Interpretable Low-Resource Legal Decision Making

Rohan Bhambhoria[1,2,3]**, Hui Liu** [1,2]**, Samuel Dahan** [1,3,4]**, Xiaodan Zhu** [1,2,3]

[1] Ingenuity Labs & [2] ECE , Queen's University
[3] Conflict Analytics Lab, Queen's University
[4] Cornell Law School
{r.bhambhoria, hui.liu, samuel.dahan, xiaodan.zhu}@queensu.ca

Abstract

Over the past several years, legal applications of deep learning have been on the rise. However, as with other high-stakes decision making areas, the requirement for interpretability is of crucial importance. Current models utilized by legal practitioners are more of the conventional machine learning type, wherein they are inherently interpretable, yet unable to harness the performance capabilities of data-driven deep learning models. In this work, we utilize deep learning models in the area of trademark law to shed light on the issue of likelihood of confusion between trademarks. Specifically, we introduce a model-agnostic interpretable intermediate layer, a technique which proves to be effective for legal documents. Furthermore, we utilize weakly supervised learning by means of a curriculum learning strategy, effectively demonstrating the improved performance of a deep learning model. This is in contrast to the conventional models which are only able to utilize the limited number of expensive manually-annotated samples by legal experts. Although the methods presented in this work tackles the task of risk of confusion for trademarks, it is straightforward to extend them to other fields of law, or more generally, to other similar high-stakes application scenarios. Our code is available here[1].
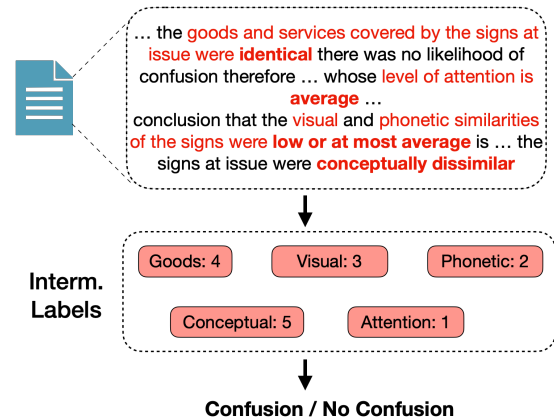
Figure 1: For a given case law document, important factors involved are—the similarity of goods & services, phonetic similarity, conceptual similarity, visual similarity, and the level of attention of the relevant public. These factors are sufficient to decide the outcome of a case, and we denote them as "intermediate labels" in this work.

Introduction

In many legal tasks, a thorough review of case law is required to understand the underlying elements involved in the legal decision-making process. The use-case of an automatic system capable of addressing the aforementioned varies, and ranges from promoting access to justice for self-represented litigants to determining precedents which may act favourably for legal professionals (Lawlor 1963).

In this work, for purposes of exploring a real-world scenario, we utilize cases from the likelihood of confusion for trademarks, a branch of intellectual property law. Likelihood of confusion exists between trademarks when the marks are so similar and the goods and/or services for which they are used are so related that consumers would mistakenly believe they come from the same source (Bone 2012). Accordingly, this task involves analyzing the goods and services provided by each company in order to identify the overlap. In addition, the marks of the companies are analyzed from several different perspectives—how they differ visually, phonetically, and conceptually to influence the judgment as shown in Figure 1. Finally, other less important factors such as the level of attention to the relevant public are considered. The case law hence provides a detailed discussion along with reasons and conclusions which favor the judgment process (Dahan 2020).

Assessing the likelihood of confusion for trademark is a crucial component of a company's existence, especially for IP-intensive small to medium sized companies (SMEs) that need to protect their creation. Firms with registered intellectual property (IP) rights perform better than those without, yet few SMEs in the EU actually exercise their IP rights: only 9% of SMEs have registered IP rights, compared to 36% of large companies. However, research suggests that SMEs generally do not use the IP system because they do not see the benefit of it, they lack the necessary expertise, and/or they find the procedure too costly (Dahan 2020). As a result, many of these firms often incur unforeseen costs as-

[1]https://github.com/RohanVB/Interpretable-Low-Resource-Legal-Decision-Making

sociated with litigation to defend their mark, either before or after it is registered.

Considering that innovative SMEs are a primary driver of growth, it is worth exploring new ways to democratize IP legal help. As a result, research in this space aims to contribute to this endeavour, and recent work (Dahan 2020; Dahan et al. 2022) discusses the potential implications of this field of research in a recent expert opinion for the European Union Intellectual Property Office, the higher regulatory authority in Europe for IP regulation. In particular, in consultation with regulators and practitioners, they look at a new generation of direct-to-public (DTP) AI tools that can improve access to justice by providing the minimal level of legal help necessary for meeting basic legal needs. The authors review how a combination of online dispute resolution (ODR) and predictive analytics employing machine learning and deep learning can positively affect EU trademark dispute resolution. The system operates as a legal aid tool for small to medium-sized enterprises (SMEs) to determine potential IP rights violations.

Towards meeting the requirements set by experts in the legal field, strong predictions are required to be made from case law. Generally, the computation of high-stakes legal problems requires expensive annotation efforts by domain-knowledge experts. As a result, the cost and complexity of annotating this data makes it difficult to obtain a high number of intermediate labels—for instance, visual or verbal similarity in the case of trademark confusion. However, final judgments are public and thus easy to obtain, and end-to-end text based encoders serve well in making these predictions, which may nonetheless be due to spurious correlations (Bhambhoria, Dahan, and Zhu 2021). One of the main objectives of this work is to predict intermediate outcomes. For instance whether trademarks are visually similar—insofar as it constitutes key features for interpretability. The other main objective is to utilize these intermediate outcomes towards drawing final conclusions for the right reasons i.e. not due to spurious correlations.

Limited historical case law data reflects the reality across many branches of law, specifically due to the fact that different jurisdictions may have variations in regulations (Reese 1978). Mosbach, Andriushchenko, and Klakow (2021) provide a study on fine-tuning models for better generalization on small datasets. In order to establish stronger baselines, we explore and leverage various methods commonly used with deep learning models in order to account for the small amount of annotated data available. Multi-task learning (Kendall, Gal, and Cipolla 2018), albeit beneficial for several tasks which may improve from parameter sharing of neural networks does not benefit our intermediate-label prediction task or final outcome due to the gap between the relations for the two tasks (Bingel and Søgaard 2017; Liu, Yin, and Wang 2018). Modern deep neural networks outperform conventional methods when trained on a large amount of data. However, in many high-stakes decision making domains, such as law, leveraging a large amount of data has shown limited performance gains (Chalkidis et al. 2020; Bhambhoria, Dahan, and Zhu 2021). In fact, models tend to deliver reliable and strong results when trained on high-cost,

structured, small amounts of sparsely annotated data (Ustun and Rudin 2015; Rudin 2019). In order to maximize the capabilities of deep learning models, we must use methods such as transfer learning (Yosinski et al. 2014), and perform some careful fine-tuning due to a sensitive hyperparameter optimization process. Weak or distant supervision of models is another method by which we can introduce supervision from a large number of noisily labelled data (Go, Bhayani, and Huang 2009) obtained from heuristics and rules. In this work, we create an augmented dataset which is utilized to provide weak supervision for our models. Unfortunately, naively introducing these noisy samples leads to poor performance. For this reason, we explore adopting strategies such as curriculum learning (Bengio et al. 2009; Wang, Chen, and Zhu 2020) to achieve performance gains, while maintaining fidelity to the original model.

The contributions of this work are three-fold:

- We are the first to tackle the challenge of risk of confusion for trademark by predicting interpretable intermediate scores which can be utilized to improve the final outcome, i.e. whether there is an infringement of IP law.

- We empirically show the effects of utilizing curriculum learning with weakly supervised learning to ensure generalizability, robustness, and stability of our model which can be used for other similar legal decision-making tasks.

- We release the first trademark case law dataset for tackling the challenge of assessing risk of confusion.

Related Work

Recent research has explored blanket terms of explainability and interpretability in great detail. Some works use these terms interchangeably (Jain and Wallace 2019; Wiegreffe and Pinter 2019), whilst others highlight their intricate differences (Gilpin et al. 2018). In this work, we opt to use the definitions of "interpretability" as per Gilpin et al. (2018) to highlight the fact that our model as a function is interpretable, and the explanations produced have high fidelity aswell as transparency. Previous work dealt with interpretability by means of introducing post-hoc methods which are model agnostic and may be reliant on perturbations (Ribeiro, Singh, and Guestrin 2016, 2018) or grounded deeper in axioms (Sundararajan, Taly, and Yan 2017; Srinivas and Fleuret 2021). Other methods explored interpretability on various degrees, i.e. on the level of individual neurons (Dhamdhere, Sundararajan, and Yan 2018; Bau et al. 2020), or based on the role of layers (Binder et al. 2016). Another method of providing interpretability is by means of counterfactuals (Mothilal, Sharma, and Tan 2020). As these methods involve providing explanations for black boxes, they have been met with criticism (Rudin 2019), in favor of inherently interpretable models (Rudin et al. 2021).

Several papers focus on interpretability, with a narrower focus on the task of legal judgment prediction trained on large annotated datasets (Ye et al. 2018; Jiang et al. 2018; Zhong et al. 2020). Unlike the present study, past work did not rely on costly and time-consuming annotations by subject-matter experts. For these reasons, these studies do
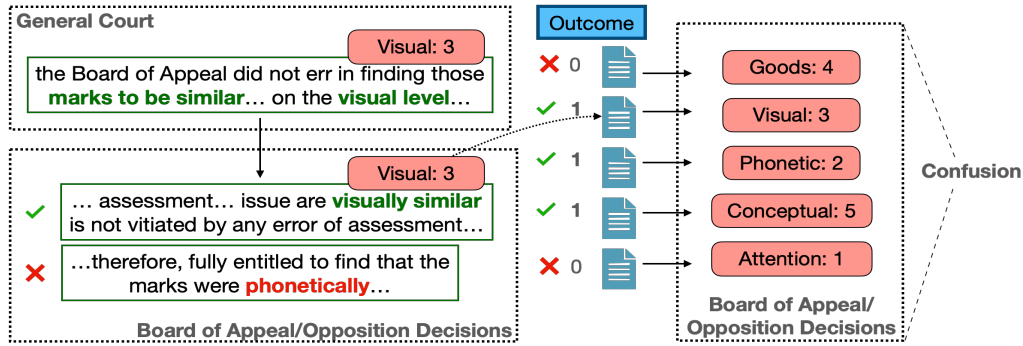
Figure 2: Augmented dataset labels (Board of Appeal/Opposition Decisions) are assigned the intermediate labels from the clean dataset (General Court) which was used for extracting the relevant sentences based on cosine similarity. The outcome of the final judgment (shown as 0 for "No Confusion" and 1 for "Confusion") is assigned by taking the majority of the publicly available labels from the board of appeal/opposition decision cases.

Subset	Num. Cases	Avg. Char/S
Clean	525	1081
Augmented	2852	941

Table 1: Clean subset represents cases annotated by lawyers, obtained from the CJEU. Augmented subset represents cases which are extracted from the Board of Appeal/Opposition Decisions. S denotes the input text comprising of factors which influence the likelihood of confusion.

not adequately reflect the real-world scenario for the application of deep learning towards high-stakes decision making. Phang, Févry, and Bowman (2019) utilized intermediate predictions in order to predict outcomes, although the main motive is not towards providing interpretability, and again, these intermediate predictions do not reflect real-world challenges. Moreover, intermediate labels in the aforementioned work are abundantly available in contrast to the main task.

Wei and Zou (2019) explored simple methods to augment data for text classification tasks. However, in high-stakes decision making we must take into consideration data augmentation methods which enables generalization of our model, especially in the legal domain wherein data from different levels of courts may be available. Naively introducing these noisy augmented samples will lead to a reduction in performance. (Wang, Chen, and Zhu 2020) provides a detailed overview of various curriculum learning strategies through which augmented samples can be introduced. In this work we opt for a modified self-paced learning strategy. Wu, Dyer, and Neyshabur (2021); Zhou, Wang, and Bilmes (2021) detailed the scenarios in which noisy samples of data provide benefits towards the training process. In this work, we utilize the traditional method of curriculum learning which introduces noisy samples from easy-to-hard for training.

Dataset

The datasets we created for the purpose of this study are comprised of two parts—firstly, cases obtained from the General Court, a constituent of the Court of Justice of the European Union (CJEU)[2]. The CJEU is the highest judicial authority in the field of trademark in the European Union. The size of the data sample is significant from a legal standpoint insofar as we were able to extract a large pool of decisions related to trademark confusion. As a result, the CJEU body of case law constitutes a meaningful and relevant source of training data for our models. The cases obtained from the CJEU were annotated by a team of lawyers with a training in trademark law. An annotation guideline was designed by legal professionals, taking into consideration the various factors which are involved in deciding the outcome of a case. Despite the guideline comprising of 32 fields, a condensed set of 5 factors were deemed important towards analyzing the risk of confusion for the likelihood of trademarks, shown in Table 5 of the Appendix. These factors are "Similarity of Goods & Services", "Visual Similarity", "Conceptual Similarity", "Phonetic Similarity", and "Level of Attention". Few cases are more nuanced and require "split decisions" to be made for annotations. We exclude the consideration of these outlier cases in this work. Due to the presence of missing values in our annotated cases under the selected important features, we have a set of 426 cases, from which 213 samples are reserved for training, 53 for validation, and 160 for testing. All cases are extracted from 2008 onwards, as major changes[3] in trademark law have been made since. The resulting cases are hence comprised of 525 annotated documents in its entirety. Secondly, we weakly labelled an additional 2852 cases from a larger pool of "Opposition decisions", and "Board of Appeal decisions"[4]. The average length of these document are comparable, as shown in Table 1. These additional cases are similar in terms of semantics to our expert-annotated cases, with the exception that from a legal perspective, they were decided by a lower authority and thus do not bear the same legal value as the cases of the CJEU. We refer to these cases as "aug-

---

[2]https://curia.europa.eu/

[3]For the sake of consistency, we have annotated cases decided after Regulation (EC) No. 207/2009, which brought forward various changes in trademark law.

[4]https://euipo.europa.eu/eSearchCLW/

mented data", the entirety of which is utilized for training under different settings. The detailed construction process of these datasets will be elaborated below.

Obtaining Expert-Annotations

Factors that are relevant to determine the existence of confusion among trademarks were drawn from the Regulation (EC) No 207/2009), the foundational legislation in the field of trademark. Lawyers were trained on identifying these factors. Sentences were extracted corresponding to these important features, and were labelled on a predefined scale. In this work, the resulting dataset is regarded as a "clean dataset". Next, a control-group study was created, wherein the independent variable was set to be the final outcome i.e. confusion or no confusion. Factors influencing the judgment were presented to the control group consisting of lawyers. The participants were asked to determine the outcome when presented with only the labelled factors. A human-annotated Macro F1 score of $\approx 86\%$ was achieved in determining confusion for this task. The factors responsible for the inability in obtaining perfect assessment are as follows—1) minor concluding statements by courts may shift the judgment, 2) courts may have paid more attention to specific features in a few cases, 3) human annotator error, 4) case contained varying claims, 5) split-decisions in cases. Participants involved in this control group were also asked to identify rules involved in assessing the likelihood of confusion. These rules were utilized to establish the labels for the cases from "Opposition Decisions", and "Board of Appeal Decisions", and are detailed in Table 6 of the Appendix.

Obtaining Augmented Cases

The capability of a simple process to obtain augmentation for this task is possible due to the repetitiveness in trademark case law. We first retrieve ≈ 52000 cases from the "Opposition Decisions", and "Board of Appeal Decisions" dating from 1998 to 2020. This serves as a large pool of initial cases, from which we follow the extraction, randomization, filtering, and assigning steps, shown in Figure 2 detailed as follows:

1. **Extraction:** We utilize cases from our annotated clean training dataset and create TF-IDF matrices for all sentences. We also repeat this process for the lower court (augmented, pre-filtered) cases, of which we have a larger pool of cases. Then, we calculate cosine similarity and extract the top sentences corresponding to the sentences obtained from our training set.

2. **Randomization**: As extracted sentences may correspond to one of five features (Similarity of Goods & Services, Visual Similarity, Conceptual Similarity, Phonetic Similarity, or Level of Attention), we randomly concatenate sentences corresponding to these features to create concluding paragraphs, which albeit noisy, contain a large number of samples.

3. **Filtering:** By utilizing a few basic keywords which are known to be present in sentences attributing to confusion (e.g. 'conceptually' for 'Conceptual Similarity'), we obtain basic filtered textual inputs. We then utilize the rules
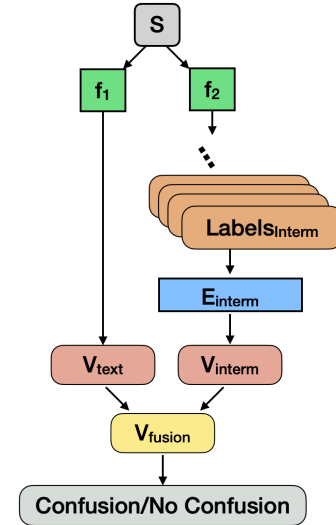

Figure 3: Our proposed model which takes into consideration intermediate labels for interpretability, and addresses the challenge of decision-making due to spurious correlations. S denotes the input excerpts extracted from case law. f_1 and f_2 are initialized by fine-tuning on independent loss functions. V_{interm} is produced by first calculating intermediate labels, after which an encoder denoted by E_{interm} enables us to fuse it with the representation obtained from f_1. V_{text} and V_{interm} are fused to create V_{fusion}.

devised by the control group study to further filter our augmented samples.

4. **Assigning:** We utilize the annotated (clean) samples to assign pseudo-labels to these augmented samples by maintaining the original labels from which they were extracted for intermediate labels. We also make use of the final outcome to assign a label for the final judgment of the case by consideration of the majority number of samples attributing to a label of "Confusion" or "No Confusion". For example, if a majority of the cases from which the statements are derived consist of the final judgment of confusion, the pseudo-label for the document will also be confusion. In this work, the resulting cases obtained from this stage will be referred to as our "augmented dataset".

Methodology

The motivation for designing the model in this work is to introduce interpretability in a model-agnostic setting which can be utilized for legal decision-making tasks. As shown in Figure 3, this is done by introducing an abstracted layer derived from text, which we call the components of as "intermediate labels". By predicting these labels, we are able to induce interpretability in our model. Furthermore, by utilizing these intermediate labels to make judgments of the final outcome, we enable the model to make decisions for the right reasons. In other words, we tackle the challenge of spurious correlations wherein decisions may be made correctly due to incorrect reasons by introducing these intermediate labels

for producing the outcome. As our "clean dataset" contains only a few samples, we utilize our "augmented dataset" to promote generalization of our model through a curriculum learning strategy which is detailed in Figure 4.

Baselines

End-to-End Transformer-based models such as BERT (Devlin et al. 2019) have shown strong performance on various text classification tasks. We utilize RoBERTa (Liu et al. 2019) for the task of predicting the likelihood of confusion for trademarks. This model is tested on the various splits of training data—1) clean data, 2) augmented data, and 3) mix data which is simply a combination of the shuffled clean, and augmented data. The experiments under these settings are performed separately for both, likelihood of confusion classification task, and for intermediate-label prediction regression task as detailed in Figure 6 of the Appendix. As a consequence of this disjoint training, the two tasks are unable to benefit each other, and hence decisions for final labels may be made due to spurious correlations.

Multi-Task As our task is targeted to tackle the optimization of intermediate labels, and the prediction of likelihood of confusion, it is reasonable to consider a multi-task structure wherein parameters are shared across two identical encoders as shown in Figure 7 of the Appendix. As with our end-to-end model, we use RoBERTa (Liu et al. 2019) for the multi-task setting in order to ensure consistency across our results. However, we still face the challenge of spurious correlations due to the inability of this method to build a connection between the two tasks.

Fusion Model

To overcome the aforementioned weaknesses in the baseline methods, wherein decisions may be made due to spurious correlations, and to address the need for interpretability, we propose a fusion model shown in Figure 3.

For a given set, $S = \{s_1, s_2, ..., s_n\}$ $\forall s_i \in S'$ we have S' which corresponds to factors in the form of text influencing the likelihood of confusion either extracted manually in the case of the clean dataset, or automatically obained in the case of the augmented dataset.

We utilize an encoder, f_1 which is initialized from the best performing checkpoint of the end-to-end model for the binary prediction task to determine the likelihood of confusion in a case. This is done by utilizing cross entropy loss shown in Equation 1 wherein m are the number of samples, y_i is the target label, and $\hat{y}_i$ is the output obtained from $f_1(S)$.

$$L = -\frac{1}{m} \sum_{i=1}^{m} y_i \cdot \log(\hat{y}_i) \qquad (1)$$

For the intermediate task, encoder f_2 is initialized by a similar method after being fine-tuned by optimizing Equation 2, for the value of variables detailed in Equation 3. We use Smooth L1 Loss here as it accounts for outliers, and leads to better performance in lieu of utilizing mean absolute error, or mean standard error loss. The unreduced loss is given by:

$$L = \{l_1, \ldots, l_I\}^T, \qquad (2)$$

where l_i can be formatted as:

$$l_i = \begin{cases} 0.5 \left(\hat{y}_i - y_i\right)^2 / \beta, & \text{if } |\hat{y}_i - y_i| < \beta \\ |\hat{y}_i - y_i| - 0.5 * \beta, & \text{otherwise} \end{cases} \qquad (3)$$

where $\hat{y}_i$ and y_i denote the output of $f_2(S)$ and the target respectively. β is a hyperparameter, wherein if $\beta \to 0$, smooth L1 Loss will converge to L1 Loss, and if $\beta \to +\infty$, will converge to a constant 0 loss. We set β to 1, to account for outliers.

A feedforward layer which applies a linear transformation is utilized to obtain the label values, after which we utilize an embedding layer. A transformer block made up of self-attention as shown in Equation 4 and a feedforward network is then utilized as an encoder. A final feedforward connection is used to obtain the representation for the intermediate labels. We do not directly consider the hidden representation of f_2 here due to the fact that we require our model to produce the intermediate labels, a property which is not obtainable otherwise.

$$Attention = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \qquad (4)$$

where Q is a query matrix, i.e. embedding of s_i. K are the keys, i.e. the embedding of all words in S. V are values, and $\sqrt{d_k}$ is the dimension of K, by which scaling is performed as per Vaswani et al. (2017).

The fusion layer is a modified version of Cai et al. (2019) and is detailed in Equation 5.

$$V_{Fusion} = s1 * V_{text} + s2 * V_{interm} \qquad (5)$$

where $s1$ and $s2$ are scores calculated by using hyperbolic tangent function denoted in Equation 6 after a feedforward layer stemming from the concatenated vector representation of V_{text} with V_{interm}, and conversely V_{interm} with V_{text} respectively:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \qquad (6)$$

where x can be set to the concatenated vector representations to obtain $s1$, and $s2$ respectively.

Curriculum learning

As our augmented dataset contains several noisy samples, it is natural to consider a structured method of introducing samples in order of easy-to-difficult for training our model. We adopt a curriculum learning strategy to be utilized in introduction of samples wherein the base classifier is utilized to set a threshold for difficulty on the basis of softmax probabilities as detailed in Figure 4. The model, M_{init} is trained on only clean data. M_{base} is an untrained encoder with the same architecture as M_{init}. By dividing the augmented training dataset into bins, $B = \{b_1, b_2, ..., b_n\}$, we select the bins containing confident samples to be introduced as training samples with clean data for M_{base}. The updated M'_{base} is further trained on only clean data and serves as our model for the curriculum learning strategy.
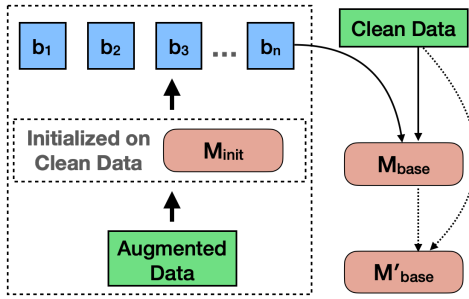
Figure 4: Curriculum learning strategy by which augmented data is utilized to improve predictive capabilities of a model.

Experimental Setup

In all experiments, the f_1 and f_2 architectures utilized are RoBERTa-base models. In our model, the embedding layer in E_{interm} has a dimension of 300, and is arbitrarily set. The transformer block in E_{interm} for the intermediate labels consists of 6 layers with 6 heads, and a feedforward dimension of 512. In our experiments, we utilize a batch size of 8 for all experiments on a single RTX 2080Ti machine with 11 GB of memory, taking into consideration the limited computational resources practioners in the legal field may have. For curriculum learning, we arbitrarily set the number of bins to be 3 wherein the first bin contains the most confident samples with a probability threshold of ≥ 0.99, the second bin contains the remaining correct samples, and the final bin contains all samples from the mixed dataset for training which were not contained within the previous bins. For the calibration experiments, we arbitrarily set the number of bins to be 5. PyTorch-based libraries are utilized for all experiments.

Results and Discussion

Our objective with the experiments is to demonstrate with strong empirical evidence that our model is capable of producing stable, reliable results with excerpts obtained from case law. The results are shown on our carefully curated dataset for the likelihood of confusion for trademarks in Table 2. However, the findings in this work can easily be utilized for other similar legal decision-making tasks.

End-to-end models may suffer from issues such as making decisions due to spurious correlation, as they are in no way constrained by the intermediate labels. On the other hand, the multi-task baseline, and our model is capable of utilizing the intermediate labels to make decisions. As shown in Table 2, there is no benefit in utilizing the end-to-end model as a predictive solution for this task. Firstly, the performance is much worse than our model, and the multi-task model for the clean dataset. For the mixed dataset, it is unable to benefit from training on the larger number of noisy samples. Secondly, the standard deviation between different runs is very high for random starts of the model for training. The usecase of the end-to-end model is only as initializing M_{base} for the curriculum learning strategy.

The multi-task model, despite incorporating the intermediate labels during training, is unable to take advantage of

Model	Data Split	
	Clean	**Mix**
Multi-Task	75.52 ± 6.13	60.51 ± 7.34
End-to-End	75.22 ± 6.89	62.47 ± 7.47
Ours	$\mathbf{79.16} \pm 2.34$	65.60 ± 3.01
Ours + curric.	-	$\mathbf{83.05} \pm 2.54$

Table 2: Macro F1 scores reported for the binary prediction task of likelihood of confusion ($\pm$ standard deviation). Multi-task denotes fine-tuned jointly-trained roberta-base encoders. End-to-End denotes a fine-tuned roberta-base model. Results shown are averaged across 5 runs. The data split "Mix" is the randomly shuffled concatenation of the clean and augmented datasets. "Curric." is an abbreviation for our curriculum learning straegy shown in Figure 4.

Model	MAE	MSE
Random	1.81 ± 0.03	5.09 ± 0.12
Clean		
Multi-Task	0.75 ± 0.03	1.05 ± 0.06
Ours	$\mathbf{0.75} \pm 0.03$	$\mathbf{0.99} \pm 0.07$
Mix		
Multi-Task	0.84 ± 0.04	1.27 ± 0.13
Ours	0.83 ± 0.05	1.24 ± 0.13

Table 3: Mean Absolute Error (MAE) and Mean Standard Error(MSE) reported for regression task of predicting intermediate labels (lower value is better). The loss function utilized for training is Smooth L1 Loss which is more stable for training on lesser number of samples.

these labels in improving the performance of likelihood of confusion. For the clean dataset, the performance is comparable to that of the end-to-end model. On the other hand, for the mixed dataset, the performance is slightly worse than the end-to-end model. Results from both data splits show high standard deviation, and hence unstable results between runs. However, as this model utilizes a shared encoder wherein the weights are updated from both, intermediate, and final outcome prediction tasks, it is a step in the right direction towards making decisions for the correct reasons, and not due to spurious correlations.

Our model is able to incorporate the intermediate labels, much like the multi-task model, and hence also tackles the challenge of decisions being made due to spurious correlations. However, the major benefit is seen on both splits of data, wherein the standard deviation is drastically reduced between runs, hence delivering stable results. Furthermore, the performance shows great improvement under the clean dataset data split when compared with the end-to-end, and multi-task models. Nonetheless, the benefits of utilizing the augmented dataset are only apparent when we utilize our curriculum learning strategy to introduce these samples from the order of easy-to-difficult. Here, we observe a drastic improvement over the previous "Mix" data splits, and better performance than our previously best-performing clean data-split from our model.
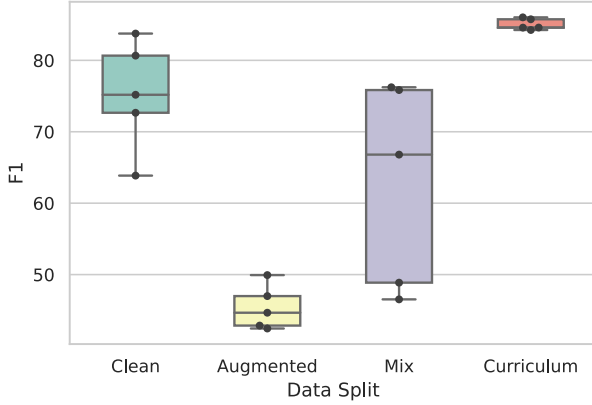
Figure 5: Stability and improved performance of our model is showcased when utilizing curriculum learning. Naively utilizing the mixed dataset leads to poor performance.

Model	ECE
End-to-End	0.1452
Multi-task	0.1520
Ours	**0.0733**

Table 4: Expected Calibration Error (ECE) of our fusion model out-performs the End-to-End model, and Multi-task models (lower values are better).

Calibration

As first shown in Naeini, Cooper, and Hauskrecht (2015), and later compared in Guo et al. (2017) for neural networks, we utilize Expected Calibration Error (ECE) to showcase a statistic regarding the calibration comparison of our models on making the final conclusions for assessment of likelihood of confusion. The importance of calibration metrics of models in high-stakes decision making is to ensure that the confidence is taken into consideration for the choice of a model, and this selection is not dependant only on the predictions.

By creating M bins, where B_m are samples belonging to bins in confidence intervals $I_m = \left(\frac{m-1}{M}, \frac{m}{M} \right]$, we can calculate the ECE value from Equation 7 which are dependent on the predictions as in Equation 8 and confidence of these values shown in Equation 9.

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left| \text{acc}\left(B_m\right) - \text{conf}\left(B_m\right) \right| \qquad (7)$$

$$\text{acc}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\left(\hat{y}_i = y_i\right) \qquad (8)$$

where $\hat{y}_i$ denotes the model output, and y_i denotes the ground truth.

$$\text{conf}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \qquad (9)$$

where $\hat{p}_i$ denotes the probabilities of the predictions.

The results shown in Table 4 suggest that our model is significantly well-calibrated in constrast to end-to-end, and multi-task models.

Conclusion

In this work, we explore the application of deep learning towards the task of predicting the likelihood of confusion for trademarks which can easily be extended to other branches of law. We address the data limitations by creation of an expertly-tailored augmented dataset. We observe that training on these cases produces stable results only when coupled with a curriculum learning strategy. We also address the challenge of interpretability by introducing intermediate labels—such as predicting whether trademarks are visually similar, which in turn acts as an interpretable mechanism to constrain decisions being made due to spurious correlations. We show that our fusion model outperforms baseline methods by not only producing improvement in the macro-F1 metric, but by also improving on the ability to predict intermediate labels. Finally, we show a drastic improvement

As computational resources available to legal decision-makers may be limited, it is crucial that stability of models is obtained from deep neural network architectures without excessive efforts for fine-tuning. From Figure 5, we observe that the results when utilizing the clean dataset are unstable, and requires multiple runs to achieve desirable performance. For the augmented dataset alone, due to dataset shift between the augmented dataset and clean dataset, very poor performance is showcased. This poor performance is expected, and accentuates the requirement of a few clean samples during the training process. The benefit of utilization of the mixed dataset is only apparent when we introduce it by means of curriculum learning, from which we are able to achieve stable and strong performance.

For the intermediate task, we utilize Smooth L1 Loss, as it leads to stable training with small batches which is a product of our small datasets for training. We report both MAE and MSE as metrics to showcase the best performing model for intermediate-label predictions in Table 3. Expectedly, randomly setting these intermediate labels shows poor performance. As with our models for prediction of likelihood of confusion as the ground truth, we observe no benefit in utilizing the multi-task model for predicting the intermediate labels. Furthermore, we omit the results of curriculum learning strategies in predicting the intermediate labels as they did not provide any benefit for intermediate-label prediction performance due to the synthetic nature of the augmented dataset. Utilizing the mixed dataset also did not show any improvement in performance.

Through experimental results, we observe no benefit in initializing our fusion-based model from models which are not finetuned. Instead, we utilize the best fine-tuned checkpoints for each respective data split. Inclusion of these intermediate labels provide transparency of the fusion model, wherein results are explainable to a degree that they can be utilized by lawyers to assess the risk of confusion for trademarks.

of calibration of our model in comparison to the baselines by reporting the Expected Calibration Error.

References

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Lapedriza, A.; Zhou, B.; and Torralba, A. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 41–48. New York, NY, USA: Association for Computing Machinery. ISBN 9781605585161.

Bhambhoria, R.; Dahan, S.; and Zhu, X. 2021. Investigating the State-of-the-Art Performance and Explainability of Legal Judgment Prediction. *Proceedings of the Canadian Conference on Artificial Intelligence*. Https://caiac.pubpub.org/pub/hj1na0xf.

Binder, A.; Montavon, G.; Bach, S.; Müller, K.-R.; and Samek, W. 2016. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. arXiv:1604.00825.

Bingel, J.; and Søgaard, A. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. arXiv:1702.08303.

Bone, R. 2012. Taking the Confusion Out of 'Likelihood of Confusion': Toward a More Sensible Approach to Trademark Infringement. *Legal History eJournal*.

Cai, Y.; et al. 2019. Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2506–2515. Florence, Italy: Association for Computational Linguistics.

Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; and Androutsopoulos, I. 2020. LEGAL-BERT: The Muppets straight out of Law School. arXiv:2010.02559.

Dahan, S. 2020. AI-powered Trademark Dispute Resolution. In *European Union Intellectual Property Office (EU-IPO) [Expert Opinion]*, p. 39.

Dahan, S.; Zhu, X.; Bhambhoria, R.; and Townsend, S. 2022. Analytics and EU Courts: The Case of Trademark Disputes. In *The Changing European Union: A critical view on the role of the Courts. Forthcoming*. Hart Publishing.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dhamdhere, K.; Sundararajan, M.; and Yan, Q. 2018. How Important Is a Neuron? arXiv:1805.12233.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *CoRR*, abs/1806.00069.

Go, A.; Bhayani, R.; and Huang, L. 2009. Twitter sentiment classification using distant supervision. *Processing*, 150.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. *CoRR*, abs/1706.04599.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. *CoRR*, abs/1902.10186.

Jiang, X.; Ye, H.; Luo, Z.; Chao, W.; and Ma, W. 2018. Interpretable Rationale Augmented Charge Prediction System. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 146–151. Santa Fe, New Mexico: Association for Computational Linguistics.

Kendall, A.; Gal, Y.; and Cipolla, R. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. arXiv:1705.07115.

Lawlor, R. C. 1963. What Computers Can Do: Analysis and Prediction of Judicial Decisions. *American Bar Association Journal*, 49(4): 337–344.

Liu, H.; Yin, Q.; and Wang, W. Y. 2018. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. *CoRR*, abs/1811.00196.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Mosbach, M.; Andriushchenko, M.; and Klakow, D. 2021. On the Stability of Fine-tuning {BERT}: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations*.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 607–617. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Naeini, M. P.; Cooper, G. F.; and Hauskrecht, M. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 2901–2907. AAAI Press. ISBN 0262511290.

Phang, J.; Févry, T.; and Bowman, S. R. 2019. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. arXiv:1811.01088.

Reese, W. L. M. 1978. Legislative Jurisdiction. *Columbia Law Review*, 78(8): 1587–1608.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. arXiv:1811.10154.

Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; and Zhong, C. 2021. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. arXiv:2103.11251.

Srinivas, S.; and Fleuret, F. 2021. Rethinking the Role of Gradient-based Attribution Methods for Model Interpretability. In *International Conference on Learning Representations*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 3319–3328. JMLR.org.

Ustun, B.; and Rudin, C. 2015. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3): 349–391.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762.

Wang, X.; Chen, Y.; and Zhu, W. 2020. A Comprehensive Survey on Curriculum Learning. *CoRR*, abs/2010.13166.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6382–6388. Hong Kong, China: Association for Computational Linguistics.

Wiegreffe, S.; and Pinter, Y. 2019. Attention is not not Explanation. *CoRR*, abs/1908.04626.

Wu, X.; Dyer, E.; and Neyshabur, B. 2021. When Do Curricula Work? In *International Conference on Learning Representations*.

Ye, H.; Jiang, X.; Luo, Z.; and Chao, W. 2018. Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1854–1864. New Orleans, Louisiana: Association for Computational Linguistics.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How Transferable Are Features in Deep Neural Networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, 3320–3328. Cambridge, MA, USA: MIT Press.

Zhong, H.; Wang, Y.; Tu, C.; Zhang, T.; Liu, Z.; and Sun, M. 2020. Iteratively Questioning and Answering for Interpretable Legal Judgment Prediction. In *AAAI*.

Zhou, T.; Wang, S.; and Bilmes, J. 2021. Robust Curriculum Learning: from clean label detection to noisy label self-correction. In *International Conference on Learning Representations*.