ANLP Project Report

Ajit Srikanth*
IIIT-Hyderabad
ajit.srikanth@research.iiit.ac.in

Jhalak Banzal*
IIIT-Hyderabad
jhalak.banzal@students.iiit.ac.in

Pranav Agrawal*
IIIT-Hyderabad
pranav.agrawal@students.iiit.ac.in

Sidhi Panda
IIIT-Hyderabad
sidhi.panda@research.iiit.ac.in

Manish Srivastava
IIIT-Hyderabad
m.shrivastava@iiit.ac.in

* *denotes equal contribution*

1 Introduction

The paper *ContractNLI* introduces a novel method and dataset for document-level natural language inference (NLI) within the domain of legal contracts. The primary goals are twofold:

1. **Natural Language Inference (NLI):** Given a document and a hypothesis, determine if the hypothesis is *entailed*, *contradicted*, or *not mentioned* in the document.
2. **Evidence Span Extraction:** Identify relevant evidence spans from the document that support the inference decision.

2 EDA

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

There is an average of 2254 tokens per document. This poses a challenge, as the majority of documents exceed the 512-token limit for models like BERT.

3 Challenges

3.1 Joint Loss Dependency

To address both objectives simultaneously, the paper proposes a BERT-based model with two classifiers (operating on CLS tokens and SPAN tokens) and a joint training loss, as illustrated in the following equations.

$$\ell_{\text{span}} = \sum_i \left(-s_i \log \hat{s}_i - (1 - s_i) \log(1 - \hat{s}_i)\right) \quad (1)$$

$$\ell_{\text{NLI}} = \begin{cases} -\sum_{L \in \{E,C,N\}} y_L \log \hat{y}_L, & \text{if } \exists s_i = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

$$\ell = \ell_{\text{span}} + \lambda \ell_{\text{NLI}} \quad (3)$$

Training models with a joint loss is sensitive to the choice of the λ parameter, which requires extensive experimentation to optimize.

3.2 Context Size Limitations

Most encoder-only transformer models struggle with context size limitations, making it challenging to process entire documents. While the paper addresses this by merging adjacent spans and aggregating predictions, this approach has limitations.

3.3 Discontinuous Evidence Spans

A significant 28% of document-hypothesis pairs require reasoning over discontinuous spans (spans that may be pages apart). Additionally, 81% of documents contain at least one hypothesis needing such discontinuous spans.

4 Proposed Approach

To mitigate the issues above, we propose splitting the task into a two-step process:

1. Evidence Inference (**EI**): Identify all relevant evidence spans from the document.
2. Natural Language Inference (**NLI**): Use the extracted evidence spans to classify the hypothesis

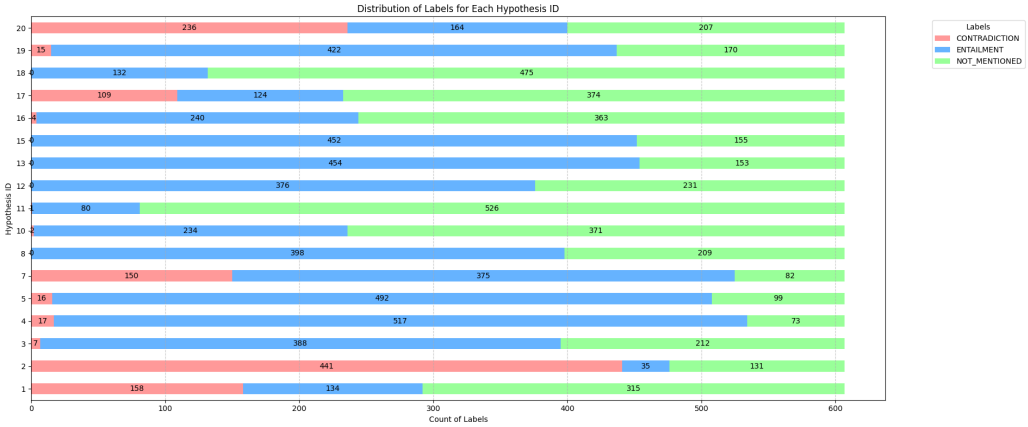

Figure 1: Distribution of labels vs hypotheses

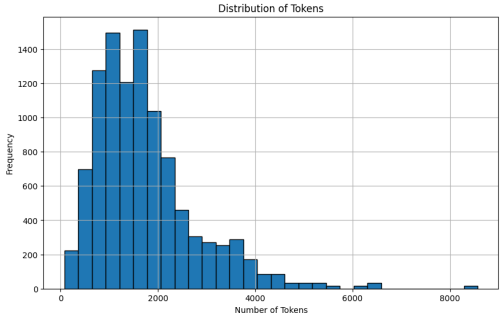

Figure 2: Token Distribution

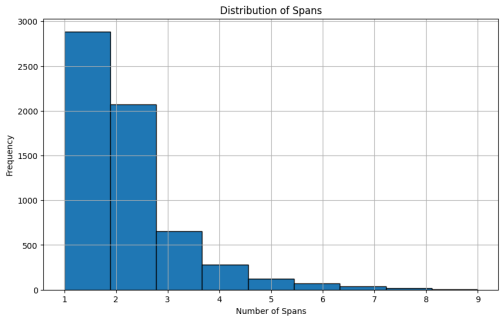

Figure 3: Evidence Span Distribution

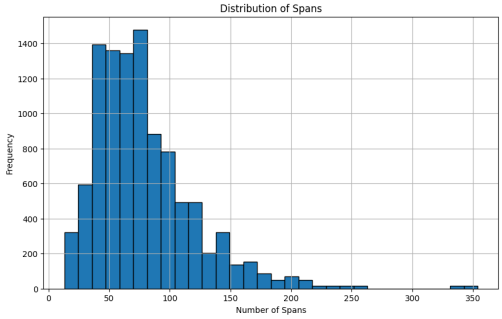

Figure 4: Span Distribution

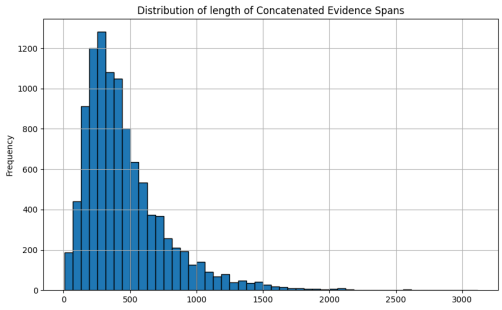

Figure 5: Concatenated Evidence Span Lengths

into one of the three labels (*Entailment*, *Contradiction*, *Not Mentioned*).

As shown in Table 1, we tested various models with different training objectives.

5 Evidence Inference

5.1 Model Architecture

We tested multiple models by passing the input [Span] $< SEP >$ [Hypothesis] through different architectures mentioned earlier. Sentence-transformer models such as all-MiniLM-L6-v2 were benchmarked due to their sentence-level embedding capabilities.

5.2 Results

Using all-MiniLM-L6-v2, we classified spans as evidence based on cosine similarity with the hypothesis.

We want a high degree of recall for evidence identification as it is important that any piece of relevant information is not left out. We can increase recall at the expense of precision since ultimately the evidence has to go for manual review by a human regardless, for whom it is easier to judge if a particular span is relevant or not.

While accuracy and recall were high (due to correct classification of non-evidence spans), precision was poor. Many non-evidence spans were incorrectly classified as evidence.

A precision below 0.1 highlights limitations in evidence selection. However, false positives are acceptable

Model	Training Objectives
BERT	Masked Language Model (MLM), Next Sentence Prediction (NSP)
DistilBERT	Knowledge Distillation, MLM Only
RoBERTa	MLM with dynamic masking
ALBERT-base-v2	MLM, Sentence Order Prediction (SOP)
DeBERTa	MLM with disentangled attention, Replaced Token Detection (RTD)
BigBird	Block sparse attention instead of normal attention, 4096 seq len
ST:MiniLM-L6	Contrastive Learning Objective, distilled from RoBERTa
ST:MPNet	MLM with permutation-based pre-training for enhanced bidirectional context

Table 1: Model Training Objectives, where ST denotes Sentence Transformers' model with additional fine-tuning using contrastive learning objective primarily for sentence similarity.

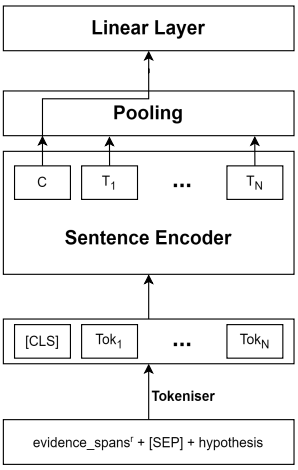

Figure 6: Our proposed architecture
r: We pass span for EI, and concatenated spans for NLI

Model	Accuracy	Recall
*ST:MPnet (CS)	0.921	0.426
AlBERT	0.942	0.907
RoBERTa	0.954	0.919
BERT	0.958	0.917
DeBERTa	0.966	0.929
ST:miniLM	0.974	0.926
ST:MPnet-base	**0.988**	**0.932**

Table 2: EI only model total accuracy and Evidence Recall;
*denotes Evidence precision< 0.4

to an extent since the NLI model that will follow can classify these spans as *Not Mentioned*. The priority is high recall to ensure no evidence is missed.

6 Natural Language Inference

6.1 Ensemble Model Design

Instead of a single model, we implemented an ensemble approach:

1. Entailment Classifier: Determines whether a hypothesis is entailed by the evidence spans.

	Evidence	Not Evidence
Evidence	1612	1325
Not Evidence	117	124126
Precision	0.549	0.993
Recall	0.932	0.989
Acc. 98.87%		**mAP** 0.771

Table 3: EI Results
Columns are actual values, rows are predicted values

2. Contradiction Classifier: Determines whether a hypothesis is contradicted by the evidence spans.

This ensemble approach mitigates issues with imbalanced datasets and overfitting in single-model architectures.

If both classifiers return false, the hypothesis is classified as Not Mentioned. If both return true (collision), a secondary threshold or MLP model resolves the conflict based on pre-argmax values.

6.2 Results

Concatenation of evidence spans is feasible because 94.12% of concatenated evidence spans fit within the token limit.

Model	Acc(E)	Acc(C)	Pr(E)	Pr(C)
3-label	0.859	0.876	0.88	0.69
distilBERT	0.815	0.842	0.86	0.70
AlBERT	0.852	0.867	0.88	0.72
BERT	0.899	0.886	0.90	0.75
ST:miniLM	0.881	0.871	0.91	0.77
RoBERTa	0.898	0.902	0.90	0.77
BigBird	0.928	0.941	0.92	0.78
ST:MPnet	0.946	0.953	0.94	0.80
DeBERTa	**0.951**	**0.968**	**0.95**	**0.82**

Table 4: NLI model with noised spans

	Independent			Combined		
	E	C	NM	E	C	NM
E	909	12	35	524	26	61
C	28	197	16	16	79	25
NM	26	11	858	12	23	129
P	0.95	0.82	0.96	0.95	0.62	0.60
R	0.94	0.90	0.94	0.86	0.66	0.79
F1	0.95	0.85	0.95	0.90	0.64	0.68
Acc.	93.88%			81.79%		

Table 5: NLI Results
Columns are actual values, rows are predicted values

Model	**Accuracy**	**mAP**
3 label classifier	73.36%	0.815
Paper	78.28%	0.853
Ours	81.79%	0.912

Table 6: Comparison with main paper

7 Related Works

"**DocInfer**: Document-level Natural Language Inference using Optimal Evidence Selection" (Mathur et al., EMNLP 2022) presents a novel approach to Document-level Natural Language Inference (DocNLI). DocInfer builds a hierarchical document graph enriched through inter-sentence relations (topical, entity-based, concept-based), performs paragraph pruning using the novel SubGraph Pooling layer, followed by optimal evidence selection based on REINFORCE algorithm to identify the most important context sentences for a given hypothesis.

Legal NLI has been explored in case law, statute law, and contracts. For case and statute law, tasks like COLIEE-2020 (Rabelo et al., 2020) involve determining entailment between court decisions and legal texts, while Holzenberger et al. (2020) addressed entailment between statements and statutes. These domains typically feature consistent language and one-to-one hypothesis-document relationships, unlike the many-to-many relationships in ContractNLI.

ContractNLI distinguishes itself from prior work on contract clause annotation (e.g., Lippi et al., 2019) and information extraction (Leivaditi et al., 2020; Hendrycks et al., 2021) by requiring reasoning over diverse hypotheses. It also poses unique challenges compared to claim verification tasks (Thorne et al., 2018; Jiang et al., 2020), given the linguistic complexity of contracts. These characteristics emphasize the distinct and advanced nature of ContractNLI.

8 Conclusion

Our two-step approach overcomes some limitations of the ContractNLI joint model:

1. We are able to achieve better scores for NLI over discontinuous spans, i.e. from possibly different contexts.
2. By decoupling evidence inference and NLI tasks, we eliminate the need for λ parameter tuning.
3. Our ensemble-based NLI method provides improved performance compared to single classifiers.