ANLP Project Interim Submission

Natural Language Inference(NLI) is a task in natural language processing with the goal of classifying the relationship between two separate pieces of text: the premise and the hypothesis. It is generally classified into 3 categories:

1. Entailment: The hypothesis follows from the premise
2. Contradiction: The hypothesis contradicts the premise
3. Neutral: There is no clear connection between the two

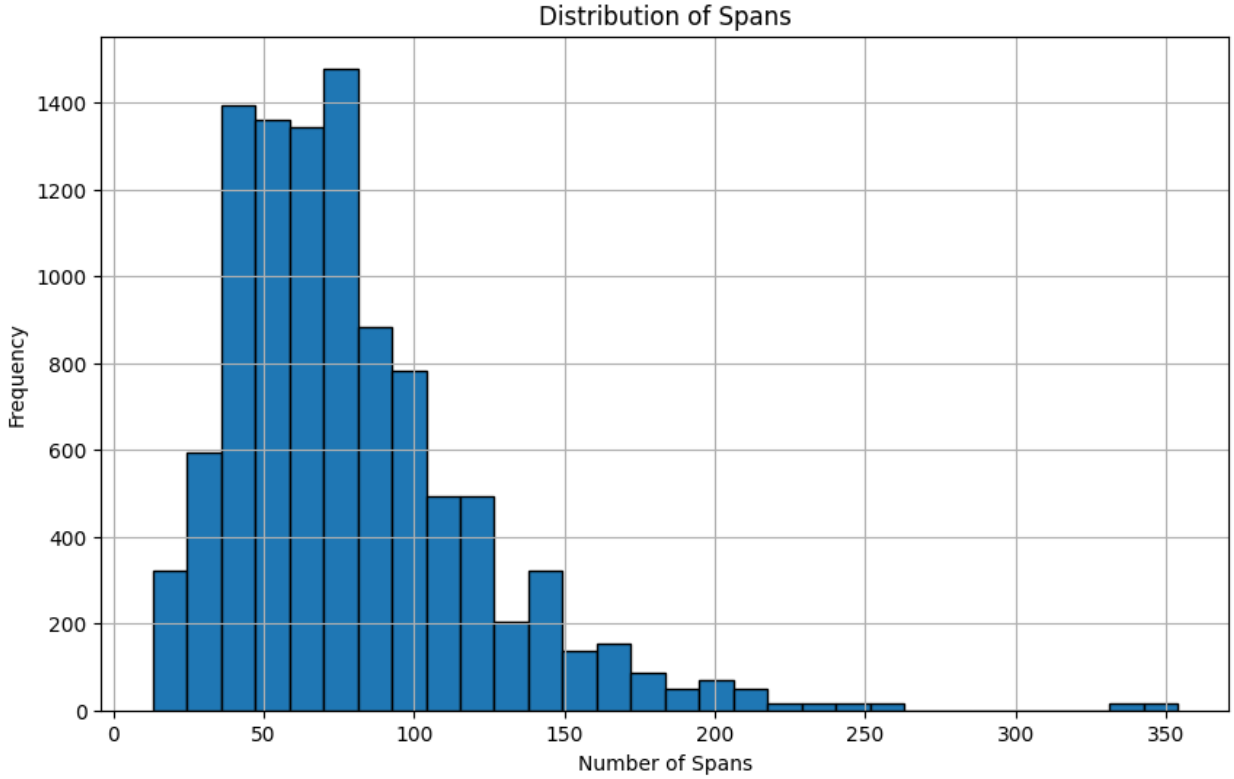
Previous research into NLI for legal documentations led to results which help in extracting information from the document but do not help in drawing conclusions or making inferences from that information.

The problem in applying NLI for contracts is first, negation by exception. For example, in “Recipient shall not disclose Confidential Information to any person or entity, except its employees or partners”, the first half of the sentence forbids information sharing but then it states the complete opposite under certain conditions. This is hard for language models to grasp, especially when such negations may occur far away from the initial premise.

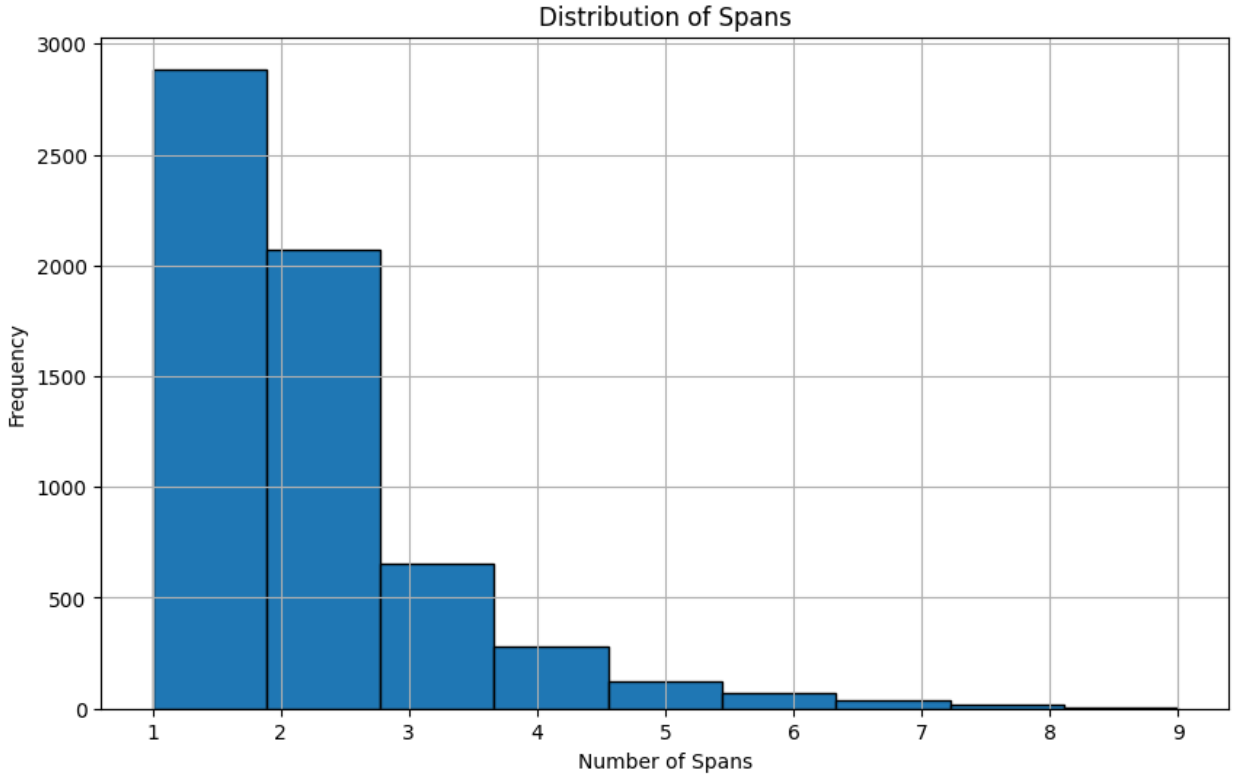
Second is the issue of discontinuous spans of evidence. This means that evidence which supports or contradicts a premise can be discontinuous and these discontinuities can be pages apart. Thus, it is harder for the model to gather necessary and sufficient evidence especially when dealing with shorter context windows.

Exploratory analysis (using the given dataset)

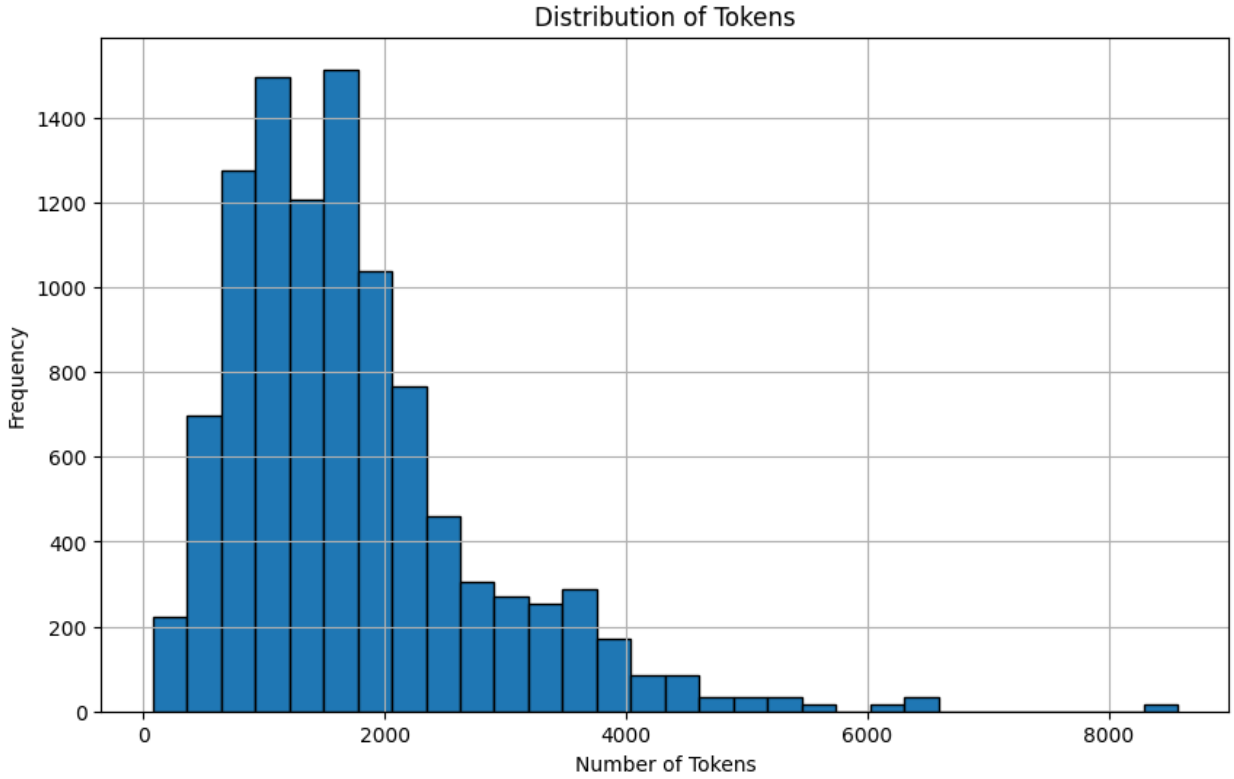
- The document corpus has an average of 79.2 spans.



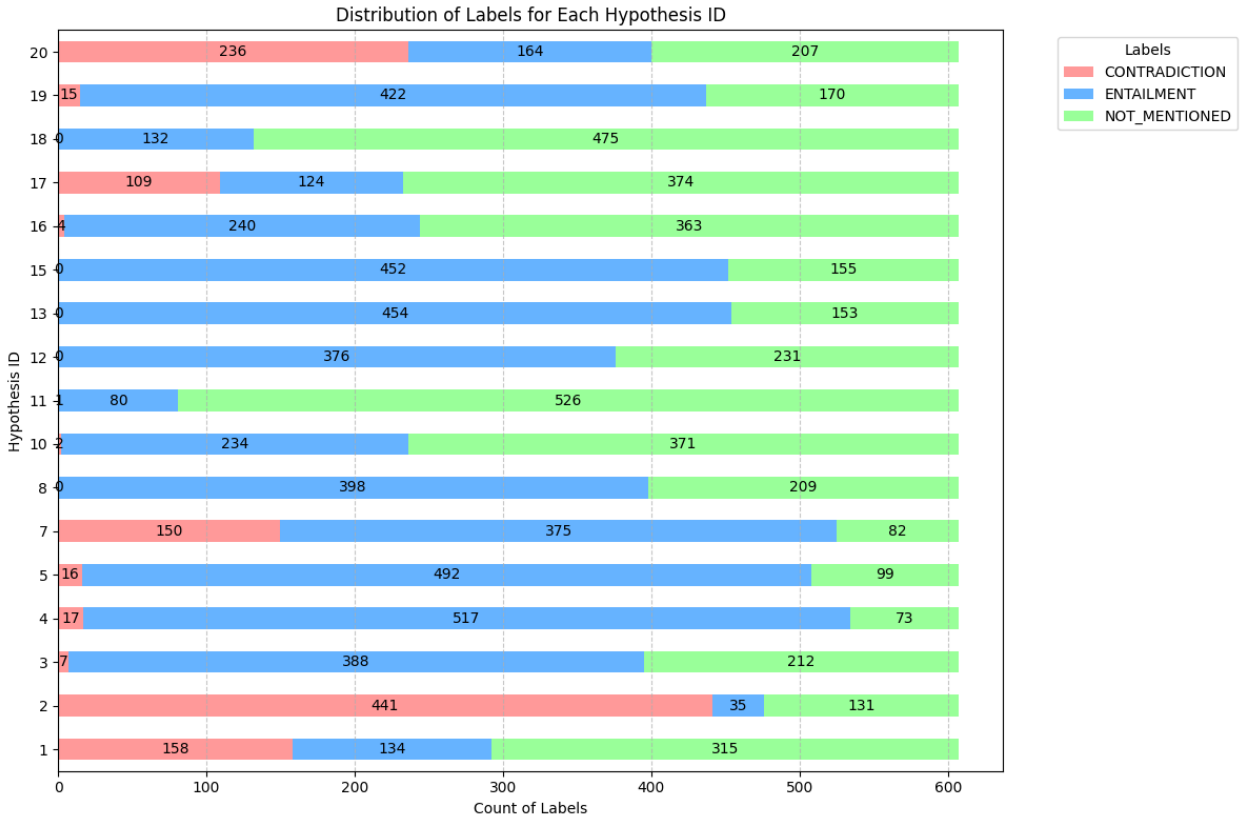
- The most of entailed/contradicting hypotheses have one or two evidence spans, but some have up to nine spans.



- There is an average of 1739 tokens per document. This poses a challenge, as the majority of documents exceed the 512-token limit for models like BERT.



- The following plot presents the distribution of NLI labels across hypotheses. ENTAILMENT and NOT_MENTIONED are dominant, but around half of the hypotheses contain both ENTAILMENT and CONTRADICTION.



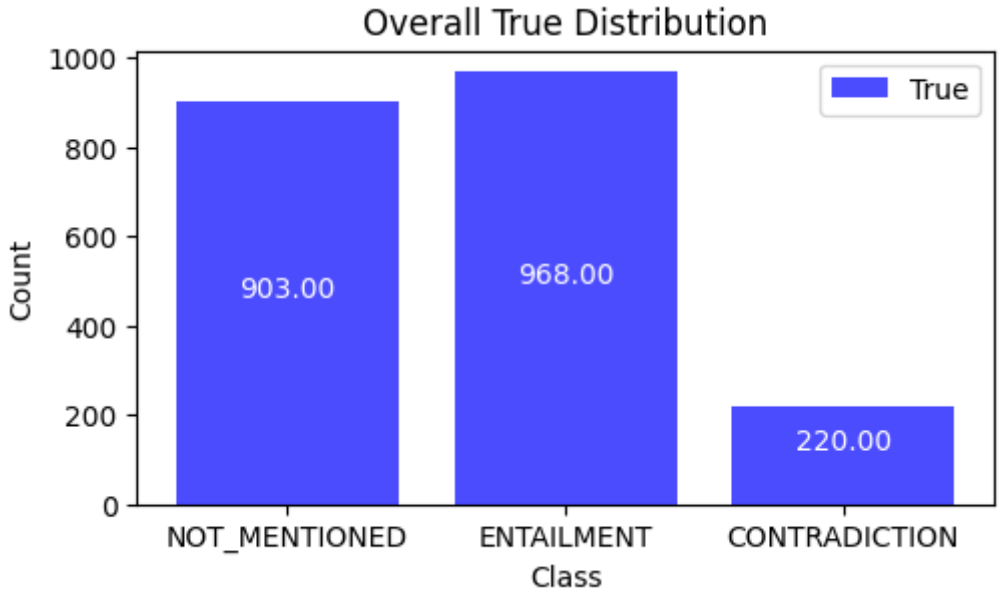
- **Basic statistics overall**

Metric	Per Document			Tokens per Instance		
	Average	Min	Max	Average	Min	Max
Paragraphs	45	9	248	50.2	1.3	416.3

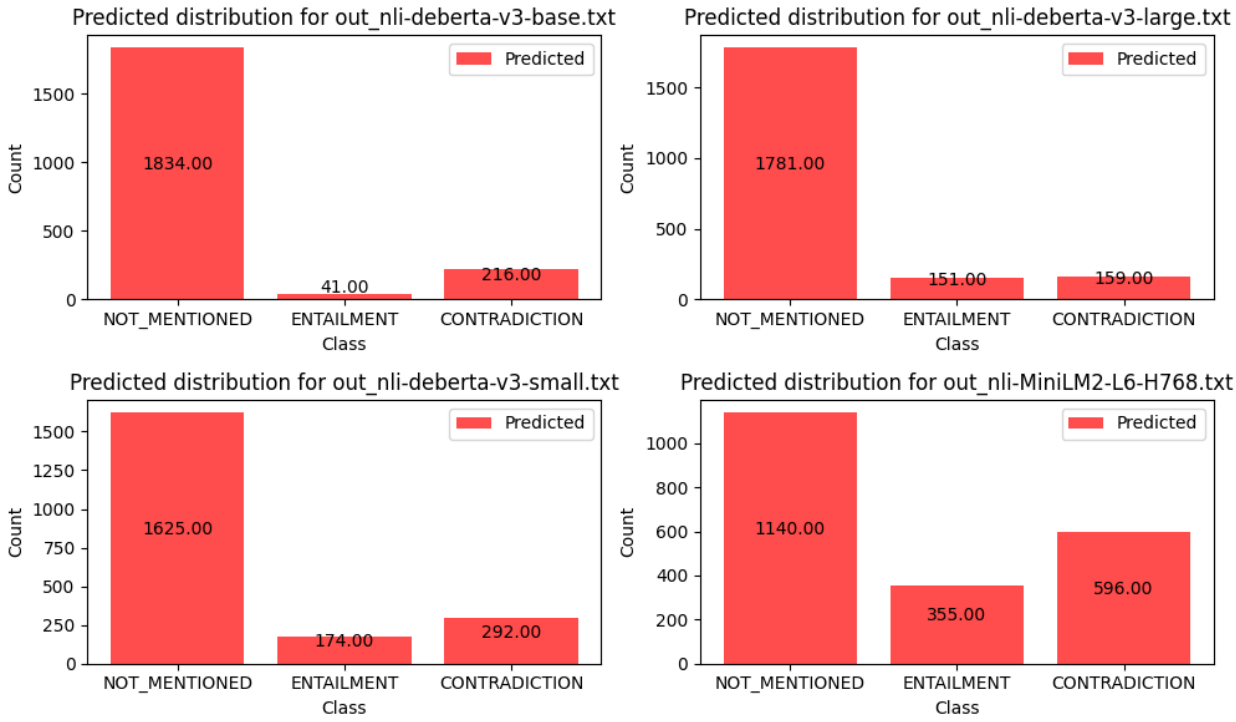
```
| Spans          | 79      | 13   | 354 | 27.9      | 0.9   | 285.7 |
| Tokens         | 1739    | 74   | 8575|           |       |       |
```

Performance against baselines

- **True distribution of the training data.**



- **Prediction of each NLI model.**



Metrics for 4 baseline models

```
apart from the main paper
```

1. nli-deberta-v3-large

Class	Precision	Recall	F1-Score	Support
NOT_MENTIONED	0.46	0.91	0.62	903
ENTAILMENT	0.81	0.13	0.22	968
CONTRADICTION	0.15	0.11	0.13	220
Accuracy			0.46	2091
Macro Avg	0.47	0.38	0.32	2091
Weighted Avg	0.59	0.46	0.38	2091

---

2. nli-deberta-v3-base

Class	Precision	Recall	F1-Score	Support
NOT_MENTIONED	0.46	0.93	0.61	903
ENTAILMENT	0.90	0.04	0.07	968
CONTRADICTION	0.19	0.19	0.19	220
Accuracy			0.44	2091
Macro Avg	0.52	0.38	0.29	2091
Weighted Avg	0.63	0.44	0.32	2091

---

3. nli-deberta-v3-small

Class	Precision	Recall	F1-Score	Support
NOT_MENTIONED	0.46	0.82	0.59	903
ENTAILMENT	0.69	0.12	0.21	968
CONTRADICTION	0.14	0.19	0.16	220
Accuracy			0.43	2091
Macro Avg	0.43	0.38	0.32	2091
Weighted Avg	0.53	0.43	0.37	2091

---

4. nli-MiniLM2-L6-H768

Class	Precision	Recall	F1-Score	Support
NOT_MENTIONED	0.47	0.59	0.52	903
ENTAILMENT	0.59	0.22	0.32	968
CONTRADICTION	0.15	0.40	0.22	220

Class	Precision	Recall	F1-Score	Support
Accuracy			0.40	2091
Macro Avg	0.40	0.40	0.35	2091
Weighted Avg	0.49	0.40	0.39	2091

(from the "Contract-NLI-project-ANLP\outputs\output_metrics.txt")

```
out_nli-deberta-v3-base.txt
               precision    recall  f1-score   support

NOT_MENTIONED       0.46      0.93      0.61       903
   ENTAILMENT       0.90      0.04      0.07       968
CONTRADICTION       0.19      0.19      0.19       220

     accuracy                           0.44      2091
    macro avg       0.52      0.38      0.29      2091
 weighted avg       0.63      0.44      0.32      2091


out_nli-deberta-v3-large.txt
               precision    recall  f1-score   support

NOT_MENTIONED       0.46      0.91      0.62       903
   ENTAILMENT       0.81      0.13      0.22       968
CONTRADICTION       0.15      0.11      0.13       220

     accuracy                           0.46      2091
    macro avg       0.47      0.38      0.32      2091
 weighted avg       0.59      0.46      0.38      2091


out_nli-deberta-v3-small.txt
               precision    recall  f1-score   support

NOT_MENTIONED       0.46      0.82      0.59       903
   ENTAILMENT       0.69      0.12      0.21       968
CONTRADICTION       0.14      0.19      0.16       220

     accuracy                           0.43      2091
    macro avg       0.43      0.38      0.32      2091
 weighted avg       0.53      0.43      0.37      2091


out_nli-MiniLM2-L6-H768.txt
               precision    recall  f1-score   support

NOT_MENTIONED       0.47      0.59      0.52       903
   ENTAILMENT       0.59      0.22      0.32       968
```

```
CONTRADICTION       0.15      0.40      0.22       220

    accuracy                            0.40      2091
   macro avg        0.40      0.40      0.35      2091
weighted avg        0.49      0.40      0.39      2091
```

Evaluation

- We evaluate the performance of various pretrained Natural Language Inference (NLI) models utilizing the same test dataset outlined in the original paper. This evaluation enables a consistent comparison of model effectiveness and provides insights into their relative strengths and weaknesses (as shown above)
- The selected metrics include Precision, Recall, and F1-score for the three classes, as well as Macro Average and Weighted Average for these metrics, along with Accuracy.
- For the ease of calculating new metrics, we are saving the predicted and true values for each sample.

The original model was trained with 100 different parameter sets (on multiple A100s) to achieve their scores, however due to computational constraints, we only trained a small subset (4) of those, due to which our attained accuracy was only around 50%.