorrow we die!", but the cost may be a lifetime of poverty. The reverse also applies. The pain of getting a degree and student loan, can result in the lifetime pleasure of a satisfying career. Or of course, it may also result in even more misery: not worth it at all! It is so hard to be sure (**Epicureanism**)

D PCA Visualization of Embedding Data

Figure 2 is the dimensionality-reduced embedding data. In interactive 3-dimensions, this data is fairly separable – signifying that the embedding model is able to distinguish the semantic meaning behind the different texts. Note that the datapoints included in this analysis are source texts, Wikipedia summaries, Philosophize This transcripts, and GPT-4o summaries.

PCA Visualization of Source Texts

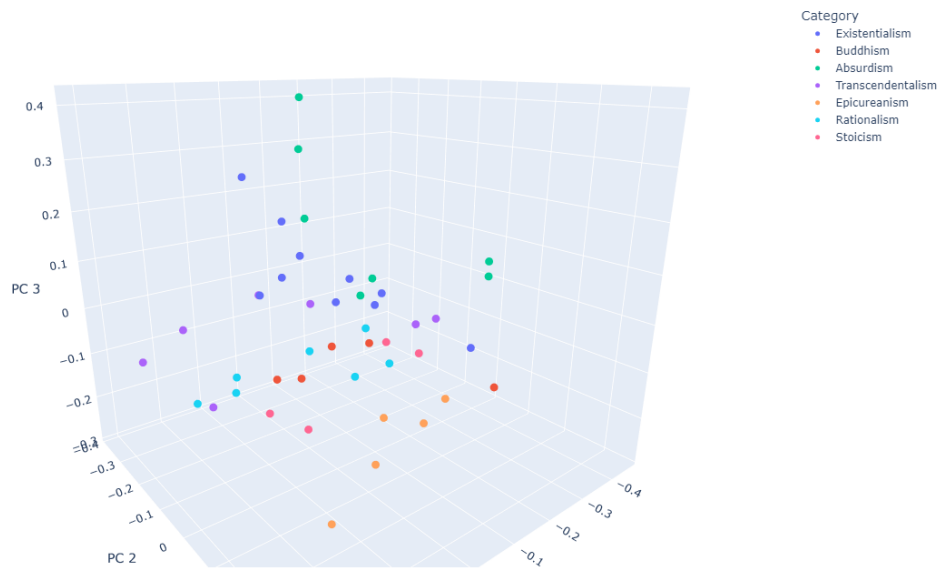


Figure 2: PCA on mean embedding of training data