



# 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 those who use AI in assignments vs those who do not.



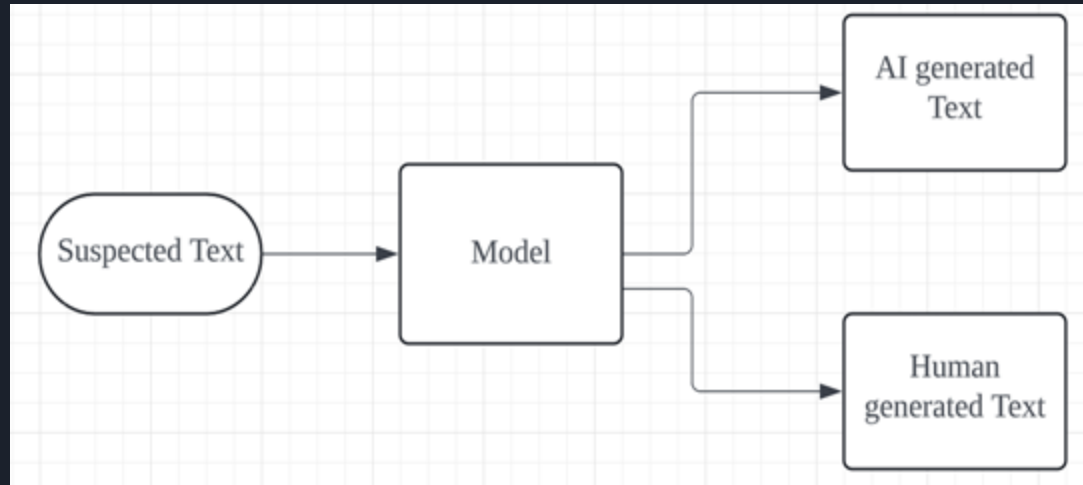
# This is a BIG problem!

- 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)

A text	# label	A prompt_n...	A 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)

Δ 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.



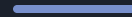
**Layer Normalization**

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

Model: Final Model ▼

Input: I am taking a course on neural networks this summer with the course code APS360. It is a very interesting course that will allow me to learn about a subject matter that is relatively new to me. I am excited to learn more about this through a group project that we get to choose. The topic that we ended up choosing is determining if text is AI or human generated. Hopefully it does well.

Analyze

Human Wrote this text. (Accuracy: 88.23%)

Accurate classification of human generated text

Model: Final Model ▼

Input: APS360 is a deep dive into the world of neural networks, where students explore the theory and applications of these powerful tools. In this course, you'll work on a group project focused on a real-world challenge: developing a model to classify whether text is AI-generated or human-generated. This hands-on experience is designed to enhance your understanding of neural network architectures, training processes, and evaluation metrics. Through collaborative problem-solving and practical implementation, APS360 equips you with the skills necessary to navigate the complexities of machine learning and contribute to advancing the field.

Analyze

AI generated text. (Accuracy: 91.10%)

Accurate classification of AI generated text

# Qualitative Results: Modified AI Generated Input

Model: Final Model

Input: and applications of this powerful field. In this course, we're given the opportunity to work on a group project focused on a real world problem. The real world problem chosen was developing a model to classify whether text is AI-generated or human-generated. This hands-on experience is going to improve my understanding of neural network architectures, training processes, and evaluation metrics. Through collaborative problem-solving and practical implementation, APS360 will give the skills necessary to understand the complexities of machine learning and hopefully contribute to the advancing the field.

Analyze

AI generated text. (Accuracy: 81.27%)

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

Model: Final Model

Input: APS360 is a deep dive into the world of neural networks, where students are exploring the subject and applications of this powerful field. In this course, we're given the opportunity to work on a group project focused on a real world problem. The real world problem chosen individually this term was developing a model to classify whether text is AI-generated or human-generated. APS360 will give the skills necessary to understand the complexities of machine learning and hopefully contribute to the advancing the field.

Analyze

Human Wrote this text. (Accuracy: 66.85%)

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

# Qualitative Results: Variation in Text Length and Personal Pronouns

Model: Final Model ▼

Input: APS360 covers an introduction to neural networks.

Analyze

AI generated text. (Accuracy: 59.39%)

Human generated text incorrectly classified as AI generated.

Model: Final Model ▼

Input: In APS360, we're diving headfirst into neural networks. It's all about getting our hands dirty with the nitty-gritty of how these things work. I'm part of a team project where we're tackling a pretty cool challenge: building a model that can sniff out whether a piece of text was written by a human or an AI. We're learning all about different network architectures, how to train these bad boys, and how to measure if we're hitting the mark. It's a great mix of theory and practice, and it's prepping us to make some real waves in the machine learning world.

Analyze

Human Wrote this text. (Accuracy: 97.12%)

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