A NLP Approach to Understanding Patent Acceptance Criteria

Ryan Kearns; Sauren Khosla; Benjamin Wittenbrink
Stanford University

Abstract

Patent applications and acceptances are a useful domain for assessing the state of innovation across various fields, including biomedical sciences, artificial intelligence, and software services.

Patent filings per year number in the hundreds of thousands (650,000 patent filings in the 2020 fiscal year) and this number has nearly doubled since 2000. Until now, no large-scale corpus of patent filings existed for ML and NLP practitioners to leverage. The Harvard USPTO Patent Dataset (HUPD) is the first example of such a corpus.

Here, we vary the metadata inputs to a number of NLP models to conduct an ablation study on the binary classification of filed patents (i.e. acceptance or rejection). Depending on the metadata inputs to our models, at its best, our model achieves 63.32% accuracy on the binary classification problem.

Data

- We are using the Harvard University Patent Dataset (HUPD),[1] which contains approximately 4.5 million patent applications from 2004-2018.
- Each application contains 20 metadata fields.
- The main text fields are the abstract (average 132 tokens), claims (1,272), background (627), summary (918), and description (11,856).
- Figures 3 & 4 show correlation between metadata and patent acceptance.
- We use a subsample of patents corresponding to "information retrieval," the largest individual category of patents in the data.

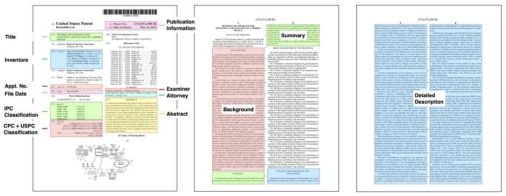

Figure 1. Example of a patent application in the dataset.

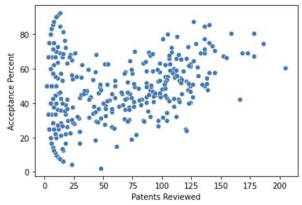

Figure 3. Each patent examiner's acceptance rate, correlated with the number of patents they reviewed.

Methods and Approach

- Our primary approach uses a BERT architecture to encode the patent abstract and claims text stream and combines it with a neural network embedding of the patent metadata.
- We perform an ablation study, introducing additional metadata one-by-one, including the year of filing, the patent examiner, and the IPC code.
- We implement RoBERTa, BERT, DistilBERT, and Longformer[2] language models.
- Figure 2 shows the architecture of our metadata-augmented model.

Results

Model	Batch Size	Validation Accuracy (Abstract)	Validation Accuracy (Claims)
DistilBERT	64	60.82%	61.83%
RoBERTa	32	60.74%	61.76%
Logistic Regression	32	59.31%	57.48%
Naive Bayes	1	61.54%	**64.37%**
Metadata-Augmented DistilBERT	64	**63.08%**	63.32%

Table 1. Comparison of our augmented models to the baseline LMs explored in Suzgun et al.[1]

NLP Model	Metadata	Validation Accuracy (Abstract)	Validation Accuracy (Claims)
DistilBERT	None	60.82%	61.83%
DistilBERT	Examiner ID	62.01%	62.38%
DistilBERT	Examiner ID + Year	62.11%	63.30%
DistilBERT	Examiner ID + Year + Imputed Mean	**63.08%**	**63.32%**
None	Examiner ID + Year	57.59%	57.59%
None	Examiner ID + Year + Imputed Mean	58.32%	58.32%

Table 2. Ablation study showing the incremental effects of our metadata embedding layers.

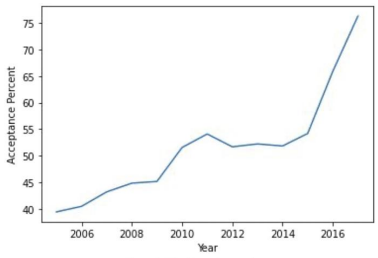

Figure 4. Patent acceptance rate per year.

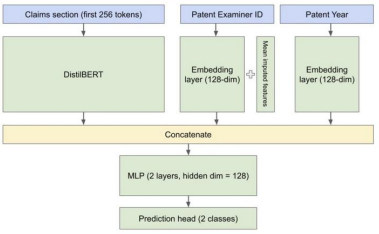

Figure 2. Architecture of our Metadata-Augmented Model

Discussion

Bag-of-words models, trained on the Claims sections of patents, were able to outperform language model-based approaches even with metadata augmentation. These results can be explained by some of the qualitative linguistic features of patent applications:

- Bag-of-words models can outperform pretrained language models in cases where there exists a plethora of technical jargon
- BERT's masking pre-training task does not prepare it well for language of shortened form (i.e. lists and sentence fragments)
- The lengths of many of the data fields are longer than BERT's maximum sequence length of 512. Bag-of-words models, by contrast, can support much larger "sequence" sizes by ignoring sequences altogether and just taking as many words as the maximum vocabulary size allows

Model	Abstract	Claims	Description	Summary	Background
DistilBERT Examiner ID + Year + Imputed Mean	63.08%	**63.32%**	62.46%	61.28%	62.34%

Table 3. Accuracy of DistilBERT with augmented metadata on varying sections of patent applications.

Conclusions

We use the newly created HUPD to fulfill the binary classification task on patent filings. We establish benchmarks for multiple NLP models using metadata from the dataset, taking into account varying data fields as augmentation to increase validation accuracy. We hope that the usage of the dataset in conjunction with these preliminary results will inform future NLP tasks on patent filings and help individuals prepare more well-informed patent applications during the filing process.

References

[1] Mirac Suzgun, Suproteem K Sarkar, Luke Melas-Kyriazi, Scott Kominers, and Stuart Shieber. The Harvard USPTO Patent Dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. 2021.
[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. CoRR, abs/2004.05150, 2020.