

# Argument Move Classification

God exists.

Pros

Cons

Philosophical arguments support the existence of God.

The classical definition of God is contradictory or incoherent; thus, a classical God cannot, in principle, exist.

Testimonial evidence of divine revelations and religious experiences supports the existence of God.

The available scientific evidence supports the non-existence of God.

## 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

The screenshot displays a digital debate platform. At the top, a statement reads: "Philosophical arguments support the existence of God." Below this, a table organizes arguments into two columns: "Pros" (green header) and "Cons" (red header). Each argument entry includes a progress bar, the text of the argument, and a comment icon with a count.

Pros	Cons
<u>Cosmological arguments</u> support the existence of God.	Philosophy is not a reliable means to test the existence of God.
<u>Teleological arguments</u> support the existence of God.	<u>Most academic philosophers</u> (69.7%) don't believe a God exists.

# 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

The screenshot shows a web application interface for analyzing arguments. At the top, a text box contains the statement: "Philosophy is not a reliable means to test the existence of God." Below this, the interface is divided into two columns: "Pros" (green header) and "Cons" (red header). Each column contains two argument cards. The "Pros" column has two cards with green progress bars and 5 and 7 comments respectively. The "Cons" column has two cards with red progress bars and 4 and 1 comment respectively. Each card contains a specific argument related to the main topic.

Pros	Cons
Philosophical arguments are mostly theoretical, and not empirical.	By this logic, we can dismiss any argument for any competing view of reality (such as materialist atheism).
Philosophical evidence is down to interpretation depending on whether you believe in the evidence or not. The existence of God is purely down to the individual.	All human knowledge is grounded, at some level, on philosophical arguments. To dismiss philosophical argumentation is to reject the search for knowledge and embrace fideism.

# 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.	<b>82.87</b>
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)

## Method:

4 logical and theory informed mechanisms :

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

## Results:

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

	Normative Arguments						Non-normative Arguments					
	ACC	AUC	F1	F1 <sub>sup</sub>	F1 <sub>att</sub>	F1 <sub>neu</sub>	ACC	AUC	F1	F1 <sub>sup</sub>	F1 <sub>att</sub>	F1 <sub>neu</sub>
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 <sup>‡</sup>	73.8 <sup>‡</sup>	52.1 <sup>‡</sup>	47.0 <sup>‡</sup>	43.6 <sup>‡</sup>	65.7 <sup>‡</sup>	57.0 <sup>‡</sup>	76.0 <sup>‡</sup>	54.0 <sup>‡</sup>	50.1 <sup>‡</sup>	42.6 <sup>‡</sup>	69.3 <sup>‡</sup>
5 \ Fact	55.1 <sup>‡</sup>	74.3 <sup>‡</sup>	52.4 <sup>‡</sup>	47.1 <sup>‡</sup>	41.6 <sup>‡</sup>	68.4 <sup>‡</sup>	58.6 <sup>‡</sup>	77.1 <sup>‡</sup>	55.1 <sup>‡</sup>	50.5 <sup>‡</sup>	42.2 <sup>‡</sup>	72.7 <sup>‡</sup>
6 \ Sentiment	<b>62.1<sup>‡</sup></b>	77.6 <sup>‡</sup>	57.5 <sup>‡</sup>	49.1 <sup>‡</sup>	45.8 <sup>‡</sup>	<b>77.7<sup>‡</sup></b>	61.3 <sup>‡</sup>	77.8 <sup>‡</sup>	56.7 <sup>‡</sup>	50.3 <sup>‡</sup>	44.1 <sup>‡</sup>	75.7 <sup>‡</sup>
7 \ Causal	54.4 <sup>‡</sup>	73.1 <sup>‡</sup>	52.3 <sup>‡</sup>	45.4 <sup>‡</sup>	45.4 <sup>‡</sup>	66.0 <sup>‡</sup>	57.6 <sup>‡</sup>	76.1 <sup>‡</sup>	54.3 <sup>‡</sup>	48.7 <sup>‡</sup>	43.4 <sup>‡</sup>	70.7 <sup>‡</sup>
8 \ Normative	51.8 <sup>‡</sup>	68.6 <sup>‡</sup>	49.4 <sup>‡</sup>	44.3 <sup>‡</sup>	40.4 <sup>‡</sup>	63.4 <sup>‡</sup>	54.7 <sup>‡</sup>	70.3 <sup>‡</sup>	51.4 <sup>‡</sup>	47.0 <sup>‡</sup>	40.3 <sup>‡</sup>	66.8 <sup>‡</sup>
9 \ Sentiment + Chain	61.9 <sup>‡</sup>	<b>77.7<sup>‡</sup></b>	<b>57.7<sup>‡</sup></b>	<b>49.3<sup>‡</sup></b>	<b>46.2<sup>‡</sup></b>	77.6 <sup>‡</sup>	<b>61.5<sup>‡</sup></b>	<b>78.0<sup>‡</sup></b>	<b>57.2<sup>‡</sup></b>	<b>50.8<sup>‡</sup></b>	<b>44.7<sup>‡</sup></b>	<b>76.1<sup>‡</sup></b>

(a) Kialo

	Normative Arguments					Non-normative Arguments				
	ACC	AUC	F1	F1 <sub>sup</sub>	F1 <sub>att</sub>	ACC	AUC	F1	F1 <sub>sup</sub>	F1 <sub>att</sub>
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)	<b>63.9*</b>	<b>68.3*</b>	<b>63.9*</b>	63.8	64.0 <sup>‡</sup>	73.0	76.1	73.0	74.2	71.7
5 \ Fact	63.4*	67.1	63.4*	<b>64.0</b>	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	<b>65.5*</b>	<b>74.5</b>	<b>78.7</b>	<b>74.5</b>	<b>75.4</b>	<b>73.6</b>
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:**  $\text{Support}(S, I) \wedge \text{Support}(I, C) \rightarrow \text{Support}(S, C)$ ,

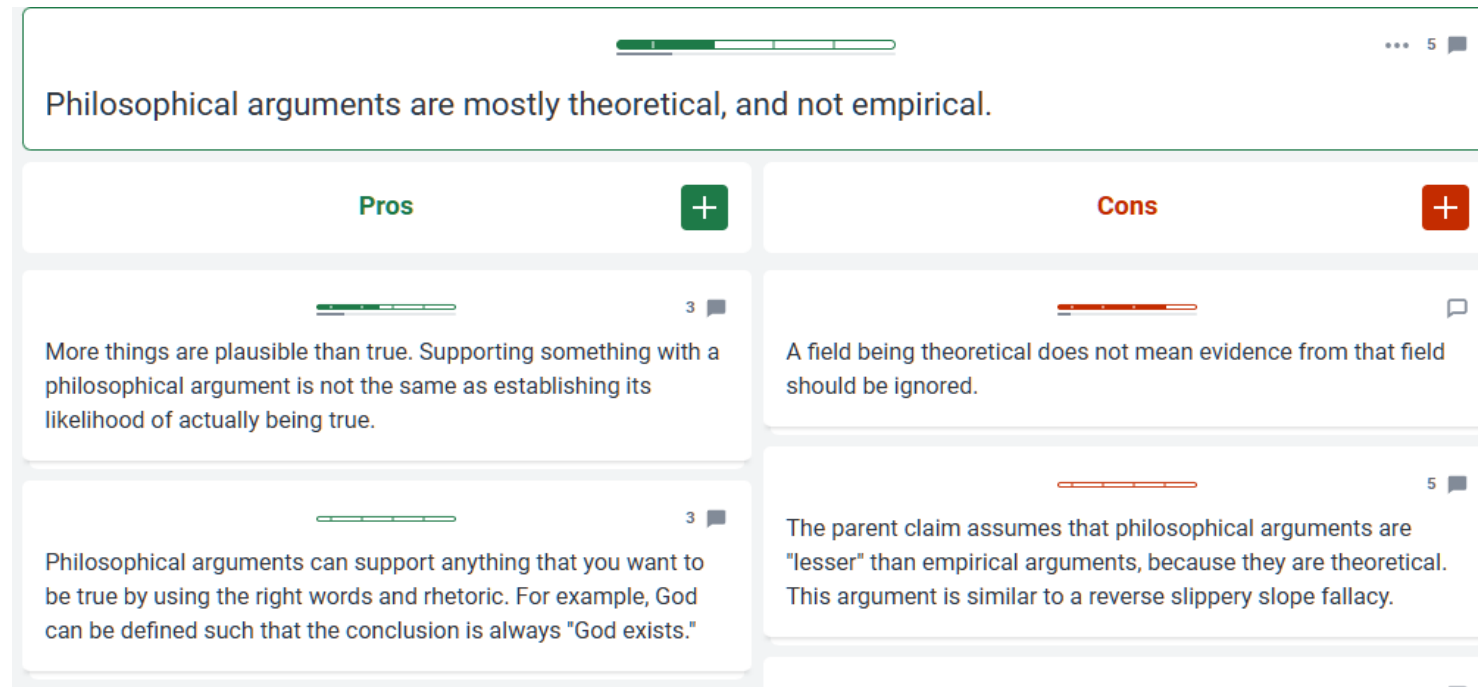
**R15:**  $\text{Attack}(S, I) \wedge \text{Attack}(I, C) \rightarrow \text{Support}(S, C)$ ,

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

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

# 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/](https://netsys.surrey.ac.uk/datasets/graphnli/orig_data/)



The screenshot displays a Kialo debate interface. At the top, a central claim is presented: "Philosophical arguments are mostly theoretical, and not empirical." Below this claim, the interface is divided into two columns: "Pros" (green header) and "Cons" (red header). Each column contains a list of arguments. The "Pros" column has two visible arguments, each with a green progress bar and a comment icon. The "Cons" column has two visible arguments, each with a red progress bar and a comment icon. The interface also includes a progress bar and a comment icon at the top right of the central claim area.

Philosophical arguments are mostly theoretical, and not empirical.

**Pros**

More things are plausible than true. Supporting something with a philosophical argument is not the same as establishing its likelihood of actually being true.

Philosophical arguments can support anything that you want to be true by using the right words and rhetoric. For example, God can be defined such that the conclusion is always "God exists."

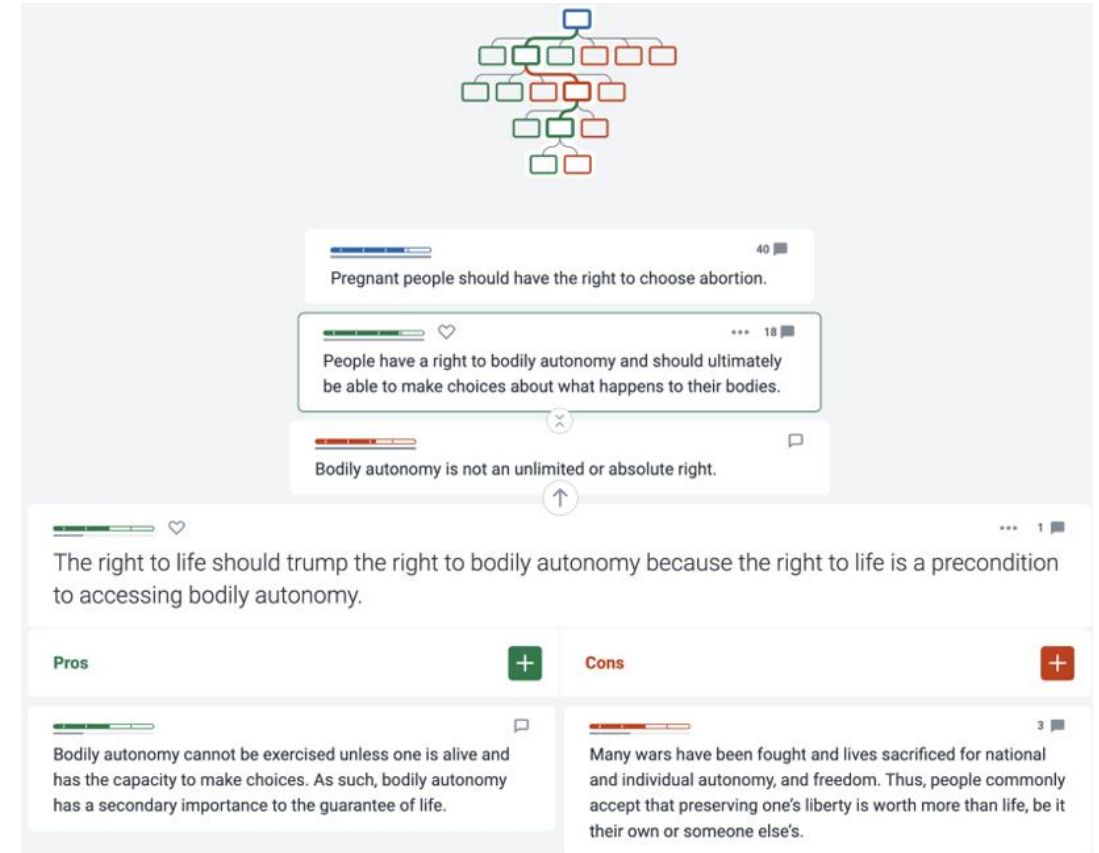
**Cons**

A field being theoretical does not mean evidence from that field should be ignored.

The parent claim assumes that philosophical arguments are "lesser" than empirical arguments, because they are theoretical. This argument is similar to a reverse slippery slope fallacy.

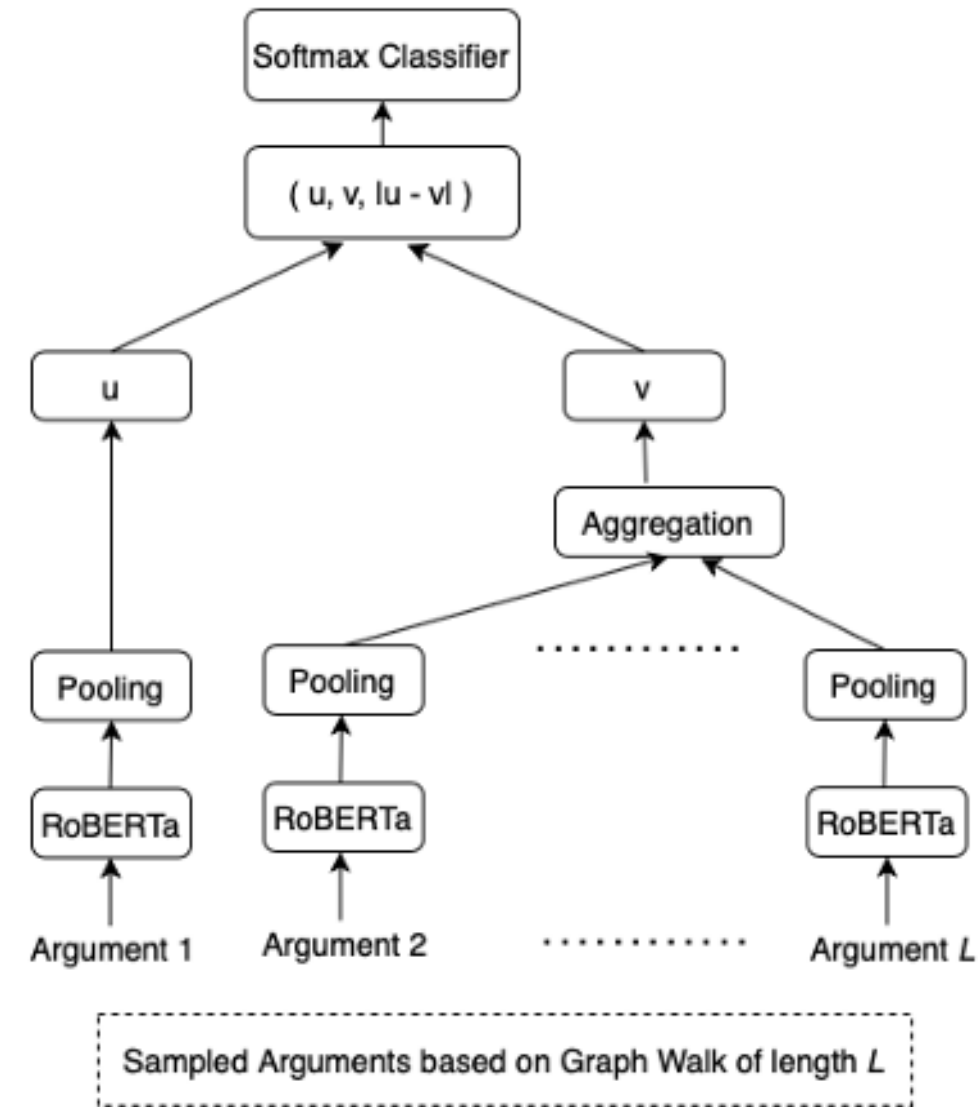
# Data Analysis

- 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

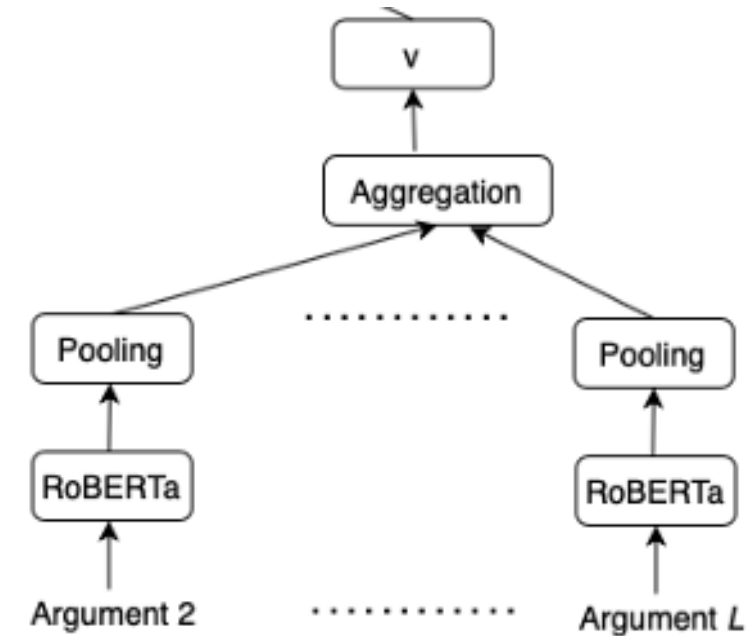
- **Sentence Transformer model:**  
Encodes the raw text of each argument node into a sequence of token embeddings.
- **Pooling:**  
Applies mean pooling to the token embeddings to produce a single fixed-size vector
- **Aggregation:**  
Merges vectors from parent nodes using an aggregation function. (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.



The fact that dishonest arguments exist does not preclude the duty to deal seriously with honest arguments.

...

4

Pros

+

Cons

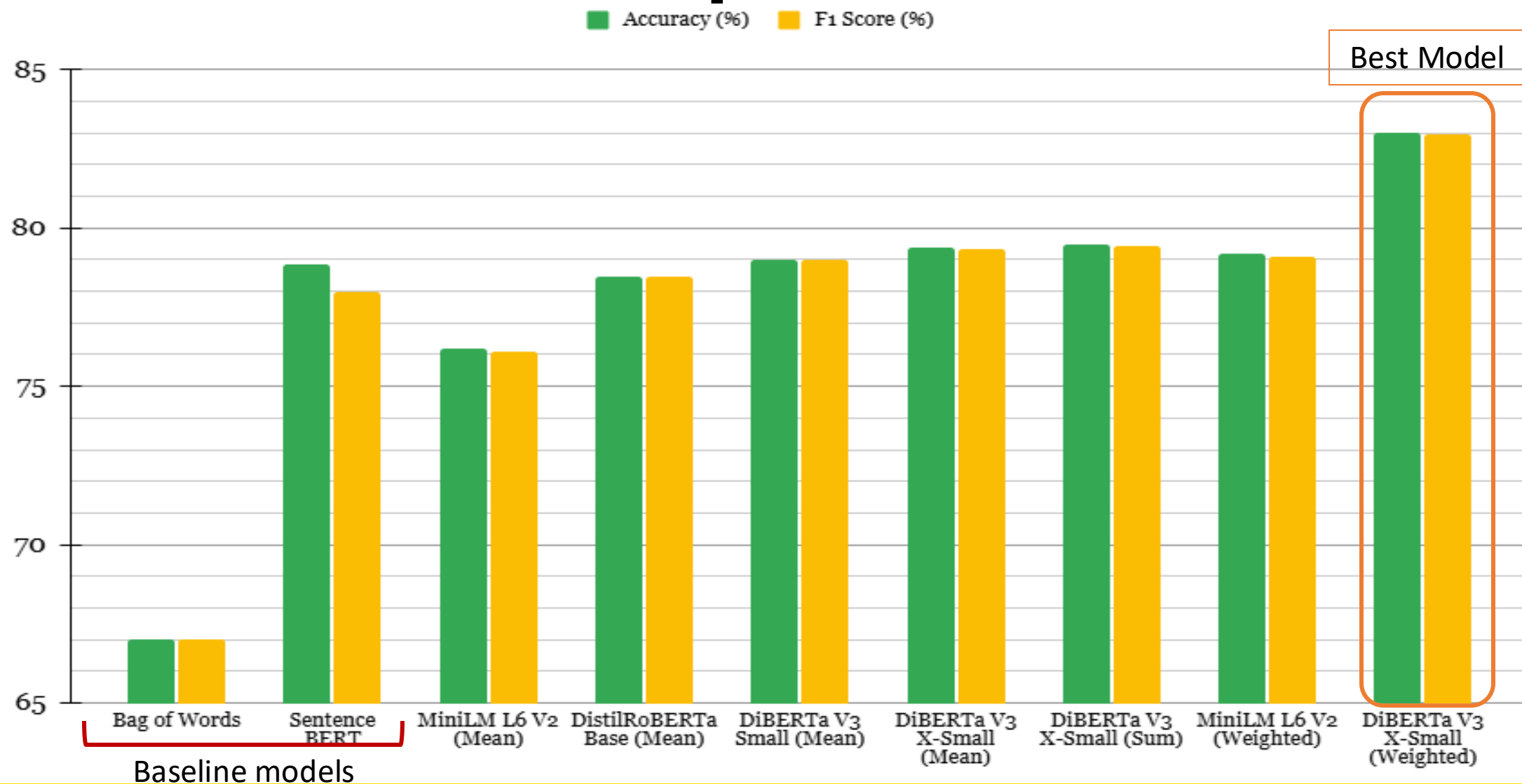
+

# 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

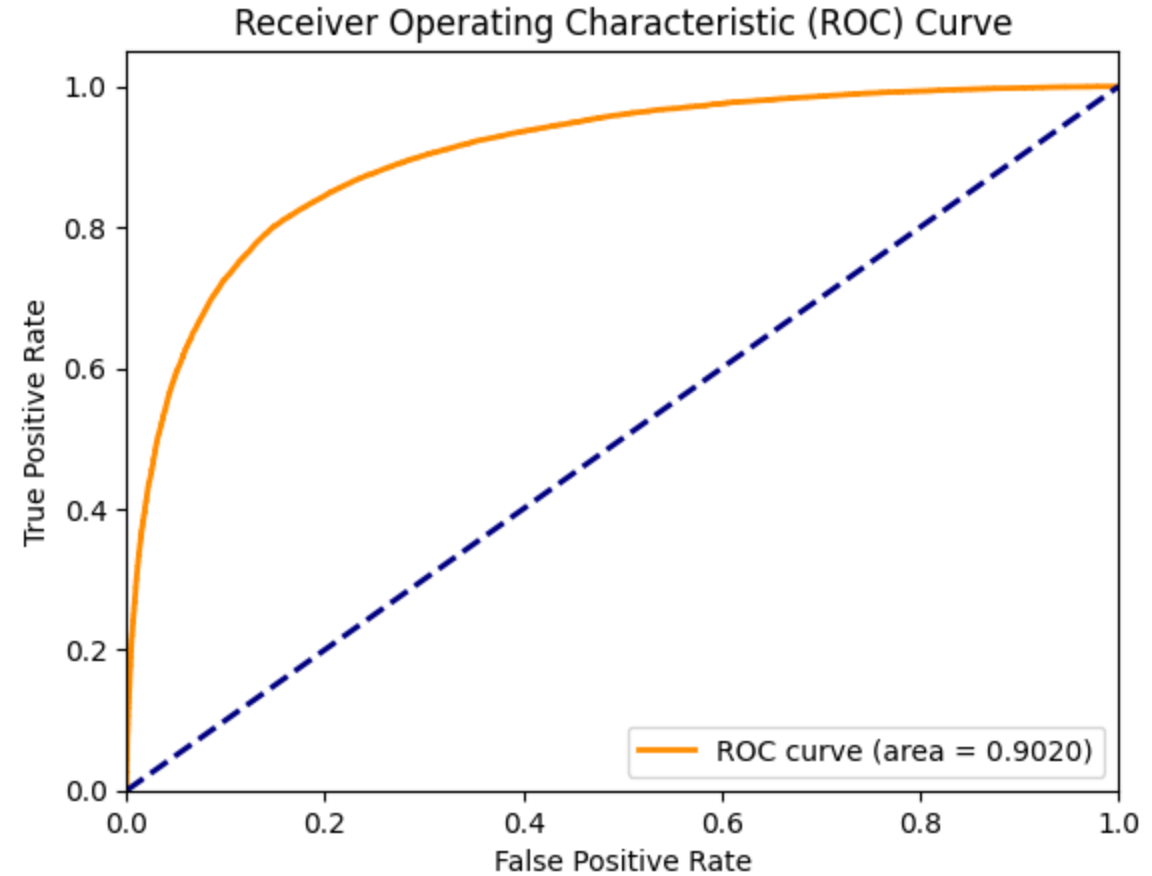
# Best Results and Discussion

- **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

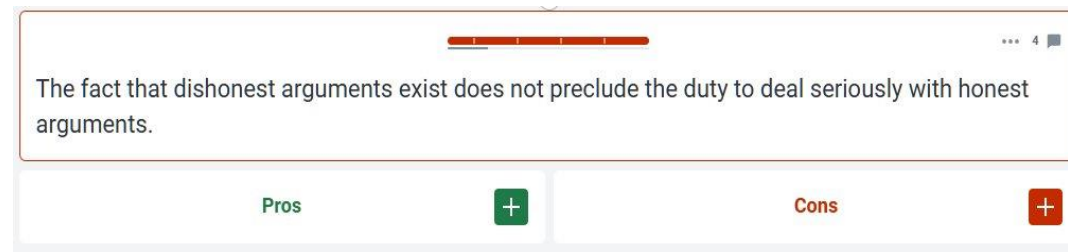


# Final Discussion

- 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

