THE UNIVERSITY OF CHICAGO

Transformers for Human vs. GPT Text Identification

Wesley Janson Piper Kurtz Sam Pavlekovsky

Can NLP models classify human and GPT-created text?

We have seen the introduction of generative machine learning to the public lead to an explosion in widespread use, seen in photo generation on social media, deep fake videos and audio, and the application of large language models for text generation in a variety of contexts, personal and professional. Implementation of these models for malicious purposes, or in ways that can lead to unintended consequences, is an important risk for policy makers to consider and proactively create legislation preventing. Our goal is to construct a language model that is able to accurately classify a string of text as human or computer generated. We deploy the following methods:

1. **Bag-of-Words:** The BOW approach presented a baseline for machine analysis of the task, with the hopes of interpretability. It performed relatively well, getting results as high as 79% accuracy.
2. **Convolutional Neural Network (CNN):** We chose to implement a CNN as they are extremely skilled at local feature extraction, which allowed us to determine if specific sequences of k-grams identify machine-generated text. We found the model highly successful, returning up to 98% accuracy after 20 epochs.
3. **Bidirectional Encoder Representations from Transformers (BERT):** We implemented BERT for classification to capture global context, as well as contextualized word embeddings, which allowed for analysis of how different combinations of k-grams might indicate machine-generation. We found the model highly successful, returning up to 98% accuracy after 5 epochs.

Our Data

We utilize a dataset of 150,000 paired Wikipedia article and GPT-generated introduction paragraphs via Hugging Face. GPT was fed the first seven words of real articles with the prompt "200 word Wikipedia style introduction on <Article Title>". Our preliminary examination and cleaning of the data revealed two concerns:

- **Poor responses:** GPT was prone to factual errors, struggled with some types of prompts e.g. disambiguation pages, and occasionally had repetition. However, we deemed that this does not seriously interfere with the classification task, since it is based more on identifying words and phrases.
- **Inconsistent lengths:** Both human- and GPT-written introductions varied widely in length, so we limited model inputs to random short excerpts from the samples.

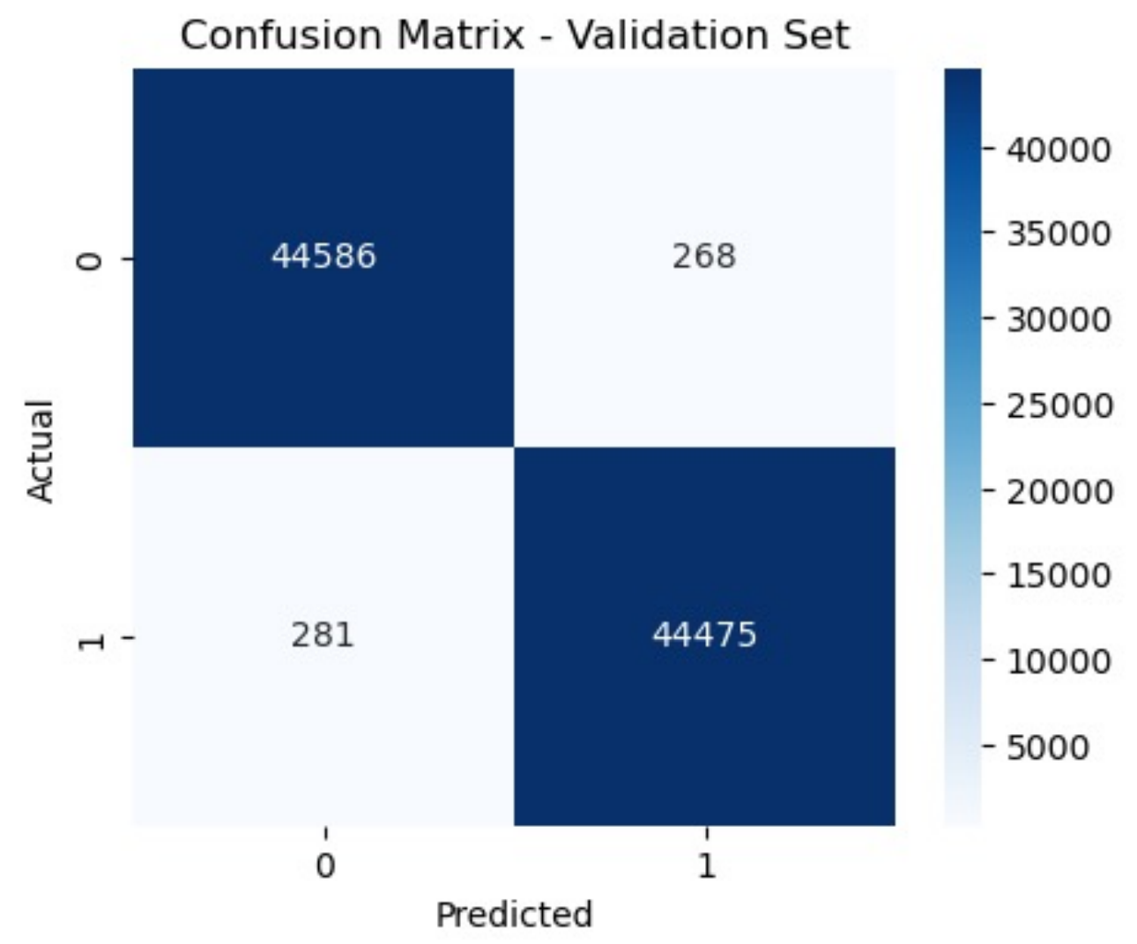
Baseline Comparisons: Human and Bag of Words

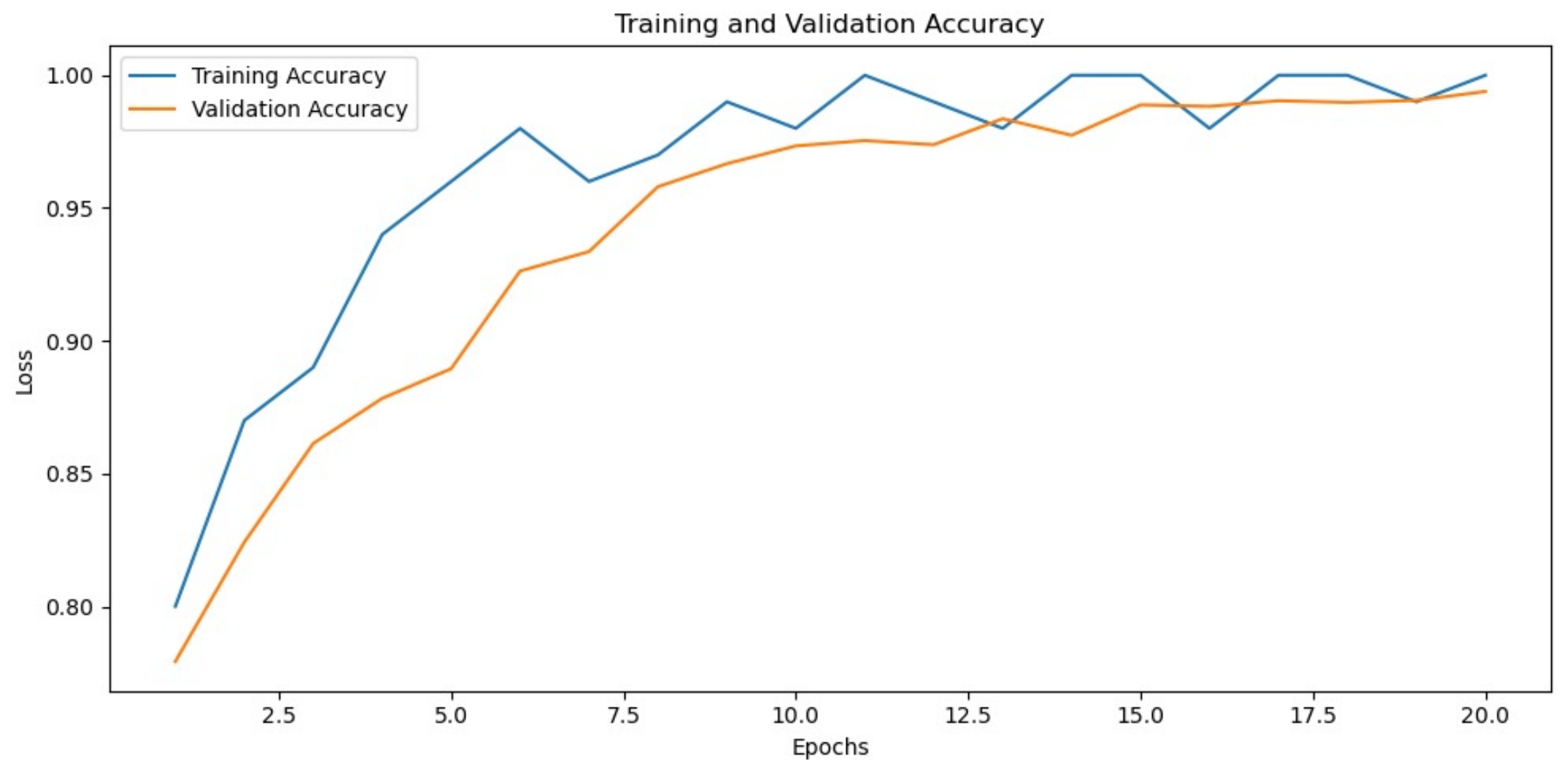
We have two "baseline" models to act as benchmarks against our CNN and BERT models:

- **Human:** We assessed our peers on a random sample of 20 paragraphs, evenly split, and got 33 responses. The accuracy was around 56%, and precision was 5% higher than recall, suggesting that readers were more likely to assume human authorship but might be slightly more confident when they have identified GPT. On the whole, people were not much better than random guessing.
- **Bag of Words:** The BOW model improved on the human baseline with an accuracy, precision, and recall of 78 to 79%. However, it quickly reached this level and did not show improvement over the course of 10 epochs. This is not surprising: we expect GPT to use similar words to human authors from its training, so just the frequency of individual words will have limited usefulness.

CNN

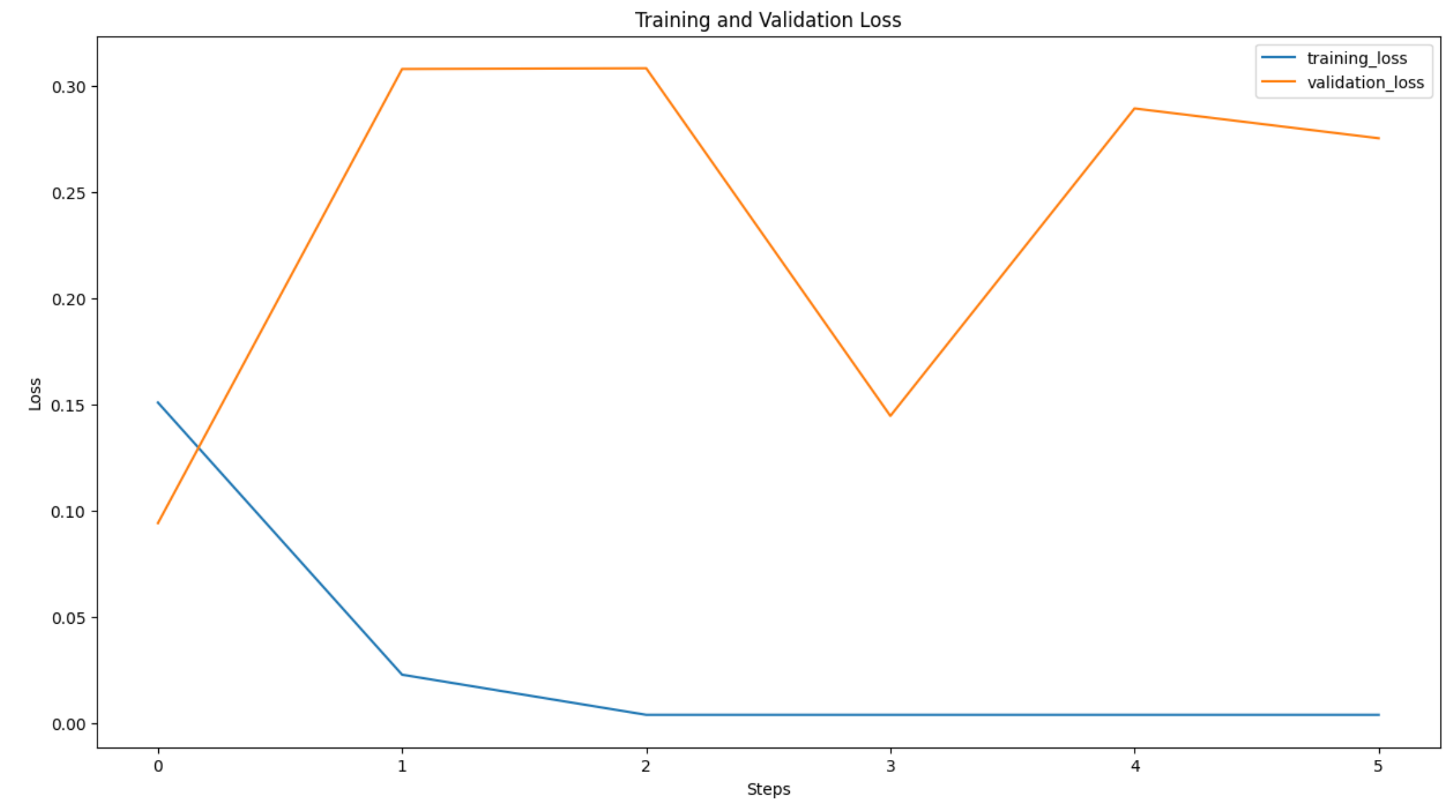
The model structure incorporated an embedding layer, three convolutional layers and respective max pooling layers, and a fully connected neural net for output. The embedding layer transformed word IDs into dense vector representations, while the convolutional layers detected local patterns with a kernel of size 3 and identified 256 key k-grams to consider as indicators. This number was originally higher, starting at 4000, but through hyperparameter tuning found we were able to achieve equivalent accuracy with a much smaller model. The max pooling layers extracted the most relevant features, and fed into the neural net that provided the prediction. The model achieved promising results, showing results above 98 percent accuracy, with little bias in precision vs recall. We have found that the model performed at this level with various size k-grams, ranging from 10 to 23.





BERT

In order to implement the Biderectional Encoder Representations from Transformers (BERT) model, we utilized the pre-trained backbone found on Huggingface, specifically "distilbert-base-uncased". We chose to implement this specific backbone as it is smaller and faster than BERT, which was a major concern for us, given that it still took seven hours to run on our dataset. We utilized a similar preprocessing methodology to our CNN, in which we pulled random 23 word sequences from an entry and tokenized the sequence. We chose to do this to eliminate any larger trends found in entries, given that BERT is known to easily identify global context in an entry; we specifically wanted to analyze the ability to recognize machine-generated text from a small sequence of words. We initiated a learning rate of 0.00002 and a decay in the learning rate of 0.01, leading to our results. Our model was able to achieve an accuracy rate of around 98% and a loss very close to 0 within a small number of epochs.



Conclusions

Between BERT and our CNN, we believe we have successfully constructed models capable of differentiating between human and machine generated text at a very high accuracy, both of which did so at great improvement over our human and traditional bag of words approach. Our next steps would primarily include trying to identify what specific patterns or sequences facilitate this distinction, to determine what is inherently different, as our human survey indicates it may be invisible to the human eye.