Capstone Three – Project Proposal

Project Title:
Detecting AI-Generated Text Using Natural Language Processing

---

1. Problem Statement

The rise of generative AI models like ChatGPT has made it difficult to distinguish whether a piece of writing is created by a human or an AI. This creates serious challenges in areas like education (plagiarism detection), content moderation, hiring, journalism, and public trust in online content.

This project will explore whether we can build a model to detect AI-generated vs. human-written content using natural language processing (NLP) techniques and linguistic features.

---

2. Context

AI-generated writing is now widely used for blog posts, resumes, academic essays, and marketing. As this content becomes harder to detect, people and organizations need better tools to verify authorship. The ability to classify content origin is becoming a key concern in education, journalism, and digital security.

---

3. Criteria for Success

- Develop a model that can predict if a given text was written by a human or by AI with **better-than-random accuracy (≥60%)**
- Present interpretability on what features (e.g., grammar errors, sentence length) help distinguish AI content
- Visualize and explain how predictions are made

---

4. Scope of Solution Space

This project will focus on:

- Supervised classification using NLP (e.g., TF-IDF, sentiment, linguistic markers)
- Traditional models (Logistic Regression, Random Forest), with potential extensions to more complex models (XGBoost, BERT)

It will not aim to classify the **specific** AI model used — just whether the text is AI or human.

---

5. Constraints

- AI and human-written text are often very similar, so performance might be limited
- Dataset size is ~1,300 samples, which restricts deep learning approaches
- Some linguistic features may not generalize across topics

---

6. Stakeholders

- **Educators**: Verify if student essays are original
- **Content Platforms**: Moderate fake/spam submissions
- **Employers**: Detect AI-generated resumes
- **General public**: Build trust in written content

---

7. Data Source

- **Dataset Name:** AI vs Human Content Detection – Kaggle (2025)
- **Fields include:**
    - `text_content` (actual writing)
    - Readability, grammar, burstiness, lexical diversity, etc.

- `label`: 0 = Human, 1 = AI

---

8. Method Overview (Brief Outline)

- **Text preprocessing** (cleaning, stopword removal)
- **Feature extraction** using TF-IDF and built-in linguistic metrics
- **Train/Test split** and modeling using Random Forest and Logistic Regression
- **Evaluation** using accuracy, precision, recall, confusion matrix
- **Next steps**: Try advanced models like BERT or use additional metadata

---

9. Deliverables

- A **Jupyter Notebook** with EDA, modeling, and visualizations
- A **PDF project proposal** submitted to GitHub
- A **slide deck** summarizing insights (for mentor review)
- A **GitHub repository** with all project files