Flagging AI Written Text

by Spencer Holley

Dangers

of AI written text

This technology is amazing, but in the wrong hands it can be dangerous!

- Anyone can publish massive amounts fake news and articles on the internet.
- AI can generated text can be very convincing, despite holding no truth.
- Students can cheat on their essays without plagiarizing, therefore unlikely to get caught.

WHAT'S GPT2 ?

GPT2 is a Language generator that was released by OpenAI in 2018. can generate any kind of text when finetuned on texts of similar length

- uses an advanced architecture that outperforms the current models traditionally used for NLP
- OpenAI, the group that discovered GPT2, was slow to release it due to safety concerns
-

here's an example of text written by GPT2. Given the prompt "Recycling is good for the world. NO! YOU COULD NOT BE MORE WRONG!!" the "Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming."

My Solution

To help combat this issue, I trained a Neural Network to classify text as real or AI generated.

Inspiration

I was interested in creating something that could help Wikipedia users and contributors flag AI written articles.

- Wikipedia articles are very easy to create
- They are often the first thing that come up in a google search

How I did it...

Generation

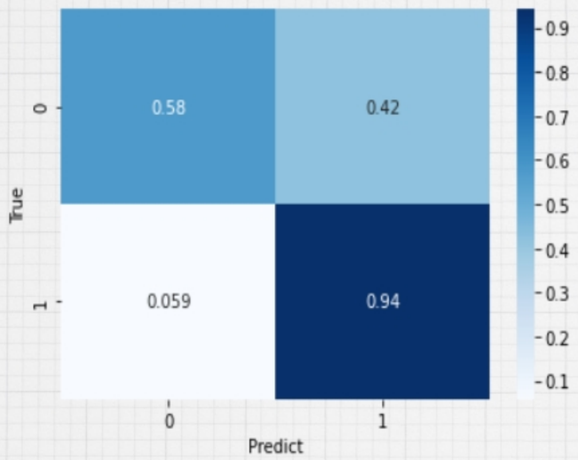
1. I retrained GPT2 on 1,000 Wikipedia articles on random topics, including Shanghai, Johnny Depp, and Psychiatry
2. The tuned model generated 180 articles averaging around 1600 to 2000 words

Classification

1. I put the generated articles in a dataset with 180 randomly selected real articles into a dataset
2. I trained a Bidirectional LSTM model to classify the articles appropriately

FINDINGS

The model's performance on testing data



Overall, the model achieved 76% accuracy
- the model identified 94% of AI generated text correctly
- but only 58% of real text

frequently used words from the AI Generated articles

frequently used words from the real articles

Conclusions

Here are some ways I recommend using AI written text classifiers

Education

Teachers and Professors can flag AI written Essays

Wikipedia

Wikipedia users an contributors can flag AI generated text

News

News companies can make sure that all the stories they get were written by their reporters

Future Works

Google Cloud Platform

A great continuation of this project would be to do it bigger

1. generate more articles and create a bigger dataset overall.
2. Use Google Cloud Platform or AWS to help with the extra computation.


THANK YOU