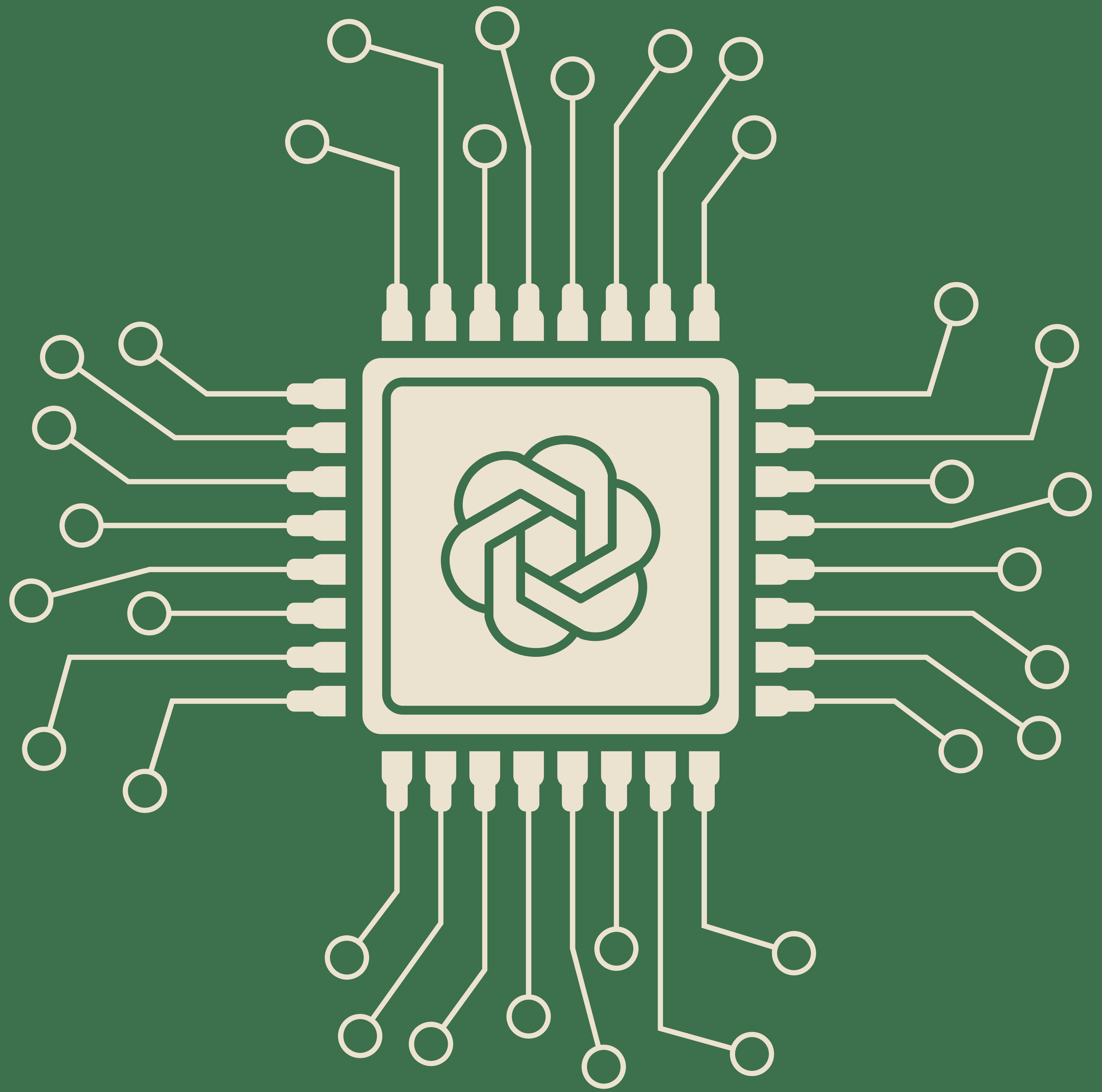
Distinguishing Between AI and Human-Generated Content



INSY 669 Text Analysis Final Project

Presented by: Vivi Li, Jennifer Liu, Wenya Cai, Hongyi Zhan, Kazuya Hayashi, Rodrigo Castro

Table of Contents

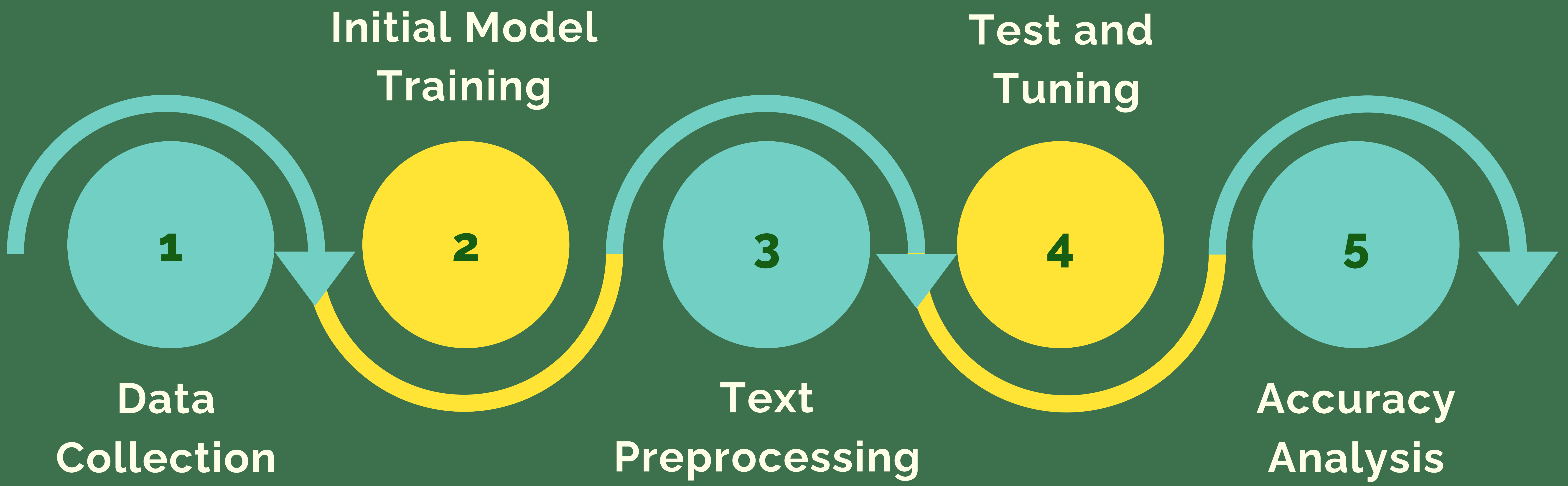
Introduction and Problem Statement

01 **Generative AI** increasingly mimics human writing, **complicating** the distinction between AI and human-produced content. This convergence highlights the demand for effective “fake” text detection to preserve information integrity and ethical AI use.

02 Our project focuses on developing a model to **differentiate** between **AI-generated** and **human-written** texts, addressing the critical challenge of maintaining authenticity in the digital realm.

03 By conducting this project, we thereby **protect** against misinformation, **maintain** academic honesty, **ensure** the authenticity of online communications, and **foster** trust in the digital ecosystem.

Analytics Approach: Our Steps



Analytics Approach: Preprocessing

Data Overview

- We gathered a dataset that contains
  - human-written, and
  - AI-generated samples

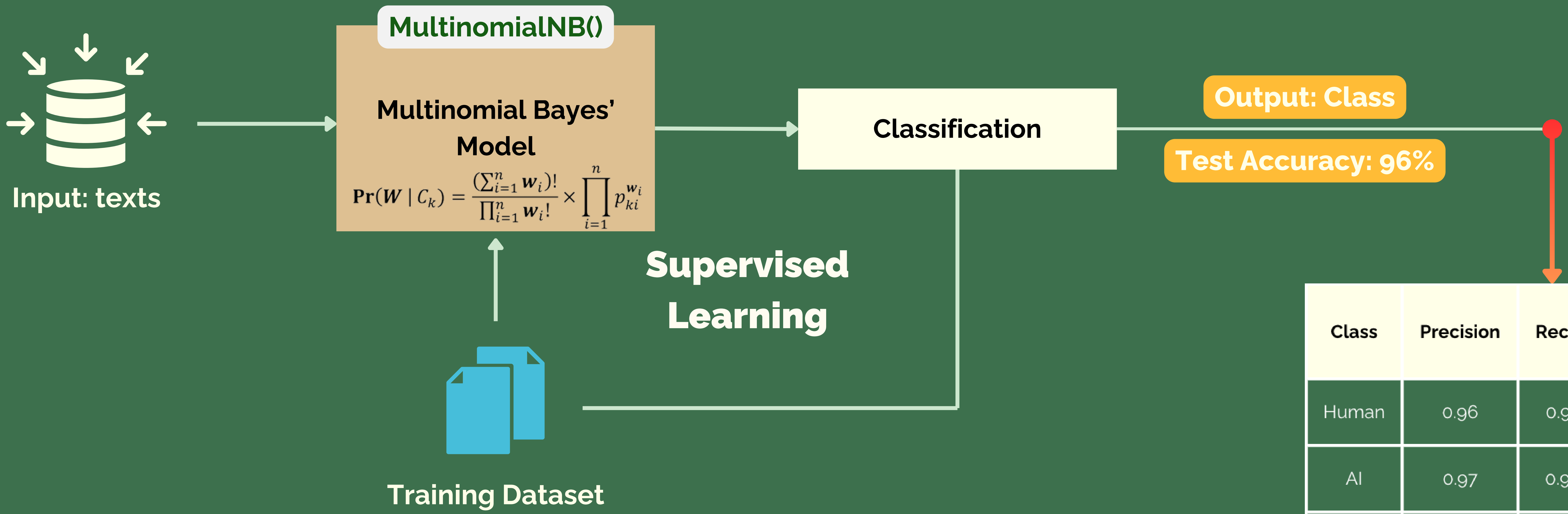
- We sampled the data so that it contains 50% human and 50% AI (for unbiased model training)
  - Large Language Model(LLM) used to generate samples: llama, Falcon, GPT, etc.

Text Preprocessing

Our approach used:

- Tokenization
- Lowercasing
- Filtering stop words and punctuations

Analytics Approach: Multinomial NB - V1



Class	Precision	Recall	F1 Score	Support
Human	0.96	0.97	0.97	1999
AI	0.97	0.96	0.97	1999
Overall	0.97	0.97	0.97	3998

NB: assume independence of features

countvec: high-dim sparse matrix with occurrences

Analytics Approach: Multinomial NB - V2

Hyperparameter Tuning
(GridSearch)

1

2

Achieve Optimal Alpha = 0.0001

Accuracy: 0.958

“Optimal alpha value indicate minimal smoothing preference, which maintained the same initial accuracy level.”

Additional text preprocessing
Lemmatization, Stop Words Removal

Analytics Approach: Logistic Regression

Optimized Parameters

{'C': 1000, 'class_weight': None, 'dual': False, 'fit_intercept': True, 'intercept_scaling': 1, 'l1_ratio': None, 'max_iter': 100, 'multi_class': 'auto', 'n_jobs': None, 'penalty': 'l2', 'solver': 'lbfgs', 'tol': 0.0001, 'verbose': 0, 'warm_start': False} w/ TF-IDF (weighted importance x feature dependence and regularization)

Test Data Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	1974	25
Actual Positive	28	1971

Accuracy: 0.996

Model Evaluation: Latent Dirichlet Allocation

Key word in each topic:

1. **vote, electoral , college, state, president**
2. **car, people, would, driving, venus**
3. **student, school, would, people, help**

lda_topic: 1 accuary: 0.99978
lda_topic: 2 accuary: 0.99855
lda_topic: 3 accuary: 0.99532

Model Evaluation: Utilizing GPT-4

Our model adeptly identified the AI-generated essays with **high** successes .

However,

When we tell GPT to act "**more like human**", its accuracy decreased to **60%**.

"This underscores the evolving challenge in distinguishing advanced AI-generated content from human writing, highlighting the necessity for continuous sophisticated detection methodologies."

Expected Impact of Our Text Analysis Model

Impact

Crucial in Journalism

1. Detects AI-generated text to ensure accurate and reliable news.
2. Prevents the spread of fake news by verifying content sources.
3. Maintains trust in journalistic integrity.

Enhance Online Content Authenticity

1. Ensures reviews and comments on platforms are written by real users.
2. Boosts the authenticity and trustworthiness of online community interactions.

Vital for Academic Integrity

1. Identifies AI-generated assignments to prevent academic plagiarism.
2. Helps uphold standards of original student work and scholarship.

Corporate Sector Safeguard

1. Protects against the creation of false AI-generated endorsements.
2. Prevents the generation of misleading business documents.
3. Preserves brand integrity and consumer trust.
4. Challenges with "More Human" AI Content:

Reflections on High Accuracy: Data Problem

- Different Topics
- Length of Docs
- Mispelling
- Level of Words

More Model Evaluation Is Needed

1. We tell GPT to act more like human (done) with specific prompting

2. Test our model on research paper (aligning on topics and level of words

3. Human generated but use Grammarly (misspelling)

4. We use LLM to convert human generated contents (when ideas are human generated)

Future Consideration

Each use case tries to capture different sets of differences,

So, we need to train model for each use case to improve reliability

Thanks

We’re open to answer all your questions!