Patent Acceptance Prediction With Large Language Models

Stanford CS224N Custom Project
Project proposed & mentored by Mirac Suzgun

Akayla Hackson
Department of Electrical Engineering
Stanford University
akayla@stanford.edu

Miguel Gerena
Department of Computer Science
Stanford University
miguelg2@stanford.edu

Abstract

Small businesses and inventors often face significant challenges in securing patent approvals for their inventions. This paper proposes the development of a Language Model capable of predicting patent approval outcomes based on linguistic analysis. The current SOTA model obtained 57.96% accuracy on their best model, while our best achieved an accuracy of 64.37%. While our fine-tuned model has surpassed the current SOTA model (Suzgun et al., 2022), it has hit a performance plateau fine-tuning a variant of the BERT (Devlin et al., 2019) architecture. Subsequent efforts have focused on fine-tuning more capable models like Mistral-7b (Jiang et al., 2023), Gemma-7b, and Gemma-2b (Banks and Warkentin, 2024). After implementing strategies like low-rank adaptation (Hu et al., 2021), model quantization LoRA (Dettmers et al., 2023), and model quantization (Jacob et al., 2017) during inference, alongside optimizations in data loading and processing, we eliminated various limitations, yet have encountered more. With current results surpassing the SOTA we are pleased, yet not satisfied. We plan on continuing our work to overcome recent hurdles once more.

1 Introduction

In the realm of intellectual property management, small businesses and individual inventors often encounter barriers in securing patent approvals for their innovations. The intricacies of patent application processes, paired with the necessity for precise technical documentation, often pose a considerable challenge, especially for people with limited resources and access to specialized legal counsel. Recognizing this critical need, this paper introduces the development of an advanced Language Model (LM) designed to predict the outcomes of patent approvals through a detailed linguistic analysis of patent application texts.

By leveraging the latest advancements in natural language processing (NLP) and machine learning, this tool aims to democratize the patent application process, offering small businesses and independent inventors a valuable resource to assess the viability of their patent applications before going through the lengthy and costly formal review process. The predictive capabilities of the proposed LM provide actionable insights into the strengths and potential weaknesses of their applications, based on patterns learned from datasets of historical patent applications and their outcomes.

Such a tool not only eases the path to securing patent protections but also levels the playing field for smaller entities in the innovation ecosystem. By affording a clearer understanding of the factors that contribute to successful patent applications, businesses and inventors can make informed decisions on how to refine their submissions, potentially increasing their chances of approval and reducing the risk of costly rejections. Ultimately, this approach stands to create a more vibrant, inclusive environment for innovation, where an idea, rather than the depth of one's resources, dictate the likelihood of securing patent protection.

Our proposed model outperformed baselines and state-of-the-art (SOTA) models detailed in Suzgun et al. (2022), achieving an accuracy of 64.37%. Experiments performed on various models and configurations deemed a variation of the Bert (Devlin et al., 2019) LM to outperform the rest. While these results are exciting, there is still work to be done on further enhancing performance.

2 Related Work

This section provides an exploration of various LMs and the implementation of specific tools aimed at enhancing the development of a system for predicting patent approval outcomes. Through a detailed examination of models like BERT, Gemma, and Mistral, alongside the introduction of efficiency-enhancing techniques like QLoRA, the paper outlines an approach designed to utilize linguistic analysis in aiding the development process, with a more particular focus on improving accessibility and effectiveness for users.

2.1 HUPD

HUPD, or The Harvard USPTO Patent Dataset paper (Suzgun et al., 2022), provides detailed breakdowns of text based sections of patent documents, ranging from titles to detailed descriptions, along with metadata about the average number of tokens in each section. This detail is valuable for tasks such as long-sequence summarization and language modeling. The dataset also enhances research capabilities by including continuation information for each patent application. This facilitates studies on the progression and outcomes of patent filings. The authors also present various tasks and evaluation metrics for the datasets use. These tasks range from patent acceptance prediction and language modeling to abstractive summarization, among others. The corresponding evaluation metrics include accuracy, perplexity, and ROUGE/BLEU scores. Additionally, benchmarks are provided for future research utilization. Overall, it presents a rich repository of information for a diverse spectrum of analytical applications, serving as the primary data source for our project.

2.2 Bert

The BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) architecture represents a significant advancement in the field of NLP, introducing a method for pretraining LMs on large text corpora in a bidirectional manner. BERT is able to understand the context of words based on their surrounding text, greatly enhancing its ability to interpret the complex and technical language typical of patent applications. By leveraging BERT's capabilities, our project can benefit from deep contextual insights. This enables more accurate classification of patents based on the nuanced information contained within their applications.

2.3 Gemma 2B and 7B

The Gemma models, developed by Banks and Warkentin (2024), offer lightweight, SOTA open models that are built on the foundation of the Gemini technology (Team et al., 2023). Comprising two primary sizes, Gemma 2B and Gemma 7B, these models are designed to deliver great performance for their scale, outperforming larger models. This achievement is attributed to their pretraining and instruction tuning, which ensures that they excel in a wide variety of tasks. For focused on tasks like patent acceptance prediction, the Gemma models offer an exceptional mixture of sophistication and practicality.

2.4 Mistral 7B

Mistral 7B (Jiang et al., 2023) is a 7-billion-parameter LM, distinguishable by its plentiful capabilities. Mistral 7B outperforms the Llama 2 13B model (Touvron et al., 2023) across a suite of benchmarks, as well as surpassing the Llama 1 34B (Touvron et al., 2023) in domains requiring reasoning, mathematical abilities, and code generation. The core innovation behind Mistral 7B's performance is in grouped-query attention, a technique that greatly accelerates inference times. This is complemented by the sliding window attention mechanism, which allows Mistral 7B to manage sequences of arbitrary lengths, reducing the inference time without compromising the model's effectiveness. The model's ability to excel in reasoning indicates a good understanding of complex, structured data, an essential trait for analyzing the detailed, technical language found in patent applications.

2.5 LoRA and QLoRA

Low-Rank Adaptation (LoRA) (Hu et al., 2021) and its extension, QLoRA (Dettmers et al., 2023), signify great strides in the efficient fine-tuning of prerained LMs, for which they are known for enhancing AI applications while reducing computational demands. LoRA is a technique that integrates trainable low-rank matrices into each layer of the Transformer architecture, greatly reducing the number of parameters that need to be trained during fine-tuning. This approach lessens the computational load and aims to enhance the model's performance on specific tasks. QLoRA builds upon the foundation of LoRA, incorporating advanced quantization methods such as 4-bit NormalFloat quantization and Double Quantization to further boost parameter efficiency during the fine-tuning process. QLoRA is able achieve a remarkable reduction in GPU memory requirements, enabling the fine-tuning of large models. Implementing QLoRA allows for the application of LLMs, for which possess a deeper understanding of complex technical language and nuanced textual patterns, without the barrier of extensive hardware resources.

3 Approach

This section dives deeper into the intricacies of the proposed approach, offering detailed insights and establishing clear distinctions among the baselines utilized for this study.

3.1 Our Method

Our research aims to predict patent acceptance by employing the Harvard USPTO Patent Dataset (HUPD) (Suzgun et al., 2022), thus offering a novel application of advanced linguistic models in the patent domain.

The HUPD paper utilized a series of LMs to conduct this classification task. Our methodology extends the work initiated by the HUPD study, refining the models used (BERT and its variant DistilBERT (Sanh et al., 2020)), incorporating new, more advanced, models (Mistral (Jiang et al., 2023)) and investigating the adaptation of these models to enhance patent acceptance prediction. This project distinguishes itself by implementing newer and more advanced models, as well as adapting these LMs with methods such as linear probing (LP), fine-tuning, LoRA, and QLoRA. The latter two techniques were not used for fine-tuning the models in the HUPD paper and the authors also did not perform any LP studies. To the best of our knowledge, there has been no adaptation-centric paper on the task of classifying the approval of a patent. The architectures used in our approach are the base-non-instruct LLM's described herein, with a 2-class classification head on top.

3.2 Baselines

We have two baselines we are working with. One is the base model without any adaptation and the other is the current SOTA model published in Suzgun et al. (2022). These SOTA models are based on the BERT (Devlin et al., 2019) architecture. The code provided in Suzgun et al. (2022) was used for determining the SOTA baselines. For later studies, we completely rewrote the codebase, added additional models, LoRA, and QLoRA capabilities in addition to gradient clipping, loss weighting, and masking options, among other features. In addition, we adapted the HUPD hugging dataset scripts to serve our experimentation.

4 Experiments

We have developed a data loading and training framework to address the challenge of BERT embedding dimension limits; specifically, the maximum dimension is 512, whereas the dimensions for abstract and claims sections significantly exceed this, often reaching into the thousands. To navigate this limitation, our approach involves tokenizing only those words that appear at least three times, highlighting the necessity for models with larger capacity. Therefore, we created a new pipeline that employs Google's Deepmind Gemma 2B model (Banks and Warkentin, 2024) and the Mistral7B model (Jiang et al., 2023). We have fine-tuned the DistilBERT and Mistral7B models, as well as performed sensitivity studies on, learning rate, batch size, dataset size, dataset sections, combining dataset sections, epochs, and loss function weighting due to the skewed dataset. The tokenizer and model max length for Mistral-7b was reduced from 4096 to 512, so we could have a one-to-one comparison with our DistilBERT model.

4.1 Data

For this project, the HUPD dataset (Suzgun et al., 2022) forms the backbone of our analysis. This is a public dataset encompassing more than 4.5 million English-language utility patent applications filed to the United States Patent and Trademark Office (USPTO) from January 2004 through December 2018. The data is divided into 34 fields, including filing date, fine-grained classification codes, examiner information, and many others. The complete data set is 360gb. We have pre-processed the data and changed all the accepted, pending, and rejected patents from their label to index mapping. This aids us in producing hot-encodings of the outcome. We are focusing on the claims and abstract sections for each patent application in the dataset as a starting point. Note that these sections contain an average token length of 1271.5 and 132.0, respectively, which the claims surpass the Bert context length.

4.2 Evaluation Method

To evaluate our models performance the metric of accuracy is used. We use confusion matrices to help acquire the accuracy, which can be more formally described as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where: TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives

We are also utilizing the area under the receiver operating characteristic curve (ROC AUC) metric being that the nature of the data is unbalanced.

4.3 Experimental Details

All experiments were run on either a Nvidia RTX 3090 or 4090 GPU. We ran various experiments to test out the best configurations. Our experiments found that best model had a configuration of epoch 2, batch size 8, learning rate 2E-5, weight decay 0.01, dropout 0.1, dimension size 768, hidden dimension count 3072, head count 12, and number of layers 6. It is important to note that the maximum embedding size was 512, therefore, 512 was the number used.

The parameters highlighted in Table 1 do not represent all the experiments that we performed, but a subset of the parameters used for most of the experiments. Batch sensitivities were done for batches of size 8, 16, 32, 48, and 64. The best results obtained were with a batch size of 8. A Learning rates sensitivity study was performed for values of 2e-5, 3e-5, 4e-5, 5e-5, 1e-4, 8e-4, and 5e-4. A learning rate of 2e-5 was best for our model.

Run ID	Model	Batch Size	LR	LoRA Rank	LoRA Alpha	Trained Modules	Quantization	Train Time [1]
1	Distilbert	8	$2e^{-5}$	N/A	N/A	N/A	N/A	6 hr
2	Distilbert	48	$2e^{-5}$	N/A	N/A	N/A	N/A	6 hr
3	Mistral-7b	8	$1e^{-4}$	64	16	q_proj, v_proj	4bit	14 hr
4	Mistral-7b	8	$1e^{-4}$	64	16	All[2]	4bit	16 hr
5	Mistral-7b	8	$1e^{-4}$	64	128	All[2]	4bit	16 hr
6	Mistral-7b	8	$1e^{-4}$	64	128	All[2]	8bit	16 hr

Table 1: Hyper Parameters for Highlighted variants.

4.4 Results

We evaluated the current SOTA model against a test dataset specifically curated by our team, ensuring this dataset had not been used in training either our model or the SOTA model. This process validated the SOTA model at a 57.96% accuracy rate, while our model excelled, achieving a 64.37% accuracy.

Our experiments were conducted on a select portion of the full dataset, focusing exclusively on patent applications from 2008 to 2014 for model fine-tuning, with a 2015 subset utilized for testing. This strategic choice was informed by the desire to manage the dataset's scope effectively.

In addressing data imbalance, particularly in our training data, we developed and calibrated specific weights for the loss functions, ensuring the validation data remained balanced. This adjustment was pivotal in enhancing model performance, which, when coupled with reduced batch sizes and optimal settings identified through sensitivity analyses, led to a significant performance leap. Our preliminary findings, highlighting a 60.16% accuracy by prioritizing claim data over abstract data for training, are documented in Table 2.

Fine-tuning literature (Kumar et al., 2022) dictates that Linear Probing followed by Fine-Tuning yields the best results when adapting a model for out-of-distribution data. Our exploration into fine-tuning methods, including traditional fine-tuning, linear probing, and their combination, revealed traditional fine-tuning as the superior approach for our use case, as detailed in Table 3.

An innovative aspect of our methodology was the development of a specialized loss function. Comparative experiments showcased its effectiveness, outperforming the standard model's loss function by 3%, achieving a 63.06% accuracy, as shown in Table 4. This is due to the class imbalance in the data.

Our final experiment was a head-to-head comparison of high-performing models with advanced models like fine-tuned DistilBERT and Mistral-7B. Other models, such as Gemma-2b and Gemma-7B, were also experimented with, however, we dedicated most of our time to the DistilBERT and Mistral7B models due to their well-known performance and the project's condensed schedule. Yet, Gemma-2b and Gemma-7b are still implemented in our codebase. Our DistilBERT model surpassed the SOTA accuracy, setting a new benchmark at 64.37%, with detailed results presented in Table 5. It's crucial to underline that patent classification can leverage various data segments, not restricted to claims data alone, broadening the applicability of our findings.

Model Name	Abstract Accuracy	Claims Accuracy
distilbert-base-uncased (untrained)	48.05	48.05
H01L (Abstract)	54.42	54.44
H01L (Claims) (SOTA)	48.50	**57.96**
G06F (Abstract)	54.64	54.24
G06F (Claims)	48.05	52.56
Ours-Preliminary (Claims)	56.32	**60.16**

Table 2: Correct patent classification utilizing Abstract or Claims section. The dataset the model is trained on is denoted by model name (data).

[1] One year worth of data

[2] All corresponds to the following modules "q_proj", "o_proj", "k_proj", "v_proj", "gate_proj", "up_proj", "down_proj"

Fine tuning method	Accuracy
Fine Tuning (FT)	62.64
Linear Probing (LP)	57.76
LP + FT	62.19

Table 3: Fine tuning method techniques comparison

Loss Method	Accuracy
Model Loss	57.02
Our Loss	63.06

Table 4: Loss method comparison

4.5 Analysis

The unexpected performance of the Mistral7B model, which did not surpass the DistilBERT model despite its significantly higher capacity, took us by surprise. The training duration for the Mistral7B was nearly triple that of the DistilBERT model, as detailed in Table 1. We believe the sub-par performance of Mistral-7b is due to our tokenizer configurations and future work should be focused on overcoming these issues, as Mistral-7b is a highly capable model. Intriguingly, the less complex and faster-trained DistilBERT model not only demonstrated commendable performance but also exceeded the SOTA model. Looking at the mean train and validation loss of the DistilBERT model, we saw the model hitting a performance limit was the performance hit a plateau. This suggested that further improvements might be constrained by the model's inherent limitations.

Curious to understand the areas of strength and weakness within the DistilBERT model's predictions, we analyzed its performance across various patent application features. Figure 1 contrasts the United States Patent Classification (USPC) classes where the model shows proficiency against those it finds challenging. The x-axis labels correspond to specific patent classifications, with the model demonstrating superior prediction accuracy in categories such as Multiplex Communications, Data Processing (including financial, business practice, management, or cost/price determination), and Stock Material or Miscellaneous Articles. On the other hand, it struggles more with predictions in the realms of Computer Graphics Processing and Selective Visual Display Systems, Land Vehicles, and Flexible or Portable Closures, Partitions, or Panels, without showing a clear pattern of performance across these categories. This nuanced performance distribution, was also observed in subclass classifications. This likely reflects the limits of DistilBERT's modeling capabilities.

Additionally, our analysis revealed a slight bias of the model towards predicting local over foreign patent applications, as illustrated in Figure 2. Due to the fact that the difference is minimal we are unable to conclude that we are completely certain there is bias and more testing must be done because this could simply be due to an unbalanced training sample.

5 Conclusion

This study embarked on addressing the significant hurdles small businesses and inventors encounter in obtaining patent approvals, through the lens of a language model tailored to predict patent approval outcomes via linguistic analysis. Achieving a notable accuracy of 64.37%, our model outperformed the existing SOTA benchmark of 57.96% accuracy, as established by (Suzgun et al., 2022). Despite this advancement, our model, built upon the BERT architecture (Devlin et al., 2019), has reached a performance ceiling, prompting us to explore the potential of more sophisticated models such as Mistral-7b (Jiang et al., 2023), Gemma-7b, and Gemma-2b (Banks and Warkentin, 2024).

Our journey involved the incorporation of cutting-edge strategies such as LoRA (Hu et al., 2021), model Quantization LoRA (QLoRa) (Dettmers et al., 2023), and model Quantization (Jacob et al., 2017) during the inference phase, coupled with enhancements in data handling and processing. These efforts have substantially mitigated several constraints but have also unveiled new challenges.

As we surpass current SOTA performance milestones, we are not completely satisfied. We plan on continuing this work, assessing our implementation of Mistral7B and further exploration with other more capable models.

Model	Accuracy
H01L (SOTA)	57.96
Ours-Preliminary	60.16
Ours-Final	**64.37**
Ours-Mistral7B	53.70

Table 5: Best performing models. Note all were fine-tuned on the claims section and the models not named are based on Distilbert

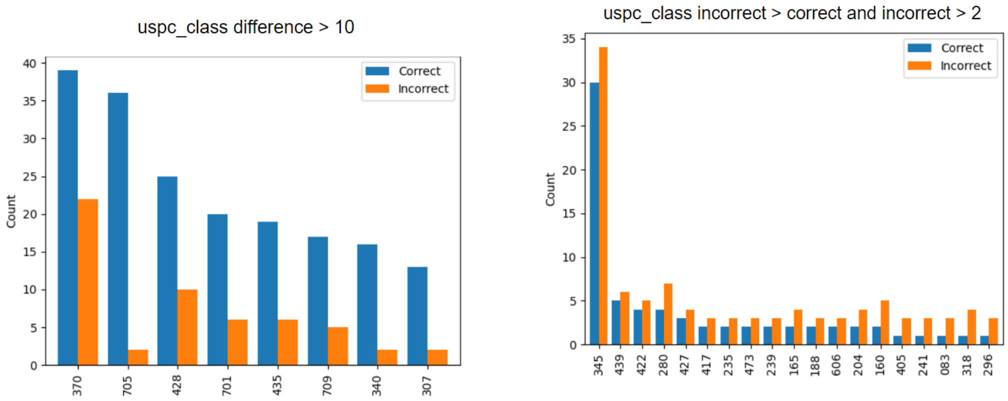

Figure 1: Left: Illustrates the classes that the model predominantly predicts with high accuracy. Right: Depicts the classes where the model encounters challenges in achieving correct predictions, relative to its successful predictions. *Note: The numbers along the x-axis represent various classes deemed by the United States Patent Classification (USPC). To search for the corresponding written class name see: https://www.uspto.gov/web/patents/classification/*

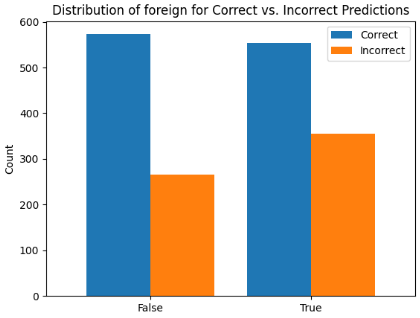

Figure 2: Displays the slight bias in the Distilbert model to local patent applications compared to foreign ones.

References

Jeanine Banks and Tris Warkentin. 2024. Gemma: Introducing new state-of-the-art open models.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2017. Quantization and training of neural networks for efficient integer-arithmetic-only inference.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution.

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

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.