**Problem title: Predicting Effective arguments in an essay**

The goal of this project is to classify argumentative elements in student writing as "effective," "adequate," or "ineffective." We created a model trained on data that is representative of the 6th-12th grade population in the United States in order to minimize bias. The model developed will help pave the way for students to receive enhanced feedback on their argumentative writing. With automated guidance, students can complete more assignments and ultimately become more confident, proficient writers.

.

There are numerous automated writing feedback tools currently available, but they all have limitations, especially with argumentative writing. Existing tools often fail to evaluate the quality of argumentative elements, such as organization, evidence, and idea development. Most importantly, many of these writing tools are inaccessible to educators due to their cost, which most impacts already underserved schools.

**Exploratory Data Analysis:**

We did EDA to get to know the relation between different features, to find any outliers, any missing values and some feature engineering.

Reading the text from the text files and adding it as a column in the existing dataframe.

Creating the training data by separating the target class and the training data.

**Text Processing:**

We clean the text, convert it to lower case, removing the stop words, expand all the contractions, remove any unnecessary characters.

**One hot encoding:**

We have 3 different values for the target class. So, we perform one hot encoding on the target class and convert the single class classification problem into multi output regression problem.

Here, we are trying to predict how much percent of the essay is adequate, inadequate and effective. So, this now turns into a multioutput regression problem.

**TFIDF:**

Next, we use TFIDF vectorizer to get the importance of a word in a document relative to a collection of documents.

Next, we split the dataset in 80:20 ratio for training and testing.

**Model:**

Next, we used the KNeighborsRegressor with the MultiOutputRegressor Wrapper to perform the Multioutput regression.

KNEighborsRegressor is a regression algorithm based on k-nearest neighbors. It predicts the target variable for a new data point by considering the average or weighted average of the target values of its k-nearest neighbors in the feature space.

MultiOutputRegressor is a wrapper class that extends regression models to handle multiple target variables simultaneously. It fits one regressor per target variable.

KNeighborsRegressor is used as the base estimator within the MultiOutputRegressor wrapper, allowing it to handle multiple target variables.

We will use the trained model to predict the output and compute the Mean Squared Error.

**Transfer Learning:**

We have also tried using the Transfer learning approach using the DeBERTa model.

We used the pre trained the DeBERTa tokenizer for performing the encoding on the dataset.

We will use the encodings and create a custom dataset with the encodings and the labels.

Next, we use the pre trained DeBERTa model for sequence classification for training.

DeBERTa (Decoding-enhanced BERT with disentangled attention) is a transformer-based language model that builds upon BERT with enhancements to its attention mechanism. Like BERT, DeBERTa can be used for various natural language processing (NLP) tasks, including sequence classification.

We use the trained model to predict the output.

**What did not work in the project?**

The DeBERTa model is a heavy model with the pretrained weights. Even for 3 epochs, it took almost 24 hours to train. Also, the dataset contains only single target variable. But converting the classification problem into multi output regression problem was a challenge but we have overcome it. We plan on increasing the efficiency and the accurate predictions of the model.

**Potential buyers:**

Writing is crucial for success. In particular, argumentative writing fosters critical thinking and civic engagement skills, and can be strengthened by practice. However, only 13 percent of eighth-grade teachers ask their students to write persuasively each week. Additionally, resource constraints disproportionately impact Black and Hispanic students, so they are more likely to write at the “below basic” level as compared to their white peers. An automated feedback tool is one way to make it easier for teachers to grade writing tasks assigned to their students that will also improve their writing skills.

Will be accessible to educators for free due to their cost, which most impacts already underserved schools.