What the text??

Using deep learning to identify AI-generated text content

R. Amzi Jeffs

Salil Singh

Junichi Koganemaru

Ashwin Tarikere Ashok Kumar Nag

Motivation: AI-generated text is everywhere, but not always easy to spot

We want to combat misinformation, false sentiment, and plagiarism in various contexts.

- Social media
- News websites
- Product reviews
- Public comments on policy
- Homework submissions
- Etc.

Our dataset:
10,000 human-generated and 10,000 AI-generated text snippets in various contexts from various models

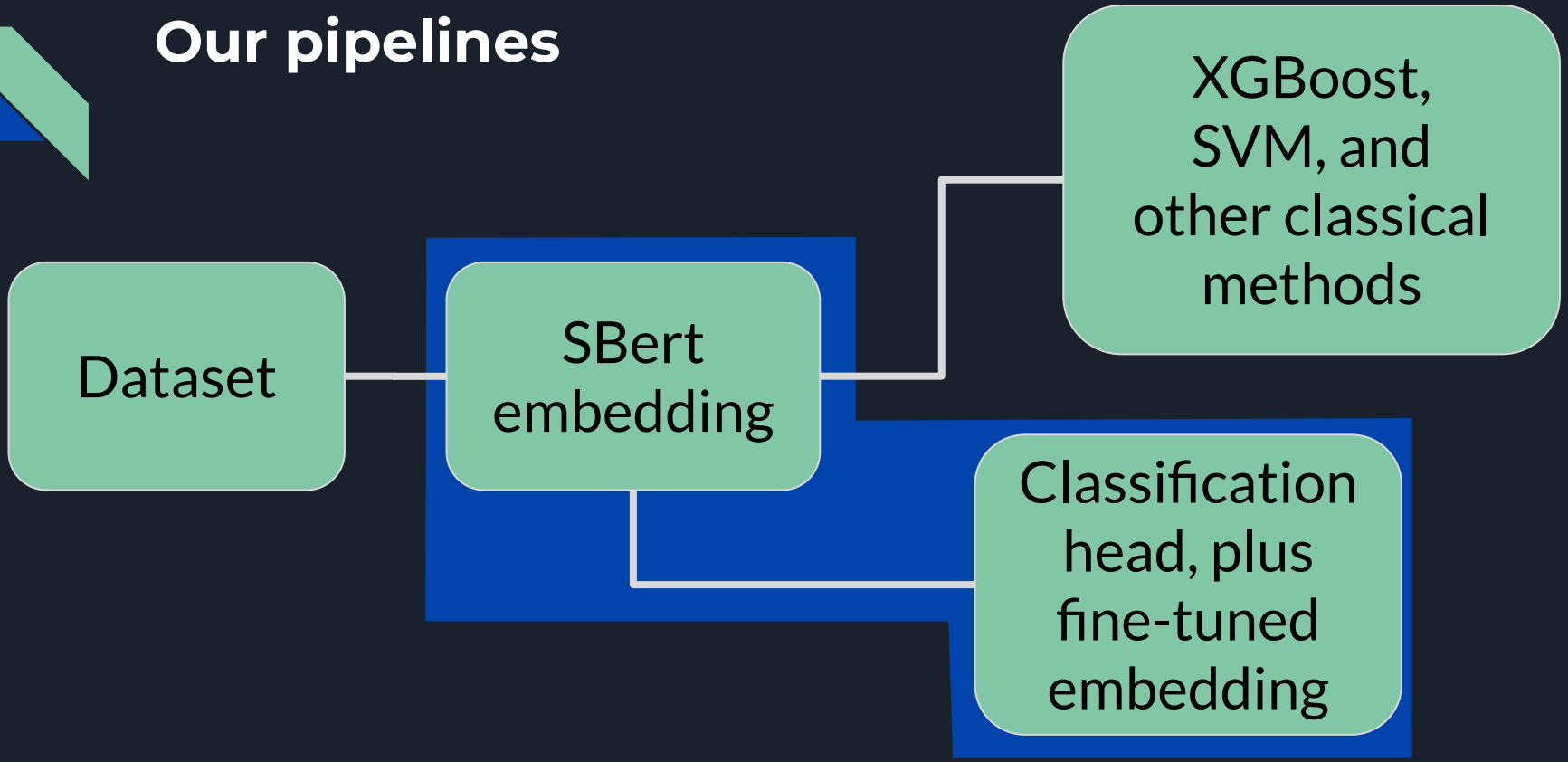
Product reviews (GPT-2)
https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset

Wikipedia intros (GPT-3 Curie)
https://huggingface.co/datasets/aadityaubhat/GPT-wiki-intro

News articles (Grover)
https://github.com/rowanz/grover/tree/master

Essays (Various models)
https://www.kaggle.com/datasets/thedrcat/daigt-v2-train-dataset

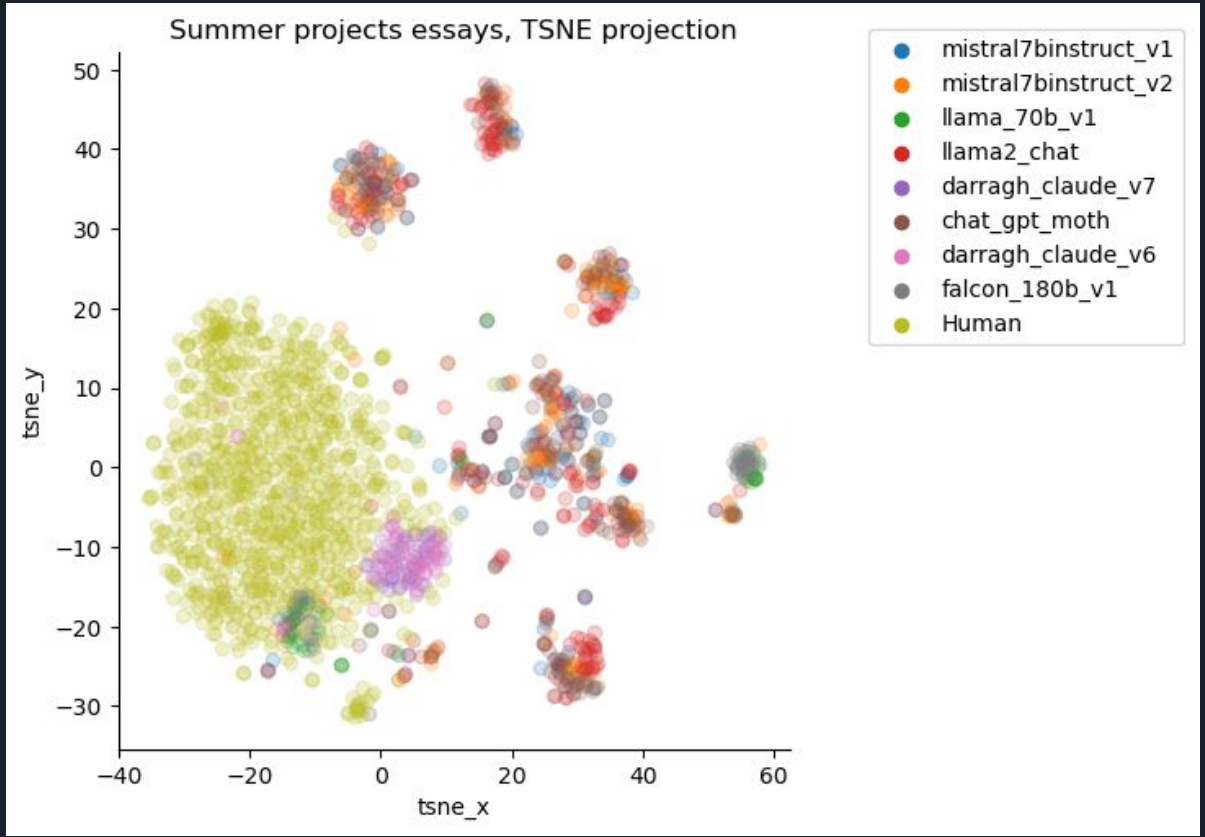
Our pipelines



Red flag!

Baseline model had 95%+ accuracy on essays data.

Through EDA, discovered issue was repetitive prompts, leading to artificial clustering of embeddings by model.



Our dataset:
40,000 human-generated and 40,000 AI-generated text snippets in a variety of contexts from a variety of models

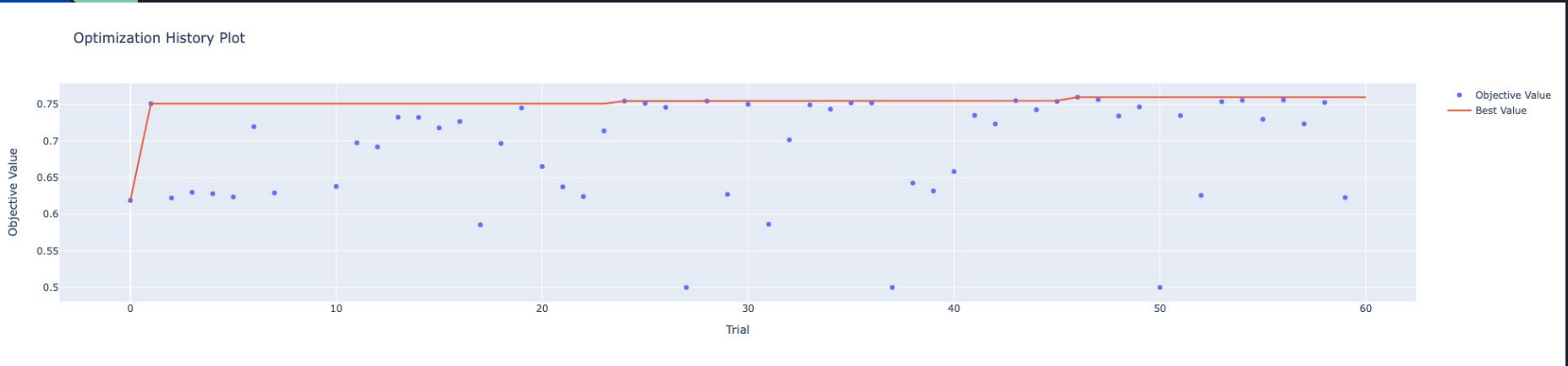
Product reviews (GPT-2)
https://www.kaggle.com/datasets/mexwell/fake-reviews-dataset

Wikipedia intros (GPT-3 Curie)
https://huggingface.co/datasets/aadityaubhat/GPT-wiki-intro

News articles (Grover)
https://github.com/rowanz/grover/tree/master

~~**Essays (Various models)**
https://www.kaggle.com/datasets/thedrcat/daigt-v2-train-dataset~~

XGBoost with hyperparameter optimization (Optuna)

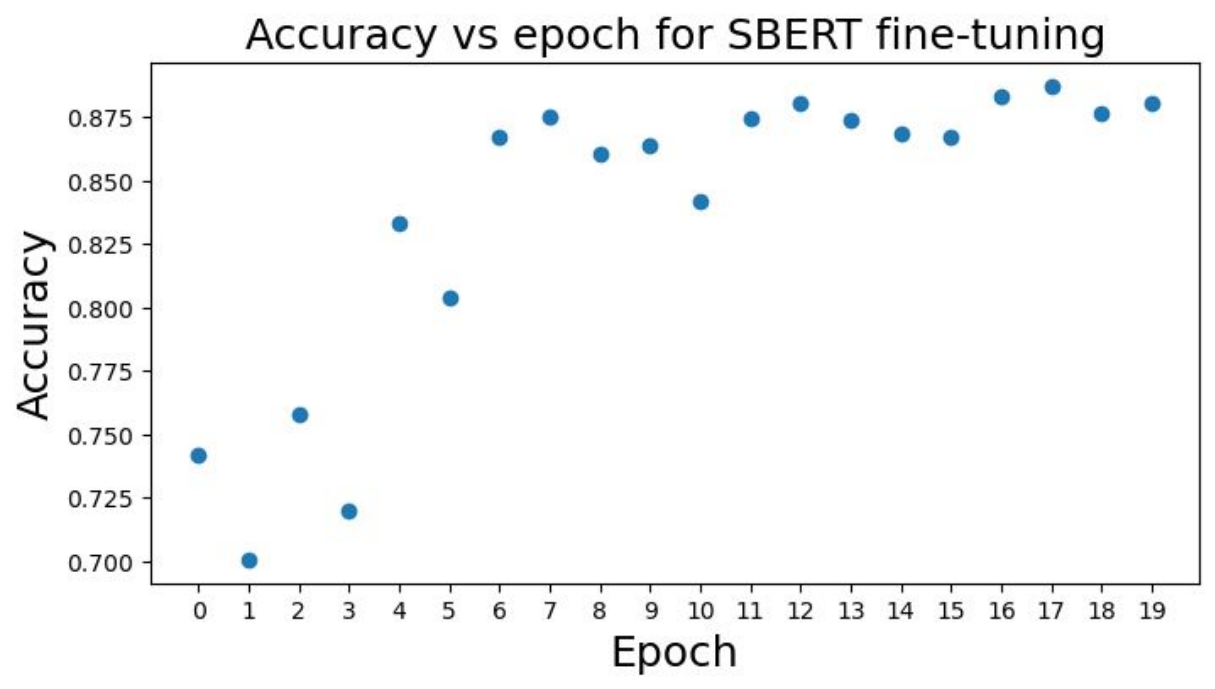


After dozens of hours of hyperparameter tuning, XGBoost maxed out at 75% accuracy. Unstable hyperparameter results.

Fine-tuned SBERT + classifier head

Frozen SBERT + optimized linear classification layer: 69% accuracy after 100 epochs (worse than XGBoost)

Fine tuned SBERT + classification layer: 92% accuracy.



Conclusion

- Fine-tuning the entire model (including SBERT) is necessary for optimal performance
- Broader data is likely needed to improve performance
- Challenges to collecting clean and useful data
- Would need to update our models to account for new LLM models as they are released

Future work

- SBERT has embedding models that can handle both text and image data - incorporating images can potentially be more useful
- Building software widgets to incorporate detection models
- Few shot methods can also be attempted to speed up training
- Should test how well our models can generalize

Thank you Erdos Institute mentors, teachers, and collaborators!

https://github.com/jkoganem/fakereview

R. Amzi Jeffs
Seeking: remote job opportunities in machine learning & data science
Contact: amzijeffs0@gmail.com; www.linkedin.com/in/amzi-jeffs/

Junichi Koganemaru
Seeking: job opportunities in machine learning & data science
Contact: jkoganem@andrew.cmu.edu; www.linkedin.com/in/junichi-koganemaru/

Salil Singh
Seeking: job opportunities in technology and quantitative analysis (data science and other modalities)
Contact: salils@andrew.cmu.edu; www.linkedin.com/in/salil-singh/

Ashwin Tarikere Ashok Kumar Nag
Seeking: job opportunities in statistical analysis and machine learning
Contact: ashwintan1@gmail.com; www.linkedin.com/in/ashwin-tarikere/