CUSTOM ENTITY EXTRACTION: DOCUMENT PROCESSING TAILORED TO CUSTOMER DATA

EFFECTS OF PROPOSED CHANGES TO COPYRIGHT LAW IN RESPONSE TO GENERATIVE ARTIFICIAL INTELLIGENCE

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Madelyn Khoury

October 27, 2023

Technical Team Members: None

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Rider Foley, Department of Engineering and Society

Briana Morrison, Department of Computer Science

Introduction

Around 80% of all business data is in an unstructured format, such as in PDFs, emails, or documents (Automation Anywhere, 2023). Because such data doesn't conform to a standard format – the way that data in a database does, for example – a different method must be used to process each type of information, making aggregation difficult (ibid). Additionally, some of this data (particularly that from PDFs) is not interpretable by computers, which means it cannot be used in any application in which a computer would need to understand the content of the data. To work with unstructured data, a human employee must manually convert it into a structured format, which is time intensive, tedious, and prone to human error. To avoid this issue, many companies have begun to use intelligent document processing (IDP), a way of automatically extracting structured information from documents that does not require human labor (Martínez-Rojas et al., 2023). Instead, IDP depends on machine learning models, or computer programs that undergo a training process to learn how to complete a task, to extract information (ibid). Figure 1 shows a typical situation in which IDP might be used: after documents are uploaded, IDP extracts information from them, a human can optionally verify the results of IDP, and the results are written to a database (Appian, 2023a).

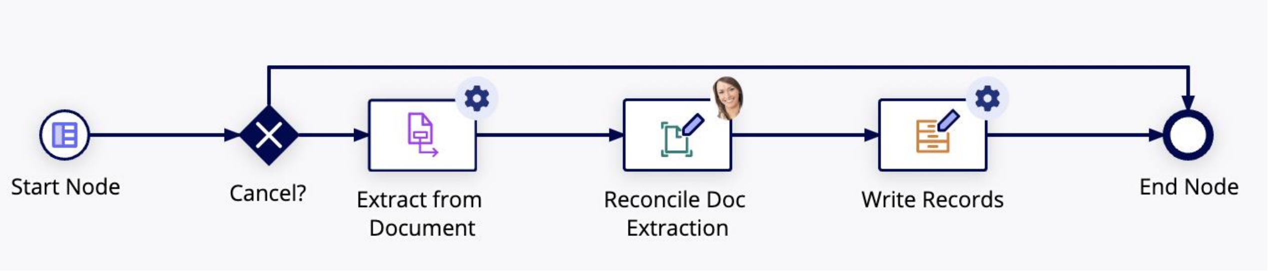
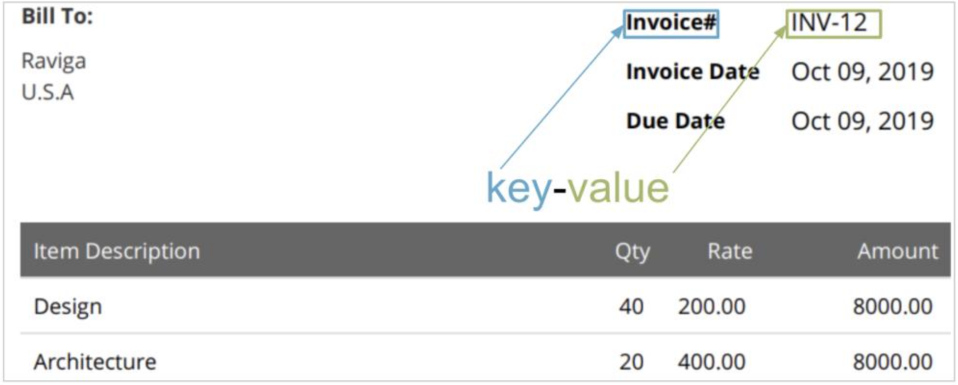


Figure 1. Typical usage of intelligent document processing (Source: Appian, 2023a).

Appian, a producer of enterprise software, provides its customers with this kind of IDP capability. Soon, Appian hopes to improve its IDP accuracy, particularly on documents that are uniquely formatted. To this end, Appian has considered the tradeoff between training a single, general purpose model and training a unique model for each customer, a tradeoff that many software companies consider (Hadar, 2021). General models may make use of more training data, finding overall patterns among all customers' data and performing well as a result (Mazzanti, 2023). However, training with inaccurate data decreases model performance. Therefore, when one customer's data differs from most of the data used to train the general model – as is the case with Appian's clients who have unique documents – a specialized model trained exclusively on that customer's data may be the best option (Hadar, 2021). This paper focuses on a modification to Appian's IDP system that aims to provide better accuracy and reduced need for human verification by using a specialized model for each customer.

Custom Machine Learning Model for Entity Extraction

This modification to Appian's IDP system is called the "custom entity extraction" feature by Appian's engineering department. To understand what custom entity extraction entails, one must first understand Appian's current IDP system, which aims to extract "key-value pairs" from documents, or the name of a piece of information coupled with the piece of information itself. Figure 2 shows one such key-value pair on an actual document.



Bill To:

Raviga
U.S.A

Invoice#	INV-12
Invoice Date	Oct 09, 2019
Due Date	Oct 09, 2019

Item Description	Qty	Rate	Amount
Design	40	200.00	8000.00
Architecture	20	400.00	8000.00

Figure 2. A key-value pair to be extracted from a PDF with IDP (Source: Appian, 2023b).

When documents are uploaded, they are first passed to a model which identifies their text contents. Next, the text is passed to a model that has learned which piece of text corresponds to each piece of information in which users are interested. This type of model, called an entity extraction (EE) model, pulls out the relevant information from the documents and stores it in key-value pairs. See Figure 3 for a diagram description of this process. Using a trained model to perform a task in this way is called "inference".

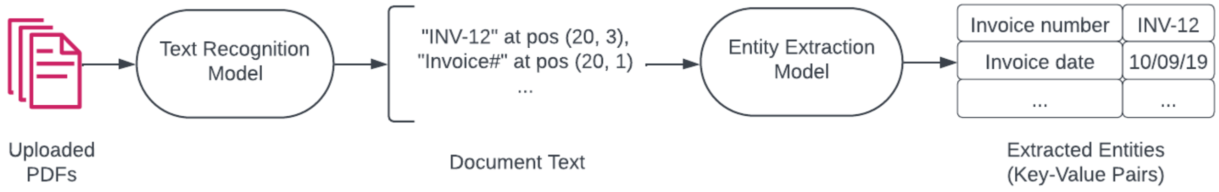


Figure 3. Appian's Current IDP System.

This technical project involved replacing the EE model in the current IDP system with a specialized model for *each* customer. Each of Appian's customers will have the opportunity to train an EE model on their most used documents, tailoring the model to their own data. However, this project was not focused on training custom models for each customer, but rather performing inference with the already-trained models.

The process of performing inference with a machine learning model involves several steps, each of which I implemented in its own program. In addition to writing these programs, I also used a Python framework called Kedro to combine all the steps into a single pipeline that passed information from one program to the next. My team elected to use Kedro because it allows for easy loading in and exporting of data (Alam et al., 2023). Figure 4 shows the full pipeline, which includes a step to load in a custom trained model. Thus, unlike the current system in Figure 3, the new IDP system will not use the same model every time it is used to perform inference.

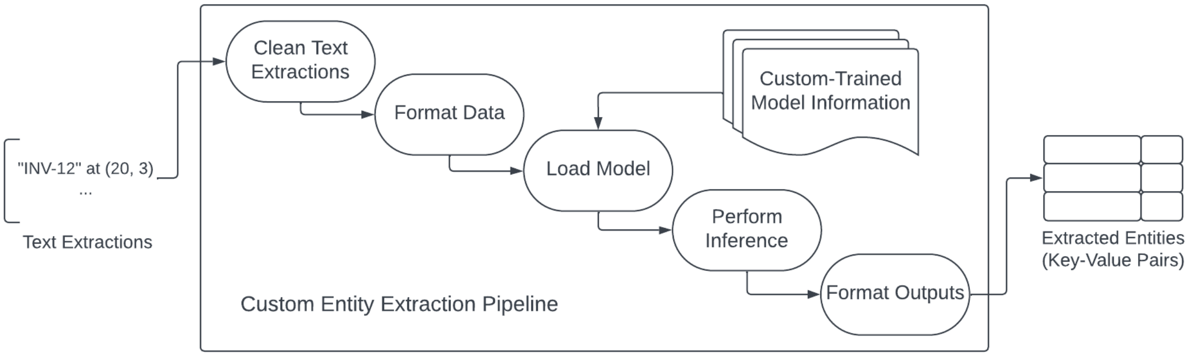


Figure 4. Custom entity extraction pipeline

Not only did this project involve accessing a separate model for each customer, but it also involved using a different *type* of EE model altogether. The current EE model that Appian uses was produced internally at Appian. By contrast, custom EE models will all be trained instances of the Bert Relying on Spatiality (BROS) model (Hong et al., 2022), which I streamlined and refactored as part of my project. With the BROS model's excellent performance in preliminary tests and with the performance boost of custom-trained models, the custom entity extraction feature is expected to improve the accuracy of Appian's IDP system.

Interpretive Flexibility of Generative AI

In my technical project, the data fed into each customer model belonged to the customer itself. In contrast, many public models scrape the web to get their training data, which creates privacy and intellectual property considerations (Argento, 2023). This is especially true for generative artificial intelligence (AI), where the model creates new content. As generative AI has become more prevalent in the last year, individuals, academics, and media sources have engaged in discourse around the dangers it poses (Tredinnick & Laybats, 2023). A particularly polarizing topic is generative AI that creates images, or image-generative AI.

After conducting sentiment analysis on Tweets from the last four years, Miyazaki et al. (2023) found that most professions view image-generative AI favorably. Some graphic designers are open to working alongside it (Weiming et al., 2023), but artists and illustrators take a strong negative stance on the topic. In fact, a group of artists sued several companies that make image-generative AI in *Andersen et al. v. StabilityAI, Midjourney, DeviantArt*, claiming that the AI infringed on their copyright because it used their art as training data without permission (UNESCO, 2023). The AI companies, of course, hold a different perspective.

The Social Construction of Technology (SCOT) framework is useful for analyzing these perspectives and how image-generative AI interacts with humans and social attitudes. SCOT, created by Pinch & Bijker (1987), emphasizes interpretive flexibility, or the idea that a technological artifact will be viewed differently by different people. Particularly, SCOT recognizes that a set of individuals who all share a similar problem, called a social group, will view an artifact through the same lens (ibid).

Creators of generative AI may be split into a couple social groups. Some want ownership over their models' outputs: for example, Stephen Thaler, the CEO of Imagination Engines Incorporated, has attempted to register the products of his AI system, "Creativity Machine", as copyrighted to himself (Feldman, 2022). Conversely, some AI producers have undertaken initiatives to give credit to the original human creators of the data used to train AI systems. The creators of the Stack, a training dataset for models that write computer programs, have devised methods of acknowledging human authors of code that is similar to model outputs (Li et al., 2023).

Users of image-generative AI have diverse perspectives, too. First, copyright attitudes in the U.S. have tended towards desire of access to information for practical uses (Sinnreich et al., 2021). Many users seem untroubled about the copyright concerns for image-generative AI (Miyazaki et al., 2023), perhaps because they believe that AI models must have access to information in order to provide practical results. On the other hand, some users may be hesitant to use generative AI because they might be found liable for any copyright infringement that the model commits (Zirpoli, 2023). Yet another group of users may view their use of image-generative AI as a creative endeavor, comparing the process of creating prompts with the "trial and error" process of traditional art, and thus believe they should have some ownership over the model's outputs (Hayes, 2023).

Finally, academics and policymakers express their own interpretations of image-generative AI. Abbott and Rothman (2022) argue that the purpose of copyright is to protect the public good, and that attributing ownership to AI systems will benefit the public good the most. Meanwhile, policymakers have already proposed several bills that will allow people to opt out of their data being scraped by generative AI models (Busch, 2023).

Research Question and Methods

There are multiple social groups with vested interest in how copyright law that pertains to generative AI will soon change, and each group advocates for different changes to occur. These conflicting attitudes prompt me to ask: how will proposed changes to copyright law due to generative AI impact various social groups? I will answer this question by conducting a policy analysis in the style of Bardach & Patashnik (2023). However, I will execute my analysis with the goal of determining effects on social groups rather than deciding upon the best policy option, as Bardach and Patashnik's method intends (ibid). I will list the policy alternatives, specifying which social groups argue for each proposed policy, before making an outcomes matrix to record the expected outcomes of each proposal. To populate this matrix, I will conduct email or phone interviews of experts in the fields of art, copyright, and artificial intelligence. My list of potential interviewees includes James Hutson, a professor of art history at Lindenwood University who has written about generative AI for creating art, officials at the U.S Copyright Office, and Christopher Zirpoli, a UVA alumnus that studies copyright law in his role as a legislative attorney. Further, I will create surveys to be distributed broadly with the goal of collecting responses from people belonging to the social groups in question. The survey will first consist of a self-identification section, which will collect demographic information as well as the relevant social groups that the respondent belongs to. Then, the survey will ask respondents to predict the outcomes of – and the actions they personally would take as a result of – each of the proposed policies. I will use natural language processing or manual methods to group similar responses to each question, then synthesize these responses to construct a prediction of the outcome of each policy. The questions regarding a respondent's actions will allow me to use Bardach and Patashnik's (2023) "Other-guy's-shoes Heuristic" to imagine unintended consequences of the policies. I will then use SCOT to examine how the potential outcomes will differently affect various social groups.

Conclusion

To reap the benefits of IDP – such as reduced need for human input – it is important that IDP accurately extract information from documents. Thus, this technical project involved implementing Appian's new custom entity extraction feature, which allows each client to use a machine learning model trained on their unique documents. Custom-trained models are anticipated to provide more accurate IDP results.

Additionally, this paper synthesized several stakeholder groups and their desires to change U.S. copyright law in response to generative AI. My STS research project will provide insight on how social groups will be affected by proposed changes to copyright law.

References

Abbott, R., & Rothman, E. (2022). Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence. *Florida Law Review, Forthcoming*. https://doi.org/10.2139/ssrn.4185327

Alam, S., Chan, N. L., Couto, L., Dada, Y., Danov, I., Datta, D., DeBold, T., Gudaniya, J., Holzer, J., Kaiser, S., Kanchwala, R., Katiyar, A., Kumar Pilla, R., Koh, A., Mackay, A., Merali, A., Milne, A., Nguyen, H., Nikolic, V., … Theisen, M. (2023). *Kedro* (0.18.12) [Computer software]. https://github.com/kedro-org/kedro

Appian. (2023a). *Build a Doc Extraction Process with AI Skill*. Appian Corporation. https://docs.appian.com/suite/help/23.3/doc-extraction-tutorial.html

Appian. (2023b). *Document Extraction Suite*. Appian Corporation. https://docs.appian.com/suite/help/23.3/Appian_Doc_Extraction.htm

Argento, Z. (2023, August 9). Data protection issues for employers to consider when using generative AI. *The Privacy Advisor*. https://iapp.org/news/a/data-protection-issues-for-employers-to-consider-when-using-generative-ai/

Automation Anywhere. (2023). *What is Intelligent Document Processing - (IDP)? | Automation Anywhere*. https://www.automationanywhere.com/rpa/intelligent-document-processing

Bardach, E., & Patashnik, E. M. (2023). *A Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving*. CQ Press.

Busch, K. (2023). *Generative Artificial Intelligence and Data Privacy: A Primer* (Congressional Report R47569). Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R47569

Feldman, J. (2022, April 22). The art of artificial intelligence: A recent copyright law development. *Reuters*. https://www.reuters.com/legal/legalindustry/art-artificial-intelligence-recent-copyright-law-development-2022-04-22/

Hadar, Y. (2021, August 18). *Should I train a model for each customer or use one model for all of my customers?* Medium. https://towardsdatascience.com/should-i-train-a-model-for-each-customer-or-use-one-model-for-all-of-my-customers-f9e8734d991

Hayes, C. M. (2023). *Generative Artificial Intelligence and Copyright: Both Sides of the Black Box* (ssrn:4517799). SSRN. https://doi.org/10.2139/ssrn.4517799

Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., & Park, S. (2022). *BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents* (arXiv:2108.04539). arXiv. https://doi.org/10.48550/arXiv.2108.04539

Li, R., Allal, L. B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Akiki, C., Li, J., Chim, J., Liu, Q., Zheltonozhskii, E., Zhuo, T. Y., Wang, T., Dehaene, O., Davaadorj, M., Lamy-Poirier, J., Monteiro, J., Shliazhko, O., … de Vries, H. (2023). *StarCoder: May the source be with you!* (arXiv:2305.06161). arXiv. https://doi.org/10.48550/arXiv.2305.06161

Martínez-Rojas, A., López-Carnicer, J. M., González-Enríquez, J., Jiménez-Ramírez, A., & Sánchez-Oliva, J. M. (2023). Intelligent Document Processing in End-to-End RPA Contexts: A Systematic Literature Review. In S. Bhattacharyya, J. S. Banerjee, & D. De (Eds.), *Confluence of Artificial Intelligence and Robotic Process Automation* (pp. 95–131). Springer Nature. https://doi.org/10.1007/978-981-19-8296-5_5

Mazzanti, S. (2023, September 15). *What Is Better: One General Model or Many Specialized Models?* Medium. https://towardsdatascience.com/what-is-better-one-general-model-or-many-specialized-models-9500d9f8751d

Miyazaki, K., Murayama, T., Uchiba, T., An, J., & Kwak, H. (2023, May 16). *Public Perception of Generative AI on Twitter: An Empirical Study Based on Occupation and Usage*. arXiv.Org. https://arxiv.org/abs/2305.09537v1

Pinch, T., & Bijker, W. (1987). The Social Construction of Facts and Artifacts. In B. Wiebe, T. Hughes, & T. Pinch (Eds.), *The Social Construction of Technological Systems* (pp. 17–50). MIT Press.

Sinnreich, A., Aufderheide, P., Clifford, M., & Shahin, S. (2021). Access shrugged: The decline of the copyleft and the rise of utilitarian openness. *New Media & Society*, *23*(12), 3466–3490. https://doi.org/10.1177/1461444820957304

Tredinnick, L., & Laybats, C. (2023). The dangers of generative artificial intelligence. *Business Information Review*, *40*(2), 46–48. https://doi.org/10.1177/02663821231183756

UNESCO. (2023, July 20). *Navigating intellectual property rights in the era of generative AI: The crucial role of educating judicial actors | UNESCO*. https://www.unesco.org/en/articles/navigating-intellectual-property-rights-era-generative-ai-crucial-role-educating-judicial-actors

Weiming, L., Zhoa, Y., Wang, G., & Yang, L. (2023). Exploration of AI Aided Design Path for Graphic Designers. In F. Rebelo & Z. Wang (Eds.), *Ergonomics In Design* (Vol. 77, pp. 288–296). AHFE International.

Zirpoli, C. (2023). *Generative Artificial Intelligence and Copyright Law* (Congressional Report LSB10922). Congressional Research Service. https://crsreports.congress.gov/product/pdf/LSB/LSB10922/6