Patentability: Improving Acceptance Prediction of US Patent Applications using Ensemble Modeling

Tommy Bruzzese, Alex Lerner, Oscar O'Rahilly
{tbru, alerner1, oscarfco}@stanford.edu

Problem & Summary of Work

Patents certify and protect new ideas. They are an essential part of modern innovation and are a primary driver of economic growth.

However, current patent processes are inefficient and imprecise:

- 650k annual applications to USPTO, beginning to overwhelm the office
- Only 10% of patent inventors are women; micro-entities and women are meaningfully less likely to have their patents approved

Better acceptance prediction and model understanding can reduce strain on USPTO, save governments money, and reduce bias in granting applications

Our work improves acceptance accuracy and model understanding:

Accuracy Improvement in Patent Acceptance Prediction

For the first time, analyzes **saliency** of Prediction Model

Confirms that **BERT** models cannot yet outperform **Naive Bayes**

Background

Patents are highly-structured. They are also more complex, contextual, and technical than other natural language:

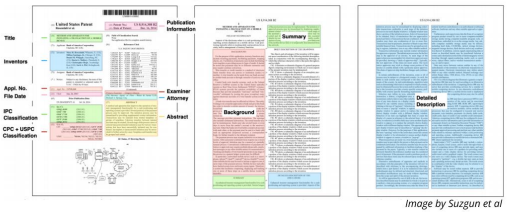


Image by Suzgun et al

Baseline for acceptance accuracy set by Naive Bayes (not BERT models):

- Prior work trained on only one section at a time *(only Abstract, only Claims)*
- State-of-the-art baseline set with Bernoulli Naive Bayes

We train and test on #1 International Patent Classification (IPC) subclass, *G06F: Electric Digital Data Processing*, which constitutes 10.4% of applications

Technical Methods

Naive Bayes (top performing baseline)

- Assumes independence of tokens when classifying

$$p(Acceptance|w_1,\ldots,w_{|V|}) \propto p(Acceptance)\prod_{i=1}^{|V|} p(w_i|Acceptance)$$

Naive Bayes Model

DistilBERT (second-best performing baseline) by Sanh et al. @ HuggingFace

- Lighter, Cheaper, Faster: 40% decrease in model size and 60% faster pretraining, retains 97% of performance of BERT model
- Uses knowledge distillation to minimize loss with "teacher" BERT model

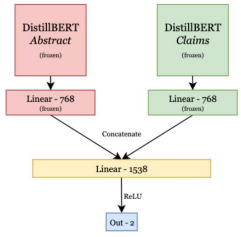
Prior work acknowledges limits of only training on one section

Our Custom Ensemble Architectures

Ensembling is better than training one larger model:

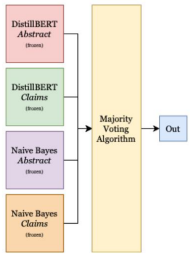
- Abstract and Claims average 1403 tokens combined (too big for BERT)
- Each section is semantically different, preserve nuance
- Can achieve full advantage of different data representations

Ensemble Model #1
DistilBERT: Multiple Patent Sections



Froze weights of both models, including their final linear layer, and learned the weights for the final two layers which were a linear layer and ReLU

Ensemble Model #2
Naive Bayes + DistilBERT: Multiple Sections & Models



Froze weights of the four models, and then used a majority voting algorithm to determine the final classification of the ensemble model

Experiments & Results

Dataset

- Using largest, richest patent dataset: Harvard USPTO Patent Dataset
- Trained on 2011-2013 subset for efficiency — **664569 patent applications**
- Validation set is balanced between rejected/accepted, i.e., true baseline of 50%

Model Evaluation: Overall Decision Accuracy

Bernoulli Naive Bayes	DistilBERT	Ensemble #1	Ensemble #2
Abstract — 60.33	*Abstract — 58.63*	*61.76*	*62.91*
Claims — 62.17	*Claims — 59.59*		

Analyzed saliency to understand how BERT models think

- Use integrated-gradient saliency analysis for DistilBERT baseline
- We look at words that have strong impact on the overall decision prediction
- **Technical words** like *circuit, semiconductor, device* and adjectives that stress **novelty** like *first* have strong **positive impact** in the classification decision
- **Action words** like *introduced, and, method, controls, connected* are **penalized** as they are generic and do not stress the concept's novelty

[CLS] a semiconductor memory device for reducing ripple noise of a back - bias voltage , and a method of driving the semiconductor memory device include a word line driving circuit and a delay logic circuit . the word line driving circuit enables a sub - word line connected to a selected memory cell to a first voltage , and di ##sable ##s the sub - word line of a non - selected memory cell to a second voltage and a third voltage , in response to a sub - word line enable signal , a first word line driving signal , and a second word line driving signal . the delay logic circuit controls the semiconductor memory device so that an amount of charge of the sub - word line that is introduced to the third voltage is greater than an amount of charge of the sub - word line that is introduced to the second voltage by changing a transition point of time of the sub - word line enable signal with respect to a transition point of time of the first word line driving signal , during the di ##sa ##bling of the sub - word line . [SEP]

Confirmed: BERT Models Alone Cannot Yet Outperform Naive Bayes

- We confirm in our subset of models that BERT models underperform Naive Bayes
- BERT models seemingly cannot do more than word-level extraction

Discussion

For the decision classification task, ensemble modeling was a lightweight but powerful improvement on our baseline accuracy. Furthermore, saliency proved to be a useful method for understanding what BERT models learned for prediction. Finally, there still are significant avenues for improvement in the patent domain, with several other tasks that can be explored in future work.

References & Acknowledgements

1. Mirac Suzgun, Luke Melas-Kyriazi, Suproteem Sarkar, Scott Duke Kominers, and Stuart M. Shieber. The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent Applications. (in review), 2022
2. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: Smaller, Faster, Cheaper and Lighter, 2019

Thank you to our great mentors, Mirac Suzgun and Michihiro Yasunaga for their invaluable guidance on our project