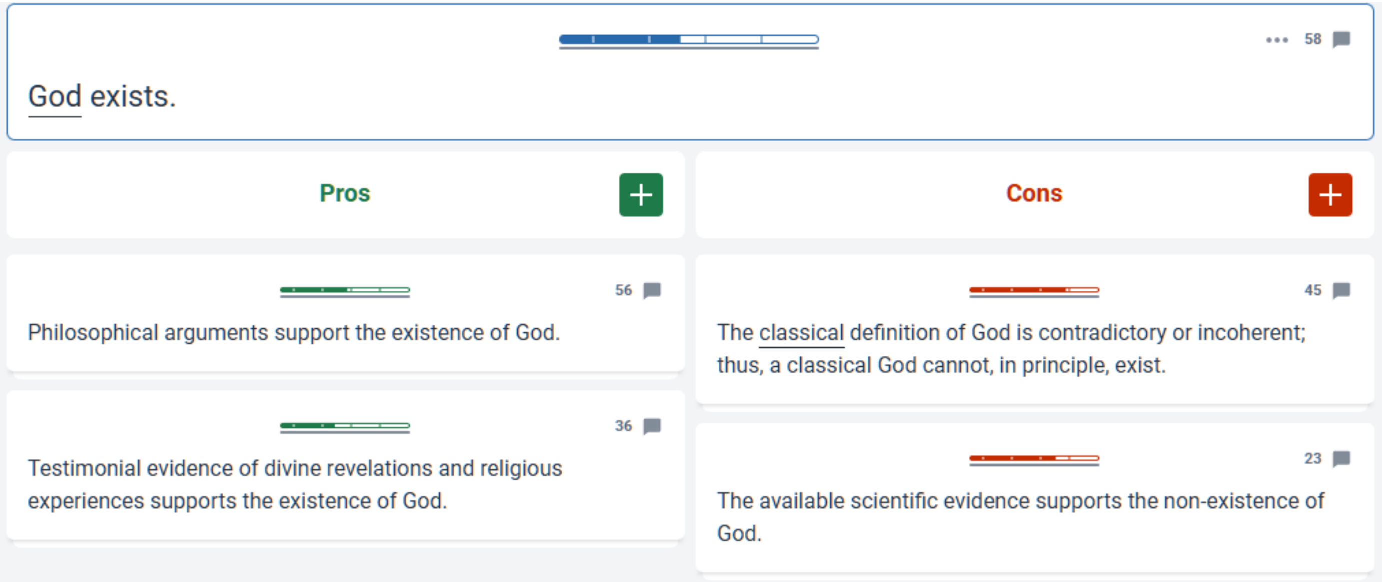
Argument Move Classification



Clankas

Lachlan Wallbridge (z5359327)

Dhruv Joshi (z5448031)

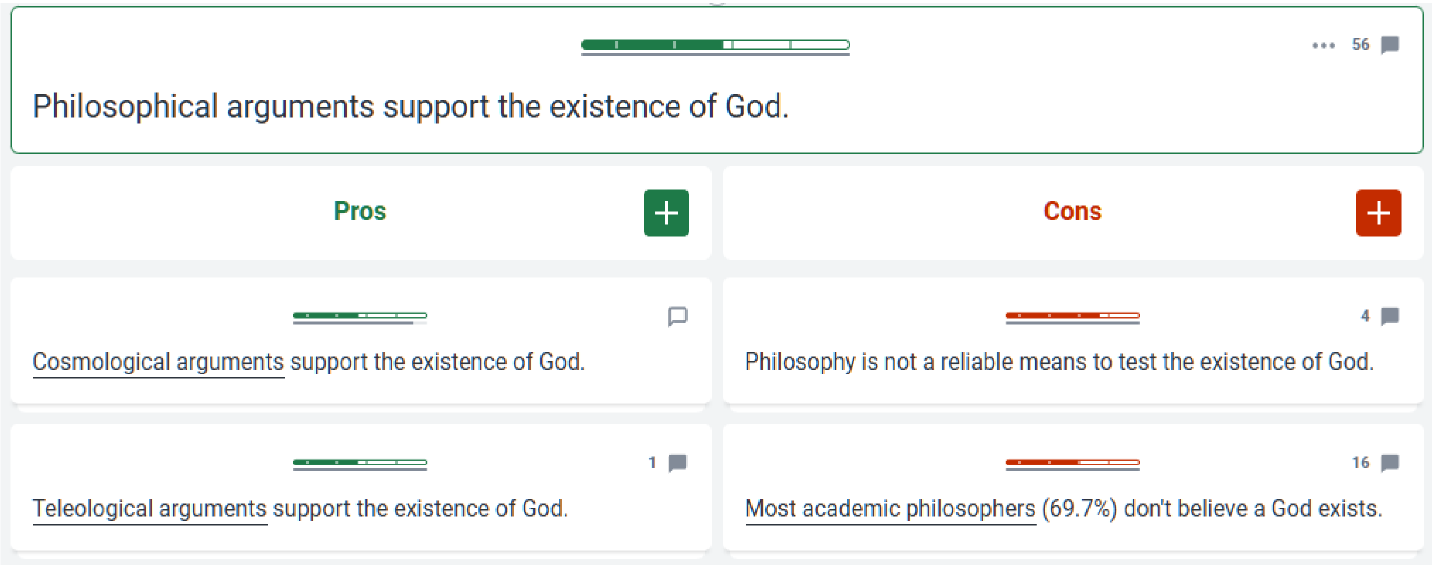
Arkar Myo (z5459275)

Myles Benyon (z5363307)

Farhan Raonak (z5481648)

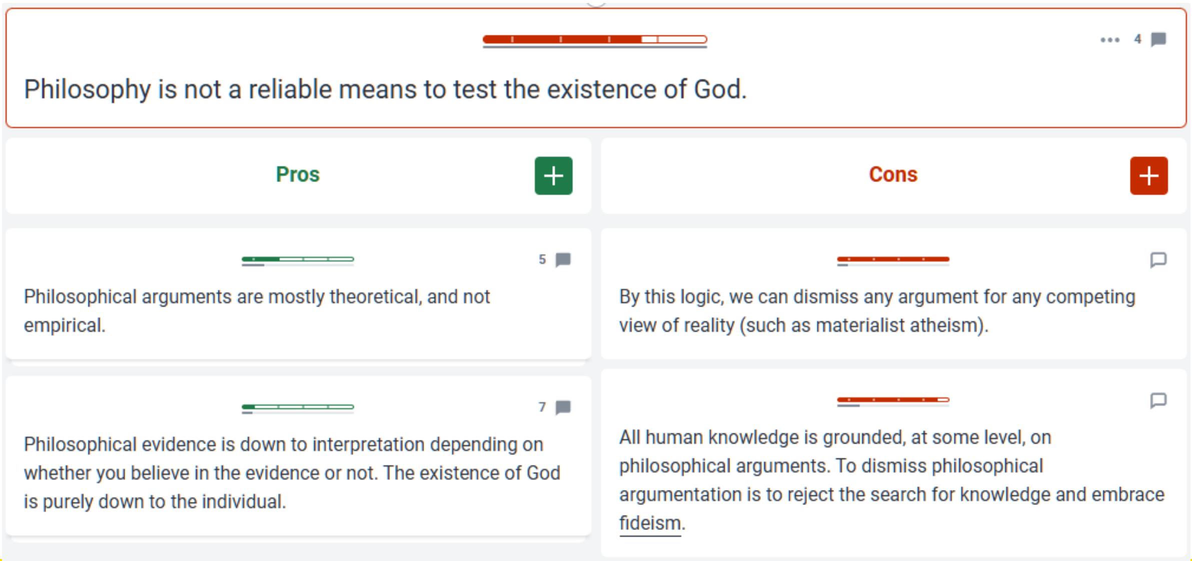
Motivation

- Polarity Prediction: Identifying if the given text agrees or disagrees with a point made previously.
- Natural Language Processing Model
- Applications for such a model:
  - Live debates
  - Polling
  - Generalising opinions
  - Providing detailed summaries
  - Separating hate speech
  - Discerning ambiguity



Problem Statement

- Argumentative discussions involve opinionated content.
- In said discussions, participants try to convince each other of their point of view
- Correctly understanding each other's points is key to achieving a healthy conversation
- The aim is to create a model that can correctly predict the polarity of a given argument



Literature Review – ACM Digital Library

Graph NLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates

Dataset:

- Kialo (1,560 discussion threads, mean of 204 arguments per topic with standard deviation of 463)

Method:

- Converts Kialo arguments into tree graphs
- Utilises advanced context by using graph walks to further understand discussions

Model	Accuracy (%)
Bag-of-Words + Logistic Regression	67.00
Prompt Embeddings + Logistic Regression	61.20
Sentence-BERT with classifier layer	79.86
BERT Embeddings: Root-seeking Graph Walk + MLP	70.27
GraphNLI: Root-seeking Graph Walk + Sum	80.70
GraphNLI: Root-seeking Graph Walk + Avg.	81.96
GraphNLI: Root-seeking Graph Walk + Weighted Avg.	**82.87**
GraphNLI: Biased Root-seeking Random Walk + Sum	79.95
GraphNLI: Biased Root-seeking Random Walk + Avg.	80.44

Results:

Outperforms relevant baselines. Overall accuracy of 82.87% using the Facebook AI / roBERTa-base model that has 125M parameters

Literature Review – MIT Press Direct

Classifying Argumentative Relations Using Logical Mechanisms and Argumentation Schemes

Dataset:

- Kialo (1,417 topics split into 2 subsets. Normative and Non-normative arguments)
- Debatepedia (508 topics with 15K pro and con responses)

	Normative Arguments						Non-normative Arguments					
	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	$F1_{neu}$
1 Random	33.5	50.2	32.6	27.8	30.1	39.9	33.4	49.9	32.5	28.7	28.8	40.0
2 Sentiment	40.8	64.1	40.7	40.6	39.1	42.4	43.7	61.1	42.2	40.0	35.2	51.5
3 Text Entail	51.8	61.8	36.7	12.8	30.4	67.0	52.1	62.8	38.6	18.4	31.0	66.4
4 PSL (R1–R13)	54.0‡	73.8‡	52.1‡	47.0‡	43.6‡	65.7‡	57.0‡	76.0‡	54.0‡	50.1‡	42.6‡	69.3‡
5 \ Fact	55.1‡	74.3‡	52.4‡	47.1‡	41.6‡	68.4‡	58.6‡	77.1‡	55.1‡	50.5‡	42.2‡	72.7‡
6 \ Sentiment	**62.1**‡	77.6‡	57.5‡	49.1‡	45.8‡	**77.7**‡	61.3‡	77.8‡	56.7‡	50.3‡	44.1‡	75.7‡
7 \ Causal	54.4‡	73.1‡	52.3‡	45.4‡	45.4‡	66.0‡	57.6‡	76.1‡	54.3‡	48.7‡	43.4‡	70.7‡
8 \ Normative	51.8‡	68.6‡	49.4‡	44.3‡	40.4†	63.4‡	54.7‡	70.3‡	51.4‡	47.0‡	40.3‡	66.8‡
9 \ Sentiment + Chain	61.9‡	**77.7**‡	**57.7**‡	**49.3**‡	**46.2**‡	77.6‡	**61.5**‡	**78.0**‡	**57.2**‡	**50.8**‡	**44.7**‡	**76.1**‡

(a) Kialo

Method:

4 logical and theory informed mechanisms :

1. Factual Consistency
2. Sentiment Coherence
3. Causal Relation
4. Normative Relation

	Normative Arguments					Non-normative Arguments				
	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$	ACC	AUC	F1	$F1_{sup}$	$F1_{att}$
1 Random	47.7	49.4	50.2	49.0	51.4	53.0	54.6	52.4	53.7	51.1
2 Sentiment	59.3	63.9	59.2	61.0	57.4	69.1	73.4	68.5	72.7	64.3
3 Text Entail	52.2	55.8	49.4	37.6	61.2	70.6	74.2	70.5	69.0	72.0
4 PSL (R1–R13)	**63.9***	**68.3***	**63.9***	63.8	64.0†	73.0	76.1	73.0	74.2	71.7
5 \ Fact	63.4*	67.1	63.4*	**64.0**	62.7*	71.8	75.6	71.7	73.2	70.3
6 \ Sentiment	63.1*	67.2	63.1*	62.7	63.5*	70.9	74.0	70.9	71.6	70.2
7 \ Causal	62.4*	66.3	62.1*	58.6	**65.5***	**74.5**	**78.7**	**74.5**	**75.4**	**73.6**
8 \ Normative	61.0	64.7	61.0	60.3	61.6*	68.2	72.4	68.2	68.3	68.1

(b) Debatepedia

R14: $\text{Support}(S, I) \wedge \text{Support}(I, C) \rightarrow \text{Support}(S, C)$,

R15: $\text{Attack}(S, I) \wedge \text{Attack}(I, C) \rightarrow \text{Support}(S, C)$,

R16: $\text{Support}(S, I) \wedge \text{Attack}(I, C) \rightarrow \text{Attack}(S, C)$,

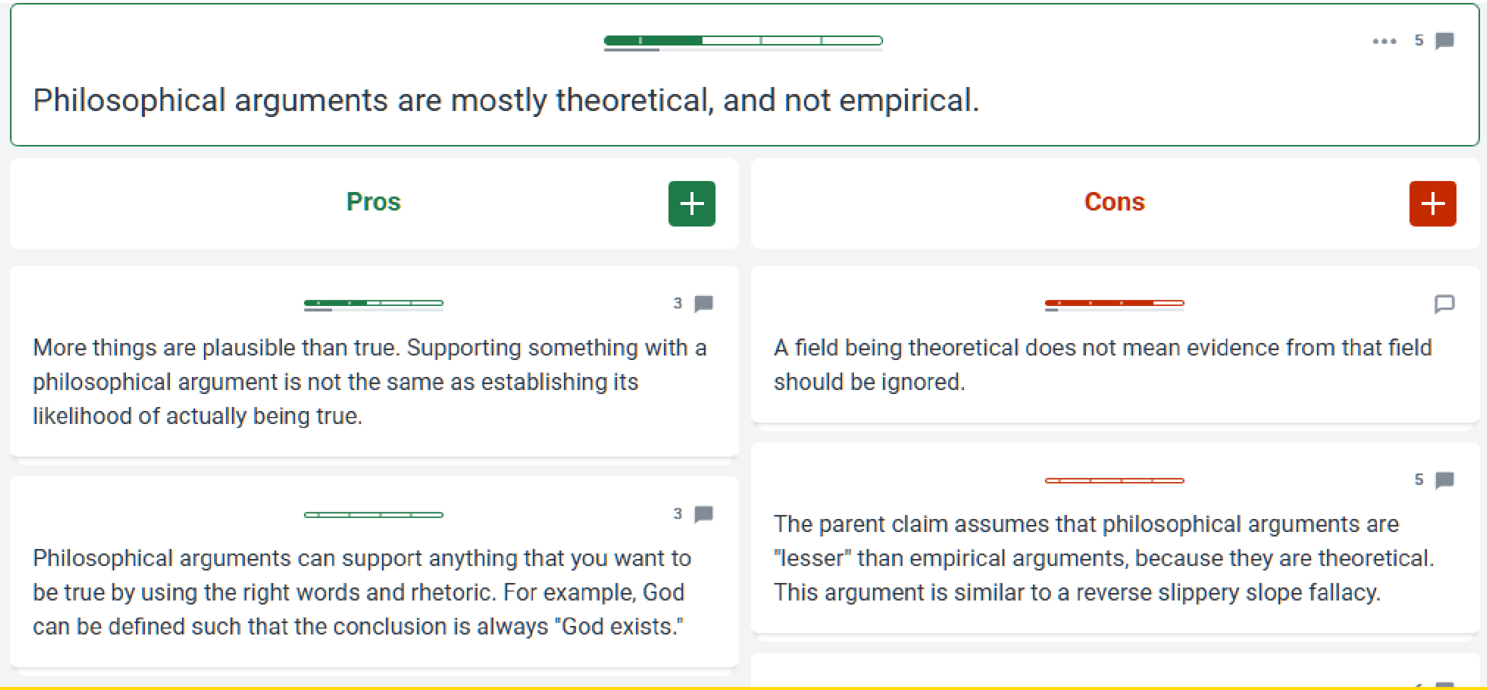
R17: $\text{Attack}(S, I) \wedge \text{Support}(I, C) \rightarrow \text{Attack}(S, C)$.

Results:

Significant improvement over other unsupervised models, highest accuracy achieved was 74.5%

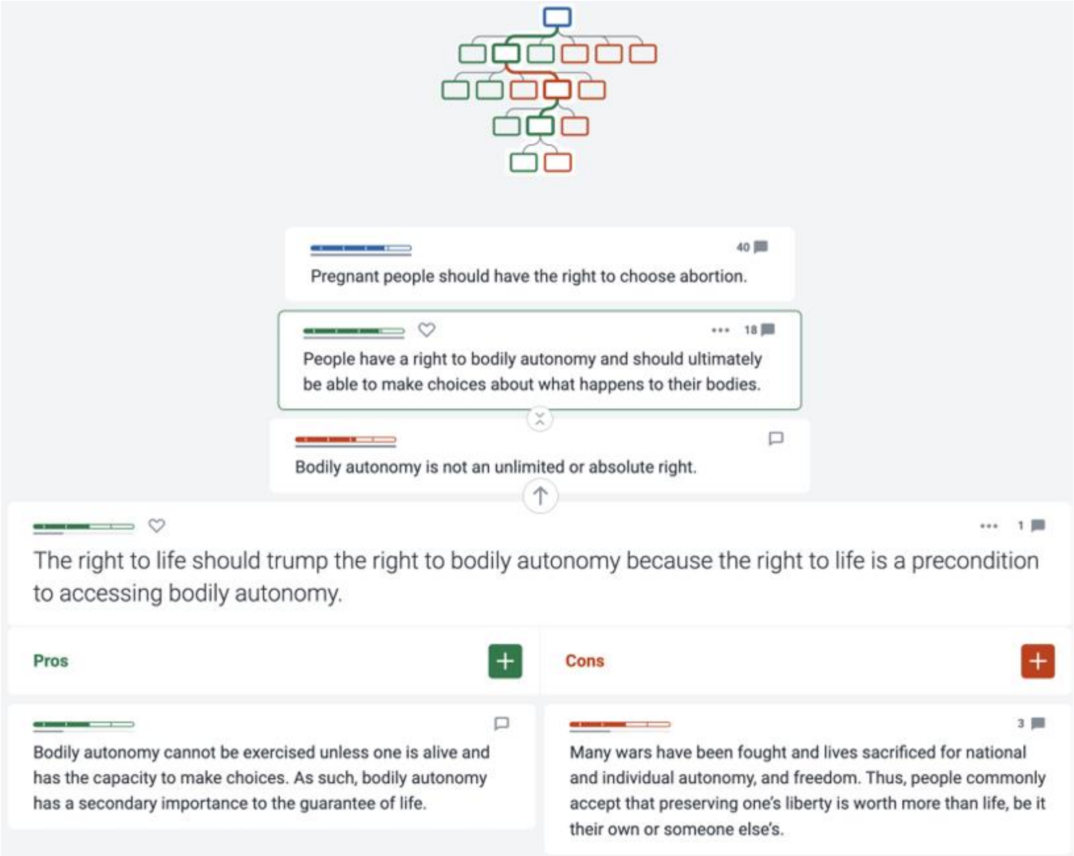
Dataset

- Sourced from Kialo
- Contains 33,000 files of type .pkl, each of which are transcripts of conversations and debates of different topics
- https://netsys.surrey.ac.uk/datasets/graphnli/orig_data/



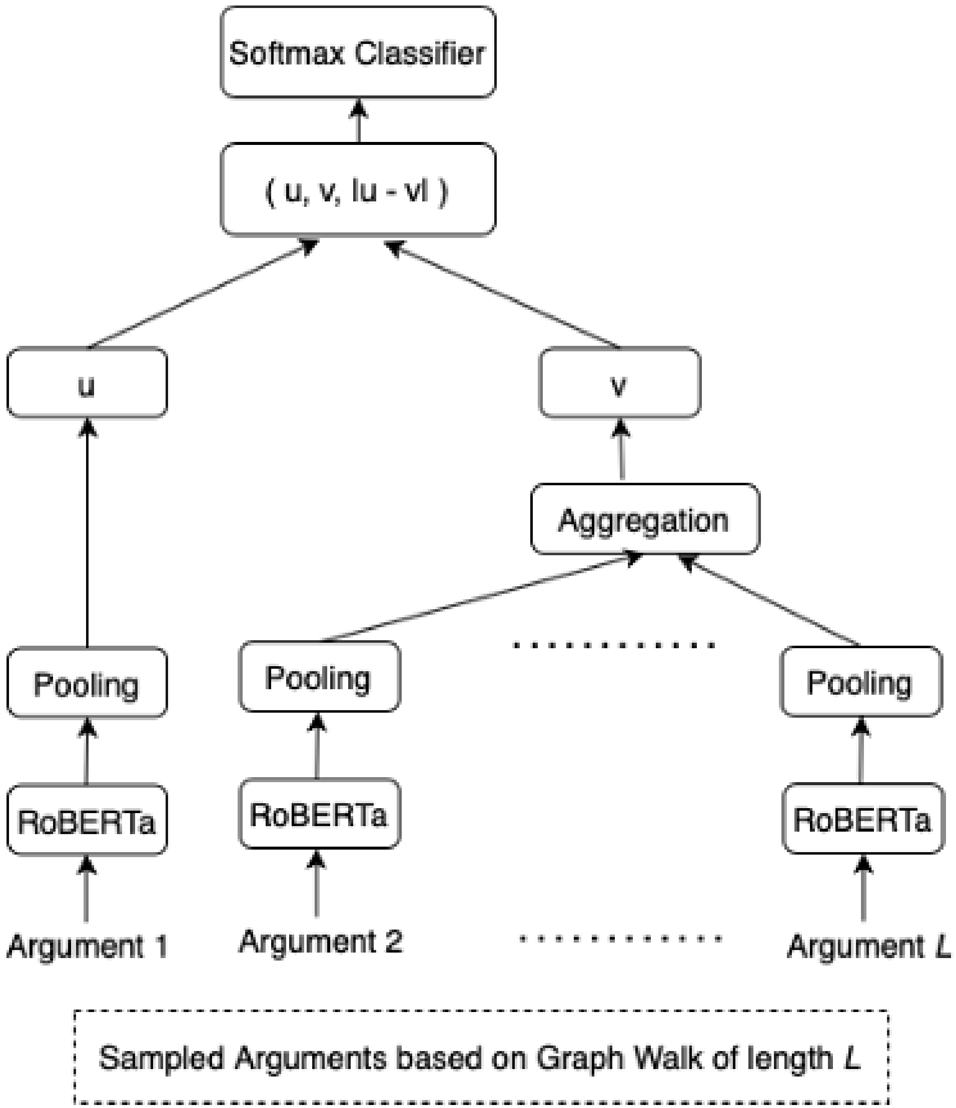
Data Analysis

- Each .pkl file represents a Kialo graph about a certain topic
- Using the graph walk approach, we generate a sequence of nodes for our model.
- There are multiple walk strategies available:
  - Random walk
  - Biased Root Seeking Random Walk
  - Weighted Root Seeking Graph Walk (works best)
- Higher walk length increased training time but provides more argument context.



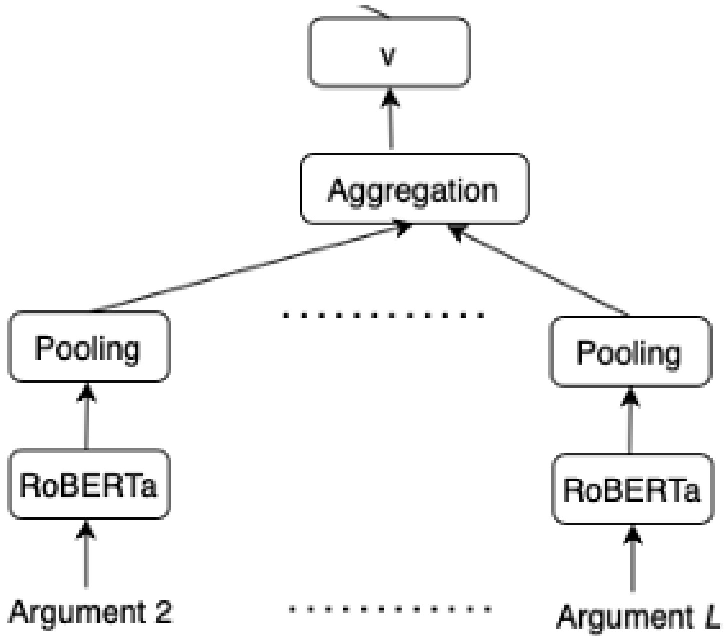
Model Architecture

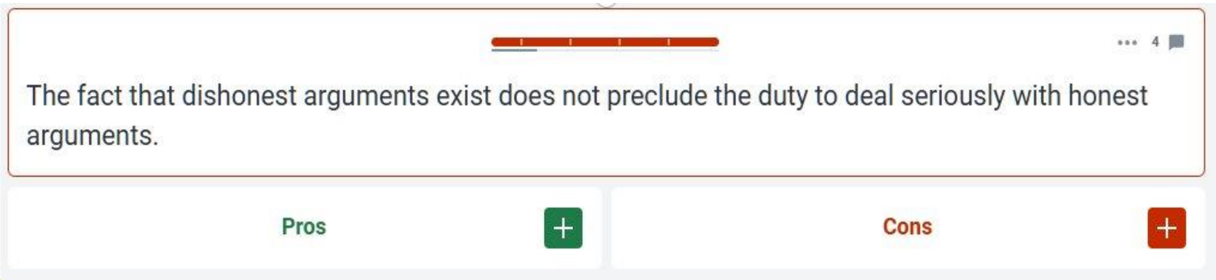
- **Sentence Transformer model:**
  Encodes the raw text of each argument node into a sequence of token embeddings.
- **Pooling:**
  Applies mean pooling to the token embeddings to produce a single fixed-size vector
- **Aggregation:**
  Merges vectors from parent nodes using an aggregation function. (e.g. Mean, Weighted, Sum)
- **SoftMax Layer:**
  Uses the concatenation of 2 vector inputs:
  1. The sentence embedding of the current argument
  2. The aggregated vector from its parent arguments.



Method – Aggregation

- Combines arguments into a single sentence embedding
- Aggregation methods:
  - **Mean**: Highlights the average context of the argument chain. Equal weight for all nodes.
  - **Sum**: Aggregates argument embeddings by accumulating their content.
  - **Weighted**: Implements decaying importance of "far" arguments. A loss of 0.75 at each node was used.





Method – Sentence Transformer model

- Compared to the baseline Sentence Transformer
  - **distil-roberta** (82.8M params) (50K Vocab)

- We tested the following:
  - **all-MiniLM-L6-v2** (22.7M params) (30.5K Vocab)
  - **DeBERTa-V3-small** (44M params) (128K Vocab)
  - **DeBERTa-V3-xsmall** (22M params) (128K Vocab)

- Constants for evaluation:

Batch Size: 16

Sequence Length: 256

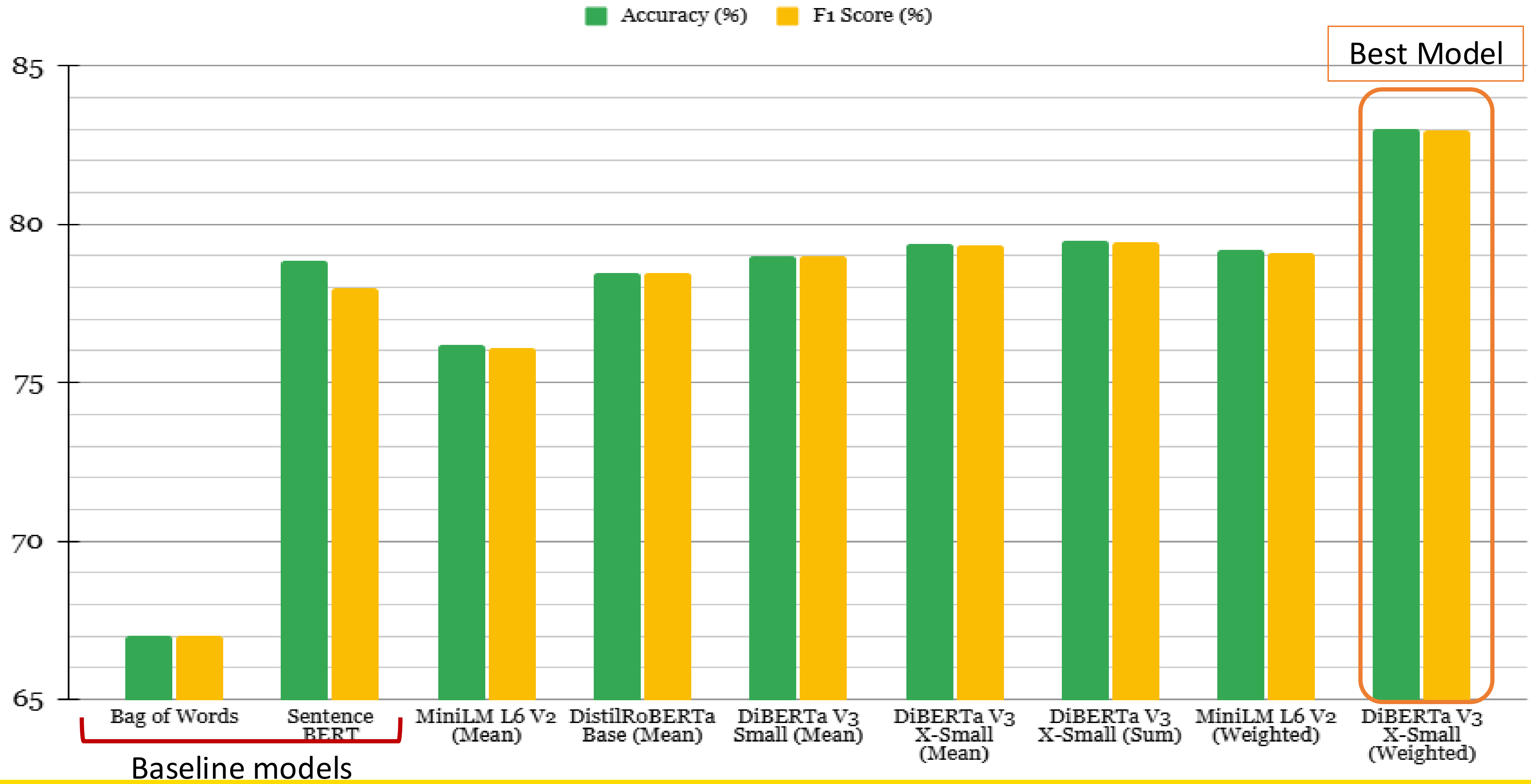
Walk Length: 4

Fixed graph walks + dataset

Mean Pooling

Model	Vocabulary(K)	Backbone #Params(M)	SQuAD 2.0(F1/EM)	MNLI-m/mm(ACC)
RoBERTa-base	50	86	83.7/80.5	87.6/-
XLNet-base	32	92	-/80.2	86.8/-
ELECTRA-base	30	86	-/80.5	88.8/
DeBERTa-base	50	100	86.2/83.1	88.8/88.5
DeBERTa-v3-large	128	304	91.5/89.0	91.8/91.9
DeBERTa-v3-base	128	86	88.4/85.4	90.6/90.7
DeBERTa-v3-small	128	44	82.8/80.4	88.3/87.7
DeBERTa-v3-xsmall	128	**22**	**84.8/82.0**	**88.1/88.3**
DeBERTa-v3-xsmall+SiFT	128	22	-/-	88.4/88.5

Results – Model Comparison



Results – Score Comparison

Models	Accuracy (%)	F1 Score (%)	Precision	Recall	ROC AUC
Bag of Words	67	67	67	67	0.75
Sentence BERT	78.86	78	78.43	78.86	0.855
MiniLM L6 V2 (Mean)	76.18	76.08	76.08	76.18	0.84
DistilRoBERTa Base (Mean)	78.48	78.46	78.47	78.48	0.86
DeBERTa-v3-small (Mean)	79.01	78.98	78.99	79.01	0.863
DeBERTa-v3-xsmall (Mean)	79.36	79.32	79.31	79.36	0.8706
DeBERTa-v3-xsmall (Sum)	79.48	79.45	79.44	79.48	0.8723
MiniLM L6 V2 (Weighted)	79.19	79.11	79.12	79.19	0.8667
DeBERTa-v3-xsmall (Weighted)	83.01	82.98	82.97	83.01	0.902

Best Results and Discussion

- **Accuracy:** 0.8301
- **F1 Score:** 0.8298
- **Precision:** 0.8297
- **Recall:** 0.8301
- **ROC AUC:** 0.9020

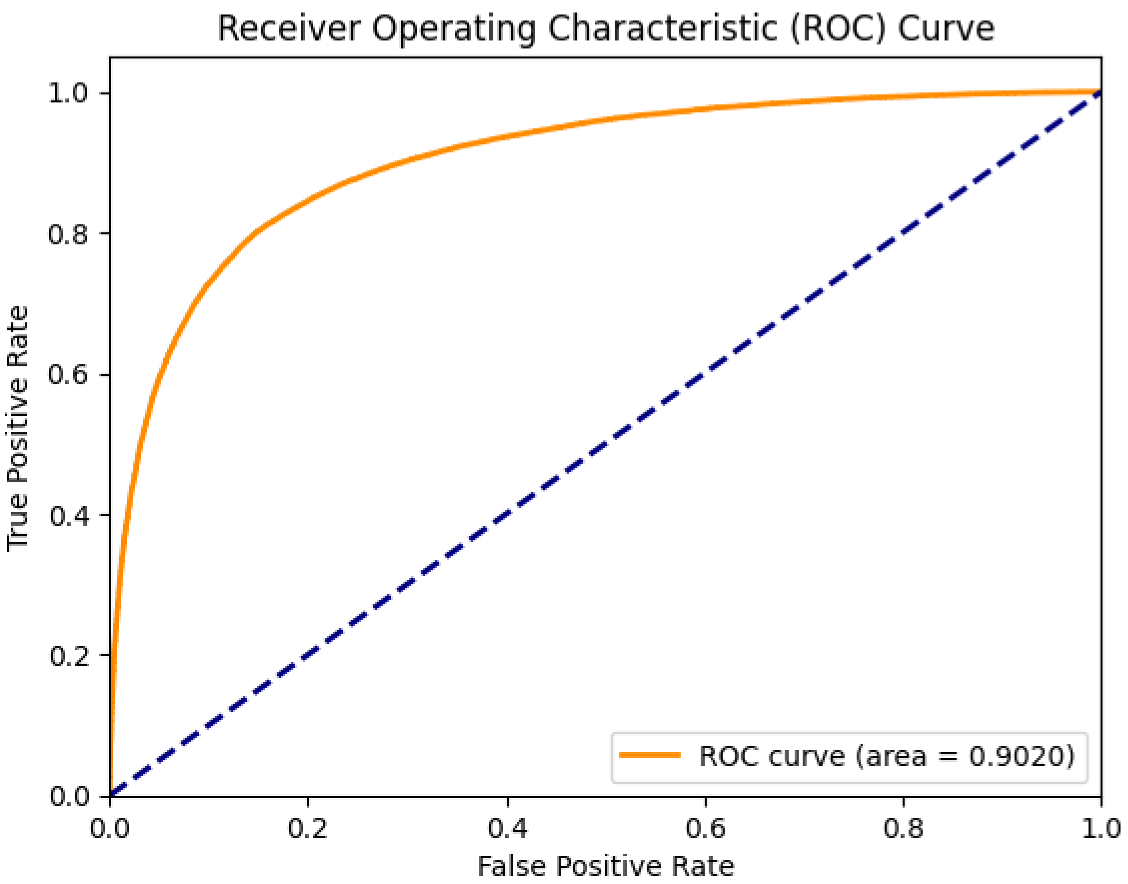
Confusion Matrix:

[[**31850** 5150]
[5873 **22001**]]

```
Classification Report:
              precision    recall  f1-score   support

           0     0.8443    0.8608    0.8525     37000
           1     0.8103    0.7893    0.7997     27874

    accuracy                         0.8301     64874
   macro avg     0.8273    0.8251    0.8261     64874
weighted avg     0.8297    0.8301    0.8298     64874
```



Final Discussion

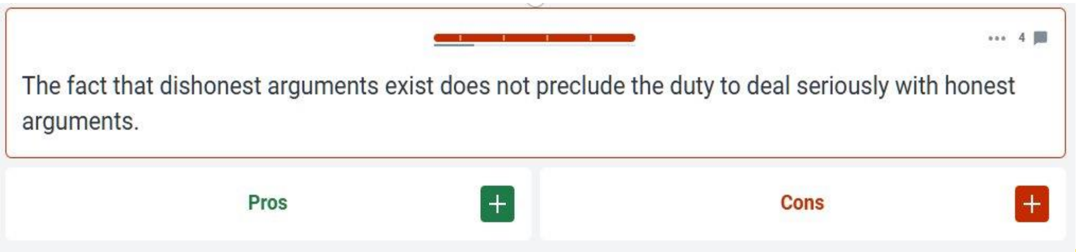
- What compromises were implemented:
  - Unable to test all walk lengths and aggregation types with all models due to timing constraints.
  - Focused on Sentence Transformer and aggregation as main areas of improvement.

- Final Model success:
  - Our model outperforms the baselines and larger models with many less parameters.

- What improvements could have been made:
  - Larger pretrained models will just be better (e.g. DeBERTa-v3-large).
  - Testing of longer walk lengths and additional aggregation methods.
  - More advanced pooling methods.
  - Full Kailo dataset.

Conclusion

- Improved **training and testing times** with higher accuracy.
- Achieved highest accuracy with weighted aggregation (83.14% Accuracy).
- A model like this can be utilised to perform a wide variety of applications.
- Further research could involve training and testing on noiser debating platforms like Reddit or Twitter.



Questions