AI Generated Text vs Human Written Text

Vivian hyuh
Amelie Smithson
Ji sung Han
Hana Truchla

Agenda

1) Introduction & Problem : Ji
2) Data Selection & Processing : Hana
3) Model & Demo: Vivian
4) Quantitative & Qualitative Results : Amelie
5) Takeaways: Amelie

Problem

- With the rapid development and increase of use of AI technology, it becomes increasingly difficult to distinguish between AI-generated text and human-written text.

This is a BIG problem!

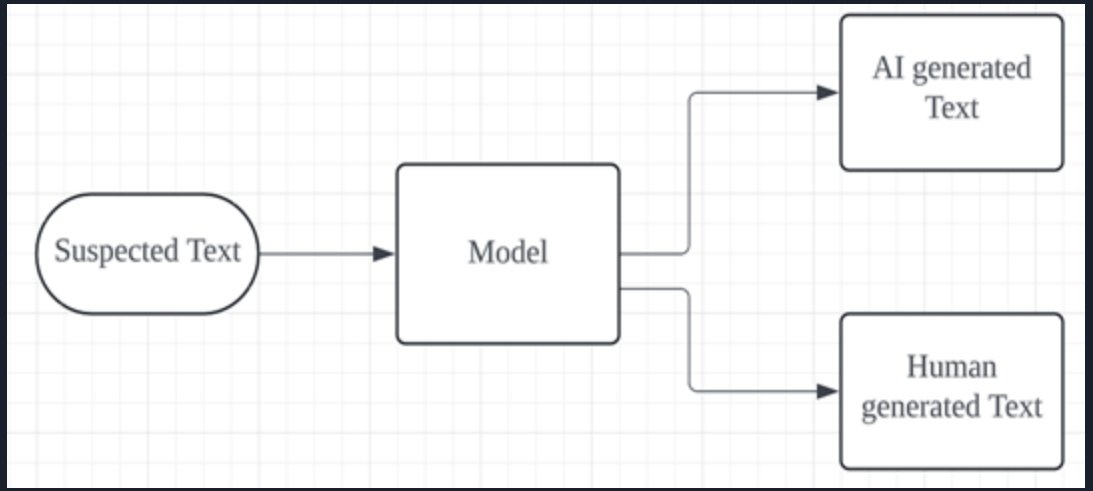
- Unfair advantages for thoses who use AI in assignments vs those who do not.

This is a BIG problem!

- AI-generated texts may be used to spread disinformation/propaganda on the Internet.

The problem to be solved

- The goal of this project is to develop a deep learning model that can distinguish between AI-generated and human-written text with high accuracy, aiming for a model accuracy of at least 70-80% or higher.



Data Selection

Data set 1: daigt-v3-traindataset (Kaggle)

text	label	prompt_n...	source	RDizzl3_s...
Phones Modern humans today are always on their phone. They are always on their phone more than 5 ho...	0	Phones and driving	persuade_corpus	False

Data Selection

Data set 2: AI vs Human Text Kaggle)

text	generated
Cars. Cars have been around since they became famous in the 1900s, when Henry Ford created and built...	0.0

Data Formatting

[['i', 'am', 'certain', 'that', 'the', 'face', 'on', 'mars', 'is', 'just', 'a', 'natural', 'landform', 'it', 'has', 'many', 'features', 'and', 'it', 'has', 'many', 'people', 'concerned', 'about', 'it', 'the', 'face', 'on', 'mars', 'is', 'a', 'natural', 'landform', 'because', 'it', 'was', 'spotted', 'with', 'shadowy', 'likeness', 'of', 'a', 'human', 'it', 'resembles', 'the', 'human', 'head', 'and', 'it', 'actually', 'shows', 'a', ..., 'to', 'represent', 'mars'], 0]

Model

- Using an Recurrent Neural Network (RNN), more specifically a Long Short-Term Memory Network (LSTM)
- Can handle more complex language/text based tasks

GloVe Embedding

Commonly used embedding method that has proven to be very effective.

Layer Normalization

Normalizes the features so there is less biases in the learning method.

Linear Layer

To consolidate the features into a relevant output.

Max Pooling

To focus the model only on the most important features.

Demo

ChatGPT Generated Text:

Neural networks are a type of artificial intelligence modeled after the human brain. They consist of interconnected nodes, or "neurons," that process information in layers. By adjusting the connections based on data, neural networks can learn patterns and make predictions or decisions. They're widely used in applications like image recognition, natural language processing, and autonomous systems.

Detected AI generated with ~ 99% accuracy

Human Written Text:

This course I'm taking is called APS360. It introduces neural networks and how to develop and train them. I have learned several types of neural networks so far. Ones such as, ANN, RNN, and CNN. Also this course requires that I complete a final project that involves creating a neural network. So for the final project my team and I developed this model to identify AI generated text.

Human Written with ~73% accuracy

Quantitative Results

Initial model training

Training accuracy: 97.89%

Validation accuracy: 74.59%

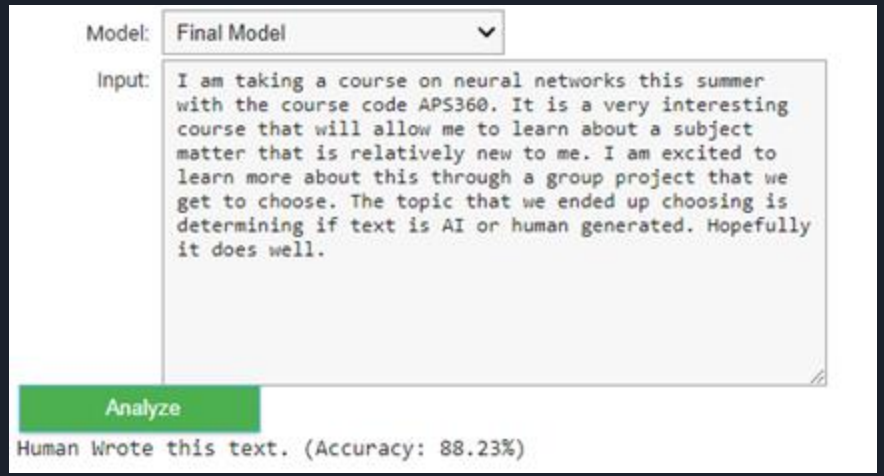
Post hyperparameter tuning

Training accuracy: 97.87%

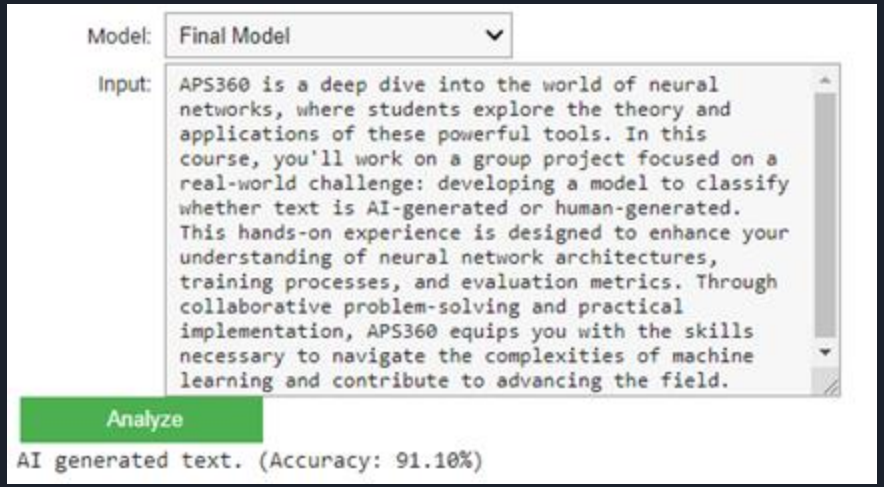
Validation accuracy: 84.94%

Test accuracy: 86.49%

Qualitative Results: An example of both AI and Human generated input

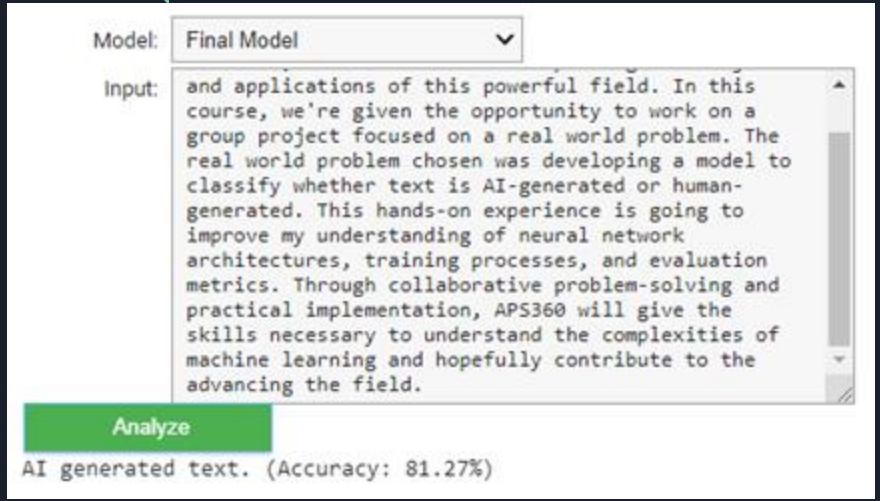


Accurate classification of human generated text

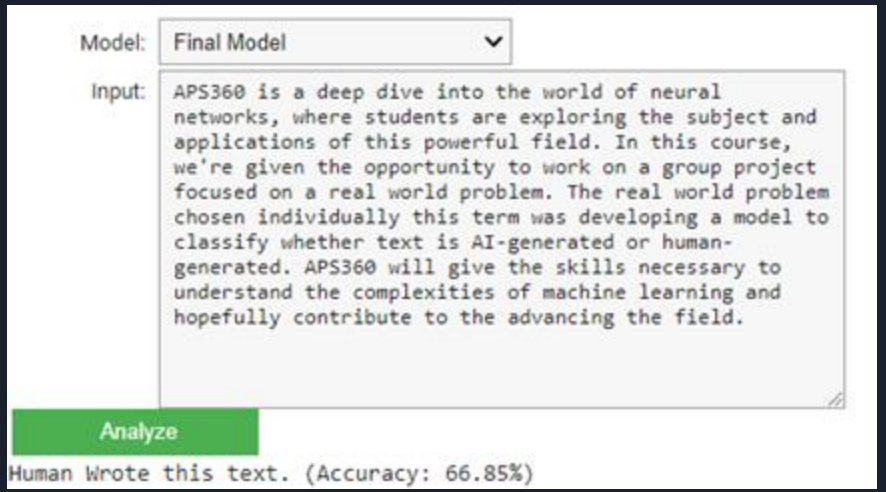


Accurate classification of AI generated text

Qualitative Results: Modified AI Generated Input

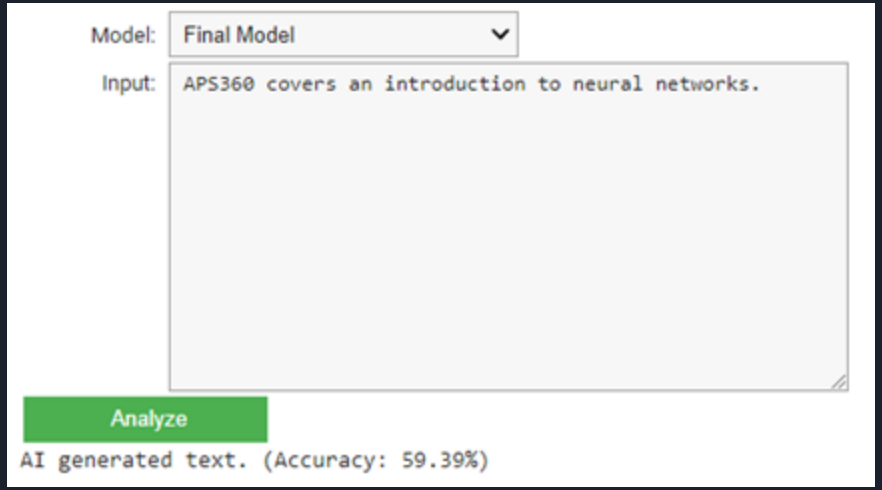


Text correctly classified as AI generated despite human tampering of exact ChatGPT result.

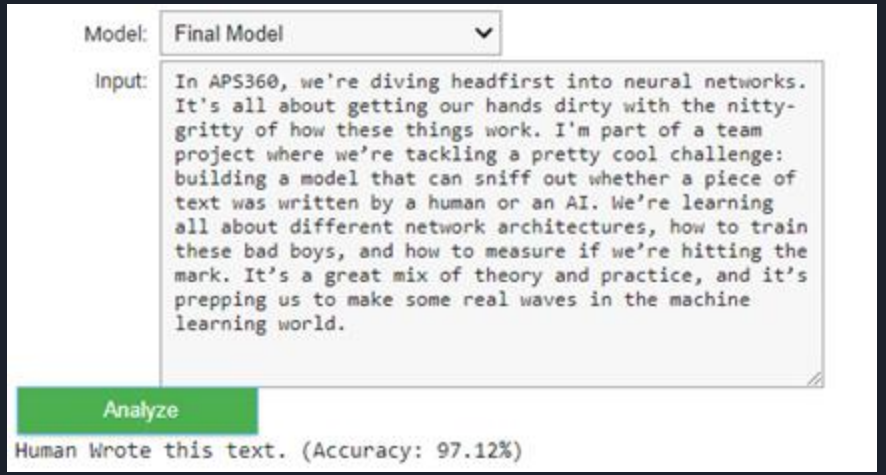


Text incorrectly classified as human generated after making further edits to AI generated text.

Qualitative Results: Variation in Text Length and Personal Pronouns



Human generated text incorrectly classified as AI generated.



Misclassified text as human generated when copying a result directly from ChatGPT with explicit instructions to use personal pronouns.