Decoding Decisions:
Unveiling Nuances in Patent Classifier Strategies

UC Berkeley W266 NLP Project
Fall 2023

Chi So, Rita Tu

Abstract

This study delves into diverse strategies for training a patent decision classifier, revealing that while all models exhibit similar overall performance, nuances emerge in their emphasis on specific words and their associated word counts. Classical classifiers, including Naive Bayes (NB) and Random Forest (RF), achieved comparable accuracy but showed susceptibility to text noise. The Logistic Regression (LR) classifier, adept at circumventing stop words, fell short of optimal performance. Contrary to expectations, models like BERT, emphasizing attention between words and sentences, did not significantly impact classification outcomes. However, our experiments underscore the potential of incorporating a deep understanding of "innovation" as a crucial feature to elevate the performance of future models.

1 Introduction

Patents are vital as they incentivize innovation by granting exclusive rights, protecting intellectual property, and contributing to economic development and global technological advancement. According to the United States Patent and Trademark Office (USPTO)[1], there were over 700,000 patent applications by October 2023 awaiting approval. The 6-18 months' timeline and \$10,000-25,000 cost associated with completing a single patent application paves way for an opportunity to create a new tool that enhances efficiency and optimization in the patent decision process.

The introduction of the Harvard USPTO Patent Dataset (HUPD)[2], encompassing over 4.5 million patent application documents from 2004 to 2018, marks a significant advancement by addressing this data deficit and offering a substantial resource for analysis. This dataset not only builds upon prior literature that employed natural language processing (NLP) approaches but also opens avenues for leveraging large language models to assess hidden criteria in classification tasks.

Previous works of literature have explored using LLMs to predict patent application acceptance and were not very successful. Our study aims to delve into the underlying reasons for this observed phenomenon, seeking to provide explanations and insights into why the performance of these advanced language models may not surpass that of traditional NLP approaches in certain contexts. This paper will assess the performance of various models, encompassing traditional approaches like Bag of Words and Naive Bayes, alongside advanced models such as BERT.

2 Background

This project drew inspiration from the HUPD[2] and a Stanford CS224N class project[3].

The HUPD introduced the patent decision classification task using text[2]. The paper explored patent acceptance predictions by comparing the Naive Bayes, DistilBERT, BERT, and RoBERTa models. None of the models' accuracy scores went above 64%.

[1] https://www.uspto.gov/dashboard/patents/production-unexamined-filing.html
[2] https://arxiv.org/abs/2207.04043
[3] https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1224/reports/custom_116615529.pdf

The Stanford CS224N class project continued the HUPD study by utilizing the pretrained DistilBERT and RoBERTa models through HuggingFace to explore different metadata provided in the HUPD dataset. The study found that using the abstract field achieved the highest accuracy in their final models. However, even their best performing model, DistilBERT (63%), was not able to surpass the bag-of-words model (64%) in accuracy due to the frequency of technical jargon, legal jargon, and non-grammatical sentences.

Prior research has indicated that incorporating Large Language Models (LLMs) such as BERT and RoBERTa did not yield significant improvements in accuracy compared to classical NLP models like Naive Bayes. This is unexpected, given that LLMs typically outperform Naive Bayes in classification due to their ability to capture attention across words and sentences. This inspired our study to also examine the binary classification task and dive deeper into understanding why the more complex LLMs are underperforming compared to classical models, like Naive Bayes.

3 Method

3.1 Dataset

We utilize the Harvard USPTO Patent Dataset (HUPD), which comprises two main components. The first part encompasses the entirety of the patent application, segmented into distinct sections such as Decision, Abstract, Claims, Summary, Description, and Title. The second part encompasses an additional 34 metadata columns, including the application number. The entire dataset is approximately 500 GB, exceeding the CPU and GPU-RAM limitations on our virtual machines, thus we took steps to whittle down the dataset in our study.

In response to the HUPD model’s decline in performance due to the wide span of time, we narrowed our focus to the period from January 2015 to December 2017 to simplify the problem and ensure substantial progress within our allotted time and resources. We filtered out all patent applications that did not have “ACCEPTED” or “REJECTED” as the decision to make this a binary classification task. We also opted to concentrate on a specific IPC sub-category, A61K, due to its relatively balanced distribution of accepted and rejected patent applications. After whittling down the dataset, we opted to focus on the abstract field due to its short length that allows BERT and RoBERTa to fully tokenize each word and significance in conveying essential patent information.

For our training and validation sets, we defined a training set range spanning from 2015 to 2016, with the validation set to 2017, maintaining an approximate 2:1 ratio. Recognizing the data imbalance between accepted and rejected applications, we employed a weighted random sampler to oversample data for the language model. Additionally, a straightforward application of class weights was used to train the Naive Bayes classifier. This approach was adopted to address the skewed distribution of classes and enhance the model's ability to generalize across different decision labels.

3.2 Natural Language Models

Term Frequency – Inverse Document Frequency (TF-IDF) + NB: Utilizing the TF-IDF vectorizer from the scikit-learn library, the cleaned text data transforms numerical features. This transformation results in a matrix of TF-IDF features, where individual rows correspond to documents and columns signify unique words within the corpus. Subsequently, a Naive Bayes classifier is trained using this TF-IDF transformed data. This model serves as our baseline, aligning with previous studies that have employed a similar approach for their analyses.

BERT and RoBERTa: We utilized the same method for the BERT and RoBERTa models. We used pre-trained bert-base-uncased for BERT and roberta-base for RoBERTa to tokenize the patent abstracts and fed the CLS tokens as inputs into the respective model. We then fine-tuned the model to fit our study. Our final model employs a maximum sequence length of 512, 3 dense layers, hidden_size= [256, 100, 10], and a learning rate of 0.00001.

Bag of Words: Applying a word count tokenizer like the TF-IDF tokenizer, the cleaned text data is transformed into numerical features, producing a matrix for the training set. Subsequently, a classifier is employed to discern the association between word count features and binary class labels. Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF) are the selected classifiers in our study for this purpose. Through the utilization of Bag-of-Words (BoW) models, our objective is to visualize the pivotal words contributing to the patent application. At the end, the list of important words should indicate what matters to a specific language model. be used for each type of text in the manuscript.

3.3 Evaluation

The assessment focused on accuracy scores to gauge the overall model performance and F1 scores for specific classes, providing insights into the model's precision and recall. In the case of TF-IDF and Bag-of-Words (BoW) models, our investigation involved identifying words that significantly influenced the model's decision-making process.

4 Results and Discussion

4.1 Baseline Model

We used the TF-IDF + NB model as our baseline model for this study. The results are detailed in the table below.

Model	Validation Accuracy	Label	F1-score
TF-IDF + NB	0.64	ACCEPTED	0.71
		REJECTED	0.51

The examination reveals that the TF-IDF Tokenizer combined with the Naive Bayes classifier demonstrates a relatively favorable F1-score for the "ACCEPTED" class but exhibits poor performance for the "REJECTED" class, aligning with existing literature.

Taking a deeper dive to understand the words that the TF-IDF+ NB model places more attention to, we examine the top 50 contributors to the positive and negative labels.

Label	Top 50 Contributors
ACCEPTED	['the' 'of' 'and' 'or' 'to' 'in' 'for' 'invention' 'compositions' 'are' 'methods' 'an' 'compounds' 'as' 'present' 'is' 'composition' 'comprising' 'thereof' 'treatment' 'relates' 'such' 'pharmaceutical' 'with' 'by' 'one' 'treating' 'provides' 'provided' 'at' 'that' 'use' 'also' 'subject' 'method' 'cancer' 'disclosed' 'be' 'acid' 'formula' 'diseases' 'using' 'which' 'herein' 'least' 'from' 'compound' 'acceptable' 'agent' 'can']

Label	Top 50 Contributors
REJECTED	['the' 'of' 'and' 'or' 'to' 'in' 'for' 'invention' 'methods' 'an' 'compositions' 'is' 'are' 'present' 'composition' 'comprising' 'as' 'relates' 'with' 'treating' 'treatment' 'thereof' 'one' 'pharmaceutical' 'at' 'such' 'use' 'by' 'method' 'that' 'provides' 'subject' 'compounds' 'least' 'provided' 'cancer' 'disclosed' 'agent' 'from' 'herein' 'acceptable' 'compound' 'also' 'pharmaceutically' 'which' 'using' 'acid' 'administering' 'disease' 'be']

A closer investigation into the model's usage of common words, such as 'the', 'of', and 'and' reveals that these frequently occurring terms act as the top 50 contributors to the positive label, potentially compromising the model's reliability. The reliance on these common stop words may not furnish substantial discriminatory information between positive and negative examples. Additionally, an interesting observation is that leading contributing words for both positive and negative labels share similarities, suggesting a limited contribution to class distinctions. Exploring lesser contributing words did not unveil clear, discernible patterns.

4.2 BERT and RoBERTa Models

We trained BERT and RoBERTa models to predict whether a patent application would be accepted or rejected based on the abstract and compared the results to our base model. The results are detailed below.

Model	Validation Accuracy	Label	F1-score
Bert-base-uncased	0.6	ACCEPTED	0.66
		REJECTED	0.50
Roberta-base	0.32	ACCEPTED	0.00
		REJECTED	0.48

Our BERT model achieved approximately 60% accuracy for abstracts. The BERT model was the only model that demonstrated comparable results to the Naive Bayes base model. Similar to previous works, we were not able to train this model to outperform the base model. We suspect the attention mechanism alone may not be able to conceptualize innovation for patent applications.

Unfortunately, our RoBERTa model was unable to learn to an accuracy above 50%, which is below the performance of flipping a coin. Our DAN model was also unable to learn over 5 epochs and therefore, was omitted in the table above.

4.3 Bag of Word Models

We also examined other Bag of Word models in a similar fashion to our base model and our results are detailed below.

Model	Validation Accuracy	Label	F1-score
Word2vec + CNN	0.64	ACCEPTED	0.73
		REJECTED	0.47
BoW + LR	0.67	ACCEPTED	0.72
		REJECTED	0.49
BoW + NB	0.68	ACCEPTED	0.70
		REJECTED	0.52
BoW + RF	0.67	ACCEPTED	0.75
		REJECTED	0.52

While TF-IDF surpasses simple word counting by incorporating normalization and noise reduction, our baseline model revealed that the suppression of stop words did not yield a substantial improvement in the top contributing list. In our investigation, we employed three bag-of-words models—Logistic Regression, Naive Bayes, and Random Forest—to explore whether the choice of classifier influences the identification of key contributing words. Opting for a word count vector over TF-IDF in this experiment was driven by the desire for simplicity and interpretability, enabling a closer examination of the raw occurrence of terms. Despite comparing all three models, no significant differences in performance emerged. Notably, the performance of BoW + NB did not deteriorate compared to TF-IDF + NB, suggesting that the local frequency of terms provides sufficient information to the classifier without the need for normalization.

Like the analysis we performed for the TF-IDF model, we examine the top 50 contributors to the positive and negative labels. (Appendix B)

While the overall performance of the three models shows a degree of similarity, a noteworthy divergence becomes apparent when scrutinizing the top contributing words. In the case of BoW + NB, the word list mirrors that of TD-IDF, with stop words prominently influencing both labels. Conversely, BoW + LR's top contributing words comprise of common technical terms such as 'efficient,' 'reacting,' and 'hbv,' while rare technical terms display negative feature coefficients. BoW + RF, on the other hand, assigns the highest feature importance to stop words but subsequently devalues rare technical terms, underscoring the classifier's sensitivity to word counts. The findings indicate that only BoW + LR effectively addresses noise in texts. However, it's crucial to acknowledge that noise reduction alone may not necessarily lead to an enhanced model performance.

4.4 Future Work

Despite utilizing only the abstract section, our examination of various classifiers unveiled distinct focuses on text characteristics. Interestingly, the additional capabilities of Large Language Models (LLMs) did not result in a significant performance boost compared to classical models. This suggests that the attention mechanism between words and sentences, a characteristic of LLMs, may not be pivotal for enhancing the patent classification task. To elevate future model performance, it is proposed that the concept of "innovation" be integrated into the model. LLMs can fully leverage the attention mechanism only when evaluating the presence of "innovation" throughout the texts.

Conclusion

In our experiments exploring different approaches to training a patent decision classifier, we observed that while all models yielded similar overall performance, they exhibited variations in their focus on specific words and their respective word counts. Classical classifiers such as Naive Bayes (NB) and Random Forest (RF) achieved comparable accuracy to other models but were susceptible to noise in the texts. The LR classifier, while capable of avoiding the consideration of stop words as important features, did not perform optimally. Models like BERT, with its attention-focused approach between words and sentences, did not demonstrate a significant impact on the classification model's performance. However, the crucial insight from these experiments suggests that developing a model with a deep understanding of "innovation" could serve as a pivotal additional feature to enhance the performance of future models.

Acknowledgments

This document has been adapted by Rita Tu and Chi So from the template for earlier ACL, EMNLP and NAACL proceedings, including those for EACL 2023 by Isabelle Augenstein and Andreas Vlachos and EMNLP 2022 by Yue Zhang, Ryan Cotterell and Lea Fremann.

References

USPTO. 2023. Patents Production, Unexamined Inventory and Filings Data October 2023.

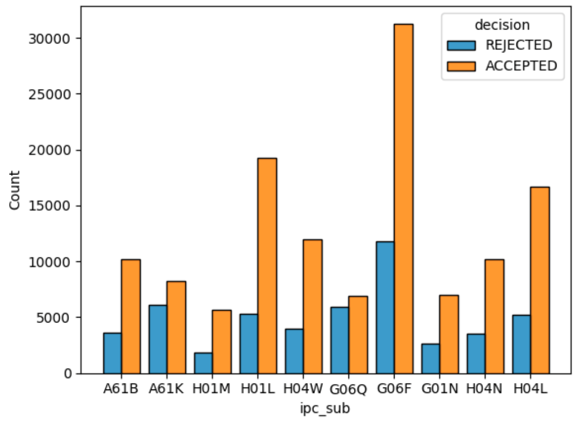
Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K. Sarkar, Scott Duke Kominers, Stuart M. Shieber. 2022. The Harvard USPTO Patent Dataset: A Large-Scale, Well-Structured, and Multi-Purpose Corpus of Patent Applications. *Computer Science Repository*, arXiv:2207.04043.

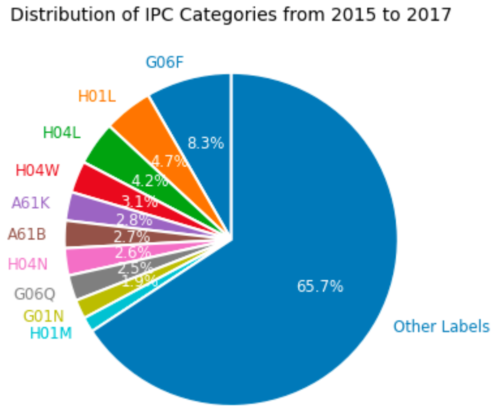
Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Ryan Kearns, Sauren Khosla, Benjamin Wittenbrink 2022. A NLP Approach to Understanding Patent Acceptance Criteria. Stanford CS224N.

A Dataset Selection

The depicted figure below illustrates the distribution of various IPC categories from 2015 to 2017 dataset, highlighting the top ten categories. Additionally, the decision labels within the data include not only "ACCEPTED" and "REJECTED" but also other statuses like "PENDING" and multiple "CONT-" statuses. To prevent duplication of patent applications, we opted to focus on applications with “ACCEPTED” or “REJECTED” statuses. The presented data reveals an imbalance across different IPC sub-categories.





B Top 50 Contributors

As mentioned in the results section, we examined the top 50 contributors for positive and negative labels.

Model	Label	Top 50 Contributors
BoW + LR	ACCEPTED	['efficient' 'reacting' 'hbv' 'airway' 'sheet' 'psoriasis' 'pancreatitis' 'hydrogels' 'sensitive' 'retention' 'capture' 'disintegrating' 'dyeing' 'bis' 'cation' 'd3' 'ecm' 'saccharides' 'short' '75' 'muscarinic' 'oligomer' 'near' 'polyvinylpyrrolidone' 'citrate' 'consensus' 'found' 'infective' 'quality' 'edible' 'viable' 'comprised' 'eclampsia' 'psma' 'cdk' 'cream' 'vaccination' 'restoring' 'envelope' 'coli' 'continuity' 'glyceryl' 'specification' 'therewith' 'expressing' 'neisseria' 'oxidizing' 'hormones' 'dihydromyricetin' 'ascorbate']
	REJECTED	['laquinimod' 'triglycerides' 'dpp' 'solubilizing' 'permeation' 'hydroxyalkyl' 'variables' 'pyrithione' 'pidotimod' 'microns' 'disintegrant' 'anesthesia' 'aerosol' 'cooh' 'glioma'

Model	Label	Top 50 Contributors
		'malignancy' 'neoplastic' 'rapamycin' 'stimulators' 'scarring' 'percent' '001' 'borrelia' '80' 'fluorine' 'nanomicelles' 'possess' 'hard' 'bifidobacteria' 'predicting' 'cyclopropyl' 'rtk' 'paclitaxel' 'kilogram' 'adsorbent' 'aspirin' 'unique' 'inositol' 'regimen' 'glycosyl' 'techniques' 'healthy' 'followed' 'adjuvants' 'reactions' 'due' 'leukocyte' 'compliance' 'shown' 'bivalirudin']
BoW+NB	ACCEPTED	['the' 'of' 'and' 'or' 'to' 'in' 'for' 'an' 'invention' 'is' 'are' 'methods' 'as' 'present' 'compositions' 'comprising' 'composition' 'with' 'by' 'one' 'such' 'thereof' 'at' 'treatment' 'relates' 'that' 'compounds' 'also' 'pharmaceutical' 'treating' 'from' 'be' 'method' 'provided' 'least' 'use' 'which' 'provides' 'acid' 'can' 'subject' 'agent' 'more' 'disclosed' 'wherein' 'cancer' 'including' 'having' 'containing' 'using']
	REJECTED	['the' 'of' 'and' 'or' 'to' 'in' 'for' 'invention' 'an' 'is' 'methods' 'are' 'as' 'present' 'with' 'compositions' 'comprising' 'composition' 'one' 'relates' 'by' 'at' 'treating' 'treatment' 'that' 'thereof' 'such' 'method' 'from' 'pharmaceutical' 'least' 'also' 'use' 'provides' 'provided' 'be' 'subject' 'which' 'agent' 'more' 'disclosed' 'acid' 'cancer' 'compounds' 'wherein' 'administering' 'cells' 'active' 'disease' 'herein']

Model	Label	Top 50 Contributors
BoW + RF	ACCEPTED	['the' 'and' 'of' 'or' 'to' 'in' 'for' 'compounds' 'are' 'an' 'present' 'methods' 'invention' 'comprising' 'compositions' 'is' 'as' 'pharmaceutical' 'with' 'composition' 'treatment' 'thereof' 'such' 'treating' 'relates' 'one' 'provided' 'provides' 'by' 'use' 'method' 'that' 'disclosed' 'herein' 'at' 'cancer' 'also' 'which' 'diseases' 'formula' 'acceptable' 'subject' 'from' 'agent' 'least' 'can' 'be' 'acid' 'including' 'disease']
	REJECTED	['ωsubstitution' 'emanating' 'ellipsoid' 'similarity' 'simmondsia' 'electroporated' 'electrophysiological' 'electromechanical' 'sinemet' 'siliqua' 'electrochemical' 'elaterium' 'sinter' 'sinuses' 'eighty' 'eighth' 'eighteen' 'eigengene' 'eicosan' 'elderberry' 'siliconate' 'silicified' 'embraces' 'sibiricum' 'encasing' 'sibo' 'sided' 'sides' 'sig' 'ena' 'emulsifies' 'signatures' 'sii3' 'emulsfiable' 'silanes' 'emphysema' 'silibinin' 'emm' 'emitted' 'emetogenic' 'embryoid' 'embrittlement' 'eicosa' 'shrinking' 'eicos' 'sir53' 'slc23a2' 'easing' 'easiness' 'easier']

357