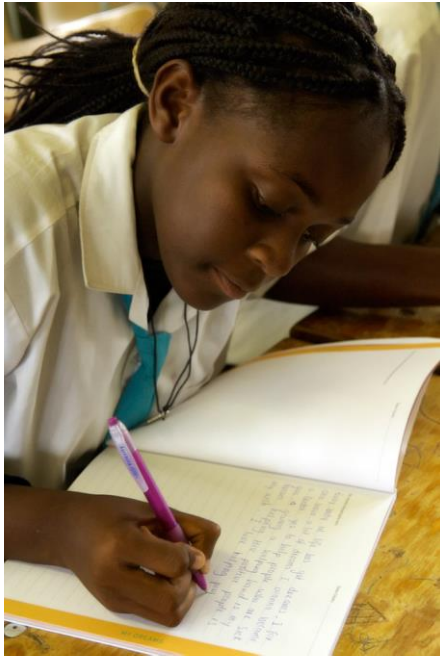
Predicting Effective Arguments

W207 Final Project

Adam, Vivek & Waqas

Problem Motivation: Grading Argumentative Essays



Problem: It is difficult to provide uniform grading of argumentative essays, on-demand and at scale.

The current state-of-the-art can provide writing feedback, but is weak at grading argumentative essays.

Feedback Prize - Predicting Effective Arguments | Kaggle. Kaggle.com. Published 2022. Accessed July 12, 2022. https://www.kaggle.com/competitions/feedback-prize-effectiveness/data

Dataset Description

Datasets

Data Source:

- The dataset contains **4161** argumentative **essays** written by U.S students in grades 6-12.
- These essays were annotated by expert raters for **7 discourse elements**

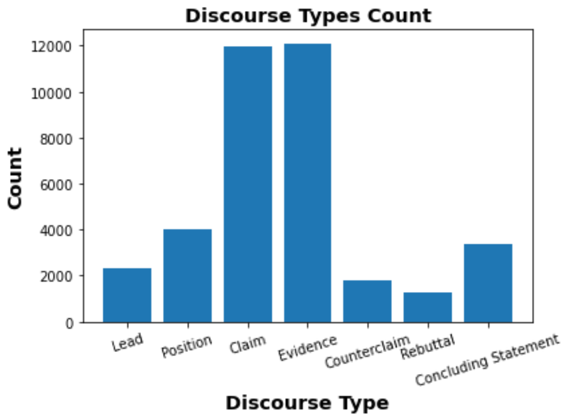
  1) Lead, 2) Position, 3) Claim, 4) Counterclaim, 5) Rebuttal, 6) Evidence & 7) Concluding Statement

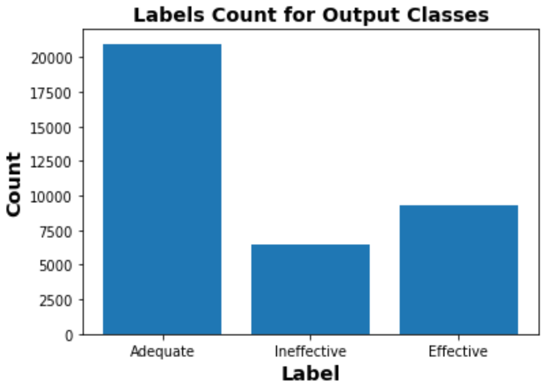
- **36765** discourse texts

Labelling:

Human readers rated each argumentative element, in order of increasing quality, as one of **3 Classes**:

1) Ineffective, 2) Adequate & 3) Effective





Inputs and Outputs

Essay: There is no life on Mars

Lead, **Adequate**

The story is about how NASA took a picture
and a face was seen on the planet. NASA doesn't
know if the landform was created by life on Mars, or if
it is just a natural landform.

Perspective, **Ineffective**

On my perspective, I think that the face is a natural
landform because I dont think that there is
Mars. In these next few paragraphs, I'll be
about how I think that is is a natural landform.

[...]

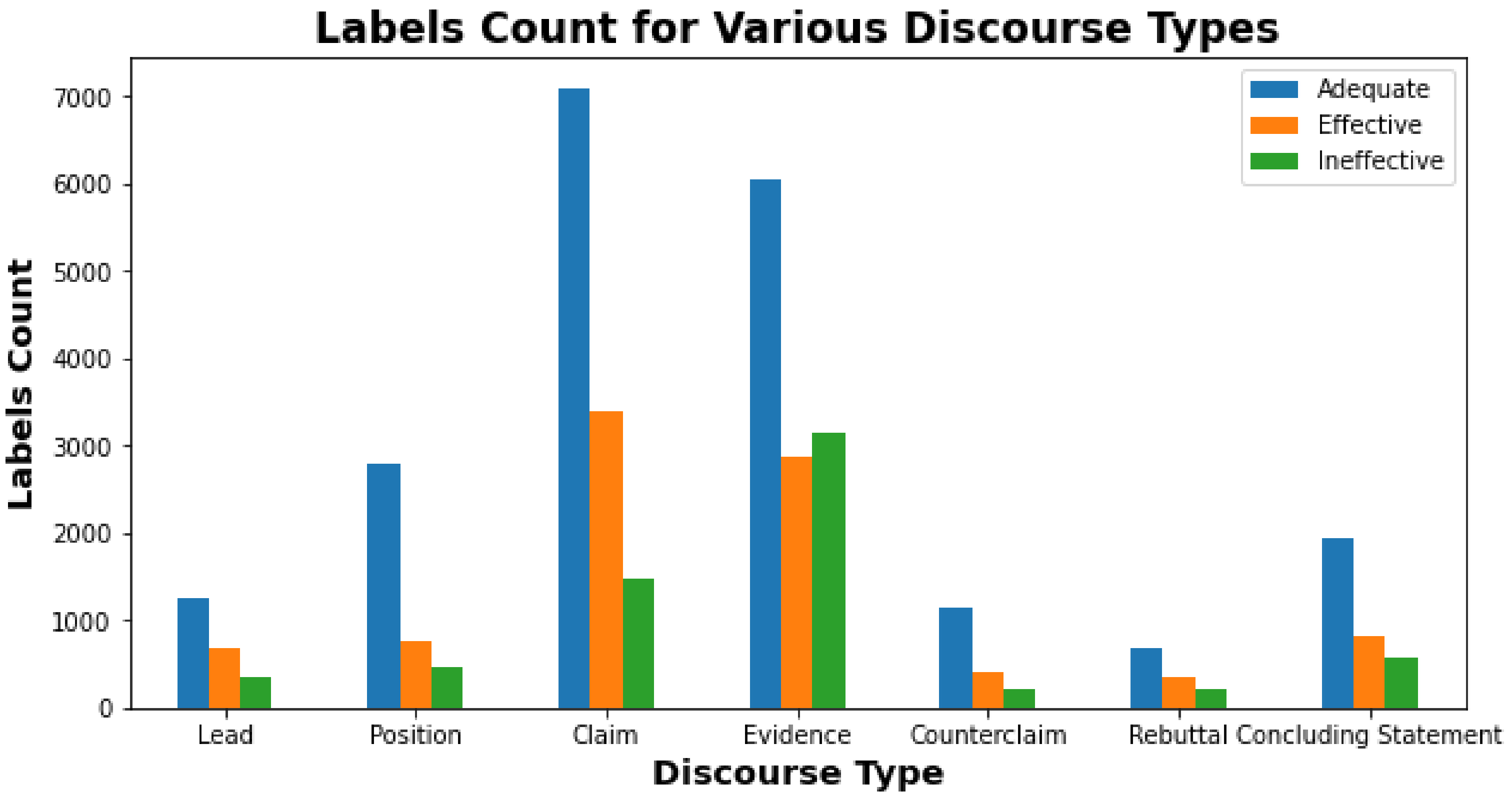
Solution Output

Discourse ID	Ineffective	Adequate	Effective
a261	0.2	**0.6**	0.2
5a8b	**0.7**	0.15	0.15

Goal → Predict the effectiveness rating for each discourse element

argumentation_scheme_and_rubrics_kaggle.docx. argumentation_scheme_and_rubrics_kaggle.docx. Google Docs. Published 2022. Accessed July 12, 2022.
https://docs.google.com/document/d/1G51Ulb0i-nKCRQSs4p4ujauy4wjAJOae/edit

Imbalance Training Data



Approach

Leveraging Expert Knowledge for Feature Selection

Grading Guidance:

The introduction begins with a statistic, a quotation, a description, or some other device to grab the reader's attention and point toward the thesis.

Classification:

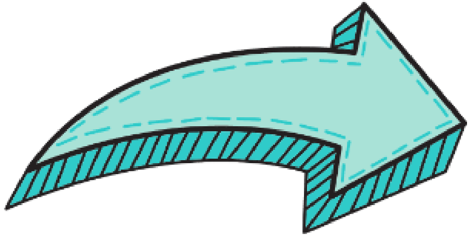
Effective

Adequate

Ineffective

Leveraging Expert Knowledge for Feature Selection

Discourse Element	Educator Response
The study the ability of humans to read subatle changes in facial expressions, thast they appiled reverse correlation technique to reveal visual features that mediate understanding of emotion expressed by the face. Suprising finding were that (1) the noise added to test face image had profound effect on the facail expression and (2) in most every istance the new expression was meaningful. [...]	I would suggest the evidence was graded as ineffective because it isn’t in direct support of a claim. Honestly, I don’t see a claim at all. This essay is really difficult to read/follow.



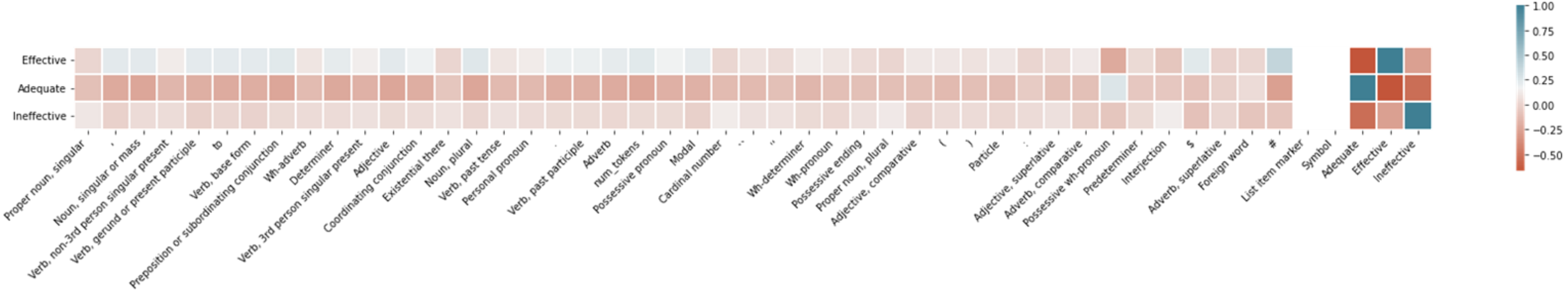
Include feature: Text Perplexity

Class Imbalance

Technique	Description
Downsampling	Reducing samples of majority class Not suited for this dataset (|Training Samples| = ~33k)
Upsampling	Replicating minority class samples. 2 * |Effective Samples| + 3 * |Ineffective Samples| + |Adequate Samples|
Focal Loss [1708.02002] Focal Loss for Dense Object Detection (arxiv.org)	Addresses class imbalance during training. Applies a modulating term to Cross-Entropy loss to focus on hard misclassified examples. (Gamma = 2) $Cross\ Entropy\ Loss(prob_t) = -log(prob_t)$ $Focal\ Loss(prob_t) = -(1 - prob_t)^{\gamma}\ log(prob_t)$
Weighted Sampling & Loss Function	Weighted mini-batch sampling. Weighted Cross-Entropy loss function [Effective: 3.93 / 1.5, Adequate: 1.75 / 1, Ineffective:5.69 / 2]

Feature Engineering

Feature Engineering: Part of speech



Part of Speech p	|Corr(p, Ineffective)|
Proper_Nouns	0.216405
misspelled_count	0.139763
Cardinal number	0.127971
Noun, Plural	0.121559
Existential There	0.074274
...	...

Not-In-Vocabulary Count

- ❖ Misspelled words are not penalized while grading Discourse elements.
- ❖ Subword Tokenizers in Large Language models - Out of Vocabulary and Misspelled Words.
- ❖ Correlation between the count of misspelled words in discourse text and effectiveness

Effectiveness Class	Effective	Adequate	Ineffective
Mean (Not in Vocabulary Words)	0.59	0.78	1.42

- ❖ Dictionary/Vocabulary is built from NLTK word corpora (words, brown and wordnet)
- ❖ |Vocabulary| - 346,423

Text Perplexity GPT-2

- ❖ Pre-trained Language Models are trained on Clean Corpora. The standard evaluation metric is perplexity:

$$Perplexity = \prod_{t=1}^{T} \left[\frac{1}{P_{LM}(x^{(t+1)} | x^{(t)}, \ldots x^{(1)})} \right]^{1/T}$$

- ❖ The lower the perplexity, the more probable (natural) the sentence is.
- ❖ We are using the Generative Pre-trained Transformer model (GPT2) to compute the perplexity score of the discourse text, and the normalized perplexity score is used as a feature.

Effectiveness Class	Effective	Adequate	Ineffective
Mean (Perplexity Score)	157.58	240.87	297.66
Median (Perplexity Score)	46.10	79.86	87.28
Standard Deviation (Perplexity Score)	1357.04	1569.65	1858.54

Model Architecture

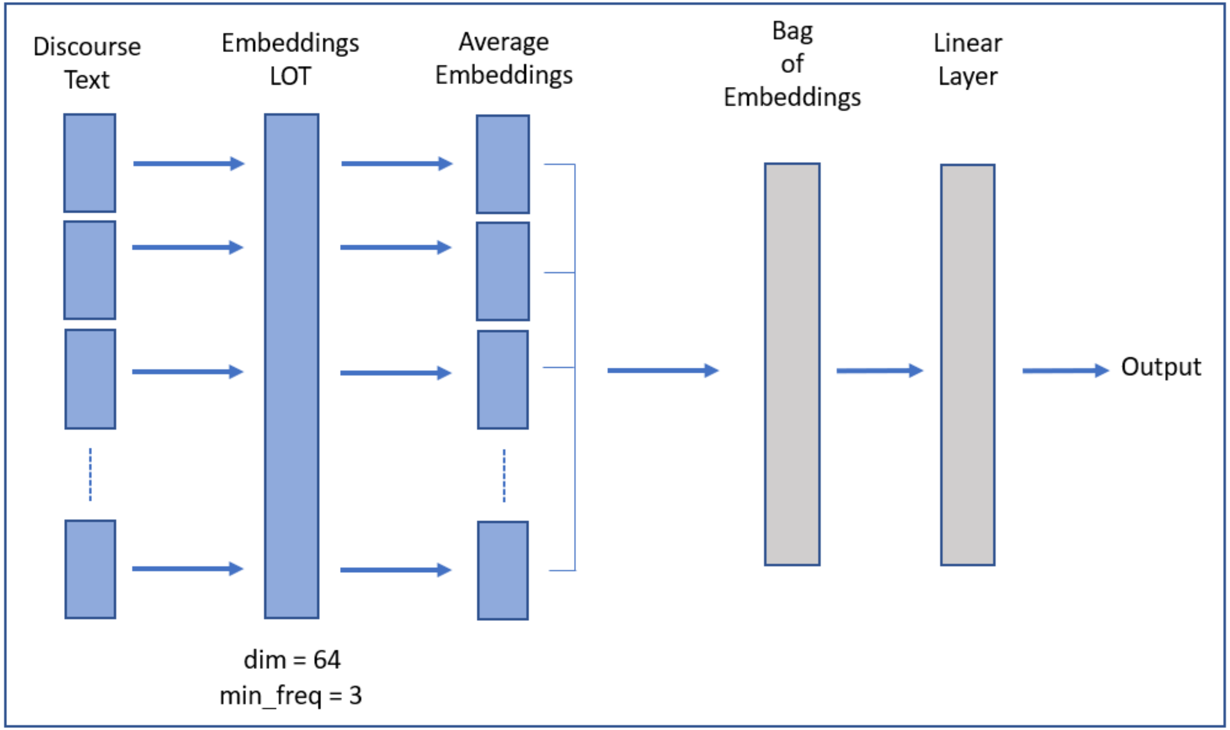
Bag of Embeddings

EPOCHS = 10

LEARNING RATE = 5

BATCH SIZE = 64

EMBEDDING SIZE = 64



Feed Forward Neural Network

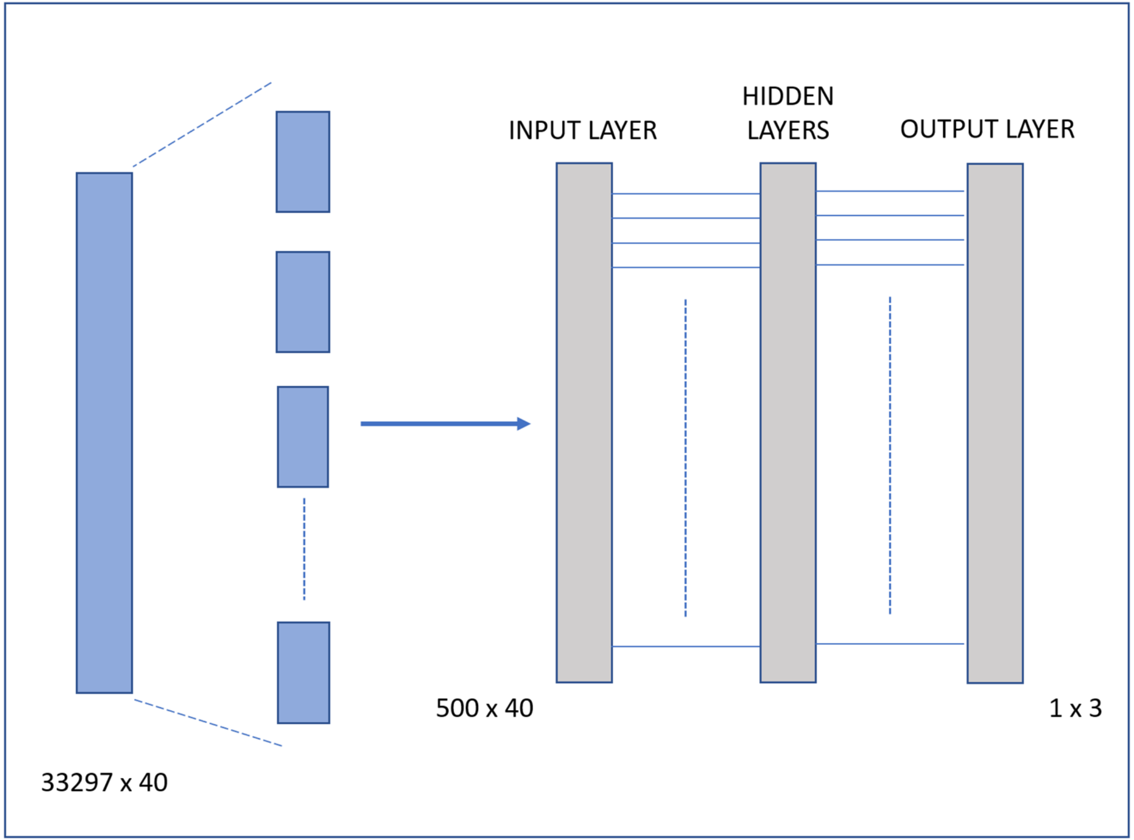
INPUT SIZE = 40

EPOCHS = 10

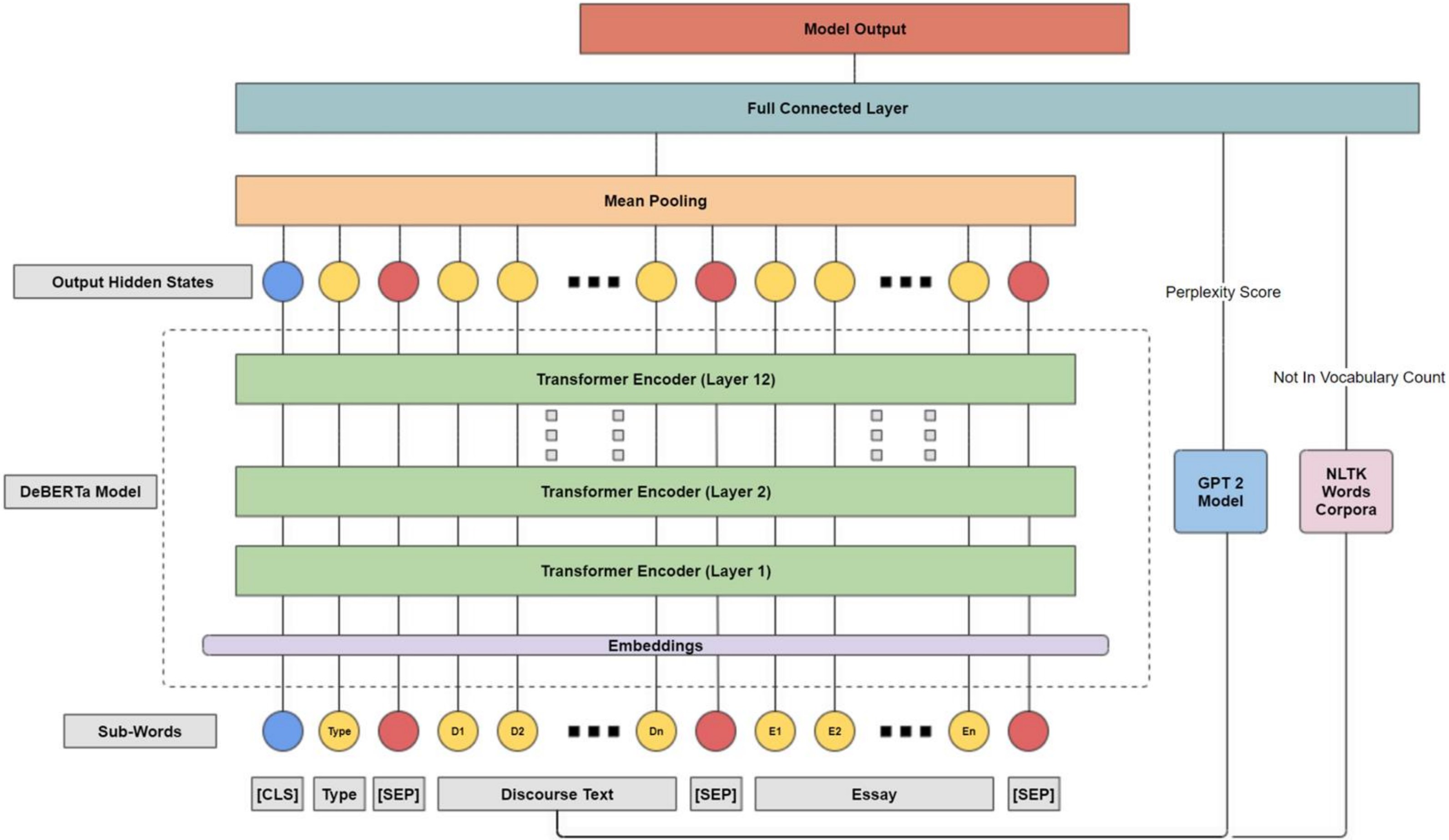
LEARNING RATE = 0.001

BATCH SIZE = 500

HIDDEN SIZE = 200



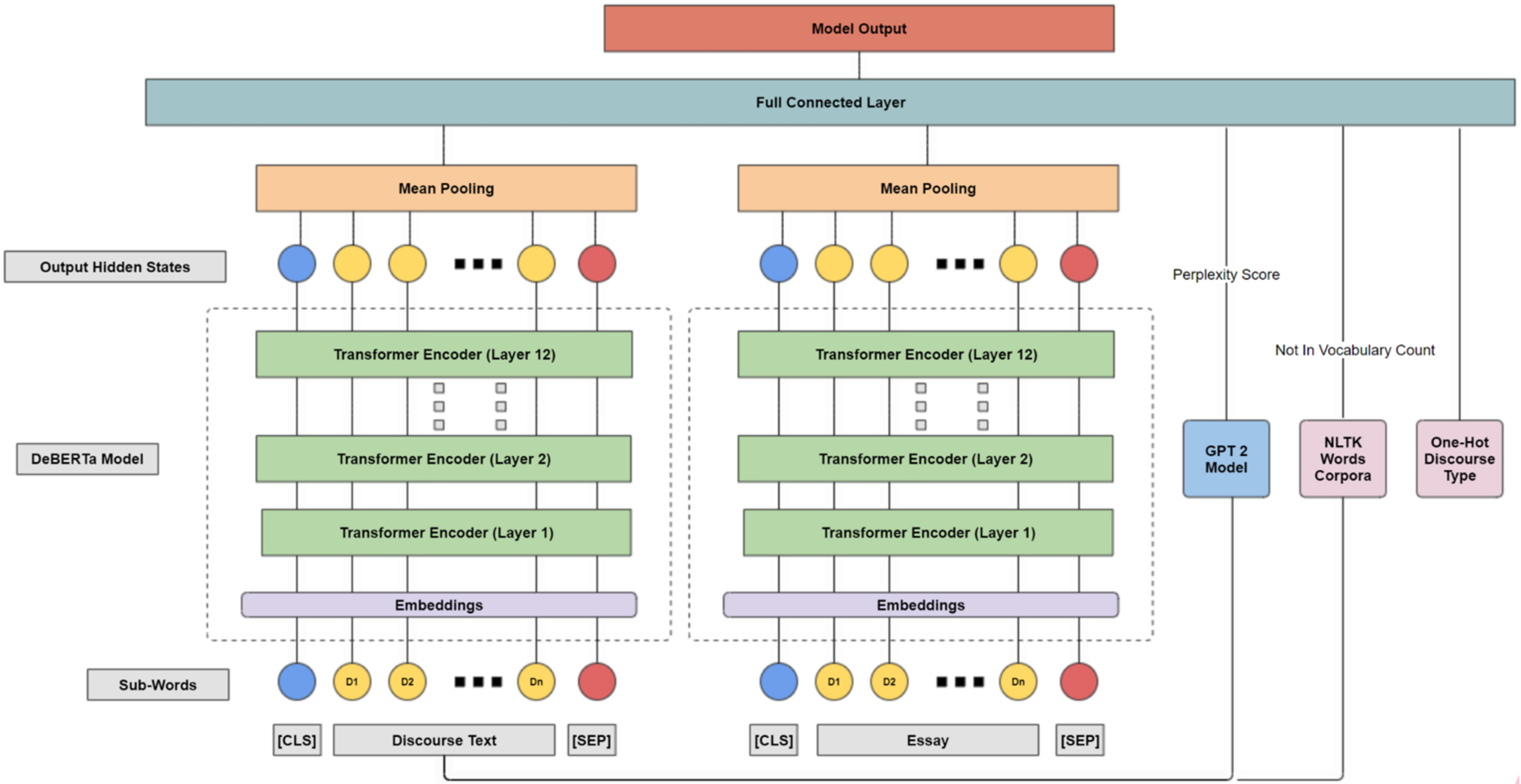
Augmented DeBERTa Network



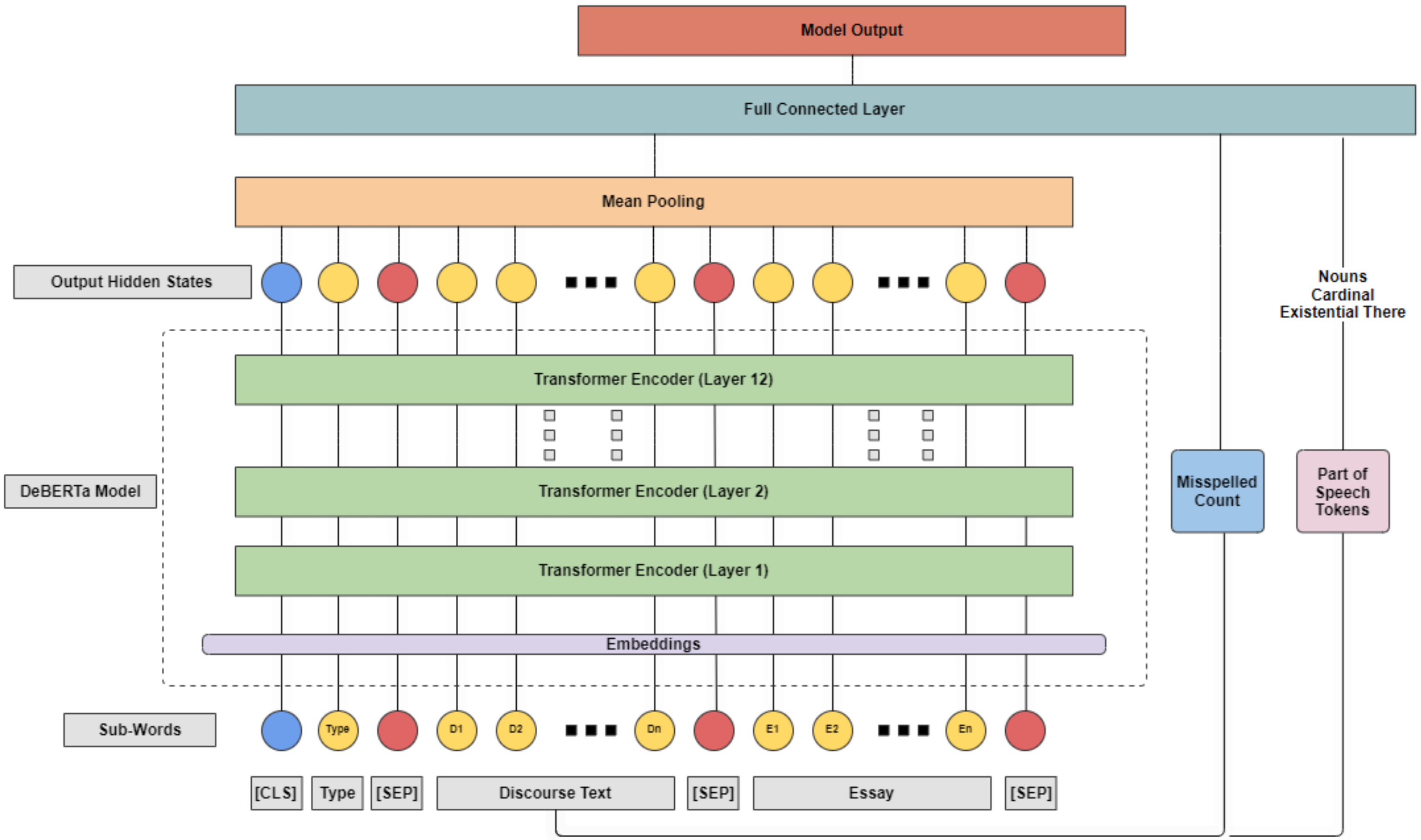
Experiment Details & Hyperparameters (Best Model)

Component / Hyperparameter	Description
Input Representation	The input text for the DeBERTa tokenizer is the concatenation of "Discourse Type", " Discourse Text", and "Argumentative Essay" separated by special tokens (start token: [CLS], separator token: [SEP]).
Input Text Length	Max token length is set to 1024, which ensures that the complete discourse text and type for all samples are represented, and a portion of the essay is covered.
DeBERTa Representation	Output hidden states for each sub-word is fed to the Mean Pooling layer, which computes the final representation considering the Mask values.
Feed Forward Input	Input to the Feed Forward Network consisted of representation from the DeBERTa model and augmented features of normalized GPT-2 Perplexity score and Not-In-Vocabulary count using NLTK word corpora.
#Hidden Units in Feed Forward Layer	128
Optimizer	AdamW
Scheduler	Linear Scheduler with Warmup (500 steps)
Loss Function	Cross Entropy Loss
Batch Size	4
#Epochs	5
Learning Rate	3e-6
Dropout	0.10
Weight Decay / L2 Regularization	0.005

Augmented 2-DeBERTa Network



Augmented 2-DeBERTa Network with PoS Counts



Results

Description	Parameters	Training Time	Log Loss	Balanced Accuracy	Test Accuracy
Bag of Embeddings [Input: Discourse Text, Embedding size: 64]	620K	50.5 sec	3.390	-	**61.0%**
Neural Network [Input: Features (POS) , Hidden layer: 200]	8.8k	4.7 sec	0.823	-	**63.5%**
BERT[Input: Discourse Text, BERT Hidden Layers: 6]	66.95M	2 hours 50 mins	0.777	56.38%	**66.70%**
RoBERTa [Input: Discourse Text, Essay, **RoBERTa** Hidden Layers: 8, Features: **Not-In-Vocabulary, Discourse Type**, 2 RoBERTa models]	193.38M	5 hours 35 mins	0.741	59.22%	**68.80%**
RoBERTa [Input: Discourse Text, Essay, RoBERTa Hidden Layers: 6, Features: Not-In-Vocabulary, Discourse Type, 2 RoBERTa models, **Weighted Loss**]	164.63M	4 hours 10 mins	0.741	62.61%	**65.66%**
DeBERTa-v3 [Input: Discourse Text, Essay, DeBERTa Hidden Layers: 12, Features: Not-In-Vocabulary, Discourse Type, 2 DeBERTa models, **Focal Loss**]	367.66M	10 hours 35 mins	0.681	59.93%	**69.98%**
DeBERTa-v3 [Input: Discourse Type, Discourse Text, Essay, **DeBERTA** Hidden Layers: 12, Features: **misspelled count**]	184.03M	6 hours 3 mins	0.657	64.63%	**70.6%**
DeBERTa-v3 [Input: Discourse Type, Discourse Text, Essay, DeBERTA Hidden Layers: 12, Features: misspelled count, **cardinal numbers, proper nouns, and existential there**.]	184.02M	4 hours 31 mins	0.660	62.98%	**70.0%**
DeBERTa-v3 [Input: Discourse Text, Essay, DeBERTa Hidden Layers: 12, Features: Not-In-Vocabulary, Discourse Type, **GPT-2 Perplexity Score**]	183.93M	10 hours 30 mins	0.660	63.05%	**71.11%**

Kaggle Submissions

5 submissions for ABC

Sort by Select...

All Successful Selected

Submission and Description	Status	Public Score	Use for Final Score
notebook108e887556 (version 1/1) 2 days ago by Vivek Bhatnagar Notebook notebook108e887556 | Version 1	Succeeded	0.684	☐
notebookf13e582216 (version 4/5) 3 days ago by Vivek Bhatnagar Notebook notebookf13e582216 | Version 4	Succeeded	0.704	☐
notebookd875aba482 (version 4/4) 4 days ago by Vivek Bhatnagar Notebook notebookd875aba482 | Version 4	Succeeded	0.749	☐
notebookd875aba482 (version 3/4) 4 days ago by Vivek Bhatnagar Notebook notebookd875aba482 | Version 3	Succeeded	0.718	☐
dummy_notebook_predicting Version 2 (version 2/2) 20 days ago by Adam Childs Notebook dummy_notebook_predicting | Version 2	Succeeded	13.948	☐

754	xuemc234	0.683	11	20d
755	**ABC**	0.684	5	2d

Your Best Entry!
Your most recent submission scored 0.684, which is an improvement of your previous score of 0.704. Great job!

Tweet this

756	Panggelia	0.684	38	5d

Conclusions

- Prediction with unconstrained content and qualitative labels is non-trivial
- Domain Experts help when Feature Engineering
- Teams: Parallel model development and verification
- Abandon features when they aren’t helping
- Ensemble models can improve accuracy

References

1. Feedback Prize - Predicting Effective Arguments | Kaggle. Kaggle.com. Published 2022. Accessed July 12, 2022. https://www.kaggle.com/competitions/feedback-prize-effectiveness/data.

2. argumentation_scheme_and_rubrics_kaggle.docx. argumentation_scheme_and_rubrics_kaggle.docx. Google Docs. Published 2022. Accessed July 12, 2022. https://docs.google.com/document/d/1G51Ulb0i-nKCRQSs4p4ujauy4wjAJOae/edit