Predicting Patent Litigation Risk Using RoBERTa and Metadata Augmentation Techniques

Stanford CS224N Custom Project

Brian Park
Department of Computer Science
Stanford University
brian02@stanford.edu

Nikita Bhardwaj
Department of Computer Science
Stanford University
nikitab7@stanford.edu

Simone Hsu
Department of Linguistics
Stanford University
shsu8@stanford.edu

Abstract

Patent litigation is labor-intensive, costly for all stakeholders involved, and liable to produce unpredictable outcomes (Mazzeo et al., 2013). Companies would therefore benefit from the ability to assess the likelihood of their patents being litigated in order to pre-emptively mitigate the risk of being taken to court. Our project introduces a novel approach to predicting whether or not a patent will be litigated using the RoBERTa model–both the base and metadata augmented variants–and two different methods of sampling data. As part of this project, we construct a brand new dataset that cross-references litigated patents from the Stanford NPE database (Stanford Law School) with all accepted patents from the Harvard-USPTO Patent Dataset (HUPD). Over several trials, we run RoBERTa models fine-tuned on data with different combinations of metadata in order to identify the model configuration with best overall performance. Our best model is finetuned on patent claims concatenated with both CPC code and examiner ID metadata, and achieves an accuracy of 0.787839 as compared to a baseline accuracy of 0.720854.

1 Key Information to include

- Mentor: Mirac Suzgun <msuzgun@stanford.edu>

2 Introduction

Patents play an indispensable role in incentivizing research and development into new areas and rewarding individuals and firms who make important innovations in traditional and emerging industries. Patent ownership is often very lucrative as the patent-holder can monopolize use and distribution of this technology. Patent litigation is therefore a pressing matter. Patents that are not original and infringe upon others' intellectual property significantly damage the integrity of innovation and the patent holder's financial stability. Meanwhile, original patents may also be mistakenly sued, which similarly results in expensive and time-consuming legal proceedings and entanglements, posing significant risk to innovation and financial stability, especially for newer companies. Ideally, stakeholders would be able to assess the likelihood of a patent being litigated and take measures to mitigate such litigation risks before legal proceedings commence.

This problem motivated us to explore whether language models could be utilized to predict if a patent is likely to be litigated. This is an interesting and difficult problem for several reasons. Foremost,

there currently is no single database that correlates accepted patents with their litigation status. In addition, patent litigation is often dependent on what industry is affected; patents filed in the pharmaceutical industry may be more prone to litigation than patents in software, due to the inherent nature of the industries themselves. Additional hurdles exist as well, such as the fact that patents are not consistently processed across individual examiners and regional offices, with some offices being historically more lax with approval. Therefore, patent litigation may not only depend on the content of a patent itself.

To date, there currently are no rigorous studies on predicting patent litigation. To address this gap, we present a pioneering dataset, merging HUPD with the Stanford NPE Litigation Database to create a new dataset that comprises information on patent litigation status. The contents of the novel dataset are elaborated upon subsequently. We will use this dataset and the transformer-based language model, RoBERTa, taking inspiration from Suzgun et al. (2022)'s use of the DistilRoBERTa model, to predict patent litigation. Understanding that patent litigation is dependent on the subject to which a patent relates, examiner information, and more, we further incorporate feature engineering to fine-tune our model to achieve higher accuracy in predictions. Our work hopes to set the stage for future applications of language models to the field of patent litigation.

3 Related Work

HUPD contains over 4.5 million US patent applications with 34 data fields per patent, a significant expansion on its predecessor BIGPATENT (Sharma et al., 2019), which included 1.3 million patent applications, four meta-data fields, and a notable lack of a patent's claims section, where the bulk of a patent's technical detail lies (Suzgun et al., 2022). The construction of a highly structured, information-rich dataset enabled the proliferation of NLP methods for performing patent acceptance prediction. The HUPD paper mentions patent novelty evaluation as another potential use case for NLP in conjunction with the dataset, but does not explicitly address patent litigation. Our project uses a HUPD pretrained DistilROBERTa model to run baselines. Additionally, we implement the concatenation technique as used by HUPD and described in Raffel et al. (2023) to augment the input data with metadata in later trials.

The first mention of NLP for patent litigation dates back to 2007, predating HUPD. Indukuri et al. (2007) tokenize a small set of randomly selected patent claims and run a rudimentary algorithm to measure syntactic and semantic similarity between clusters of patents. While our computational approach differs entirely from theirs given over a decade of improvements in NLP technology, we similarly identified the relevance of the patent claims section to litigation likelihood. More recently, Kim et al. (2021), Juranek and Otneim (2021), and Liu et al. (2018) attempt various other methods for predicting patent litigation, but none make use of the most recent ML models utilized by HUPD. This project is the ideological descendant of this small set of computational patent litigation papers, but draws data and technological inspiration from the more recent HUPD work.

4 Approach

4.1 Prediction Modeling

Our research framework combines natural language processing, data sampling, and metadata augmentation to predict the likelihood of patent litigation.

4.1.1 Natural Language Processing with RoBERTa and DistilRoBERTa

In our project, we utilized RoBERTa and DistilRoBERTA, both variants of the BERT (Bidirectional Encoder Representations from Transformers) architecture pioneered by Devlin et al. (2019), as these models excel at capturing subtleties in large text corpuses, such as patents.

The BERT architecture stands out because it bidirectionally trains its Transformer models, leveraging a multi-headed self-attention mechanism to contextually understand a given word by learning from preceding and succeeding words. This self-attention mechanism is formulated as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V represent the query, key, and value matrices respectively, and d_k is the dimension of the key. For BERT, and by extension, RoBERTa and DistilRoBERTa, this mechanism is expanded into a multi-headed version to capture various aspects of information:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O$$

where each $\text{head}_i = Attention(\text{QW}_i^Q, KW_i^K, VW_i^V)$ and W^O, W_i^Q, W_i^K, W_i^V are parameter matrices. This allows the model to focus on different parts of the sentence simultaneously, enhancing its understanding of language context and relationships within the text.

RoBERTa (Liu et al., 2019) expands on BERT by training on 160GB of text corpuses (as opposed to the 16GB of BERT) and optimizing training with adjustments like dynamic masking, which randomly masks tokens across training epochs. RoBERTa also removes the Next Sentence Predicting (NSP) task during pretraining and instead focuses on more extensive sequence lengths, allowing it to better understand the relationships between sentences and paragraphs.

DistilRoBERTa (Sanh et al., 2020), on the other hand, is a streamlined version of RoBERTa that retains much of the original performance while optimizing resource efficiency. DistilRoBERTa retains approximately 97% of RoBERTa's performance but is 40% smaller in size, and 60% faster at inference. Specifically, DistilRoBERTa has fewer transformer layers (6 compared to RoBERTa's 12) and fewer attention heads, with a total parameter count reduction of around 43 million. This resource efficiency makes DistilRoBERTa suited for situations where compute is limited or faster processing is required.

One limitation we navigated is the maximum sequence length constraint inherent to these models–both RoBERTa and DistilRoBERTa have a maximum sequence length of 512 tokens. This limitation means that a significant portion of the patent claims are truncated, since claims typically exceed the 512-token mark. To mitigate the impact of this limitation, we incorporated metadata augmentation into our methodology, as discussed below.

4.1.2 Metadata Augmentation

Recognizing that raw patent texts contain dense and highly technical language, we supplemented patent inputs with metadata features. The patents in the HUPD-NPE combined dataset bring 34 metadata fields per patent, which include attributes from administrative details to technical classification. However, not all fields contribute equally to the litigation prediction task.

We found that certain metadata fields such as CPC (Cooperative Patent Classification) and examiner ID's hold significant predictive power for litigation prediction. The CPC classification represents the technical domain of the patent and is an important factor in understanding a patent's likelihood for litigation. Examiner ID information is relevant as there is substantial variability in patent approval rates across different examiners, and certain examiners might be associated with more litigated patents due to varying standards of scrutiny. By embedding this information, our model gains a deeper understanding of the nuances that could influence litigation likelihood.

To embed these metadata features within our model, we concatenated the CPC classification and examiner ID features as contextual prefixes to the patent's text. This metadata augmentation primes the model with the context of the patent's domain and the examiner's historical tendencies prior to processing the claims text (Raffel et al., 2023).

4.2 Data Sampling Techniques

We also explored different data sampling techniques to optimize the balance between litigated and non-litigated patents. We used two distinct sampling methods. The first method was simple random sampling, establishing a ratio of one litigated patent to every two non-litigated patents, randomly selected. The second approach was matched sampling, which aimed to achieve a more refined dataset by choosing non-litigated patents according to year and CPC criteria. We established a 1:2 ratio of litigated to non-litigated patents, where each litigated patent was matched with two non-litigated patents based on filing year and CPC subclass code. In this approach, for every litigated patent, we matched two non-litigated patents from the same filing year, ensuring one shared the same CPC class while the other belonged to a different class. If this criteria could not be met due to a lack of available patents in the same class or year, the selection is then expanded to randomly chosen non-litigated patents from the entire pool, ensuring no repeats in the dataset. We chose this method to ensure

consistency and comparability within the dataset. In implementing these sampling techniques, our goal was to mitigate data biases and improve the models performance.

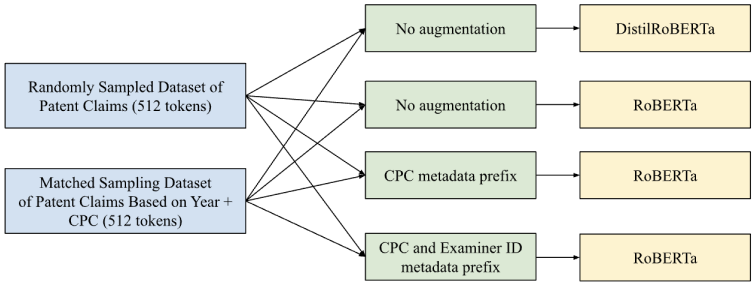

Figure 1: Experiment architecture detailing data sampling and metadata augmentation techniques

4.3 Baseline

For our baseline model, we utilized the DistilRoBERTa model fine-tuned on HUPD, which is pre-trained using the Masked Language Modeling (MLM) technique. This model was made publicly available for use by Suzgun et al. (2022). In MLM, a fraction of the words and tokens in the input are randomly masked, and the model is trained to predict these elements by using the surrounding context. This objective is a foundational approach for pretraining language models, since it allows them to predict missing words within a given text sequence, thereby learning context and an understanding of language semantics. The HUPD pre-trained model serves as our benchmark to evaluate the efficacy of our metadata augmentation strategy, and our data sampling methods. Through a comparison of these methods with our baseline, we aim to quantify the improvements these techniques offer.

5 Experiments

5.1 Data

We constructed a novel dataset consisting of both litigated and non-litigated patents. This dataset integrates two pre-existing datasets: (1) the Harvard-USPTO dataset (HUPD) containing 4.5 million patent applications filed to the the USPTO between 2004 and 2018, and (2) the Stanford NPE Litigation Database comprising nearly 70000 patent lawsuits between 2000 and 2020. HUPD is substantially larger and contains both rejected and accepted patent applications, while the NPE Litigation database lists lawsuits and associated patents from years outside the range covered by HUPD. We removed rejected patent applications from HUPD, removed duplicates across HUPD and the Stanford NPE database, and added litigation status as metadata information, therefore creating a new database merging all accepted patents and their litigation status.

Each patent, both litigated and unlitigated, has 34 associated metadata fields including filing year, examiner information, CPC code, and litigation status, which were preprocessed to serve as contextual features for our model.

5.2 Evaluation method

The primary evaluation metric we employed to evaluate our model's prediction is accuracy. We compare our model's predictions with the actual litigation status of the patents as represented in tables 1 and 2. However, recognizing that accuracy, while robust, may not fully capture model performance, especially in the case of class imbalance in the dataset, we supplemented our evaluation with F1 scores and Area Under Curve (AUC) metrics to account for cases in which classes were unevenly distributed in the dataset.

F1 scores represent the harmonic mean of precision and recall. It provides us with a single metric that represents our model's precision, or ability to accurately identify patents that will be litigated, and recall, or ability to identify all relevant instances of litigation. This metric is especially relevant

considering the context of our investigation; falsely identifying a patent as likely to be litigated (a false positive) or mistakenly predicting that a patent is unlikely to be litigated (a false negative) both have significant implications in a legal context.

Area Under the Receiver Operating Characteristic Curve (AUC) evaluates the model's performance across all classification thresholds. We selected this metric because it provides a comprehensive measure of the model's ability to differentiate litigated and non-litigated patents, regardless of any potential imbalance inherent in the dataset. This is especially relevant in our study because of potential skews in our dataset in terms of the number of litigated and non-litigated patents and the distribution of patents across CPC codes.

All in all, these metrics provide a nuanced evaluation of our models, not only comprising accuracy itself but moreover addressing false positives and false negatives, as well as performance with potential class imbalances.

5.3 Experimental details

As the baseline, we adopted the HUPD DistilRoBERTa model utilized by Suzgun et al. (2022), which was fine-tuned on the Harvard USPTO Patent Dataset using a masked language modeling (MLM) objective. All RoBERTa models were fine-tuned from the pre-trained RoBERTa model publicly available through HuggingFace. We fine-tuned this model on (1) CPC classification information and (2) CPC classification information and examiner ID information.

We ran all experiments utilizing the Google Cloud setup provided in the course, using one Tesla T4 GPU with 16GB of memory. We maintained a consistent set of hyperparameters across our fine-tuning process, using a learning rate of 2e-5, a weight decay of 0.01, and a batch size of 16 across 4 epochs.

Training took place on the claims section of the patent. When experimenting to develop our baseline, we noticed that training and evaluating patent litigation using the claims section produced slightly higher accuracies than when trained on the abstract. To balance accuracy and efficiency, an epoch size of 4 was decided. Furthermore, our training set comprised patents before 2014, while the test set consisted of the patents filed from 2014 onwards, resulting in a near 80/20 split between training and testing data.

5.4 Results

Table 1: Random Sampling

	Baseline	RoBERTa	Fine-tuned RoBERTa (CPC)	Fine-tuned RoBERTa (CPC + Examiner ID)
Accuracy	0.720854	0.711558	0.718593	**0.722111**
F1 Score	0.724602	0.723773	0.711786	**0.732354**
AUC	**0.799704**	0.798880	0.797851	0.795974

Table 2: Matched Sampling

	Baseline	RoBERTa	Fine-tuned RoBERTa (CPC)	Fine-tuned RoBERTa (CPC + Examiner ID)
Accuracy	0.730486	0.718844	0.766332	**0.787839**
F1 Score	0.616996	**0.701203**	0.589585	0.645838
AUC	0.798241	0.800805	0.818513	**0.820530**

Tables 1 and 2 present performance metrics for the various models we tested.

Figures 2, 3, and 4 show the performance of our models on the patent classification task, employing both random sampling and matched sampling to address the imbalance between litigated and non litigated patents. As predicted, matched sampling to achieve a more balanced dataset generally improves performance, although F1 scores and AUC scores appear to be more volatile, with potential reasoning expanded upon later.

5.4.1 Accuracy

As predicted, accuracy increases as we fine-tune the RoBERTa model on CPC classification codes, and on CPC classification codes and examiner ID information, as compared to the base RoBERTa model. Also, when comparing identical model configurations, matched sampling yields higher accuracy than random sampling (for instance, fine-tuned RoBERTa on both CPC classification and examiner ID has an accuracy of nearly 0.788 for matched sampling, whereas it is 0.722 when patents are randomly sampled). Our baseline pre-trained distilRoBERTa model demonstrates slightly higher accuracy than the base RoBERTa model imported from Hugging Face, which was pre-trained using masked language modeling but was not fine-tuned to any downstream tasks. Base RoBERTa was not fine-tuned on any patent metainformation as compared to the baseline distilRoBERTa model, which was pretrained on HUPD, likely improving accuracy.

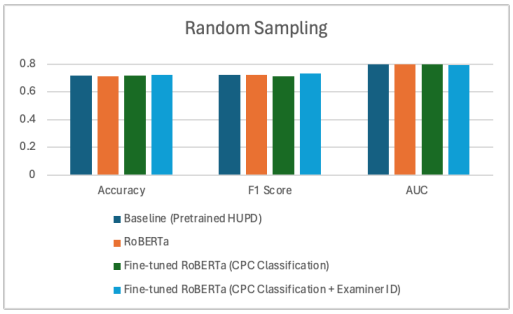

Figure 2: Comparison of different model configuration outcomes with random sampling

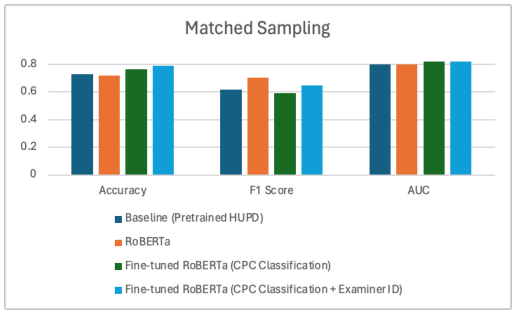

Figure 3: Comparison of different model configuration outcomes with matched sampling

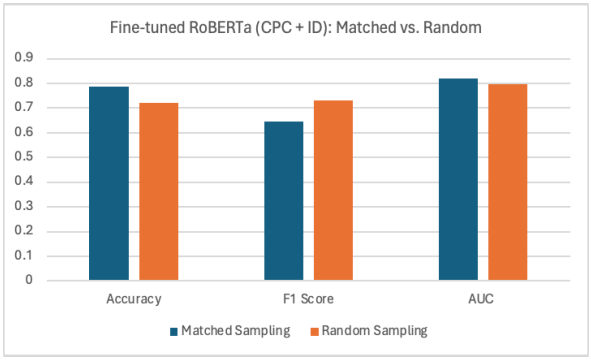

Figure 4: Matched versus random sampling scores from CPC + Examiner ID augmented model

5.4.2 F1 Score

F1 score, as aforementioned, balances precision and recall, and is more volatile across model configurations and sampling methodologies. When using the randomly sampled dataset, the RoBERTa model fine-tuned on CPC information and examiner ID not only is the most accurate, but also has the highest F1 score. However, when using the matched sampling dataset, the base RoBERTa model that is not fine-tuned on patent metadata yields the highest F1 score of 0.701, and addition of CPC classification or examiner ID does not consistently increase F1 score. This demonstrates that while additional metadata improves the accuracy of predictions, it does not universally improve the balance of precision and recall.

5.4.3 AUC

We notice as a general trend that the AUC scores of the RoBERTa models utilizing matched sampling were slightly higher than those utilizing random sampling, consistent across different model configurations. For instance, the base RoBERTa model with matched sampling had an AUC score of 0.8008 as opposed to 0.7988 utilizing random sampling. This likely underscores the class imbalance present in the randomly sampled dataset (such as a skew in CPC classification codes). It moreover indicates that introducing domain-specific features like examiner IDs improves model accuracy. However, the AUC score of the baseline DistilRoBERTa model fine-tuned on HUPD interestingly performed better with random sampling than matched sampling.

6 Analysis

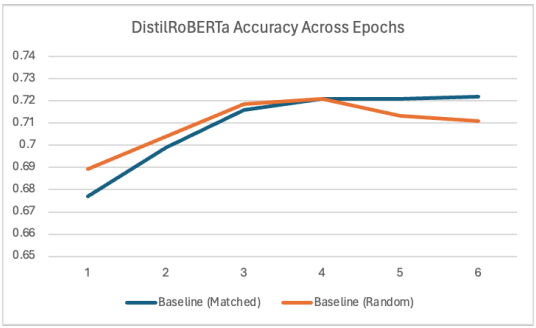

Figure 5: Comparison of matched versus random baseline (DistilRoBERTa) accuracy across 6 epochs

Our initial approach in developing a baseline involved testing our model on a range of 2 to 6 epochs. We noticed extending training past 4 epochs resulted in negligible improvements and in fact decreased accuracy, as represented in Figure 5. Therefore, we concluded that 4 epochs was the optimal balance between accuracy and efficiency, reducing the training duration by nearly 30% compared to a 6-epoch configuration.

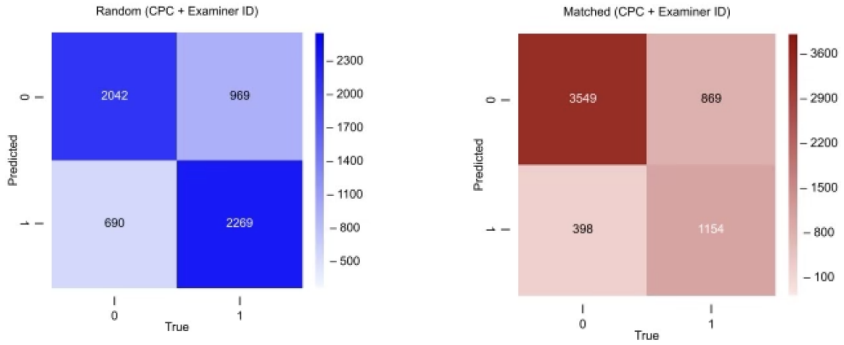

Figure 6: Confusion matrices for best-performing model (random and matched sampling)

Overall, all models, including the baseline pre-trained HUPD DistiBERT model, outperform the 66% accuracy of random guessing, with the best-performing model achieving nearly 80% accuracy. However, as aforementioned, the maximum input sequence length of 512 truncates valuable information from our input, likely reducing performance.

Despite RoBERTa being trained on a larger corpus of text than DistilRoBERTa, the base RoBERTa model has lower accuracy than the HUPD fine-tuned DistilRoBERTa model that serves as our baseline. There are several potential explanations. For one, patents are legal documents, and the claims are

structured and not free-formed. Claims include sentence fragments, enumerated lists, descriptions of tabular data, and more. This may make it harder for the masking pre-training task on base RoBERTa to effectively learn. Legal jargon and formalities may also take up valuable space while providing minimal information relevant to the patent's content itself.

We also observe that overall accuracy fails to surpass 0.8. One reason for this could be the token limit of 512, which is often far smaller than the length of a patent's claims section, meaning the input text is heavily truncated. We further noticed that while accuracy improves as metainformation is concatenated to the input sequence, it is a smaller impact than we initially hypothesized. This could relate to the fact that concatenating patent metadata to the beginning of the input string further decreases the length of the claim's text that is being processed by the model. An interesting avenue of future research could be analyzing how models supporting longer sequence lengths perform.

Analyzing the number of true positives, true negatives, false positives, and false negatives, allows us to obtain a better evaluation of our metrics, notably accuracy and F1 scores. Accuracy can simply be calculated as $accuracy = \frac{TP+TN}{total}$. As aforementioned, $F1 = 2 \cdot \frac{precision \cdot recall}{precision+recall}$ where $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$. In both equations, TP is true positive, FP is false positive, and FN is false negative.

Analyzing the model with the highest accuracy in random sampling and in matched sampling, we see that the matched sampling has a higher accuracy but a lower F1 score. This is affected by the true positives, true negatives, false positives, and false negatives amongst the examined patents, represented in Figure 6.

We notice that matched sampling, despite having fewer false positives, also has fewer true positives and more false negatives compared to random sampling. This decrease in true positives and increase in false negatives reduces its F1 score. However, its higher true negative count contributes to its higher accuracy. This suggests that the model, with matched sampling, can better correctly identify negative cases but overall has a less optimal balance between recall and precision. Therefore, we conclude that the ideal configuration depends on the overarching aim of patent litigation prediction. If it is more important to minimize false negatives (inaccurately predicting that a patent isn't going to be litigated when in fact it is likely) random sampling may be more optimal due to its higher F1 score, even at the expense of accuracy. On the other hand, if we seek to minimize false positives (inaccurately predicting that a rigorous patent is likely to be litigated), matched sampling with higher accuracy yet lower F1 score may be more suited.

7 Conclusion

Our study presents a pioneering approach to predicting patent litigation outcomes by combining the HUPD dataset and the NPE litigation dataset. Through our methodology, we demonstrate the efficacy of different sampling techniques, namely random and matched sampling, ultimately finding that matched sampling led to better performance. A significant enhancement in model accuracy was achieved by incorporating metadata augmentation, specifically through the addition of examiner ID and CPC information. Among the models tested, we found that the RoBERTa model demonstrated the best performance when applied to the matched-sampling dataset augmented with the aforementioned metadata. We did encounter a limitation concerning the token count, as we could only incorporate 512 tokens from the patent claims into our models. This highlights a potential avenue for future research, where exploring the capabilities of transformers with extended sequence lengths could potentially enhance predictive accuracy even further. Additionally, there remains lots of room for experimentation with sampling methods and metadata augmentation techniques.

Our work, while focused on patent litigation, lays the groundwork for broader exploration into the field of intellectual property rights. Our findings and methodologies can be adapted to numerous applications, such as patent infringement risk assessment. In addition to predicting patent litigation likelihood, we can supplement this using chain of thought prompting in large language models to provide stakeholders with information as to relevant risk factors. The findings from our model have the potential to impact the broader landscape of patent filing and litigation, providing a spectrum of tools to analyze and manage patents.

References

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

Kishore Varma Indukuri, Anurag Anil Ambekar, and Ashish Sureka. 2007. Similarity analysis of patent claims using natural language processing techniques. In *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications*, volume 2007, pages 169–175. IEEE. Available from IEEE Xplore. Restrictions apply.

Steffen Juranek and Håkon Otneim. 2021. Using machine learning to predict patent lawsuits. Discussion Paper 2021/6, NHH Dept. of Business and Management Science. Available at SSRN: `https://ssrn.com/abstract=3871701` or `http://dx.doi.org/10.2139/ssrn.3871701`.

Youngho Kim, Sangsung Park, Junseok Lee, Dongsik Jang, and Jiho Kang. 2021. Integrated survival model for predicting patent litigation hazard. *Sustainability*, 13(4).

Qi Liu, Han Wu, Yuyang Ye, Hongke Zhao, Chuanren Liu, and Dongfang Du. 2018. Patent litigation prediction: A convolutional tensor factorization approach. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5052–5059. International Joint Conferences on Artificial Intelligence Organization.

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.

Michael J. Mazzeo, Jonathan Hillel, and Samantha Zyontz. 2013. Explaining the "unpredictable": An empirical analysis of u.s. patent infringement awards. *International Review of Law and Economics*, 35:58–72.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization.

Stanford Law School. Stanford npe litigation database. `https://npe.law.stanford.edu/`. Accessed: insert access date here.

Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K. Sarkar, Scott Duke Kominers, and Stuart M. Shieber. 2022. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications.