

Essay Origin Identification: Distinguishing Student Essays from LLM-Generated Text

Your Name

2024-01-07

Introduction

In the evolving landscape of educational technology, the ability to discern between essays written by students and those generated by Large Language Models (LLMs) is becoming increasingly important. This study aims to develop a robust model to effectively perform this task, utilizing a range of linguistic and statistical features.

Methodology

Data Preprocessing

The study utilized essays written by middle and high school students as well as essays generated by LLMs.

- **Data Sources:** A dataset comprising student-written and LLM-generated essays.
- **Preprocessing Steps:**
 - Merging essays with corresponding
 - Removal of non-essential columns such as instructions, source_text, and prompt_name.
 - Conversion of prompt IDs to factors for modeling.

Feature Engineering

The study focused on several key features:

- **Word Count:** The total number of words in each essay.
- **Sentence Count:** The number of sentences per essay.
- **Type-Token Ratio (TTR):** A measure of vocabulary richness.
- **Readability Scores:** Computation of Flesch Reading Ease and Flesch-Kincaid Grade Level.

Sentiment Analysis

- Application of NRC sentiment analysis to gauge the emotional content of the essays.
- Calculation of emotion proportions (anger, anticipation, etc.) for each essay.

Analysis

The analysis involved examining various features to discern patterns distinguishing student-written essays from those generated by LLMs.

1. **Word and Sentence Count:** Initial observations indicated a tendency for LLM-generated essays to have a lower word count.
2. **Vocabulary Richness:** The TTR offered insights into the diversity and complexity of the vocabulary used.
3. **Readability Scores:** These metrics provided an understanding of the text's complexity and readability level.
4. **Sentiment Analysis:** Emotional content analysis revealed distinct patterns in student and LLM-generated texts.

Model Development

- **Neural Network Approach:** Trained a neural network model with various configurations to optimize classification performance.
- **Boosting Method:** Employed XGBoost for its efficiency in handling diverse datasets.
- **Model Evaluation:** Focused on accuracy and AUC (Area Under the Curve) as key performance metrics.

Results

- **Neural Network Model:** Demonstrated an accuracy of 49% in classifying essays.
- **Boosted Model:** XGBoost showed 0.02 mean of mse

Conclusion

The models developed in this study show promise in differentiating between essays written by students and those generated by LLMs. However, the limitations of the current approach and the potential for further improvements are acknowledged.

Future Work

Suggestions for future work include:

- Integration of more advanced linguistic features.
- Exploration of different machine learning architectures.
- Extensive validation with varied datasets.