

Adversarially Robust Neural Legal Judgement System

Rohit Raj^{1,*}, V.Susheela Devi^{1,†}

¹Department of CSA, Indian Institute of Science, Bengaluru, karnataka

Abstract

Legal judgment prediction is the task of predicting the out- come of court cases on a given text description of facts of cases. These tasks apply Natural Language Processing (NLP) techniques to predict legal judgment results based on facts. Recently, large-scale public datasets and NLP models have increased research in areas related to legal judgment prediction systems. For such systems to be practically helpful, they should be robust from adversarial attacks. Previous works mainly focus on making a neural legal judgement system; however, significantly less or no attention has been given to creating a robust Legal Judgement Prediction(LJP) system. We implemented adversarial attacks on early existing LJP systems and found that none of them could handle attacks. In this work, we proposed an approach for making robust LJP systems. Extensive experiments on three legal datasets show significant improvements in our approach over the state-of-the-art LJP system in handling adversarial attacks. To the best of our knowledge, we are the first to increase the robustness of early-existing LJP systems.

Keywords

Natural Language Processing, Legal Judgement Prediction, Robust Models

1. Introduction

Legal information is mainly in the form of text, so legal text processing is a growing area of research in NLP, such as crime classification [1], judgment prediction [2], and summarization [3]. Countries like India, which are highly populated, have many pending legal cases (approx 41 million). In Brazil, only in the financial domain, three hundred thirty-two thousand cases are in progress [4]. It is due to multiple factors, including the unavailability of judges. Here legal judgment prediction system can help in several steps like finding articles or the history of a case, deciding penalty terms, etc. Also, legal judgment prediction is critical, so a small error in the system may drastically affect judicial fairness.

Most of the researchers focused on making LJP systems by training NLP models (LSTM, BERT [5], legal-BERT [6]) on legal datasets. At the same time, very little or no attention has been given to the robustness of these models.

SAIL'23: 3rd Symposium on Artificial Intelligence and Law, 24-26 February, 2023, Hybrid Event, Hyderabad, India


*Corresponding author.


†These authors contributed equally.

✉ rohitr@iisc.ac.in (R. Raj); susheela@iisc.ac.in (V.Susheela Devi)

🌐 <https://www.csa.iisc.ac.in/~susheela/> (V.Susheela Devi)

🆔 0000-0003-1421-6286 (R. Raj)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

We summarise our contribution as follows:

- We implemented adversarial attacks on existing baseline models after fine-tuning them on legal datasets and found that their performance decreased drastically.
- We suggested an algorithm for adversarial training for making robust legal models.
- We implemented training using data augmentation and adversarial training methods to improve the model's robustness.

2. Related Work

2.1. Legal Judgement System

Earlier legal judgment prediction systems involved linear models like SVM with a bag of words as feature representation. In recent years, neural network [2] methods have been used for legal domains due to the availability of NLP models like RNN and BERT [5].

Most researchers used BiGRU-att [2], HAN [2], BERT [5] and Hier-BERT [7] architecture to predict article violation on ECtHR [8] dataset. Legal-BERT [6] is a domain-specific BERT pre-trained on legal-documents corpora of approx 11.5 GB, used for legal judgment prediction. A number of other tasks like legal summarization [3], prior case retrieval [9], and legal QA [10] have been introduced.

In legal judgment prediction, the model must predict the final decision based on case facts. Several datasets are introduced for training so that model can learn specific words (for example, 'restrictive covenant', 'promissory estoppel', 'tort', and 'novation') that are being used in legal documents which are not used for general purposes; for example, ECHR [8], a multilabel dataset containing violated articles as the label. SCOTUS [8] contains cases of the American Supreme Court, and ILDC [11] contains cases of the Indian supreme court. All of these are English datasets. However, datasets from different languages are also introduced, like Chinese [1], Swedes [12], and Vietnamese [13].

2.2. Adversarial Training

Several adversarial training methods have been explored in NLP models to increase their robustness. The models are trained on a dataset containing augmented adversarial examples with the original dataset in adversarial training. These adversarial examples are generated by applying adversarial attacks on pre-trained models such that generated examples should be similar to the original example, and the average human user cannot differentiate it from the natural one. Several adversarial attack mechanisms are being used in NLP, such as BERT-Attack [14], BAE [15], A2T [16], TextFooler [17]. In these attacks, the model finds essential words in the original text and replaces them with semantically similar words such that the label of the original text changes and generates adversarial text that looks similar to the original text.

2.3. Why adversarial training ?

To motivate the necessity of adversarial training, we implemented adversarial attacks on existing baseline models (BERT [5], Legal-BERT [6], RoBERTa [18]) to check their robustness. We found that the performance of these models decreased drastically, as these models could not handle the adversarial attack. We also implemented data augmentation using back-translation during training, but the model's performance was not improved much.

Legal judgment prediction is critical, so a slight variation in the input may affect judgment fairness. So during deployment, if someone intentionally perturbs the input sequence, the prediction may change drastically. It is the main reason for adversarial training.

3. Problem Formulation

Given a legal dataset, which contains a collection of legal documents, $L = \{(X_1, y_1) \dots (X_N, y_N)\}$, where X_i is a legal text extracted from a legal document and $y_i = \{1, 2, 3 \dots K\}$. Here the length of each X_i is very large, and y_i is a label corresponding to that text.

The task is to design an LJP model $M(\cdot)$ that can:

- Predict correct class on legal documents of even large length.
- Perform correct prediction even if data is perturbed. Let X' be a perturbed text, which may be perturbed intentionally or by mistake, then $M(X') \rightarrow y$, where y is the correct label of that legal text.

4. Methods

In this section, we present our training workflow. We implemented three methods for training. These are 1) Fine-tuning Baseline models, 2) Training baseline models with data augmentation 3) Adversarial training using augmenting adversarial examples with natural examples. At the end of each method, we tested our model's robustness with adversarial attacks.

4.1. Fine tuning baseline model

In this approach, we have taken baseline models (BERT [5], Legal-BERT [6], RoBERTa [18], Hierarchical Version of BERT [7], we have used a modified version of Hier-BERT, denoting as H-BERT) and fine-tuned them on our downstream tasks for legal judgment predictions. For BERT, Legal-BERT, and RoBERTa, we have taken the last 512 tokens of each input text for training, as this approach gave a better result. For H-BERT (modified Hierarchical Version of BERT), we have divided the text into chunks of 510 tokens such that two consecutive chunks overlapped each other, here RoBERTa is taken as encoder, shown in Figure 2, as it gives the best result. We have used cross-entropy as a loss function for updating the gradient and evaluated model performance on accuracy.

4.2. Training using data-augmentation

In this approach, we first generated data using back-translation[19] and then augmented it with training data. The algorithm for training is similar to Algorithm 2, where in place of an adversarial example generator, we are using a data augmenter.

We use the transformer model implemented by HuggingFace [20] for back-translation. We first translate English to French and then translate it back to English from French. We augment newly generated data such that it does not have any duplicate instances, and training is done similarly to approach 1.

4.3. Adversarial Training

In this approach, we generate adversarial examples from original legal document datasets, then further augment these examples with legal document datasets and train the model on this new dataset, i.e., $D_{new} = D_{nat} \cup D_{adv}$.

For generating an adversarial example from a text sample, first, we find the importance score of each word in that sample using greedy search with word importance ranking mechanism [21], where the importance of the word is determined by how much heuristic score changes when a word is deleted from the original input. i.e.,

$$I_{w_i} = \begin{cases} M_y(X) - M_y(X/w_i), & \text{if } M(X) = M(X/w_i) = y, \\ (M_y(X) - M_y(X/w_i)) + (M_{y'}(X/w_i) - M_{y'}(X)), & \text{if } M(X) = y, M(X/w_i) = y' \\ & \text{and } y \neq y' \end{cases} \quad (1)$$

Here we have followed the deletion approach for finding word importance because we are considering a common black-box setup which is usually followed in a real-world scenario. We denote sentence after deletion of word w_i as $X/w_i = \{w_1, \dots, w_{i-1}, w_{i+1}, \dots, w_n\}$ and use $M_y(\cdot)$ to denote prediction score of model for label y . Here I_{w_i} denote importance score of word w_i which is defined in Equation 1.

As shown in Algorithm 1, in lines 3-4, we find the importance score of words using Equation 1 (after removing stop-words). After that, generate ‘m’ synonyms for each word using cosine-similarity and counter-fitted-word-embedding [19]. We then replace original words with synonyms and make an adversarial example X' . Further, to find the similarity of an adversarial sample X' to the original sample X , we use Universal Sentence Encoder (USE) [22]. We ignore the examples below a certain threshold value. We have taken 0.7 as the threshold value for all of our experiments. We have implemented all of our algorithms on top of the Textattack framework. In all of our experiments, perturbation percent is below 20%.

For adversarial training, we first fine-tune the model using natural legal dataset D_{nat} for some iterations, i.e., n_{nat} ; then we generate adversarial example D_{adv} by using our adversarial example generator and augment with the natural legal dataset, i.e., $D_{new} = D_{nat} \cup D_{adv}$. Further, we train the model on D_{new} for some iterations, i.e., n_{adv} . Here adversarial loss

Algorithm 1 Adversarial Example Generation from legal Sample

```
1: Input: Legal judgement prediction model  $M(\cdot)$ , legal sample sentence  $X = (w_1, w_2, \dots, w_n)$ ,  
   Perturbation Generator  $P(X, i)$  which replace  $w_i$  with certain perturbed word using  
   counter-fitted-word-embedding  
2: Output: Adversarial legal sample  $X_{adv}$   
3: Calculate importance score  $I_{w_i}$  of each word  $w_i$  using equation 1.  
4: Rank them in decreasing order according to  $I_{w_i}$  and store them in set  $R = (r_1, r_2, \dots, r_k)$   
5:  $X' \leftarrow X$   
6: for  $i = r_1, \dots, r_n$  in  $R$  do,  
7:    $X_p \leftarrow$  perturb the sentence  $X'$  using  $P(X', i)$   
8:   if  $M(X_p) \neq y$  then  
9:     if  $sim(X_p, X) > threshold$  then ▷ Check similarity of  $X$  and  $X'$   
10:       $X' \leftarrow X_p$   
11:     end if  
12:   end if  
13: end for  
14: return  $X'$  as  $X_{adv}$ 
```

function is used to train the model.

Let L_{nat} be the loss function used for natural training, which is defined as a cross-entropy loss function, i.e.,

$$L_{nat} = L_{\theta}(X, y) \quad (2)$$

where X is the input text and y is the label corresponding to it. If $A_{\theta}(X, y)$ is the adversarial example generator, then the loss function for adversarial training is defined as a cross-entropy loss function, i.e

$$L_{adv} = L_{\theta}(A_{\theta}(X, y), y) \quad (3)$$

So our final loss function will be the combination of these two cross-entropy loss functions, i.e.,

$$L = \operatorname{argmin}_{\theta}(L_{nat} + \gamma L_{adv}) \quad (4)$$

where γ is a hyper-parameter used to change the importance of adversarial training.

Our adversarial training algorithm is shown in Algorithm 2, Lines 3-6 represent the natural training of the model. Lines 7-18 represent the adversarial training of the model, which is pre-trained in lines 3-6. In lines 10-15, adversarial examples are generated. Line 16 represents the augmentation of adversarial examples with natural data. Line 17 represents the adversarial training step.

Algorithm 2 Adversarial Training of legal Models

```
1: Input: Legal judgement prediction model  $M(\cdot)$ , Adversarial example generator algorithm  $A_\theta(X, y)$ , legal dataset  $D_{nat} = \{X, y\}_{i=1}^m$ , natural training epochs  $n_{nat}$ , adversarial training epochs  $n_{adv}$ 
2: Output: Adversarially trained model
3: Randomly initialize  $\theta$ 
4: for  $i = 1, 2..n_{nat}$  do,
5:   Train  $M_\theta(\cdot)$  on dataset  $D_{nat}$  using loss function from Equation (2).
6: end for
7: for  $i = 1, 2..n_{adv}$  do,
8:   Initialize set of adversarial legal dataset  $D_{adv} \leftarrow \{\}$ 
9:    $K \leftarrow$  fraction of adversarial samples to be generated of natural dataset
10:  for  $i = 1, 2..size(D_{nat})$  do,
11:    if  $size(D_{adv}) < K * D_{nat}$  then
12:       $X_{adv} \leftarrow A_\theta(X, y)$ 
13:       $D_{adv} \leftarrow D_{adv} \cup \{X_{adv}, y\}$ 
14:    end if
15:  end for
16:   $D_{new} \leftarrow D_{nat} \cup D_{adv}$ 
17:  Train  $M_\theta(\cdot)$  on  $D_{new}$  using loss function from Equation (4).
18: end for
```

5. Experiments and Results

5.1. Datasets and Models

5.1.1. Datasets

ECHR[2]: It contains cases of the European Council of Human Rights (ECHR). The dataset has 11.5k cases, of which 7100 cases are used for training, 1380 for development, and 2998 for the test set. The training and development set contains cases from 1959-2013, and the test set contains cases from 2014-2018. Total ECHR articles are 66; however, we have taken **binary representation** of the ECHR dataset, in which label 1 is assigned if any article is violated; otherwise, 0 is assigned.

SCOTUS [8] : It is a dataset of the US Supreme Court, which hears only complex cases not well solved by lower courts. SCOTUS is a multi-class dataset containing 14 classes consisting of broad areas like Civil Rights, Criminal Procedure, Economic Activity, etc. The SCOTUS cases are split into a 5k (1946-1982) training set, 1.4k (1982-1991) development set, and 1.4k (1991-2016) test set. We took only top-4 categories which are approximately balanced and consist of 3.6k cases for training, 969 for development and test sets each.

ILDC : Indian Legal Document Corpus (ILDC) is introduced by Malik et al. [11], which contains cases of the Supreme Court of India (SCI) from 1947 to 2020. It is a **binary classification** dataset having binary labels $\{0, 1\}$. It has two versions. **1) ILDC-single** contains cases of a single petition filed, label 1 is assigned to cases whose petition is accepted, and 0 is for not

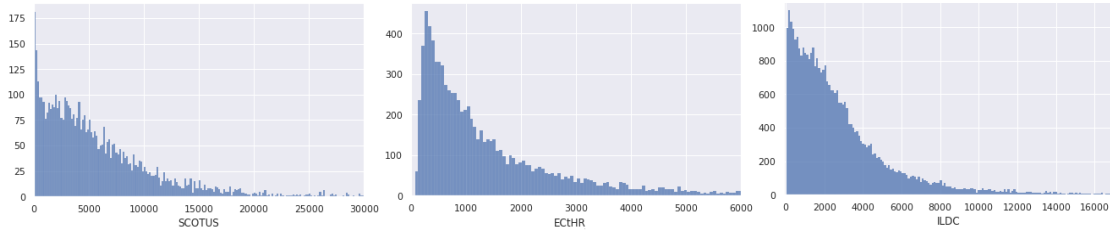


Figure 1: Text length distribution of different datasets, here horizontal axis shows the length of input texts and vertical axis show number of inputs

accepted. 2) **ILDC-multi** contains cases with multiple petitions filed. Here label 1 is assigned to cases with at least one petition accepted; otherwise, label 0 is assigned. We have taken ILDC-multi for all of our experiments.

As shown in Figure 1, legal text datasets have a substantial length. Here the average length of samples of ECHR is **1619** words, SCOTUS is **5853** words, and ILDC is **3208** words. So it is far greater than a normal BERT architecture input size. Therefore, we have implemented the modified Hierarchical Variant of BERT (H-BERT) architecture.

5.1.2. Models Used

BERT[5] is a pre-trained transformer-based language model. It is pre-trained to perform masked language modeling and next-sentence prediction.

Legal-BERT [6] is BERT pre-trained on English legal corpora, which contains legislation, contracts, and court cases. Its configuration is the same as the original BERT configuration. The sub-word vocabulary of Legal-BERT is built from scratch.

Hierarchical Variant of BERT (H-BERT) Legal documents are usually of large text length (shown in Figure 1), for example, ECHR, ILDC, and SCOTUS. Transformer-based models can handle up to 512 sub-word units. So we implemented an architecture similar to Chalkidis et al. [7] in which we divided the text into the chunk of 510 tokens such that two consecutive chunks have 100 overlapping tokens. Each chunk is sent through a BERT-Encoder to generate CLS embedding. Figure 2 shows that CLS embedding is passed to 1-dimensional convolution and max-pooling layers. A further output of the max-pooling layer is passed to Bi-directional LSTM and then the Dense layer. We have taken RoBERTa as an encoder as it gave the best result among all other BERT-based models.

5.2. Implementation Details

For all tasks, we use pre-trained transformer-based BERT models from Huggingface implementation. Each model output a 768-dimension vector regarding each input text. The batch size is set to 8. Models trained using Adam optimizer with 1e-5 learning rate for overall 10 epochs, which includes 3 epochs of natural training and 7 epochs of adversarial training. We used LSTM

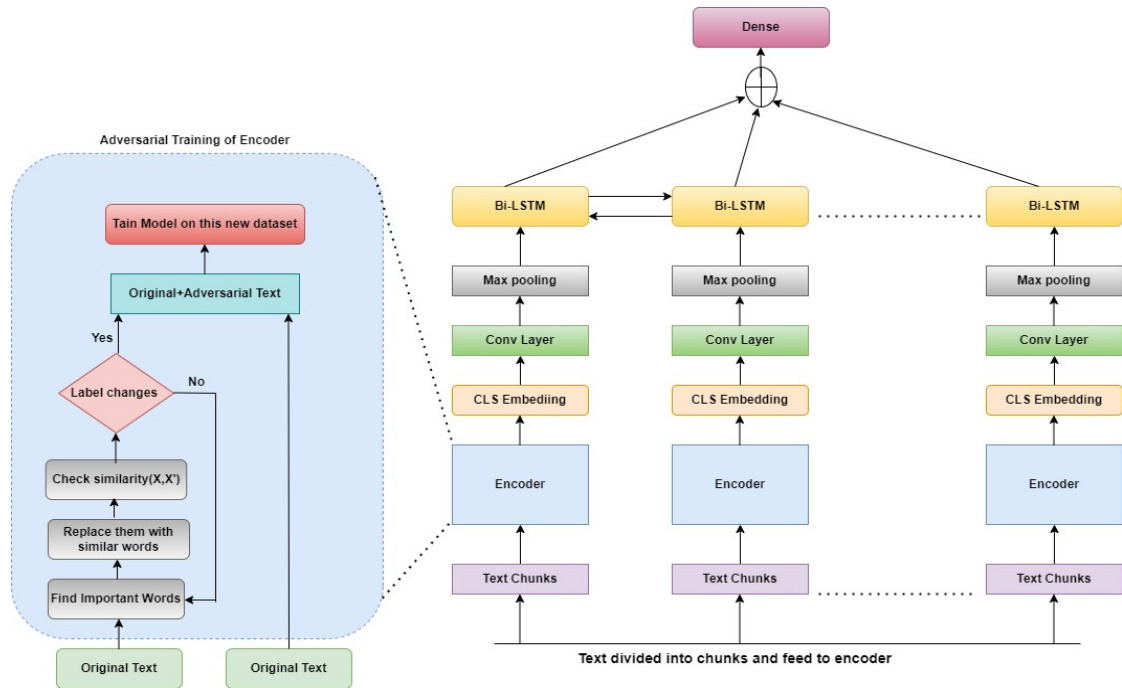


Figure 2: Robust Neural Legal Judgement Model (H-BERT model). Here training procedure of H-BERT is shown on the left side.

of 100 units and 1-D CNN with 32 filters for H-BERT.

5.3. Results

5.3.1. Results after fine tuning

We have fine-tuned models naturally, i.e., without any augmentation (shown in Table 1). For BERT, Legal-BERT, and RoBERTa, we have taken the last 512 tokens of each sample as input, and for H-BERT, we have divided the text into chunks, as mentioned in section 5.1. From empirical results, we can say that H-BERT performs better than other models as H-BERT takes whole text examples, whereas other models take only the last 512 tokens. Legal-BERT performs better on ECHR and SCOTUS datasets as it is pre-trained on legal documents of Europe and America. The performance of RoBERTa on the ILDC dataset is better than Legal-BERT because Legal-BERT is not pre-trained on the Indian-origin legal dataset, whereas RoBERTa is pre-trained on general English datasets.

5.3.2. Results after adversarial attack on naturally trained models

We feed 1000 adversarial examples generated from the adversarial examples generator to naturally trained models to check their robustness against adversarial attacks. As shown

Table 1Accuracy of Naturally trained models, (*FT*) : Fine Tuning

| Models | ECHR | SCOTUS | ILDC _{multi} |
|-------------------------|-------|--------|-----------------------|
| BERT(<i>FT</i>) | 81.21 | 68.33 | 67.24 |
| Legal-BERT(<i>FT</i>) | 83.42 | 76.47 | 63.37 |
| RoBERTa(<i>FT</i>) | 79.27 | 71.69 | 71.26 |
| H-BERT(<i>FT</i>) | 81.03 | 78.02 | 74.89 |

Table 2Accuracy of Naturally trained Models after attack, (*FT*) : Fine Tuning

| Models | ECHR | SCOTUS | ILDC _{multi} |
|-------------------------|-------|--------|-----------------------|
| BERT(<i>FT</i>) | 33.12 | 36.42 | 22.59 |
| Legal-BERT(<i>FT</i>) | 36.27 | 41.67 | 25.26 |
| RoBERTa(<i>FT</i>) | 36.05 | 41.91 | 38.92 |
| H-BERT(<i>FT</i>) | 39.18 | 43.19 | 37.21 |

Table 3Accuracy after adversarial attack , (*DA*) : Data Augmentation

| Models | ECHR | SCOTUS | ILDC _{multi} |
|---|--------------|--------------|-----------------------|
| BERT(<i>DA</i>) (Ours) | 38.03 | 41.12 | 32.56 |
| Legal-BERT(<i>DA</i>) (Ours) | 39.36 | 43.15 | 41.66 |
| RoBERTa(<i>DA</i>) (Ours) | 40.21 | 45.09 | 38.71 |
| H-BERT(<i>DA</i>) (Ours) | 46.10 | 45.02 | 42.03 |

in Table 2, naturally trained models could not handle adversarial attacks as their accuracy decreased drastically. The accuracy of BERT decreased the most because it is not pre-trained on domain-specific (legal domain) datasets, whereas, in the case of H-BERT, accuracy decreased least because H-BERT’s RoBERTa is pre-trained on general English datasets as well as during training, it is considering whole legal text documents. In contrast, other models consider only the last 512 words of each example. Legal-BERT is more robust than BERT as it is pre-trained on legal datasets. RoBERTa is pre-trained on a large corpus, so it can able to handle adversarial attacks better than Legal-BERT.

5.3.3. Results after adversarial attack on model trained using data-augmentation

We feed 1000 adversarial examples to a model trained using data augmentation to check their robustness. As shown in Table 3, the accuracy of models is less than that of naturally trained models but better than the accuracy of models after the adversarial attack on naturally trained models. This is because we are augmenting extra data, which is very similar to the original data except for a few words for training. So due to this, the model is more diverse and can handle some adversarial attacks. In most cases, H-BERT performs better than others because it considers whole text data instead of the last 512 tokens.

Original: ...He companyld number possibly have **failed** to tell Gaud that the two persons ...

Adversarial: ...He companyld number possibly have **faulted** to tell Gaud that the two persons....

Original: ...Therefore the **statement** of the appellant that accused...

Adversarial: ...Therefore the **statements** of the appellant that accused No ...

Figure 3: original and adversarial examples of ILDC dataset while training BERT.

Table 4

Accuracy after adversarial training, (*AT*) : Adversarial Training

| Models | ECHR | SCOTUS | ILDC_{multi} |
|---|--------------|---------------|-----------------------------|
| BERT(<i>AT</i>) (Ours) | 79.23 | 69.07 | 65.56 |
| Legal-BERT(<i>AT</i>) (Ours) | 82.01 | 77.02 | 61.02 |
| RoBERTa(<i>AT</i>) (Ours) | 81.73 | 70.03 | 69.97 |
| H-BERT(<i>AT</i>) (Ours) | 83.67 | 78.09 | 71.53 |

5.3.4. Results after adversarial training

We implemented adversarial training using our Algorithm 2. As shown in Table 4, sometimes, the accuracy of an adversarially trained model is better than the naturally trained model. The increase in accuracy is due to the augmentation of adversarial examples, which creates more diversity during training. The performance of the H-BERT model is best, while Legal-BERT is performing better on ECHR and the SCOTUS dataset because it is pretrained on European and American legal documents. Figure 3 shows an adversarial example on the ILDC dataset during adversarial training. As we can see, slight change perturbation in the text can change the label of an input. Due to the large length of text input, we have shown only a small snippet of an example where an example is being perturbed.

5.3.5. Results after adversarial attack on adversarially trained model

We feed 1000 adversarial examples, as earlier, to check the robustness of the adversarially trained model. The results are surprising, as shown in Table 5. Our models can handle most adversarial attacks. Accuracy is far better than accuracy after the attack on naturally trained models. This is because, during adversarial training, the model came across a diverse set of words that were not present earlier.

As shown in Table 5, H-BERT is performing better than other models because it is trained on the whole dataset. The BERT model performs worst as it is not pre-trained on legal documents. The performance of Legal-BERT is not satisfactory on ILDC because it is pre-trained on European and American legal documents, which may contain words that are different from Indian legal documents.

Table 5Accuracy after attack , (*AT*) : Adversarial Training

| Models | ECHR | SCOTUS | ILDC _{multi} |
|--------------------------------|--------------|--------------|-----------------------|
| BERT(<i>AT</i>) (Ours) | 58.96 | 52.38 | 54.46 |
| Legal-BERT(<i>AT</i>) (Ours) | 64.07 | 52.71 | 51.96 |
| RoBERTa(<i>AT</i>) (Ours) | 64.97 | 50.09 | 55.91 |
| H-BERT(<i>AT</i>) (Ours) | 69.32 | 61.53 | 58.29 |

6. Conclusion and Future work

In this work, we empirically proved that early existing legal models are not adversarially robust, which is a significant risk for deploying them in work. We also presented an adversarially robust model, which is trained on our adversarial training algorithm for legal judgment prediction, which performs better than state-of-the-art models in the presence of adversarial examples. For future work, we suggest making robust legal models which can be applied to Legal documents that are different from English. Also, one can work on zero-shot and few-shot learning in legal domains, where very few resources are available for legal documents.

References

- [1] C. Xiao, H. Zhong, Z. Guo, C. Tu, Z. Liu, M. Sun, Y. Feng, X. Han, Z. Hu, H. Wang, J. Xu, CAIL2018: A large-scale legal dataset for judgment prediction, CoRR abs/1807.02478 (2018). URL: <http://arxiv.org/abs/1807.02478>. arXiv:1807.02478.
- [2] I. Chalkidis, I. Androutsopoulos, N. Aletras, Neural legal judgment prediction in english, CoRR abs/1906.02059 (2019). URL: <http://arxiv.org/abs/1906.02059>. arXiv:1906.02059.
- [3] V. D. Tran, M. L. Nguyen, K. Satoh, Building legal case retrieval systems with lexical matching and summarization using A pre-trained phrase scoring model, CoRR abs/2009.14083 (2020). URL: <https://arxiv.org/abs/2009.14083>. arXiv:2009.14083.
- [4] Brazil: The land of many lawyers and very slow justice (????). URL: <https://www.cpr.org/2014/11/05/brazil-the-land-of-many-lawyers-and-very-slow-justice/>.
- [5] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: <http://arxiv.org/abs/1810.04805>. arXiv:1810.04805.
- [6] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. URL: <https://aclanthology.org/2020.findings-emnlp.261>. doi:10.18653/v1/2020.findings-emnlp.261.
- [7] I. Chalkidis, M. Fergadiotis, D. Tsarapatsanis, N. Aletras, I. Androutsopoulos, P. Malakasiotis, Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases, CoRR abs/2103.13084 (2021). URL: <https://arxiv.org/abs/2103.13084>. arXiv:2103.13084.
- [8] I. Chalkidis, A. Jana, D. Hartung, M. J. B. II, I. Androutsopoulos, D. M. Katz, N. Ale-

- tras, Lexglue: A benchmark dataset for legal language understanding in english, CoRR abs/2110.00976 (2021). URL: <https://arxiv.org/abs/2110.00976>. arXiv:2110.00976.
- [9] P. Jackson, K. Al-Kofahi, A. Tyrrell, A. Vachher, Information extraction from case law and retrieval of prior cases, *Artificial Intelligence* 150 (2003) 239–290. URL: <https://www.sciencedirect.com/science/article/pii/S0004370203001061>. doi:[https://doi.org/10.1016/S0004-3702\(03\)00106-1](https://doi.org/10.1016/S0004-3702(03)00106-1), *ai and Law*.
- [10] H. Zhong, Y. Wang, C. Tu, T. Zhang, Z. Liu, M. Sun, Iteratively questioning and answering for interpretable legal judgment prediction, *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (2020) 1250–1257. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/5479>. doi:10.1609/aaai.v34i01.5479.
- [11] V. Malik, R. Sanjay, S. K. Nigam, K. Ghosh, S. K. Guha, A. Bhattacharya, A. Modi, ILDC for CJPE: indian legal documents corpus for court judgment prediction and explanation, CoRR abs/2105.13562 (2021). URL: <https://arxiv.org/abs/2105.13562>. arXiv:2105.13562.
- [12] J. Niklaus, I. Chalkidis, M. Stürmer, Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark, CoRR abs/2110.00806 (2021). URL: <https://arxiv.org/abs/2110.00806>. arXiv:2110.00806.
- [13] C.-N. Chau, T.-S. Nguyen, L.-M. Nguyen, Vnlawbert: A vietnamese legal answer selection approach using bert language model, in: 2020 7th NAFOSTED Conference on Information and Computer Science (NICS), 2020, pp. 298–301. doi:10.1109/NICS51282.2020.9335906.
- [14] L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, BERT-ATTACK: Adversarial attack against BERT using BERT, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, 2020, pp. 6193–6202. URL: <https://aclanthology.org/2020.emnlp-main.500>. doi:10.18653/v1/2020.emnlp-main.500.
- [15] S. Garg, G. Ramakrishnan, BAE: bert-based adversarial examples for text classification, CoRR abs/2004.01970 (2020). URL: <https://arxiv.org/abs/2004.01970>. arXiv:2004.01970.
- [16] J. Y. Yoo, Y. Qi, Towards improving adversarial training of NLP models, CoRR abs/2109.00544 (2021). URL: <https://arxiv.org/abs/2109.00544>. arXiv:2109.00544.
- [17] D. Jin, Z. Jin, J. T. Zhou, P. Szolovits, Is BERT really robust? natural language attack on text classification and entailment, CoRR abs/1907.11932 (2019). URL: <http://arxiv.org/abs/1907.11932>. arXiv:1907.11932.
- [18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: <http://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
- [19] Q. Xie, Z. Dai, E. H. Hovy, M. Luong, Q. V. Le, Unsupervised data augmentation, CoRR abs/1904.12848 (2019). URL: <http://arxiv.org/abs/1904.12848>. arXiv:1904.12848.
- [20] Hugging-face models (????). URL: <https://huggingface.co/models>.
- [21] J. Y. Yoo, J. X. Morris, E. Lifland, Y. Qi, Searching for a search method: Benchmarking search algorithms for generating NLP adversarial examples, CoRR abs/2009.06368 (2020). URL: <https://arxiv.org/abs/2009.06368>. arXiv:2009.06368.
- [22] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, R. Kurzweil, Universal sentence encoder, CoRR abs/1803.11175 (2018). URL: <http://arxiv.org/abs/1803.11175>. arXiv:1803.11175.