Recursive Journal of Informatics
Vol. 3, No. 1, March 2025
p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

Textual Entailment for Non-Disclosure Agreement Contract Using ALBERT Method

Abdillah Azmi[1], Alamsyah[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. A Non-Disclosure Agreement (NDA) is a legal contract that binds two or more parties to confidentiality regarding shared or created sensitive information. NDAs play a crucial role in protecting intellectual property, maintaining trade secrets, and regulating the dissemination of confidential data. However, reading and analyzing contract letters is a repetitive, time-consuming, and labor-intensive process. Given the increasing volume of legal documents in the business world, automation through Artificial Intelligence (AI) has become a promising solution.
Purpose: Many researchers have explored AI applications in legal document processing, particularly in contract analysis. This study extends previous research by developing a pretrained language model specialized for contract letter comprehension using the Natural Language Inference (NLI) task. The research aims to evaluate the performance of ALBERT-base in analyzing contract documents, particularly NDAs, and compare it with other base models.
Methods/Study design/approach: The study employs a deep learning approach, training an ALBERT-base model for textual entailment using the CNLI (Contract NLI) dataset. The model is optimized using AdamW and LambdaLR for early stopping. Multiple training iterations are conducted with varying hyperparameter configurations to enhance performance.
Result/Findings: The experimental results indicate that the fine-tuned ALBERT-base model achieves an accuracy of 85% and an Exact Match (EM) score of 85.04%. Although this performance does not surpass the current State of the Art (SOTA) on the CNLI benchmark, the model outperforms other base models such as SpanNLI (BERT-base), SCROLLS (BART-base), and Unlimiformer (BART-base).
Novelty/Originality/Value: ALBERT is designed for memory efficiency with a compact parameter size while maintaining strong performance. Its ability to process long-context information with minimal hardware requirements makes it a promising solution for legal NLP applications. This research contributes to advancing AI-driven contract analysis, reducing manual effort while improving accuracy and efficiency in legal document processing.

Keywords: NLI, ALBERT, Textual Entailment, NLP, Contract, Finetuning, NDA, Deep Learning, Language Model
Received July 17, 2024 / **Revised** March 25, 2025 / **Accepted** March 27, 2025

INTRODUCTION

NLP (Natural Language Processing) is a theory that studies the relationship between computers and human language through artificial intelligence techniques [1]. NLP allows computers to perform tasks that involve language, such as translation, question-answering, or text classification [2]. Classification is a technique that can be used to determine classes of the data and can be applied to text or medical problems [3]. NLP also allows computers to perform analysis to determine the sentiment of a sentence, paragraph, or document, which is called sentiment analysis [4]. Besides its contribution to the computer science field, NLP also adds up to the knowledge of human language that benefits other fields such as linguistics, psychology, and philosophy [5]. As computing technology develops, approaches to NLP have tended towards Deep Learning, or architectures with many layers. Prior to Deep Learning, most NLP tasks were done with generative or discriminative approaches such as HMM (Hidden Markov Model), CRF (Conditional Random Fields), and SVM (Support Vector Machine). Currently, the most popular approach around NLP is to use RNN (Recurrent Neural Network) architecture, CNN (Convolutional Neural Network), or Transformer. Transformers are a dominant architecture within NLP research, exceeding the alternatives like CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) in terms of

[1]*Corresponding author.
Email addresses: azmi1752@students.unnes.ac.id (Azmi)
DOI: 10.15294/rji.v3i1.9730

performance whether in language understanding or language generation, and have boosted accuracy in almost all NLP tasks [6].

Pretrained models are NLP models that have been trained on unlabeled large corpus and can be tuned to do specific tasks. One example of a pretrained model is BERT [7]. BERT (Bidirectional Encoder Representation from Transformer) is a pretrained model that is designed for understanding context using bidirectional learning process on each layer [7]. BERT built on top of Transformer architecture, which is based solely on attention mechanisms called multihead attention [8]. BERT utilizes two methods to do bidirectional training, which are MLM (Masked Language Model) and NSP (Next Sentence Prediction). Recently many versions of models based on BERT architecture emerged. Some of them are DistilBERT [9], RoBERTa [10], DeBERTa [11], dan ALBERT [12].

ALBERT (A Lite BERT) is a model built using BERT architecture. The focus of this model is to reduce BERTs layer complexity and size while maintaining the performance with two methods, which are Cross-layer Parameter Sharing and SOP (Sentence Order Prediction) loss function. SOP loss function is suitable for understanding long text sequences especially in tasks like translation or language inference [12]. NLI (Natural Language Inference) refers to the machine ability to understand language deeply without explicitly being told earlier. There are several benchmarks that test machines on specific NLI tasks such as, question-answering, and textual entailment. Benchmarks are meant to test models' ability to solve specific AI problems in specific domains [13]. One example of benchmark is CNLI in legal domain. Contract NLI (CNLI) is a NLI benchmark that focuses on legal areas, specifically understanding NDA contract letters.

Contracts are legal documents that contain specific agreements between two or more parties. This type of legal document is normally used by law firms, companies, governments agencies etc [14]. A contract review is a process that includes verifying and clarifying the facts and provisions included in the contract, assessing the contract's feasibility, and predicting potential risks [15]. Read and understand contract documents requires a lot of time and effort. 60 to 80% business to business transactions are done with some kind of written contract, with generally 1000 companies handling 20000 to 40000 contracts every time [16]. This process was a perfect example of tasks that can be handled by Artificial Intelligence, especially NLP that can understand contract documents automatically.

There are already several research that focuses on developing models for NLI in legal areas that have been tested with CNLI benchmarks. SCROLLS [17] is a model that is based on BART and LED (Longformer Encoder-Decoder) architecture and designed for solving tasks that need long text understanding. The result was 77,4 EM score on BART version and 73,4 EM score on LED version. Another research was CoLT5 (Conditional Long T5) model [18] that takes advantage of modified T5 architecture to process long text. This model results in an 88.7 EM score on CNLI benchmarks. SLED (Sliding Encoder and Decoder) [19] was a model built with BART, LED, and T5 architecture, designed to process long text sequences. SLED managed to give 87,3 as the highest EM score. Unlimiformer [20] takes advantage of kNN (k-nearest-neighbor) to decode input on each attention layer. This model was built based on BART-base architecture with around 140M parameters, and the result on CNLI benchmark was 77.7 on EM score. UL2 (Unified Language Learning) [21] was a model that was designed to take advantage of MoD (Mixture of Denoisers) to allow it to perform multiple tasks by switching paradigm dynamically to match the desired task. This model has total parameters up to 20B and currently the state of the art of CNLI benchmark with 88.7 on the EM metrics.

Based on the previous short introduction, a few researchers focused on the accuracy and EM (Exact Match) of the model in NLI tasks. There have been limited studies concerned with the model that has fewer parameters and smaller in size, like ALBERT. Therefore, this research intends to find the result of small models like ALBERT on specific domain task like NLI in legal area. The objectives of this research are to fine-tuned ALBERT model on CNLI benchmark and evaluate the accuracy and EM scores.

METHODS

ContractNLI Dataset

CNLI (Contract Natural Language Inference) [22] is a benchmark that is designed to test models on NLI tasks, specifically textual entailment in legal domain. CNLI is a document-based dataset that consists of 607 contract files that each can have more than 2 pages. The dataset is also accompanied by 17 hypotheses that have been reviewed by legal experts. CNLIs' contract documents are a type of NDA (Non-Disclosure

Agreement), in which the purpose is to bind two or more parties to agree that some information is confidential. Each file document consists of up to 2254 tokens. The dataset initially comes in a JSON or JSONL format. To be able to be used for neural networks training, dataset needs to be turned into tabular format. The result of table structure can be seen in Table 1. The total of combination contract and hypothesis text produce 9788 pair of contract-hypothesis. All the documents are contracts that have been archived from 1996 until 2020. Most of them come from EDGAR system owned by SEC (Securities and Exchange Commission) that have been extracted with specific regular expressions query.

Table 1. CNLI Dataset Content Structure

Column	Data Type	Description
Premise	String	Contract text sequence
Hypothesis	String	Hypothesis text sequence
Label	String	Entail, Contradict, Not Mentioned
Subset	String	Train, Test, Dev

From the total of 9788 pairs of data, the label distribution was not balanced. Entailment class exceeded the number of Contradiction class with 46% from all the data. The smallest label was Contradiction class with only 11% of all the data. This shows that the labels were extremely unbalanced and needed to be evaluated with other metrics than only F1 micro score. The data have been cleaned and preprocessed in some way to only preserve the main text of the contracts. The clean data was collected from huggingface.co. After being collected, the data was splitted into 3 subsets of train, test, dev with percentage of 70%, 20%, and 10%. The splitting ratio was already determined by the dataset owner so that everyone has the same scenario to test a model with.

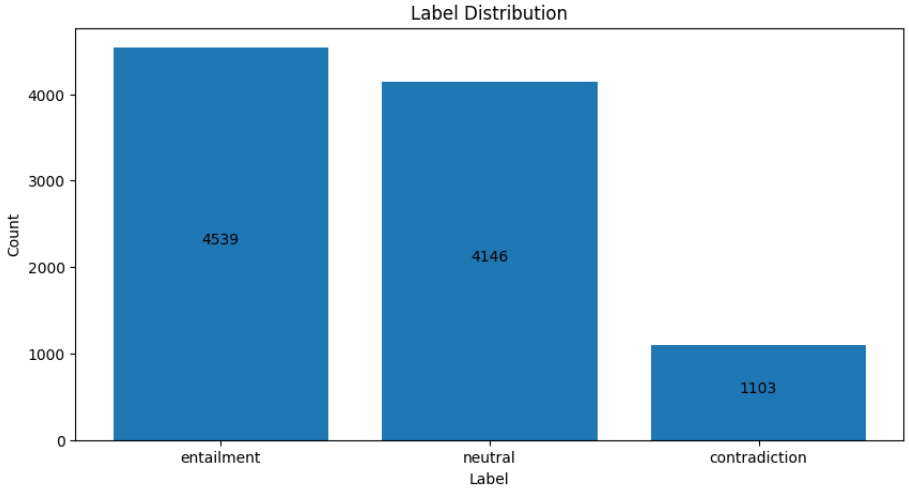

Figure 1. Label Distribution

Tokenizing for feature extraction with embedding was done using AlbertTokenizer from Transformer module by huggingface.co. The tokenizers that were used by ALBERT were based on sentencepiece library and using subword unit method like byte-pair-encoding or unigram language model to split text and create tokens. Because of the subword method, the number of tokens can be larger than the number of input text. Table 2 shows the details of the tokenized text. The contract text has the mean length of 126 tokens where the minimum length was 6 tokens, and the maximum length was 521 tokens. The standard deviation (Std) shows a high number on contract text. This means that the tokenized contract has high variance. The opposite happens with hypothesis text which has low variance.

Table 2. Detail of tokenized CNLI dataset text

	Contract	Hypothesis
Count	9788	9788
Mean	126,725276	15,587148
Std	98,125668	5,262601
Min	6	8
25%	59	12
50%	96	15
75%	162	17
Max	521	28

ALBERT for ContractNLI

ALBERT (A Lite BERT) was a model architecture that was built based on the BERT architecture with focus on size and memory efficiency [4]. It has significantly fewer parameters than BERT. As seen on Table 3, ALBERT-base has only 12 million parameters compared to 108 million. Both are using the same 12 attention head layers with the size of hidden state is 768. The focus of ALBERT was to reduce the BERT complexity while maintaining the performance. This will allow the model to run on smaller hardware. To accomplish this, ALBERT takes advantage of two methods, which is cross-layer parameter sharing factorized embedding parameterization. Factorized embedding parameterization has a purpose to reduce the layer complexity by detaching the embedding layer from the hidden layer. This way, when the vocabulary increases, the hidden layer is not increased. Meanwhile, cross-layer parameter sharing allows multiple layers to use the same weight. Whether its feed forward layers, attention, or all layers can be grouped and share the same parameters.

Table 3. ALBERT architecture compared to BERT

Model		Params	Attention	Hidden State	Embedding	Param-Sharing
BERT	base	108M	12	768	768	No
	large	334M	24	1024	1024	No
ALBERT	base	12M	12	768	128	Yes
	large	18M	24	1024	128	Yes
	xlarge	60M	24	2048	128	Yes
	xxlarge	235M	12	4096	128	Yes

Experiment Setup

To conduct the experiment, we first fine-tuned the model to perform textual entailment by modeling the layers to do classification. Textual entailment is not far different from regular multi label classification task. The model also needs to be pre-trained with the CNLI dataset. The pre-training process will be conducted several times with multiple hyperparameters to obtain the optimal result. The hyperparameters can be seen on Table 4 along with the values that been selected for the experiments. After that, the results of the experiments will be evaluated with accuracy and F1 score as metrics. The F1 macro will be used on all labels to assist the F1 micro score in evaluating the unbalanced dataset. The training process was performed using 2x T4 GPU from Kaggle. Figure 2 shows the flowchart of the experiment process.

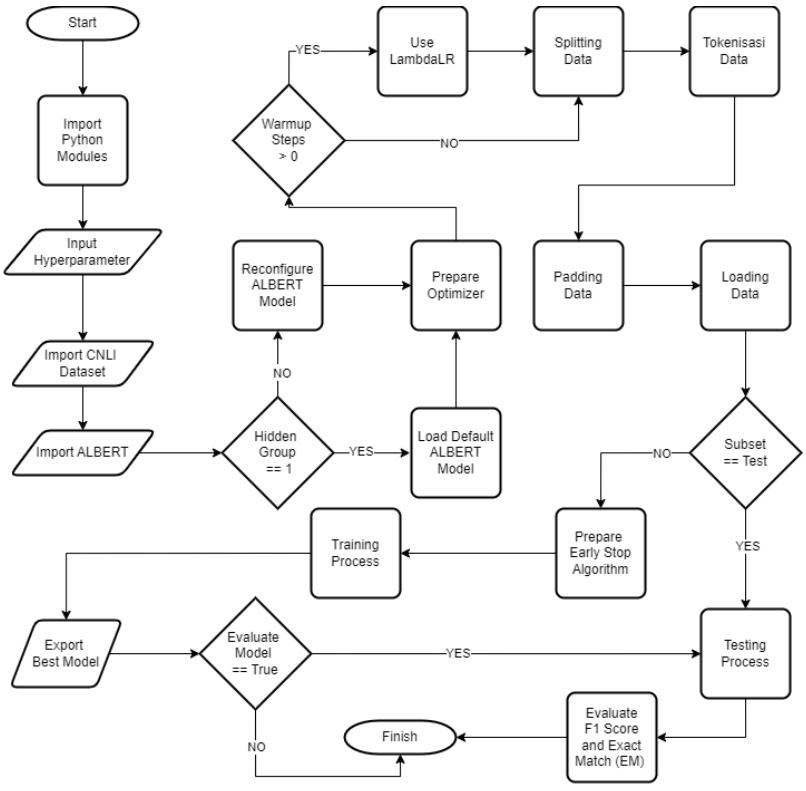

Figure 2. Flowchart of the experiment process

Deep learning library that been used to program the experiment was Pytorch, Sklearn, Numpy, and Pandas. The dataset was converted from JSON into the pandas' data frame. The ALBERT model weight and architecture were imported from huggingface.co. The model uses two kinds of configurations. The default configuration is used when the layer group was one, meanwhile the customized configuration was used when there are more layer groups. After that, the training process of the model uses several additional processes like early stopping and warmup steps to increase the training efficiency and reduce training time.

Table 4 Hyperparameters that been used on the experiments.

Hyperparameter	Value(s)
Batch Size	8, 16, 20
Learning Rate	2e-5, 2-e6
Weight Decay	0, 0.01
Warmup Steps	0, 100
Epoch	10
Layer Groups	0, 3
Early Stop	True

Early stop is a method that was used to stop the training method with certain conditions. The purpose of this method was to prevent the model to overfit. To perform early stop, we use accuracy on validation (dev) data during training process. The best accuracy was gradually saved and compared to the next accuracy score. If the current score was worse than the previously saved score, then it means the model going through local minima and the training was stopped. Warmup step was a method that allowed the learning rate to dynamically adjusted for training process. We performed warmup step with LambdaLR from PyTroch library. LambdaLR allows the learning rate to increase gradually using customized conditions by passing a lambda function. Inside the lambda function, we increased the learning rate linearly until the step reaches the upper limit that previously been set, which was 100. If the step already reaches the 100 marks, the learning rate will linearly decrease to prevent overfitting.

After the training process, because the NLI task was modelled from textual entailment into multi label classification, we can evaluate the result with classification metrices such as accuracy and F1 score. In addition, the metric EM (Exact Match) was also used so that the result can be compared to results from another research that used CNLI dataset to perform textual entailment task. EM metrics measures the percentage of predictions that match exactly with the ground truth or reference answer, in this case was the labels entailment, contradiction, and neutral.

$$F1_{micro} = 2 \times \frac{Precission \times Recall}{Precission + Recall} \quad (1)$$

$$F1_{macro} = \frac{F1_1 + F1_2 + \cdots + F1_n}{n} \quad (2)$$

$$Precission = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

The F1 score assigns equal weight to all classes, including the majority class. This was not optimal for an unbalanced dataset such as CNLI. For that matter, F1 macro score was used. F1 macro was calculated by counting F1 score for each class separately and then calculating the mean from the scores. The purpose of F1macro was to give weight that has been adjusted relative to each percentage in the dataset. This method will evaluate the unbalanced data more accurately.

RESULT AND DISCUSSION

For each hyperparameter combination, the experiment was performed twice. This meant making sure that the result was consistent. All the training processes were implemented early stop method to shorten the training time and avoid overfitting. Table 5 shows the results of training experiments with only one hidden group of layers. The experiment with BS 20 and WS 0 generates the largest validation accuracy and F1 macro score compared to the other hyperparameters combination. Meanwhile the largest testing accuracy was achieved by experiment with BS 8 and WS 0. The table also shows the increasing pattern that follow

the decreasing number on BS. This shows that the model can learn better and has better potential accuracy with BS 8.

Table 5. Experiment results with one hidden layer group

WS	BS	HG	Acc (%)		F1 micro			F1 macro	ES
			Val	Test	Entail	Contra	Neutral		
0	20	1	**84,7**	84,9	86,2	**75,7**	85,5	**82,5**	T
0	16	1	84,1	83,9	84,8	75,3	84,7	81,6	T
0	8	1	84,6	**85,0**	**86,5**	74,8	**85,8**	82,4	T
100	20	1	82,2	80,8	82,0	65,7	82,4	76,7	T
100	16	1	78,6	79,1	80,6	66,3	80,0	75,6	T
100	8	1	82,1	82,3	84,0	70,3	83,3	79,2	T

WS (Warmup Step), BS (Batch Size), HG (Hidden Groups), ES (Early Stop), Acc (Accuracy). Table 6 shows the results of the experiments with three-layer groups. Three-layer groups mean that attention or feed forward layers were divided into three groups with index 0, 1, and 2, where each group will share the same parameters. The results shows that the smaller BS give the smaller accuracy. Meanwhile the BS 20 resulting in relatively large accuracy. With three-layer groups, the best result was achieved by combination of WS 0 and BS 20. The validation accuracy reaches up to 83.5 and test accuracy reaches up to 82.6. These results are still lower than the best accuracy that were achieved by the experiments with one layer group.

Table 6. Experiment results with three hidden layer groups

WS	BS	HG	Acc (%)		F1 micro			F1 macro	ES
			Val	Test	Entail	Contra	Neutral		
0	20	3	**83,5**	82,6	84,9	72,7	0	20	3
0	16	3	83,1	**83,3**	**85,9**	**74,0**	0	16	3
0	8	3	81,8	81,0	84,0	70,8	0	8	3
100	20	3	79,1	78,9	81,7	68,2	100	20	3
100	16	3	78,3	77,9	81,3	66,9	100	16	3
100	8	3	42,4	39,0	53,2	7,8	100	8	3

WS (Warmup Step), BS (Batch Size), HG (Hidden Groups), ES (Early Stop), Acc (Accuracy). From several experiments that have been conducted, we decided to take the experiment with 8 BS, 1 HG, 2e-5 learning rate, and without warmup steps as the best model. The model was labeled as ALBERT-CNLI-base. Table 7 shows the number of correct predictions by the model on the test subset. The results show that the model could predict 93.9% of the entailment class correctly, meanwhile it could only predict 73.4% the contradiction label correctly. These numbers show that the model could predict the previously unseen data with good performance.

Table 7. The percentage of correctly predicted labels on test data

	Correct Prediction	Percentage	True Label
Entail	829	93.9 %	882
Contradict	155	73.4 %	211
Neutral	721	79.0 %	912

Span NLI-BERT is a base model that was provided by ContractNLI benchmark as a starting point for other researchers to improve. This model is designed to perform two tasks, which are textual entailment and evidence identification. The model also comes in two types, which are large and base version. On CNLI benchmark, the Span NLI-BERT-large got 87.5 accuracy and for Span NLI-BERT-base got 83.8. Table 8 shows that our model, ALBERT-CNLI-base managed to outperform the base version of Span NLI-BERT with accuracy up to 85.0, and then followed by a bigger F1 score on contradiction label up to 74.8 and 86.5 on entailment label F1 score. This shows that ALBERT-CNLI-base with only around 12M parameters could give competitive performance on NLI task specifically on legal area.

Table 8. The main result compared to CNLI base model

Model	Acc	F1(C)	F1(E)
Span NLI-BERT-base	83.8	28.7	76.5
ALBERT-CNLI-base	85.0	74.8	86.5

Lastly, Table 9 shows the ALBERT-CNLI-base performance compared to other models that also have been tested on CNLI benchmarks. The metrics that was being used was EM (Exact Match), which is to measure the similarities between two or more strings. EM metrics demand that every result need to be the exact "right" or "wrong", so there is no room for "almost right". The results below show that ALBERT-CNLI-

base managed to outperform other models that built with BART-base architecture, but still have difficulties to compete with models that built with larger architectures such are BART-large, T5, and UL20.

Table 9. The main result compared to other models on CNLI benchmark

Author	Model	EM (%)
Tay et al., 2022	UL20	88.7
Ainslie et al., 2023	CoLT5-large	88.7
Ivgi et al., 2023	SLED (BART-large)	87.3
Bertsch et al., 2023	Unlimiformer (BART-base)	77.7
Shaham et al., 2022	SCROLLS (BART-base)	77.4
	ALBERT-CNLI-base	85.04

CONCLUSION

In this research we evaluated the results of document-level natural language inference experiment using CNLI dataset. The purpose of the experiment is to understand the ability of an AI model to assist in contract review. The CNLI dataset consists of more than 10000 pairs of contracts and hypotheses that have linguistic characteristics of an NDA in a legal area. The task was designed to be textual entailment and to be trained as multilabel classification. After conducting pretraining and finetuning with multiple hyperparameter setup, the number of accuracies shows that the resulting model (ALBERT-CNLI-base) managed to exceed CNLI base model (SpanNLI BERT-base) by 1,2 percent. In addition, we compare the model with another research that uses the base version of BART architecture. Using EM as metric, ALBERT-CNLI-base also manages to exceed another model like SCROLLS BART-base and Unlimiformer BART-base. For future works, we hope researchers will experiment with another type of contract letter like lease contract and employment agreements. We believe that studying NLI with contract letters will serve as a starting point for future general AI models in the legal domain.

REFERENCES

[1] Khan, W., Daud, A., Khan, K., Muhammad, S., & Haq, R. (2023). Exploring the frontiers of deep learning and natural language processing: A comprehensive overview of key challenges and emerging trends. Natural Language Processing Journal, 4, 100026. https://doi.org/10.1016/j.nlp.2023.100026

[2] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[3] Alamsyah, A., & Fadila, T. (2021, July). Increased accuracy of prediction hepatitis disease using the application of principal component analysis on a support vector machine. In Journal of Physics: Conference Series (Vol. 1968, No. 1, p. 012016). IOP Publishing. https://doi.org/10.1088/1742-6596/1968/1/012016

[4] Larasati, U. I., Muslim, M. A., Arifudin, R., & Alamsyah, A. (2019). Improve the accuracy of support vector machine using chi square statistic and term frequency inverse document frequency on movie review sentiment analysis. Scientific Journal of Informatics, 6(1), 138-149. https://doi.org/10.15294/sji.v6i1.14244

[5] Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 82(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

[6] Just, J. (2024). Natural language processing for innovation search – Reviewing an emerging non-human innovation intermediary. Technovation, 129. https://doi.org/10.1016/j.technovation.2023.102883

[7] Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS). http://arxiv.org/abs/1706.03762

[9] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. https://doi.org/10.48550/arXiv.1910.01108
[10] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. International Conference on Learning Representations. http://arxiv.org/abs/1907.11692
[11] He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. International Conference on Learning Representations. http://arxiv.org/abs/2006.03654
[12] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. International Conference on Learning Representations. https://doi.org/https://doi.org/10.48550/arXiv.1909.11942
[13] Storks, S., Gao, Q., & Chai, J. Y. (2020). Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. https://doi.org/10.48550/arXiv.1904.01172
[14] Chalkidis, I., Androutsopoulos, I., & Michos, A. (2017). Extracting contract elements. Proceedings of the International Conference on Artificial Intelligence and Law, 19–28. https://doi.org/10.1145/3086512.3086515
[15] Leivaditi, S., Rossi, J., & Kanoulas, E. (2020). A Benchmark for Lease Contract Review. https://doi.org/10.48550/arXiv.2010.10386
[16] Exigent Group Limited. (2019). Thought Leadership Contract Management Why is contract management important? https://cdn2.hubspot.net/hubfs/4220630/How%20GCs%20Can%20Thrive%20Not%20Just%20Survive%202019.pdf
[17] Shaham, U., Segal, E., Ivgi, M., Efrat, A., Yoran, O., Haviv, A., Gupta, A., Xiong, W., Geva, M., Berant, J., & Levy, O. (2022). SCROLLS: Standardized CompaRison Over Long Language Sequences. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 12007–12021. http://arxiv.org/abs/2201.03533
[18] Ainslie, J., Lei, T., de Jong, M., Ontañón, S., Brahma, S., Zemlyanskiy, Y., Uthus, D., Guo, M., Lee-Thorp, J., Tay, Y., Sung, Y.-H., & Sanghai, S. (2023). CoLT5: Faster Long-Range Transformers with Conditional Computation. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5085–5100. https://doi.org/10.48550/arXiv.2303.09752
[19] Ivgi, M., Shaham, U., & Berant, J. (2023). Efficient Long-Text Understanding with Short-Text Models. Transactions of the Association for Computational Linguistics, 11, 284–299. https://doi.org/10.1162/tacl_a_00547
[20] Bertsch, A., Alon, U., Neubig, G., & Gormley, M. R. (2023). Unlimiformer: Long-Range Transformers with Unlimited Length Input. NeurIPS 2023. https://doi.org/10.48550/arXiv.2305.01625
[21] Tay, Y., Dehghani, M., Tran, V. Q., Garcia, X., Wei, J., Wang, X., Chung, H. W., Shakeri, S., Bahri, D., Schuster, T., Zheng, H. S., Zhou, D., Houlsby, N., & Metzler, D. (2022). UL2: Unifying Language Learning Paradigms. ICLR 2023 Conference. http://arxiv.org/abs/2205.05131
[22] Koreeda, Y., & Manning, C. (2021). ContractNLI: A Dataset for Documentlevel Natural Language Inference for Contracts. Dalam M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Ed.), Findings of the Association for Computational Linguistics: EMNLP 2021 (hlm. 1907–1919). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.164