An Efficient Active Learning Pipeline for Legal Text Classification

Sepideh Mamooler and **Rémi Lebret** and **Stephane Massonnet** and **Karl Aberer**
School of Computer and Communication Sciences, EPFL, Switzerland

Abstract

Active Learning (AL) is a powerful tool for learning with less labeled data, in particular, for specialized domains, like legal documents, where unlabeled data is abundant, but the annotation requires domain expertise and is thus expensive. Recent works have shown the effectiveness of AL strategies for pre-trained language models. However, most AL strategies require a set of labeled samples to start with, which is expensive to acquire. In addition, pre-trained language models have been shown unstable during fine-tuning with small datasets, and their embeddings are not semantically meaningful. In this work, we propose a pipeline for effectively using active learning with pre-trained language models in the legal domain. To this end, we leverage the available *unlabeled* data in three phases. First, we continue pre-training the model to adapt it to the downstream task. Second, we use knowledge distillation to guide the model's embeddings to a semantically meaningful space. Finally, we propose a simple, yet effective, strategy to find the initial set of labeled samples with fewer actions compared to existing methods. Our experiments on Contract-NLI, adapted to the classification task, and LEDGAR benchmarks show that our approach outperforms standard AL strategies, and is more efficient. Furthermore, our pipeline reaches comparable results to the fully-supervised approach with a small performance gap, and dramatically reduced annotation cost. Code and the adapted data will be made available.

1 Introduction

With the advent of pre-trained transformer-based language models (Devlin et al., 2019; Liu et al., 2019; He et al., 2021), training models from scratch has been outperformed by fine-tuning pre-trained language models for several tasks in natural language processing, including text classification (Howard and Ruder, 2018). However, fine-tuning these models still needs large labeled datasets to perform well on the downstream task (Dodge et al., 2020; Zhang et al., 2021; Mosbach et al., 2021). Collecting a large annotated dataset is a highly expensive and time-consuming process in specialized domains, where annotation can only be performed by the domain experts, such as the legal domain (Hendrycks et al., 2021).

Active Learning (AL) has been proved effective for data-efficient fine-tuning of pre-trained language models in non-specialized domains like news, emotions, and movies (Ein-Dor et al., 2020; Margatina et al., 2022). In addition, Margatina et al. (2022) have shown that the unlabeled data can be used to adapt the pre-trained language model to the downstream task, thereby improving the active learning performance with no extra annotation cost. On the specialized domains, Chhatwal et al. (2017) have evaluated multiple AL strategies in the legal domain before the emergence of pre-trained language models. Nevertheless, to the best of our knowledge, the effectiveness of active learning in fine-tuning pre-trained language models in the legal domain has been poorly studied.

In this work, we focus on efficient legal text classification with RoBERTa (Liu et al., 2019) by leveraging existing AL strategies. We identify two challenges in deploying AL strategies in the legal domain; First, legal texts contain a specialized vocabulary that is not common in other domains, including the ones on which pre-trained language models are trained. Second, the annotation of legal texts is highly expensive and time-consuming due to the necessity of specialized training for understanding these texts. For example, Hendrycks et al. (2021) reported a cost of over $2 million for the annotation of the Contract Understanding Atticus Dataset (CUAD) consisting of around 500 contracts.

To account for the specialized vocabulary, inspired by Margatina et al.'s (2022) work, we leverage the available *unlabeled* data to adapt the pre-trained language model to the downstream

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 345 - 358

task. In addition, considering the limitations of pre-trained language models like BERT and RoBERTa in capturing semantics (Reimers and Gurevych, 2019), we use knowledge distillation to further improve the task-adapted model by mapping its embedding space to a semantically meaningful space. Our experiments demonstrate that AL strategies can benefit from semantically meaningful embeddings.

Concerning the cost and time constraints, we focus on the fact that many AL strategies (Lewis and Gale, 1994; Gal and Ghahramani, 2016; Gissin and Shalev-Shwartz, 2019) require an annotated set of N positive and negative samples to start with. In practice, acquiring this set is expensive for large and skewed datasets. We propose a strategy to make the first iteration more efficient by clustering the unlabeled samples and limiting the pull of candidates to the cluster medoids. Our experiments demonstrate we can achieve comparable results with the standard initial sampling approach with up to 63%, and 25% fewer actions on the skewed Contract-NLI (Koreeda and Manning, 2021), and balanced LEDGAR benchmarks (Tuggener et al., 2020) respectively.

Our contributions can be summarized as follows:

1. We design an efficient and effective active learning pipeline for legal text classification by leveraging the available unlabeled data using task-adaptation and knowledge distillation, which obtains comparable performance to fully-supervised fine-tuning with considerably reduced annotation effort.

2. We propose a strategy to reduce the number of actions in the first iteration of active learning by clustering the unlabeled data, and collecting the samples from cluster medoids, further increasing the efficiency of our approach.

3. We evaluate our approach over Contract-NLI and LEDGAR benchmarks. Our results illustrate an increase of 0.3346, and 0.1658 in the best obtained F1-score, compared to standard active learning strategies, for Contract-NLI and LEDGAR respectively.

2 Related Work

Active learning with pre-trained language models Multiple works have studied active learning for pre-trained language models like BERT. Ein-Dor et al. (2020) have evaluated various AL strategies for fine-tuning BERT for text classification, and showed that AL can boost BERT's performance especially for skewed datasets. However, they do not leverage the available unlabeled data to adapt the pre-trained language model to the task at hand, and only focus on non-specialized domains like news and sentiment analysis that do not require experts' knowledge.

Gururangan et al. (2020) have shown that task-adaptive pre-training using the available unlabeled data leads to performance gain when using pre-trained language models like BERT. Following this observation, Margatina et al. (2022) demonstrated the importance of task-adaptation for active learning for non-specialized texts like news, movies and sentiment analysis.

Inspired by these works, we leverage the available unlabeled data to effectively adapt RoBERTa to legal text classification, where the annotation demands experts' knowledge. In addition, we propose an additional step to map the embedding space of the task-adapted RoBERTa to a semantically meaningful space using sentence transformers.

Sentence transformers Reimers and Gurevych (2019) have shown that the embedding space of off-the-shelf pre-trained language models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) is not semantically meaningful, and thus, is not suitable for common sentence comparison measures like cosine similarity. To overcome this limitation, they propose sentence transformers, obtained by adding a pooling layer on top of pre-trained language models, and fine-tuning them in a Siamese network architecture with pairs of similar sentences. In this work, we use a RoBERTa-based sentence transformer as a teacher model and distill its knowledge to the task-adapted RoBERTa to produce sentence embeddings that capture the semantics and can be compared using cosine similarity.

Active learning strategies Numerous methods have been proposed to find proper labeling candidates for active learning. Majority of them belong to one or both of two categories: diversity-sampling, and uncertainty-sampling. Diversity-based methods (Sener and Savarese, 2018; Gissin and Shalev-Shwartz, 2019; Wang et al., 2017) aim to find labeling candidates that best represent the dataset, whereas uncertainty-based methods (Gal and Ghahramani, 2016; Kirsch et al., 2019; Zhang and Plank, 2021) target candidates about which the model is uncertain. BADGE (Ash et al., 2020) is a cluster-based AL strategy that belongs to both of

these categories. It transforms data into gradient embeddings that encode model confidence and sentence feature at the same time. By applying kmeans++ on the gradient embeddings it can find samples that differ both in terms of semantics and predictive uncertainty. ALPS (Yuan et al., 2020) is another cluster-based AL strategy that leverages both uncertainty and diversity using the surprisal embeddings obtained by passing the sentences to the MLM head of the pre-trained language model, and computing the cross entropy loss for a random set of tokens against the target labels.

Existing AL strategies often require a set of labeled samples to start with, which is expensive to acquire. To overcome this high cost, we propose a clustering-based strategy to reduce the effort required to create the initial set of annotated samples.

3 Notation and Setting

In this section, we explain the structure shared between all AL strategies used in this work and fix the notation.

Active learning is an iterative process aiming to obtain a desired performance given an annotation budget. Here, we consider the annotation budget to be the number of actions performed by the annotator. In addition, we assume all annotators are legal experts, and that each annotator assigns perfect labels to text segments. Let U_0 and L_0 be the starting pool of unlabeled and labeled samples respectively. Initially, $L_0 = \emptyset$. At the first iteration, the annotator labels N sample, P positive and $N - P$ negative, to obtained L_1. Then, at each iteration i, the model is fine-tuned using L_i, and the AL strategy recommends a set of samples C_i for annotation. These samples are labeled and U_i and L_i are updated as $U_{i+1} = U_i \setminus C_i$, and $L_{i+1} = L_i \cup C_i$. The procedure is repeated until the annotation budget is exhausted, or the desired performance is achieved.

We base our work on the Low-Resource Text Classification Framework introduced by Ein-Dor et al. (2020). Following this work, we focus on binary text classification, given a small annotation budget and a potentially imbalanced dataset. This scenario matches common use cases in the legal domain, where the goal is to find phrases that correspond to a specific category, with the lowest possible number of actions, given a pool of unlabeled, imbalanced data. We perform 5 AL iterations, and assume a more restricted annotation budget compared to Ein-Dor et al. (2020), allowing only 10 annotations per iteration. For the first AL iteration, we assume that 5 positive and 5 negative samples need to be annotated.

4 Methodology

We propose an efficient active learning pipeline for fine-tuning pre-trained language models for legal text classification. Our approach leverages available unlabeled data in three phases to adapt the pre-trained model to the downstream task (Sec. 4.1), guide its embedding space to a semantically meaningful and comparable space (Sec. 4.2), and reduce the number of actions required to collect the initial labeled set (Sec. 4.3). Finally, it leverages existing AL strategies to efficiently fine-tune a classifier (Sec. 4.4). We now explain each step in detail. An overview of this pipeline can be found in Algorithm 1.

4.1 Task-Adaptation

It has been shown that fine-tuning off-the-shelf pre-trained language models with standard approaches is unstable for small training sets (Zhang et al., 2021; Dodge et al., 2020; Mosbach et al., 2021). As shown by Margatina et al. (2022), this can lead to poor performance when fine-tuning pre-trained language models with AL. In addition, existing pre-trained language models are often trained on texts that do not need specialized training to be understood. However, legal texts contain specialized words that are not common in other domains. Thus, task-adaptation is crucial for the effectiveness of active learning in legal text classification. In the first step of our proposed pipeline, we obtain the task-adapted pre-trained (TAPT) RoBERTa by continuing pre-training the model with unlabeled samples for the Masked Language Modeling (MLM) task, as suggested by Gururangan et al. (2020) and Margatina et al. (2022).

4.2 Knowledge Distillation

Previous works (Reimers and Gurevych, 2019; Li et al., 2020; Su et al., 2021) have shown that, without fine-tuning, the sentence embeddings produced by pre-trained language models poorly capture semantic meaning of sentences, and are not comparable using cosine similarity. To overcome this shortcoming, Reimers and Gurevych (2019) introduced sentence transformers by adding a pooling layer on top of pre-trained transformer-based language models, and training them in a

Algorithm 1 AL pipeline for text classification

Input: unlabeled samples U_0, PT RoBERTa, PT Sentence-RoBERTa, AL strategy α, # iterations T **Output:** text classifier CLS RoBERTa, acquired labeled dataset L_T

$L_0 \leftarrow \emptyset$
Phase 1: Task-adaptation with Masked Language Modeling (MLM)
TAPT RoBERTa $\leftarrow$ MLM(PT RoBERTa, U_0)
Phase 2: Knowledge distillation
DisTAPT RoBERTa $\leftarrow$ Distill(TAPT RoBERTa, PT Sentence-RoBERTa, U_0)
Phase 3: Initial sampling
cluster medoids $\leftarrow$ KMeans(DisTAPT RoBERTa, U_0)
$L_1 \leftarrow$ Sample(cluster medoids)
$U_1 \leftarrow U_0 \setminus L_1$
Phase 4: Active learning
for $i \leftarrow 1$ to T **do**
  CLS RoBERTa $\leftarrow$ Train(DisTAPT RoBERTa, L_i)
  $C_i \leftarrow \alpha$(CLS RoBERTa, U_i)
  $L_{i+1} \leftarrow L_i \cup C_i$
  $U_{i+1} \leftarrow U_i \setminus C_i$
end for

Siamese network architecture with pairs of similar sentences. Compared to out-of-the-box RoBERTa, a RoBERTa-based sentence transformer drives semantically comparable sentence embeddings.

As we will explain in Sec. 4.3, we cluster the normalized sentence embeddings based on their Euclidean distance to efficiently acquire the labeled samples for the initial iteration of AL. The Euclidean distance between normalized embeddings can be driven from their cosine similarity. Hence, sentence embeddings that are comparable with cosine similarity can result in clusters with higher quality. In addition, semantically meaningful sentence embeddings give a better initialization of the `[CLS]` token, thereby obtaining better classification performance with a smaller training set.

We use a pre-trained RoBERTa-based sentence transformer (PT Sentence-RoBERTa) as a teacher model, and distill its knowledge to the TAPT RoBERTa. The resulting distilled task-adapted pre-trained (DisTAPT) RoBERTa produces semantically meaningful embeddings that are comparable via cosine similarity, and, as shown by our experiments (Sec. 6.2), benefit the classification task.

4.3 Initial Sampling

Many AL strategies (Gissin and Shalev-Shwartz, 2019; Gal and Ghahramani, 2016) require an initial set of N labeled samples containing P positive and $N - P$ negative sentences, which is either assumed to be available, or obtained by randomly sampling the entire dataset until the desired number of positive and negative samples are found. This approach is highly expensive for large and skewed datasets. We propose a simple, yet effective, strategy to efficiently acquire the initial labeled set. To this end, we leverage the distilled task-adapted pre-trained RoBERTa to cluster the unlabeled samples using KMeans algorithm (MacQueen, 1967). The labeled set for the initial iteration is then driven from the cluster medoids. As a result, we shrink the pool of candidates from the entire dataset to the cluster medoids, therefore, reduce the number of actions for obtaining the initial annotated set, while achieving comparable performance with the standard approach for initial sampling.

4.4 Active Learning

In the last phase, we iteratively fine-tune the DisTAPT RoBERTa for the downstream task. The initial labeled set is used at the first iteration. Then, more samples are labeled in the following rounds using an AL acquisition strategy until the annotation budget is exhausted, or the classifier satisfies the expected performance.

Our proposed pipeline can be used with existing AL strategies and, as demonstrated by our experiments (Sec. 6.2), consistently outperforms standard AL approaches, regardless of the AL strategy used.

5 Experimental Setup

We evaluate our approach against four standard active learning strategies provided in the Low-Resource Text Classification Framework (Ein-Dor et al., 2020):

- **Random** At each iteration, this approach randomly chooses samples for annotation.
- **Hard-Mining** Selects instances that the model is uncertain about, based on the absolute difference of prediction score and 0.5.
- **Perceptron Dropout** (Gal and Ghahramani, 2016) Also selects instances for which the model is least certain. The uncertainty is calculated using Monte Carlo Dropout on 10 inference cycles.
- **Discriminative Active Learning (DAL)** (Gissin and Shalev-Shwartz, 2019) Deploys a binary classifier to select instances that best represent the entire unlabeled samples.

We consider pre-trained RoBERTa and LEGAL-BERT (Chalkidis et al., 2020) as the baselines. However, we only evaluate our strategy using the pre-trained RoBERTa as our goal is not to rely on domain-adapted models like LEGAL-BERT since they might not always be available. For example, if the data is in German, we can find a pre-trained RoBERTa in German, but the LEGAL-BERT is pre-trained on English text only.

5.1 Datasets

We evaluate our framework on Contract-NLI (Koreeda and Manning, 2021) and LEDGAR (Tuggener et al., 2020) benchmarks.

Contract-NLI (Koreeda and Manning, 2021) is a dataset for document-level natural language inference. It consists of 607 documents with 77.8 spans per document on average. Each span is checked against 17 hypotheses and classified as contradiction, entailment, or not mentioned. In this work, we adapt this dataset to the classification task by considering each hypothesis as a category. If a span is classified as contradiction or entailment for a hypothesis, we label it with the corresponding category. Following this approach, we end up with a classification dataset with 4,371 train, 614 development, and 1,188 test samples within 17 classes.

LEDGAR (Tuggener et al., 2020) is a text classification benchmark consisting of a corpus of legal provisions in contracts. The entire dataset consists of 846,274 provisions and 12,608 labels. We only consider a subset of this dataset that corresponds to provisions with labels that appeared at least 10,000 times in the corpus, resulting in 44,249 train, 7,375 development, and 12,907 test samples across 5 categories. Similar to Tuggener et al. (2020), we perform a $70\%-10\%-20\%$ random split to obtain the train, development and test sets.

The class distributions of both datasets can be found in the appendix (Sec. A.1). Compared to Contract-NLI, LEDGAR has fewer categories, is an order of magnitude bigger, and is more balanced.

5.2 Implementation Details

We base our implementation on the Low-Resource Text Classification Framework provided by Ein-Dor et al. (2020)[1], and augment it with the task-adaptation, knowledge distillation, and initial sampling steps.

As the pre-trained model, we use `roberta-base`[2] (with 125M parameters), the RoBERTa (Liu et al., 2019) language model trained on the union of 5 datatsets: Book corpus (Zhu et al., 2015), English Wikipedia[3], CC-News (Mackenzie et al., 2020), OpenWebText Corpus (Gokaslan and Cohen), and Stories (Trinh and Le, 2018), none of which belong to the legal domain.

For LEGAL-BERT, we use the `nlpaueb/legal-bert-base-uncased`[4] (with 110M parameters), trained on six datasets containing legal docments across Europe and the US.

For task-adaptation, we continue pre-training RoBERTa for the MLM task using the available unlabeled data. We train for 10 epochs with batch-size 64, and the learning rate set to $3e{-}4$. The task-adapted model has perplexity 4.9706 for Contract-NLI and 2.1628 for LEDGAR.

For model distillation, we use `stsb-roberta-base-v2` (with 125M parameters), a RoBERTa-based sentence transformer trained on the STS benchmark (Cer et al., 2017), as the teacher model, and the task-adapted RoBERTa as the student model. Mean Squared Error (MSE) is used as the loss function. The student model is trained for 10 epochs, with 10K warmup steps, $1e{-}4$ learning

[1] https://github.com/IBM/low-resource-text-classification-framework

[2] https://huggingface.co/roberta-base

[3] https://dumps.wikimedia.org

[4] https://huggingface.co/nlpaueb/legal-bert-base-uncased

rate and no bias correction. The final MSE ($\times 100$) is 6.8607 for Contract-NLI, and 7.2003 for LEDGAR.

For clustering the normalized sentence embeddings we use the KMeans implementation by `scikit-learn`. We cluster the Contract-NLI and LEDGAR sentence embeddings into 437, and 442 groups respectively. The number of clusters are chosen based on the dataset size, and the number of categories, and to make initial sampling with cluster medoids manageable for experts.

In all the active learning experiments, we perform 5 AL iterations, starting with 10 initial samples, and increasing the size of the annotated data by 10 at each iteration. Adam optimizer (Kingma and Ba, 2015) is used with learning rate set to 5e$-$5. The model is trained for 100 epochs and early stopping is used with patience set to 10. To account for randomization, we repeat each experiment three times.

To compare our approach with standard AL methods, we use F1-score as the evaluation metric as it captures both precision and recall and is sensitive to data distribution.

6 Results and Discussion

In this section, we provide the results of our experiments, and explain them in detail. We start by comparing our approach with and without the initial medoid sampling against standard AL strategies (Sec. 6.1). Then, we show the effectiveness of knowledge distillation on top of task-adaptation (Sec. 6.2). In addition, we demonstrate the efficiency of the initial sampling with cluster medoids (Sec. 6.3). Finally, we evaluate how well our approach performs for different AL strategies (Sec. 6.4).

6.1 Efficient AL Pipeline

Figure 1 compares our approach with and without the initial sampling phase (DisTAPT with IS, and DisTAPT) to standard DAL with pre-trained (PT) and TAPT RoBERTa for Contract-NLI and LEDGAR benchmarks. We report the average F1-score over all categories. DAL is chosen due to its better performance, as shown in Figure 2. The results for other AL strategies can be found in the appendix (Sec. A.2).

Our experiments show the importance of task-adaptation and knowledge distillation for pre-trained language models prior to fine-tuning with active learning. Figure 1 illustrates that, for the same size of annotated data, our pipeline

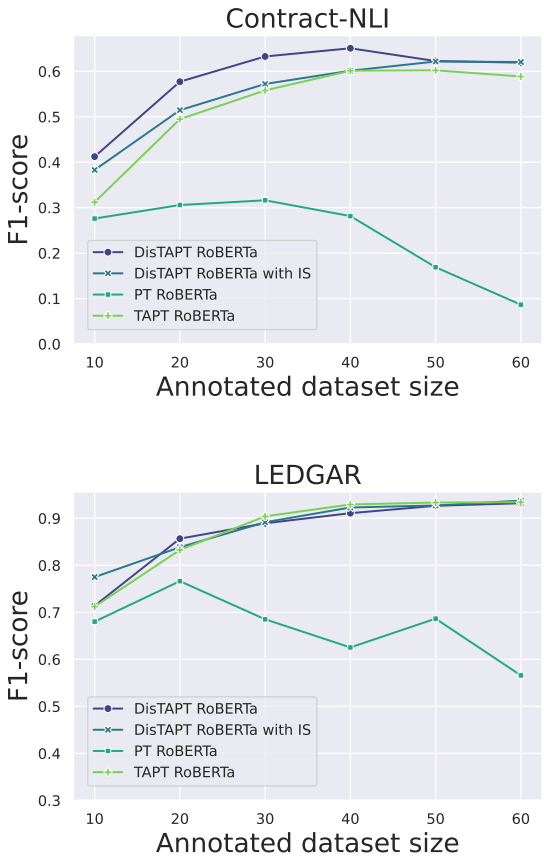


Figure 1: Test F1-score for **DAL** during AL iterations. The F1-score for the fully supervised fine-tuning is 0.6990 for Contract-NLI and 0.9538 for LEDGAR. The figure is best viewed in color.

consistently achieves better performance than standard AL approaches.

For the Contract-NLI dataset, the F1-score obtained by fully-supervised fine-tuning (with 4,371 labeled samples) is 0.6990 for `roberta-base` and 0.7152 for `legal-bert-base-uncased`. DisTAPT RoBERTa reaches a F1-score as high as 0.6508 with only 40 labeled samples. The best F1-score obtained using pre-trained RoBERTa is 0.3162 with 30 labeled samples, which is 0.3165 lower than DisTAPT RoBERTa's F1-score for the same size of annotated data.

For the LEDGAR dataset, the F1-score obtained by the fully-supervised fine-tuning (with 44,249 labeled samples) is 0.9538 for `roberta-base` and 0.9588 for `legal-bert-base-uncased`. DisTAPT RoBERTa reaches a very close performance of 0.9321 F1-score with merely 60 labeled samples. The highest F1-score that pre-trained RoBERT reaches is 0.7663 with 20 annotated samples, which

is 0.0904 lower than DisTAPT's performance with the same size of labeled data.

These results show that, for both datasets, there is only a small performance gap between our approach and the fully-supervised approach, indicating that our AL pipeline dramatically reduces the annotation cost, while achieving comparable performance with the fully-supervised fine-tuning.

In addition, It is observed that standard AL with off-the-shelf pre-trained RoBERTa is unstable. This is aligned with the previous works' observations (Mosbach et al., 2021; Zhang et al., 2021; Dodge et al., 2020). During fine-tuning, the pre-trained model should perform two tasks: adaptation to the legal domain with the new vocabulary, and classification. By performing task-adaptation and knowledge distillation before fine-tuning, we train the model in a curriculum learning approach, making the model stable even for small training sets.

6.2 Effect of Knowledge Distillation

To evaluate the effectiveness of knowledge distillation on the quality of obtained clusters, we compare the distribution of the Dunn Index of the clusters before and after knowledge distillation. For both datasets, after knowledge distillation, most of the clusters have higher Dunn Index which indicates that they are more compact and better separated than the clusters before knowledge distillation step. The results are provided in the appendix A.3 due to space constraints.

In addition, we evaluate the effect of knowledge distillation on the task-adapted pre-trained RoBERTa, and report the average F1-score over all classes for each dataset. Figure 1 shows that, for both datasets, DisTAPT RoBERTa outperforms TAPT RoBERTa at early iterations of active learning, and as the size of the labeled set increases, the two models' performance converge. This can be explained by the fact that, initially, DisTAPT RoBERTa's embeddings better capture the semantics of sentences, and thus result in better classification performance. As the labeled data grows, TAPT RoBERTa is fine-tuned and can produce semantically meaningful embeddings as well. Hence, for a highly restricted annotation budget, distilling the knowledge of a sentence transformer to the TAPT language model can lead to performance gain.

6.3 Efficiency of Initial Medoid Sampling

It was shown in Figure 1 that DisTAPT with IS obtains comparable performance with DisTAPT without IS. In this section, we evaluate the *efficiency* of the proposed sampling strategy for the initial iteration of AL.

To this end, we simulate the standard sampling strategy by randomly sampling text segments from the full dataset until 5 positive and 5 negative samples are found. The number of iterations is then considered as the number of annotations required to collect the labeled set for the initial AL iteration. Similarly, to simulate our proposed initial sampling, we randomly sample from cluster medoids until 5 positive and 5 negative samples are obtained. To account for randomness, we repeat the simulations 1000 times and report the median and the 90^{th} percentile over all runs.

Table 1 illustrates the results of our simulations for Contract-NLI and LEDGAR. Due to the high number of classes in Contract-NLI, only eight categories of this dataset are presented in this table, and the results for other categories can be found in the appendix (Sec. A.4). For each class, in addition to the median and 90^{th} percentile over 1000 runs, the difference in the 90^{th} percentile between standard approach and our strategy (in $\%$) is reported as the gain in annotation effort. For example, for the `Sharing with third-parties` class in Contract-NLI, the 90^{th} percentile is 62% less when using medoids for initial sampling, meaning that, with 90% confidence, the annotators perform 62% fewer actions to acquire the initial labeled set using our approach.

It is observed that, for the skewed Contract-NLI dataset, our proposed initial sampling strategy reduces the number of actions performed by the annotator up to 63%. For LEDGAR however, which consists of balanced categories, the highest effort gain in sampling from cluster medoids is 25%. There are also few cases where using the entire dataset is more efficient than sampling from medoids. This happens when the class' frequency is higher in the full dataset than its frequency in the cluster medoids.

Overall, our results demonstrate the advantage of using the cluster medoids for collecting the initial annotated samples for a skewed dataset like Contract-NLI, which is a realistic use-case in the legal domain. It is noteworthy that the original version of LEDGAR dataset is also imbalanced, but as explained in Sec. 5.1, due to the drastically high number of classes, and for the sake of comparison with skewed datasets, only the most dominant categories are kept in this work.

Thanks to the semantically meaningful and

comparable sentence embeddings obtained after the knowledge distillation step, the cluster medoids well represent the entire dataset, and thus sampling among them drastically reduces the annotation effort without harming the performance. As a real life scenario, consider a company with hundreds of legal contracts aiming to classify their sentences into multiple categories, under a restricted budget. Reducing the annotation effort means lowering down the financial costs of annotation, which can be highly expensive in the legal domain (over $2 million for annotating around 500 contracts according to Hendrycks et al. (2021)).

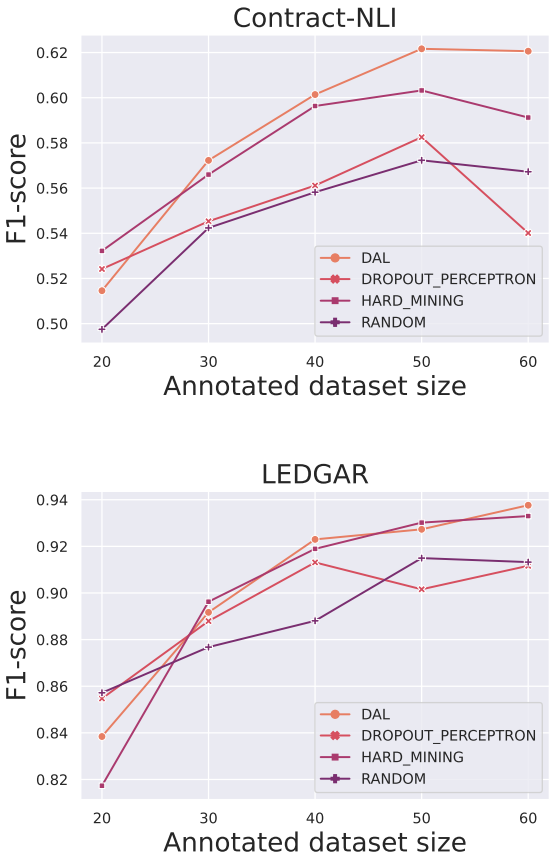


Figure 2: Comparison of four AL strategies when used with DisTAPT RoBERTa with IS.

6.4 Effect of AL strategy

Finally, we evaluate the generalizability of our approach over the four AL strategies mentioned in Sec. 5: DAL, Random, Hard-Mining, and Perceptron Dropout. As shown in Figure 2, DAL results in the best performance with at most 0.08 higher F1-score than other strategies with 60 labeled samples for Contract-NLI, and less than 0.04 higher F1-score with 40 annotated samples for LEDGAR. The small performance gap of these four AL methods in our pipeline indicates the generalizability of this approach to various AL strategies.

7 Conclusion

We propose an efficient active learning pipeline for legal text classification. Our approach leverages the available unlabeled data to adapt the pre-trained language model to the downstream task, and guide its embeddings to a semantically meaningful space before fine-tuning. We use model distillation to produce semantically comparable embeddings. A future work can study the effect of other approaches like BERT-Flow (Li et al., 2020) and whitening (Su et al., 2021) on AL with this pipeline. Moreover, we design a simple strategy to efficiently acquire a labeled set of positive and negative samples for the initial iteration of active learning.

Our experiments over Contract-NLI and LEDGAR benchmarks demonstrate the effectiveness of our approach compared to standard active learning strategies. Our results also show that our pipeline obtains very close performance to the fully-supervised approach with considerably less annotation cost. We test our methodology in the legal domain, and for four AL strategies, but we expect it to generalize to other strategies like ALPS and BADGE, and other specialized domains, like medicine. We leave this evaluation as a future work.

Limitations

In this work, we have shown the importance of task-adaptation and knowledge distillation, and that we can leverage the available unlabeled data to perform efficient fine-tuning via active learning and obtain better performance. The price to pay for this performance gain is time and computational power. The time taken by task-adaptation and distillation scales with the size of unlabeled data. On the other hand, more unlabeled samples result in more effective adaptation to the downstream task. Therefore, the user of this approach needs to find the best trade-off given their data, annotation budget, time and computational power. For, LEDGAR, the larger dataset used in this work, we performed the adaptation and distillation steps in 4 and 1 hour(s) respectively, using a single Nvidia GeForce GTX TITAN X GPU.

Moreover, we showed that by clustering the sentence embeddings produced by DisTAPT RoBERTa, the initial labeled set can be acquired

Dataset	Category	full dataset		medoids		gain(%)
		median	90^{th}%tile	median	90^{th}%tile	
Contract-NLI	Inclusion of verbally conveyed information	75.0	125.0	35.5	59.0	52.8
	No licensing	64.0	108.0	68.5	109.1	-1.0
	No reverse engineering	342.0	568.0	144.0	209.1	63.2
	Notice on compelled disclosure	74.5	122.0	99.0	155.0	-27.0
	Sharing with employees	57.0	90.0	21.0	34.1	62.1
	Sharing with third-parties	54.0	92.1	21.0	35.0	62.0
	Survival of obligations	64.0	106.0	36.0	57.0	46.2
	Return of confidential information	116.0	189.0	61.0	99.0	47.6
LEDGAR	Amendments	23.0	37.1	21.0	33.0	10.8
	Counterparts	26.0	42.0	34.0	54.1	-28.8
	Entire agreements	26.0	42.0	33.0	55.0	-30.9
	Governing laws	17.5	28.0	14.0	21.0	25.0
	Notices	29.0	49.0	26.0	44.0	10.2

Table 1: Number of actions to acquire the initial labeled set for 8 categories of Contract-NLI, and LEDGAR when sampling from the full dataset (standard approach), and sampling from the cluster medoids (our approach).

more efficiently. Nevertheless, this approach inherits the limitations of clustering. Namely, the time complexity of clustering the embeddings scales with the data, and the number of clusters should be empirically chosen. In our experiments we spent 10 minutes to cluster the 44,249 samples belonging to LEDGAR dataset into 442 groups.

Ethics Statement

Industries have hundreds of contracts with tens of thousands of sentences that belong to various topics. Labeling all of these samples is a highly expensive and time-consuming process. In this work, we aim to reduce the resources spent on this task by leveraging recent advances in natural language processing, while keeping the human expert in the loop. The goal is to reduce the human effort in annotation so that the legal experts' time and knowledge can be used in another task at which humans are better than machines.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. *ArXiv*, abs/1906.03671.

Daniel Matthew Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval@ACL*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *ArXiv*, abs/2010.02559.

Rishi Chhatwal, Nathaniel Huber-Fliflet, Robert Keeling, Jianping Zhang, and Haozhen Zhao. 2017. Empirical evaluations of active learning strategies in legal document review. *2017 IEEE International Conference on Big Data (Big Data)*, pages 1428–1437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv*, abs/2002.06305.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for bert: An empirical study. In *EMNLP*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *ArXiv*, abs/1506.02142.

Daniel Gissin and Shai Shalev-Shwartz. 2019. Discriminative active learning. *ArXiv*, abs/1907.06347.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *ArXiv*, abs/2004.10964.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *ArXiv*, abs/2103.06268.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Andreas Kirsch, Joost R. van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *NeurIPS*.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *ArXiv*, abs/2110.01799.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR '94*.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *EMNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R. Trippas, J. Shane Culpepper, and Alistair Moffat. 2020. Cc-news-en: A large english news corpus. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations.

Katerina Margatina, Loïc Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *ACL*.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *ArXiv*, abs/2006.04884.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. *arXiv: Machine Learning*.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *ArXiv*, abs/2103.15316.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *ArXiv*, abs/1806.02847.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. Ledgar: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *LREC*.

Ran Wang, Xizhao Wang, Sam Tak Wu Kwong, and Chen Xu. 2017. Incorporating diversity and informativeness in multiple-instance active learning. *IEEE Transactions on Fuzzy Systems*, 25:1460–1475.

Michelle Yuan, Hsuan-Tien Lin, and Jordan L. Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. *ArXiv*, abs/2010.09535.

Mike Zhang and Barbara Plank. 2021. Cartography active learning. In *EMNLP*.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting few-sample bert fine-tuning. *ArXiv*, abs/2006.05987.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A Appendix

A.1 Dataset Distributions

We provide the details of class distributions for Contract-NLI and LEDGAR benchmarks in Table 2. As shown in this table, LEDGAR contains considerably larger categories compared to Contract-NLI and is more balanced.

A.2 Effective Fine-Tuning

Here we present the results of standard active learning and our approach for four AL strategies discussed in Sec. 5 including Random, Hard-Mining, and Perceptron Dropout. As before, we report the average F1-score over three runs. Figure 3 corresponds to Contract-NLI and Figure 4 illustrates the results for the LEDGAR dataset.

A.3 Effect of Knowledge Distillation

Figures 5 and 6 illustrate the comparison of the Dunn Index distribution that were not presented in the main paper.

A.4 Efficiency of Initial Sampling with Medoids

In Table 3 we provide the median and 90^{th} percentile of number of actions performed to collect the initial labeled set, for the standard sampling approach, and our proposed strategy using cluster medoids,

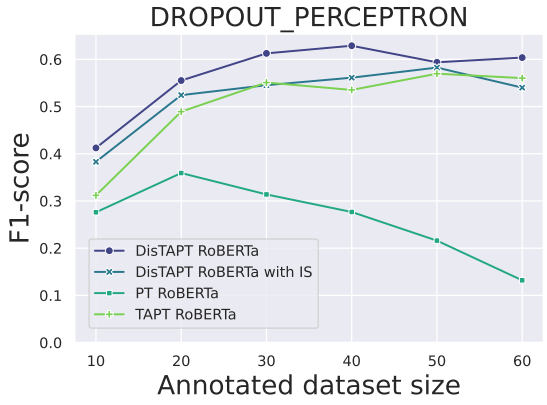
Dataset	Category	Train Size	Dev Size	Test Size
Contract-NLI	Confidentiality of Agreement	161	29	46
	Explicit identification	203	29	60
	Inclusion of verbally conveyed information	274	45	76
	Limited use	371	53	110
	No licensing	327	39	86
	No reverse engineering	60	8	13
	No solicitation	93	11	28
	None-inclusion of non-technical information	332	50	94
	Notice on compelled disclosure	276	45	77
	Permissible acquirement of similar information	311	47	96
	Permissible copy	167	17	49
	Permissible development of similar information	263	40	73
	Permissible post-agreement possession	312	25	63
	Return of confidential information	182	24	38
	Sharing with employees	358	56	94
	Sharing with third-parties	370	53	102
	Survival of obligations	311	43	83
LEDGAR	Amendments	9,132	1,515	2,615
	Counterparts	8,033	1,312	2,363
	Entire agreements	8,094	1,361	2,370
	Governing laws	11,926	1,997	3,454
	Notices	7,064	1,190	2,105

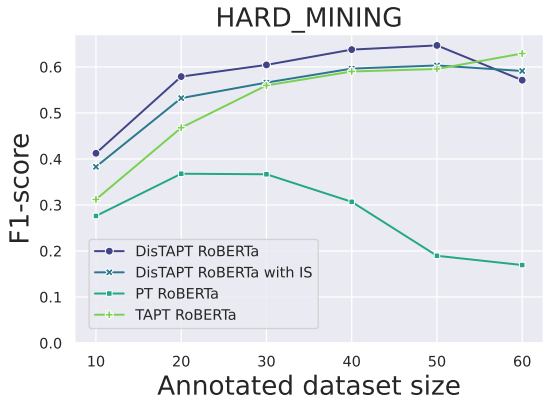
Table 2: Category frequency for Contract-NLI adapted to classification task, and LEDGAR benchmarks.

for nine categories of Contract-NLI that were not included in Table 1 in Sec. 6.3. It is observed that, for most categories, there is a considerable reduction in the number of actions performed to acquire the annotated data for the initial AL iteration.

Category	full dataset		medoids		gain(%)
	median	90^{th}%tile	median	90^{th}%tile	
Confidentiality of Agreement	125.0	215.1	120.0	178.2	17.1
Explicit identification	100.0	161.1	48.0	77.0	52.2
Limited use	56.0	90.1	37.0	58.0	35.6
No solicitation	227.0	383.0	178.0	261.0	31.8
None-inclusion of non-technical information	61.0	101.1	39.0	64.0	36.7
Permissible acquirement of similar information	65.0	107.0	91.0	145.0	-35.5
Permissible copy	121.0	197.0	68.0	108.0	45.2
Permissible development of similar information	77.0	129.1	82.0	129.0	0.1
Permissible post-agreement possession	66.0	108.1	41.0	66.0	38.9

Table 3: Number of actions to acquire the initial labeled set for 9 categories of Contract-NLI when sampling from the full dataset (standard approach), and sampling from the cluster medoids.





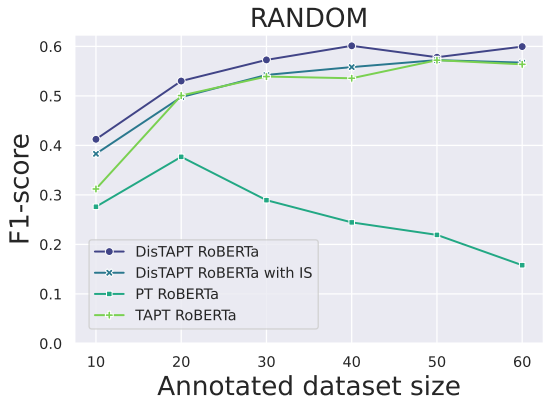
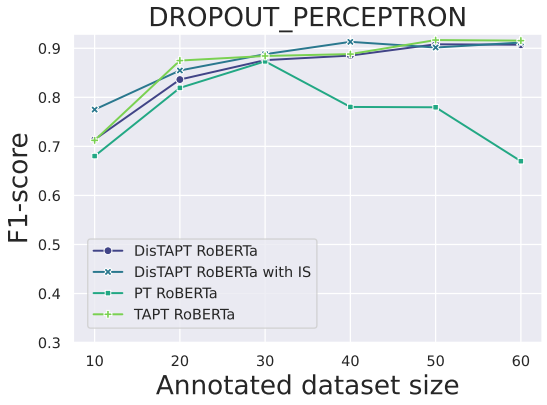
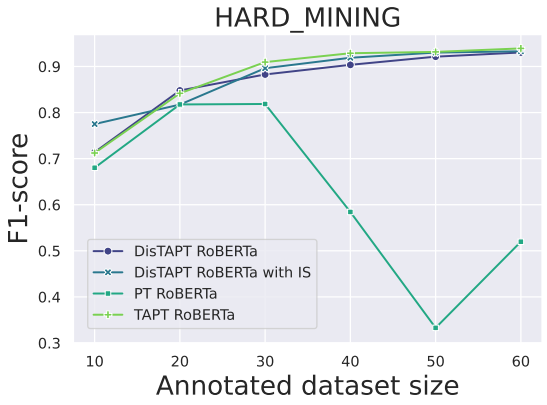


Figure 3: Test F1-score for **Contract-NLI** during AL iterations. The F1-score for the fully supervised fine-tuning is 0.6990.





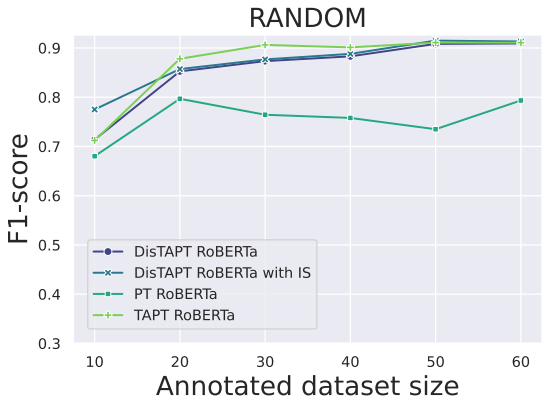
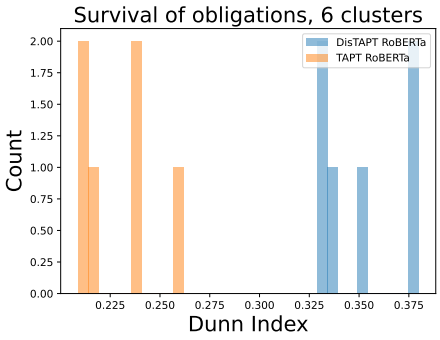
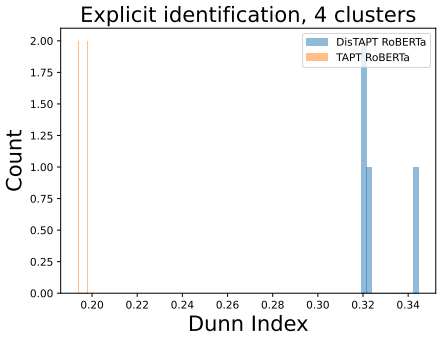
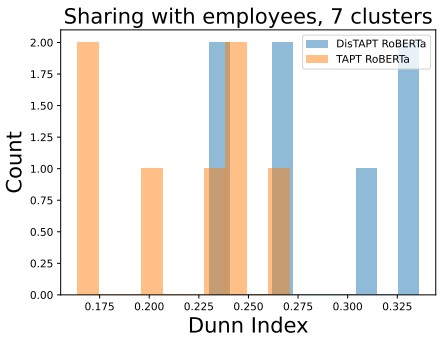
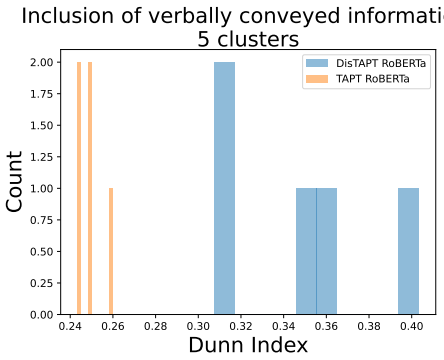


Figure 4: Test F1-score for **LEDGAR** during AL iterations. The F1-score for the fully supervised fine-tuning is 0.9538.









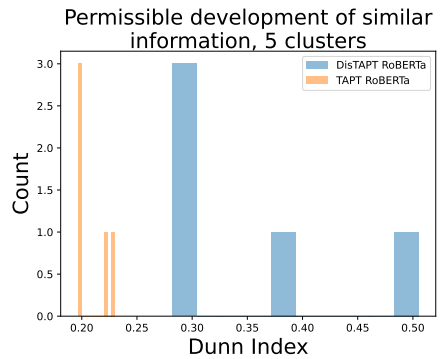
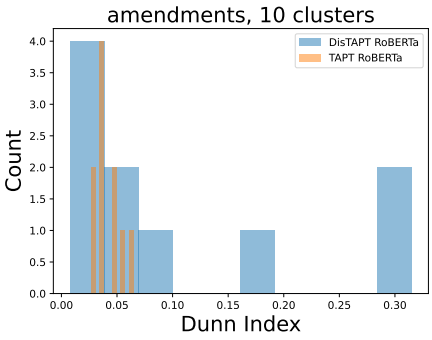
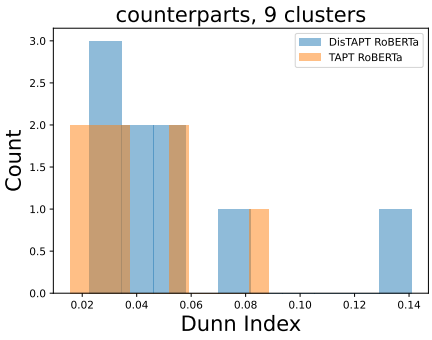
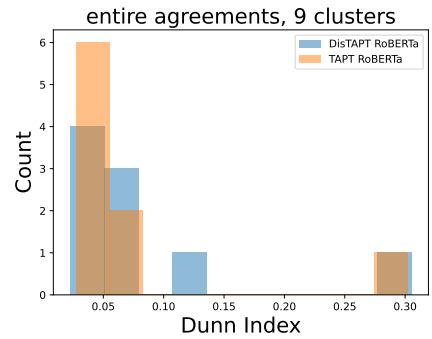
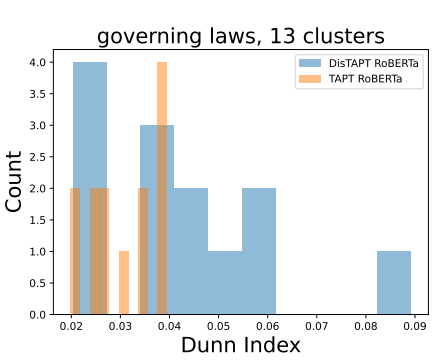


Figure 5: Comparison of the Dunn Index distribution before (TAPT RoBERTa) and after knowledge distillation (DisTAPT RoBERTa) for **Contract-NLI** dataset.









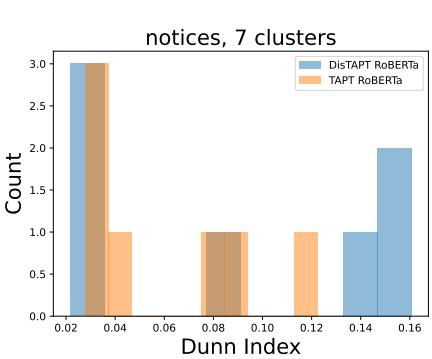


Figure 6: Comparison of the Dunn Index distribution before (TAPT RoBERTa) and after knowledge distillation (DisTAPT RoBERTa) for **LEDGAR** dataset.