IP Detective: Patent infringement detection using BERT

Buyang Li, Gokul Harindranath, Hrohaan Malhotra, Jonathan Mathai, Lakshay Vohra, Puja Gupta, Matthew A. Lanham

Purdue University, Krannert School of Management

i3947@purdue.edu;gharindr@purdue.edu; hmalhot@purdue.edu; mathaij@purdue.edu; lvohra@purdue.edu; gupta714@purdue.edu; lanhamm@purdue.edu

ABSTRACT

Patents play a significant part in innovation and help individuals and companies safeguard and retain ownership of their ideas. However, patent infringement is common, and more than 2,500 patent infringement suits are filed each year. Currently, patent infringement detection is largely done manually, and companies spend approximately $600 to identify each case of infringement. Our work provides an approach to automate this process through machine learning. Our model first vectorizes patent text using a BERT model trained on patent text, and then calculates similarity scores between competing patent claims. The overall score is then calculated by taking a weighted average of the subsection similarities, where the weights were calculated by training a logistic regression model based on historical cases of infringement. Looking at subsection scores along with the overall score, we can identify potential infringement of two competing patent claims rather accurately.

INTRODUCTION

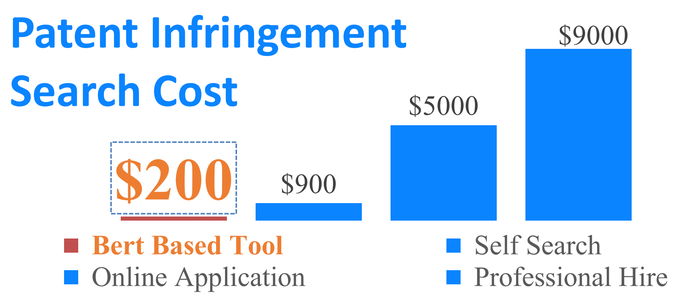
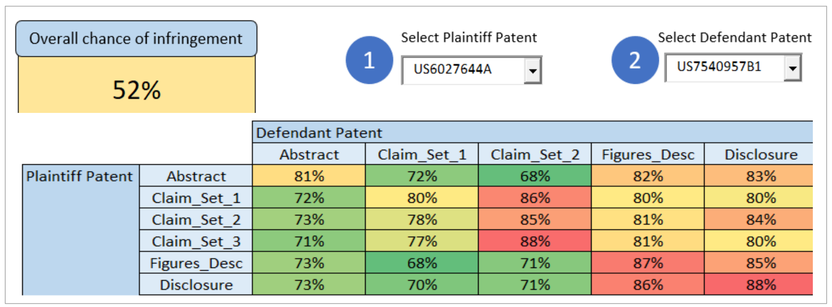


Fig 1. Patent Infringement search cost using different methods

Research Questions:

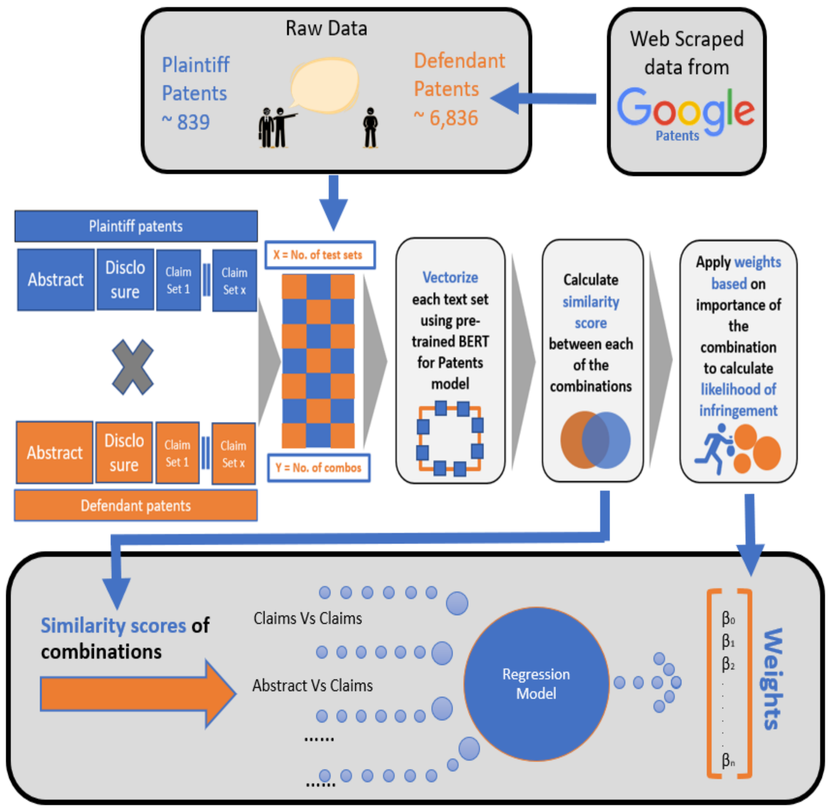
1. What are the most important components of a patent application to be considered while examining an infringement?
2. What is the appropriate methodology to score similarities between two patents?



		Defendant Patent				
		Abstract	Claim_Set_1	Claim_Set_2	Figures_Desc	Disclosure
Plaintiff Patent	Abstract	81%	72%	68%	82%	83%
	Claim_Set_1	72%	80%	86%	80%	80%
	Claim_Set_2	73%	78%	85%	81%	84%
	Claim_Set_3	71%	77%	88%	81%	80%
	Figures_Desc	73%	68%	71%	87%	85%
	Disclosure	73%	70%	71%	86%	88%

Fig 4. A UI to visualize patent infringement

METHODOLOGY



MODEL CHALLENGES

512 Max # of characters consumed by BERT. We divided the text into blocks of 512 and chose the highest similarity for the similarity matrix

Dynamic Features Number of features were different for each patent so we used the claim set with the highest similarity score for our logistic regression

EXPECTED BUSINESS IMPACT

This tool will help cut costs and save time by directing employee focus onto evaluating flagged patents rather than parsing through every patent manually. With the improved section-by-section infringement score, an expert can quickly narrow down and procure relevant evidence to present, should a business choose to enter litigation. The tool is expected to reduce patent search costs by approximately 77%.

STATISTICAL RESULTS

Metric Name	Value
AUC	0.813
F1 Score	0.760
Accuracy	0.759
Precision	0.756
Recall	0.812

Fig 2. Model Results

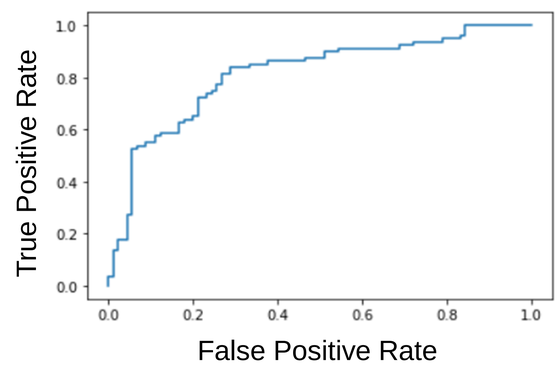


Fig 3. AUC Curve

CONCLUSIONS

- Similarity scores between plaintiff claim and defendant disclosure have, on an average, 50% higher impact on probability of infringement than any other scores
- The similarity scores on vectors, generated from the individual section text, using the BERT for patents provides great results

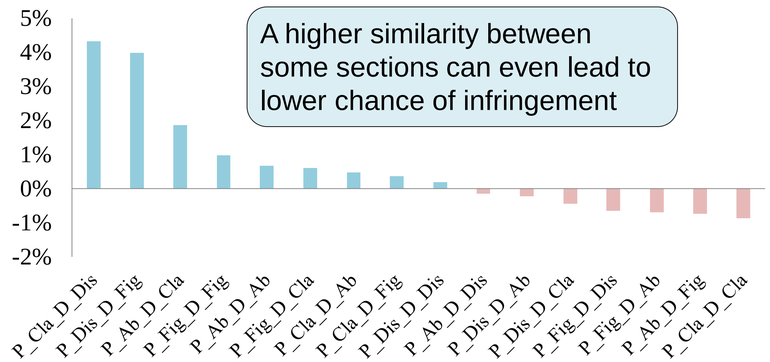


Fig 5. Percentage change in overall chance of infringement for one percent increase in similarity

ACKNOWLEDGEMENTS

We would like to thank Professor Matthew Lanham, Justin Beyers, and Geoffrey Lentner for their guidance and support on this project.

Learn More!

Share your feedback!