# CSIT6000R: Project Results on Metaphors Detection using DistilBERT

**Model**

The core model used in this project is DistilBERT, a lighter version of BERT optimized for faster performance while retaining a significant portion of its predecessor's capabilities. We fine-tuned this pretrained, base-cased model on a dataset specifically designed for metaphor and analogy detection, available at [Hugging Face Datasets](https://huggingface.co/datasets/Joanne/Metaphors_and_Analogies/tree/main/hf_datasets/Pairs_Jankowiac_random_split). This dataset includes word pairs categorized into metaphors, literals, and anomalies, which are essential for training our sequence classification model to understand nuanced language patterns.

**Database**

Our dataset is structured into multiple splits for comprehensive training and validation:

* Train Split: 50% of the data, used for training the model.
* Validation Split: 10% of the data, used for tuning the hyperparameters.
* Test Split: 40% of the data, used for evaluating the model's performance.

**Results and Visualization**

Throughout the training process, we closely monitored several key metrics such as accuracy, precision, recall, and F1 score. These metrics were plotted over various epochs to visualize the model’s learning curve and to identify any potential overfitting or underfitting issues.

* Here is a simulated graph showing the trend of these metrics:

**Discussion**

The highest F1-score observed was a result of carefully balancing the class weights, especially considering the uneven distribution of labels in our dataset. Initially, larger weights on underrepresented classes (metaphors and anomalies) helped improve model sensitivity towards these categories. To counteract this, we adjusted the training data by incorporating more examples from underrepresented classes, thus achieving a more balanced dataset and improving the F1 score.