NLP Project: Predicting Argument Effectiveness

Group 11:
Harsh Vardhan 11940460
Dhruv Deshmukh 11940380
Anupam Kumar 11940160
Abdur Rahman Khan 11940020

Introduction to the Problem Statement

- Kaggle Competition hosted by Georgia State University
- Being able to write effectively is an important skill and in that writing good arguments is important to communicate your views effectively.
- 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.

Dataset

- Set of annotated essays.
- Each essay is divided into set of discourses.
- The discourse can be one of seven types:

  Lead, Position, Claim, Counterclaim, Evidence, Rebuttal, Concluding Statement, etc.

- Discourse effectiveness can be classified into three classes:
    - Ineffective(18%)
    - Adequate(57%)
    - Effective(25%)
  - Total 36765 training examples

Evaluation

The score is evaluated as the log loss:

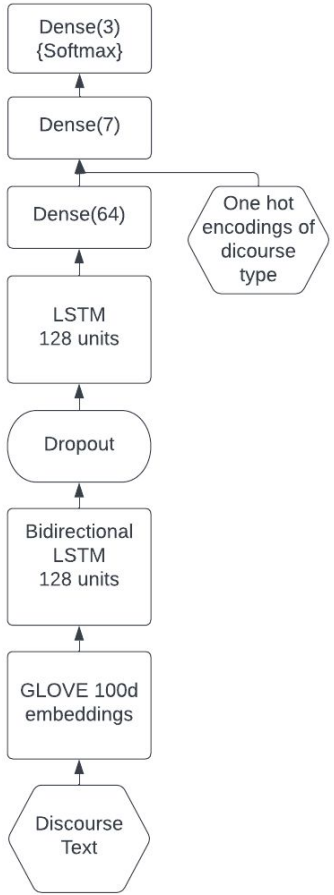
$$\text{log loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$

where N is the number of rows in the test set, M is the number of class labels, log is the natural logarithm, y_{ij} is 1 if observation i is in class j and 0 otherwise, and p_{ij} is the predicted probability that observation i belongs to class j.

Our Approaches

Approach: 1 Bidirectional LSTM

- Take each discourse text convert it into 100d embeddings
- Pass through one Bidirectional LSTM and then normal LSTM
- Pass through one dense layer and then concatenate the discourse type info in the one hot encoded format
- Finally use softmax classification to classify the effectiveness
- Batch size of 15 seemed to work the best
- Score obtained = 0.97

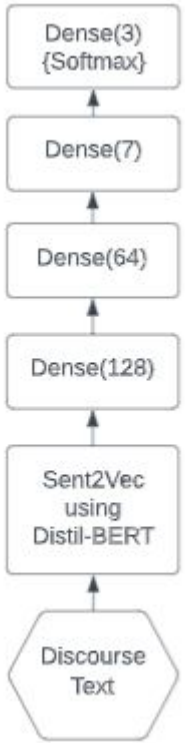


Some Observations

- The dataset is unbalanced with more almost 60% of labels as **Adequate**
- Hence the model tends to learn this pattern and output high probabilities most of the time for this class.
- One way to counter this is to add class weights = 100/(2*pi) where pi is percentage of ith class labels
- But it seems that the final test data also has similar distribution and hence using class weights is not that beneficial

Approach-2

- Use DistilBERT model to convert the discourse text to an encoding of 768 dimensions
- Due to lack of compute we truncated it to 100d encoding
- These encodings are passed through dense layers before classification.
- Here we classify both the discourse type and effectiveness separately and add both their losses to get final loss.
- This adds a regularization effect and allows the model to learn a little further
- The score obtained was 0.94

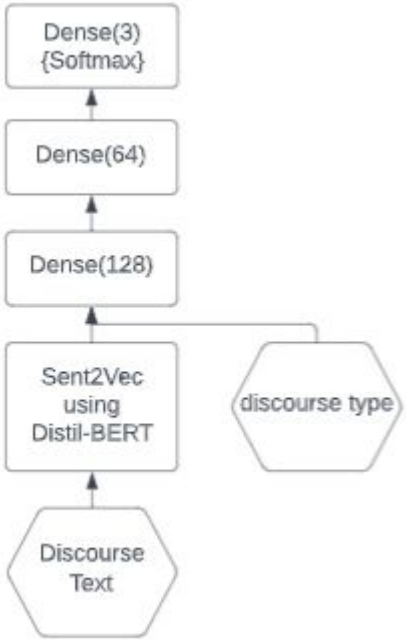


Approach-2a(Extension of Approach 2)

- Here we concatenate the current discourse with the previous and next discourse vectors.
- This adds more context and allows better learning
- Score obtained 0.86

Approach-3 Give discourse type info directly

- Similar to Approach 2 but here we removed the additional loss of classifying the discourse type as well and instead directly provided it as an integer by appending to the 100-d encoding.
- This surprisingly worked well and also upon training 5000 epochs gave good results
- Score obtained 0.77



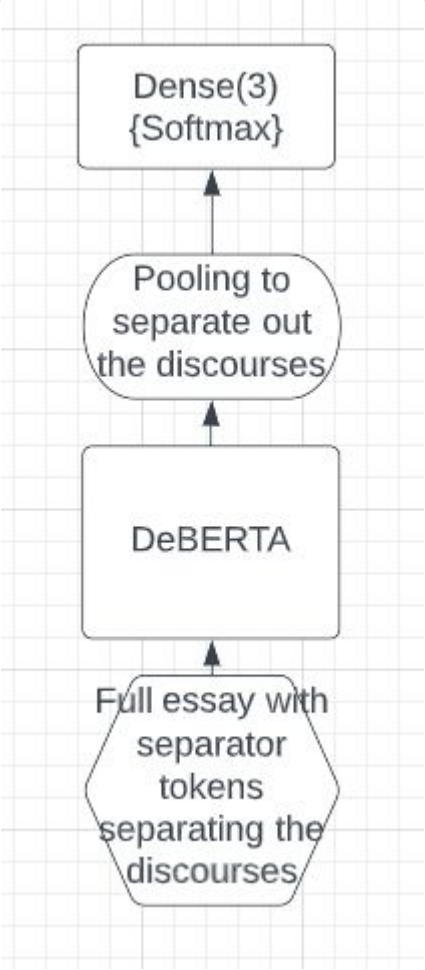
Learnings from previous approaches

1. Context helps. Concatenating previous and next discourses helps
2. Giving the discourse type directly helps. Instead of trying to learn to classify type as well f directly give it to the model.
3. DistilBERT is approximation and smaller model than BERT with less number of parameters hence there is scope of improvement if we use a better model.
4. We also had tried adding Dropout layers in our classifier networks for regularization but they didn't help much maybe because our networks weren't very deep.

Keeping all these points in minds and analysing some of the better submissions in the competitions we finally came up our 4th and final approach!

Approach-4 DeBERTa

- In this we take the essay and in it mark each discourse with starting and ending tokens containing its type also the sequence of types at the beginning.
- This whole essay is passed to DeBERTa
- The output is then pooled according to the starting and ending tokens and then is for each discourse effectiveness is classified
- Score obtained 0.70



Scope of improvement

- Train for more epochs
- Get larger dataset
- Train an ensemble of models
- Try different bert variations