Group 25 Final Project Report:
AI Essay Detector

Fei Xie, Ghena Hatoum, Haniye Hamidizadeh
{xief17,hatoumg,hamidizf}@mcmaster.ca

1 Introduction

Your report should have an introduction. You may largely copy the motivation and problem description from your previous reports. Write about previous work on your task or the most related task you can find. Note that you should include at least 7 references in your report. Did you discover anything in related work that influenced your direction after the progress report was due? References should be integrated as part of the discussion of the context of your work. They should not simply be listed as a set of relevant papers with no explanation

2 Dataset

You should describe the dataset properties and any preprocessing operations you performed. If you are annotating data yourself, describe the annotation procedure you developed and followed. Describe any changes you made to the dataset in a subsection if you have changed your dataset since the progress report. Some of you found that your dataset was too difficult or too easy. What did you have to change about your data? Did you have to augment the data or move to a completely new source? If your dataset did not change, it is reasonable to leave this mostly as is from the progress report. often a word appears in a specific document as shown in Equation 1. A higher frequency suggests greater importance within that document. Inverse document frequency (IDF) captures the term popularity as an inverse of the overall corpus as shown in Equation 2. Our TF-IDF matrix has a max of 10000 features.

3 Features

Describe your model inputs. What feature engineering or representation learning was performed? Why did you add these features? Why does it make sense to include them? Did you perform any feature selection or augmentation? One good way to vary your experiments is to try different kinds of features as inputs. How did you vary the features used for your experiments

4 Implementation

Describe the model implementation. You should have several mod- els or versions of your model that you have run on your data. You should have a simple baseline to compare it to (likely majority vote), and may have other baselines from related work. Your model should outperform your simple baseline. For most of you, your model should also outperform at least one other trained model baseline. If it does not outperform this baseline and you expected it to, why doesn't it? You should be able to provide an expla- nation. This explanation can also be part of your error analysis section. You do not have to outperform all models from previous work, but you should have an approach you implement for comparison. What was your loss function? Describe the optimization technique. If you implemented a complex model with many parts, you should consider providing ablations. This means different experiments where you only include one feature at a time so you can tell what feature is contributing more to the performance that you see.

5 Evaluation

Describe the evaluation strategy. What are your train/validation/test splits (size and label distributions)? Are you using cross-validation? What metrics are you using for evaluation? Did you find that your metrics from the progress report stage were adequate?

6 Progress

Reflect on your plan from your progress report. Did you follow-through on your plan? Did your plan change course? Why?

7 Error Analysis

Describe how you systematically examine the errors your model makes and provide supporting figures, stats, examples (e.g., confusion matrices, qualitative sample of test cases with high error margins, etc). What does your model appear to be good at? What does it seem to be bad at? How does the performance of your models differ? What patterns do you notice in the errors your model seems to make? What do you think you could do to specifically address those issues if you were to continue working on this mod

8 Team Contributions

Fei:
Haniye:
Ghena:

References

Nicholas Broad. 2023a. Daigt data - llama 70b + gpt4 + falcon180b dataset. https://www.kaggle.com/datasets/nbroad/daigt-data-llama-70b-and-falcon180b. Accessed: 2025-11-08.

Nicholas Broad. 2023b. Persuade corpus 2.0. https://www.kaggle.com/datasets/nbroad/persaude-corpus-2/. Accessed: 2025-11-08.

Darragh Hanley. 2023. 1000 essays from anthropic claude. https://www.kaggle.com/datasets/darraghdog/hello-claude-1000-essays-from-anthropic. Accessed: 2025-11-08.

Radek Osmulski. 2023. Llm generated essays. https://www.kaggle.com/datasets/radek1/llm-generated-essays. Accessed: 2025-11-08.

Alejo Paullier. 2023. Daigt external dataset. https://www.kaggle.com/datasets/alejopaullier/daigt-external-dataset. Accessed: 2025-11-08.

Muhammad Rizqi. 2023. Llm-generated essays using palm. https://www.kaggle.com/datasets/kingki19/llm-generated-essay-using-palm-from-google-gen-ai. Accessed: 2025-11-08.

Tables and figures

See Table 1

Equations

$$TF(w,d) = \frac{\text{count}(w,d)}{\text{total words in } d} \tag{1}$$

$$IDF(w,C) = \log\left(\frac{|C|}{|\{\, d \in C : w \in d \,\}|}\right) \tag{2}$$

Dataset Name	**Description**
Persuade Corpus 2.0 (Broad, 2023b)	Provides argumentative essays produced by 6 to 12 grade students. It was created by The Learning Agency and Vanderbilt University, originally pulled from the following GitHub repository.
ChatGPT (Paullier, 2023)	Contains 2.5k student written texts sourced from the FeedBack Prize 3 Competition, and 2.5k AI-generated texts using ChatGPT. The compiled dataset includes only AI-generated texts and prompts.
Llama 70b + GPT-4 (Broad, 2023a)	Contains 9k essays generated by Llama 70b and Falcon 180b. Prompts come from the Persuade Corpus and GPT4, using a total of 35 prompts for generation.
LLM Generated Essays (Osmulski, 2023)	Contains 700 essays generated by LLMs 500 from GPT3.5Turbo and 200 from GPT4.
Claude Essays (Hanley, 2023)	Contains 1000 essays generated by Claude-Instant-1 using 15 prompts from the Persuade Corpus. Prompts were sourced from the competition discussion.
PaLM Essays (Rizqi, 2023)	Contains 1384 essays generated by PaLM. Prompts were sourced from a Kaggle competition notebook.

Table 1: Summary of Datasets Used in the DAIGT V2 Train Dataset

Table 2: Performance Accross Different Feature Sets

Features Used	**Total Features**	**Test Accuracy**	**Test F1 Score**
Punction (Baseline)	18	0.83	0.7758
TF-IDF	10 000	0.9973	0.9966
Word Counts + TF-IDF	10 003	0.9973	0.9966
All features	10 021	0.9971	0.9963