An Empirical Study on Transfer Learning for Privilege Review

Haozhen Zhao
Data & Technology
Ankura Consulting Group, LLC
Washington DC, USA
haozhen.zhao@ankura.com

Shi Ye
Data & Technology
Ankura Consulting Group, LLC
Washington DC, USA
shi.ye@ankura.com

Jingchao Yang
Data & Technology
Ankura Consulting Group, LLC
Washington DC, USA
jingchao.yang@ankura.com

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9672008

***Abstract*— Protecting privileged communications and data from inadvertent disclosure is a paramount task in the US legal practice. Traditionally counsels rely on keyword searching and manual review to identify privileged documents in cases. As data volumes increase, this approach becomes less and less defensible in costs. Machine learning methods have been used in identifying privilege documents. Given the generalizable nature of privilege in legal cases, we hypothesize that transfer learning can capitalize knowledge learned from existing labeled data to identify privilege documents without requiring labeling new training data. In this paper, we study both traditional machine learning models and deep learning models based on BERT for privilege document classification tasks in legal document review, and we examine the effectiveness of transfer learning in privilege model on three real world datasets with privilege labels. Our results show that BERT model outperforms the industry standard logistic regression algorithm and transfer learning models can achieve decent performance on datasets in same or close domains.**

Keywords*— *privilege review, predictive coding, deep learning, BERT

I. Introduction

In the United States civil litigation process, a party has the right to withhold certain documents from court's production request on the basis that they contain privileged content involving communications between attorneys and clients with the purpose of obtaining legal advice. Protection of privileged information from inadvertent disclosure is a critical component of the US legal system [1]. To safeguard this information, parties are often engaged in expensive processes to review documents for production [2]. Review costs continue to rise as the number of documents that need to be reviewed in cases only gets larger and larger based on what we see in projects in the industry over the years. The high stakes involved in privilege review makes it the most expensive part of document review. The legal market has resorted to leveraging supervised machine learning approaches, or called predictive coding in legal community, to automate/semi-automate the document review process. Existing research shows that predictive coding has the potential to substantially reduce review costs [3,4].

The classic supervised machine learning approach requires significant amount of training examples to perform well in tasks. To address this bottleneck, transfer learning is a new learning paradigm that aims to leverage data from existing domains to train models that can generalize to the task at hand [5]. This is especially appealing to the privilege review task in that not only would it reduce the review costs and facilitate the document review workflow, but also that the concept of privilege may be generalizable across cases in a sense, as the definition of privileged circumstances are well defined in law.

In this paper, we empirically study three kinds of machine learning models in privilege prediction and investigate the transferability of models trained on different datasets using different machine learning methods. Main contributions of this paper include experimenting transferred models in predictive coding and conducting experiments with datasets with true labels instead on simulated labels.

The paper is organized as follows: Section II we review existing research related to machine learning for legal document review and transfer learning. Section III presents the research questions. Section IV describes the datasets used, the machine learning methods and experimental protocol, and evaluation measures. Section V discusses the results and Second VI concludes the papers with remarks on future work.

II. Related Work

A. Machine Learning for Privilege Review

Though cost pressure has driven the eDiscovery industry to adopt predictive coding solutions in cases, there is not many published research on using predict coding for privilege review. Gabriel et al. is among the first to discuss the use of predictive coding in finding privileged information in documents and showed that machine learning technology could bring significant improvement on efficiency and reliability to the privilege review workflow [6]. Gronvall et al. compared the traditional keyword search method and predictive coding approach in identifying privileged documents in cases and presented an approach to distinguish effective keywords from less efficient keywords [7]. Chhatwal et al. combined keyword search and Convolutional Neural Networks through training and predicting on potential privileged hits to identify privileged documents [8]. Oard et al. proposed a comprehensive framework called MINECORE (stands for "MINimizing the Expected COst of REview") which combines manual review and automatic classification to minimize both review costs and potential misclassification costs [9]. They experimented with a simulated data set and used surrogates for responsiveness and privilege labels.

B. Transfer Learning in Deep Learning

Over the past decade, as deep neural networks learning becomes the most popular machine learning paradigm, researchers discovered a common interesting phenomenon in the layers: first layers tend to learn features that can be generalized across domains and last layers are features more specific to the task at hand [10]. This observation paves the way to generalize very deep convolutional neural network image recognition model trained on very large image sets to image analytics applications [11]. Team in [12] has successfully taken the transfer learning approach to build an image clustering application on top of VGG16 for the legal domain. In 2018, Google made publicly available Bidirectional Encoder Representations from Transformers (BERT), a large transformers-based language representation model with 110 million parameters and pretrained on massive texts [13]. BERT immediately acquired much traction in both academia and industry due to its effectiveness in transferring knowledge learned in general domain to specific task, replicating the achievements we have seen in computer vision to natural language processing [14]. Mokrii et al. recently evaluated the transferability of BERT-based neural ranking models in question answering tasks [15]. Chalkidis et al. studied the performance of BERT in zero shot and few shot transfer learning problems [16].

III. Research questions

This paper studies leveraging transfer learning in predictive coding for privilege review. Below are the research questions:

1. How do traditional machine learning methods such Logistic Regression (LR) perform comparing to BERT-based neural networks models in identifying privileged documents?

2. How effective are pretrained machine learning models based on traditional LR, document embeddings (Doc2Vec) [18] or BERT in predicting privileged documents in a zero shot settings, i.e. the models are trained on data fully outside of the evaluation set?

IV. Experiments

We have conducted experiments on three datasets.

A. Datasets

The only publicly available evaluation dataset for privilege review is the one used in TREC 2010 Legal Track, which however has reliability issue [17]. Following our previous studies on predictive coding for privilege review [7][8], we experiment on three confidential and non-public datasets from real matters. Dataset A, and B are from two different matters related to the telecommunication industry. Dataset C is from a matter in the healthcare industry. All the documents were reviewed and coded on their privilege status by attorneys. Of the three datasets, we excluded documents with extracted text of over 1MB in size and filtered out duplicated documents. This results in three datasets of which the population size ranges from 202,235 to 566,475, and richness from 3.8% to 19.6% (Table I).

TABLE I. Datasets Statistics

Dataset	#Documents	#Privileged	#Not Privileged	Richness
A	202,235	39,653	162,582	19.6%
B	346,506	13,130	333,376	3.8%
C	566,475	100,202	466,273	17.7%

For the experiments carried out in this paper, we randomly sample a 30% subset from each of the three datasets and randomly split 70% for training and 30% for evaluation for the three subsets. Table II gives the privileged and not privileged document counts in both the training and testing splits for each dataset.

TABLE II. Experimental Dataset Splits

Dataset	Split	#Privileged	#Not Privileged	#Total
A	Train	8,340	34,129	42,469
	Test	3,566	14,635	18,201
B	Train	2,689	70,077	72,776
	Test	1,193	29,993	31,186
C	Train	21,110	97,849	118,959
	Test	9,098	41,885	50,983

B. Methodology

In our experimental protocol, we employ three machine learning algorithms: Logistic Regression with TFIDF representation of the data (LR-TFIDF), Logistic Regression with Doc2Vec representation of the data (LR-Doc2Vec) and fully connected classifier layer on BERT output (FC-BERT). For each algorithm, we train a model on the each of training splits in Table II and evaluate the performance on all the testing splits.

LR-TFIDF: Logistic regression is the current state of the art text classification algorithm in the eDiscovery industry. It's also the most effective text classification algorithm according to empirical evaluations on real world datasets [4]. In our experiment, we use the liblinear implementation of the logistic regression as packaged in the Scikit Learn library. We use the default settings and use TFIDF term weighting scheme in the classification pipeline.

LR-Doc2Vec: Doc2Vec is distributed representation of documents based on word embeddings [18] that has the potential to encode semantic meanings of the corpus. Three Doc2Vec model are trained respectively using all the data of the datasets in Table I before the privilege experiments. We used 200 as the dimension of the embedding vectors, set a max vocabulary size as 500,000 and excluded terms appeared only once. The model is trained for 40 epochs. Training of Doc2Vec models is completely unsupervised, without using the privilege labels. Then in the experiments, we converted both the training documents and testing documents into embedding vectors using the pretrained Doc2Vec models, then use those vectors to train the LR model and run the prediction. We hypothesize that the transferability of models is encoded in the Doc2Vec models trained using the underlying data.

FC-BERT: We use the 'bert-base-uncased' model for the BERT setup and set dropout rate as 0.1, then take the top-level representation and feed into a dense layer with a sigmoid activation to produce the predicted labels. BERT is fine-tuned using the training datasets. Then the fine-tuned models are used to predict the labels of the testing set of each dataset. We hypothesize that BERT from general domain fine-tuned using in domain data, can be used to encode the knowledge of privilege in documents.

In addition to the modeling and prediction within each dataset, we also use models trained on one dataset's training set to predict the labels of other two dataset's test set. This is to investigate to what extend models can be transferred across domains or matters.

We use precision, recall, F1 score and precision at 75% recall (P75R) as the measures to evaluate the results of different methods. Precision is the percentage of truly privileged documents in all "predicted" privileged documents. Recall is the percentage of "predicted" truly privileged documents out of all the truly privileged documents in the testing set. Precision and recall only characterize one aspect of the model performance, F1 score is a balanced measure to characterize model performance in a whole, which is the harmonic mean of precision and recall. In eDiscovery practice, oftentimes parties negotiate a mutually acceptable recall level in the case, which is generally 75%. Therefore, precision at 75% recall is a commonly used measure of predictive coding model performance.

V. Results and Discussion

We first analyze the performance of the three methods within each dataset, with the purpose to examine if LR-Doc2Vec and FC-BERT outperform the industrial benchmark algorithm—LR-TFIDF.

TABLE III. PERFORMANCES OF DIFFERENT METHODS WITHIN THE SAME DATASET

Dataset	Method	Precision	Recall	F1	P@75C
A	LR-TFIDF	87%	67%	76%	82%
	LR-Doc2Vec	73%	51%	60%	56%
	FC-BERT	82%	82%	**82%**	88%
B	LR-TFIDF	70%	9%	16%	17%
	LR-Doc2Vec	42%	3%	5%	9%
	FC-BERT	55%	21%	**30%**	12%
C	LR-TFIDF	85%	66%	74%	79%
	LR-Doc2Vec	71%	46%	55%	52%
	FC-BERT	83%	74%	**78%**	82%

Table III shows the results of the different methods on the three datasets and Figure 1 plots the precision and recall of LR-TFIDF and FC-BERT (omitting LR-Doc2Vec for readability). Results show that LR-TFIDF achieved best precision, but FC-BERT model achieved best recall and generally outperforms LR-TFIDF in terms of F1 score. In terms of precision at 75% recall, FC-BERT outperforms LR-TFIDF in datasets A and C, except in the lower richness dataset B where LR-TFIDF outperforms FC-BERT. In all the cases, LR-Doc2Vec is less effective than LR-TFIDF and FC-BERT. Overall, FC-BERT method is the most effective classification algorithm for these datasets, slightly better in most of the balanced measures than LR-TFIDF with difference less than or around 5%.

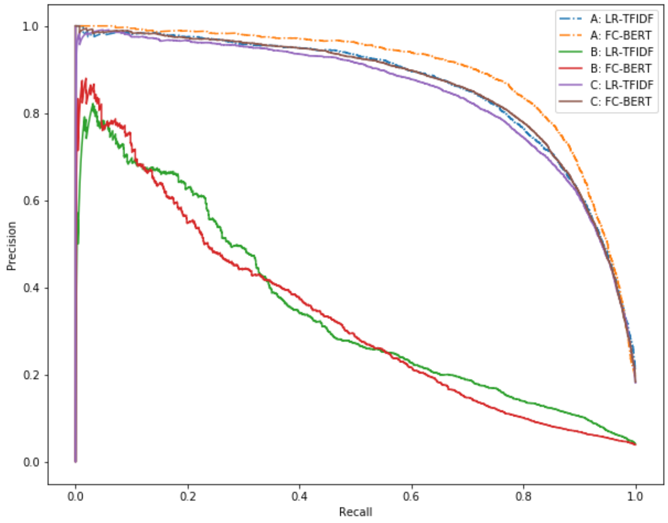

Fig. 1. Precision-Recall of LR-TFIDF and FC-BERT on Dataset A, B and C

We then study how transferable are the different models using each of the three datasets as the test datasets. We select the best model (according to F1 Score) from above as the benchmark in below comparisons.

TABLE IV. PERFORMANCE OF TRANSFERRED MODELS ON TEST DATA OF DATASET A

Method	Train Dataset	Precision	Recall	F1	P75R
Benchmark	A	82%	82%	**82%**	88%
LR-TFIDF	B	91%	14%	24%	59%
LR-Doc2Vec	B	77%	4%	8%	44%
FC-BERT	B	92%	31%	47%	65%
LR-TFIDF	C	88%	8%	15%	44%
LR-Doc2Vec	C	58%	19%	29%	39%
FC-BERT	C	64%	23%	34%	38%

TABLE V. PERFORMANCE OF TRANSFERRED MODELS ON TEST DATA OF DATASET B

Method	Train Dataset	Precision	Recall	F1	P75R
Benchmark	B	55%	21%	**30%**	12%
LR-TFIDF	A	54%	9%	15%	11%
LR-Doc2Vec	A	28%	16%	21%	8%
FC-BERT	A	32%	17%	22%	8%
LR-TFIDF	C	58%	4%	8%	8%
LR-Doc2Vec	C	23%	10%	14%	7%
FC-BERT	C	22%	6%	9%	7%

TABLE VI. PERFORMANCE OF TRANSFERRED MODELS ON TEST DATA OF DATASET C

Method	Train Dataset	Precision	Recall	F1	P75R
Benchmark	C	83%	74%	**78%**	82%
LR-TFIDF	A	77%	0%	1%	26%
LR-Doc2Vec	A	53%	3%	5%	26%
FC-BERT	A	52%	1%	2%	20%
LR-TFIDF	B	82%	0%	0%	19%
LR-Doc2Vec	B	49%	1%	1%	29%
FC-BERT	B	64%	1%	2%	19%

Table IV, V and VI show the results of transferring models on test data of new datasets. Figure 2 plots the precision and recall of the best performing zero shot transferred model of each dataset comparing with the benchmark model. For all three datasets, models trained on different datasets underperform the benchmark model which was trained on the test data from the same dataset, which is as expected. However, for Dataset A and Dataset B, transferred models perform quite well. For Dataset A the best performed transferred model is FC-BERT trained on Dataset B, which achieves 65% precision at 75% recall on A. For Dataset B, LR-TFIDF trained on A achieves 11% precision at 75% recall, close to the benchmark model's 12%.

For Dataset C, transferred models all perform badly, far behind the benchmark model. In this extreme case, LR-Doc2Vec surprisingly performs better than LR-TFIDF and FC-BERT.

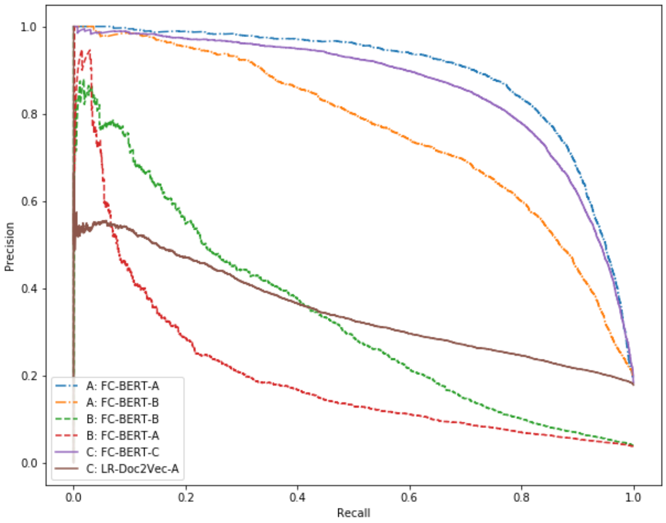

Fig. 2. Precision and Recall of best performing zero shot models and the bechmark models on Dataset A, B and C

One aspect of the results is that there seems no transitivity of transferred models. For example, models trained on training data of C perform much better on test data of A and B than that of models trained on training data of A or B on test data of C. This may be due to the size of the training data, as training set of Dataset C is larger than the training data of A or B.

The results show that domain plays a critical role in the transferability of models. Since Dataset A and Dataset B are from the same industry, i.e., telecommunications, models trained on one of them generally perform well on the other. As Dataset C is from a completely different industry, healthcare, models trained on A or B do not perform well on C.

VI. CONCLUSIONS AND FUTURE WORK

This paper empirically studies the transferability of both traditional machine learning models and deep learning models on predicting privileged documents in legal document review. We have found that that BERT-based models with fine-tuning outperform logistic regression with TFIDF document representations according to F1 score. Zero shot transferred models though underperform natively trained models, can still achieve satisfactory performance if the target dataset is from a close domain.

The above conclusions are based on the utilized portion of the three datasets. In future, we plan to test with more data points and other available data sets. For the BERT-based models, we only add a single linear layer on top of BERT. It would also be worthwhile to explore with more sophisticated architecture, such as CNN or LSTM, etc. Also due to BERT restricting the input to only 512 Word Pieces, we currently have truncated long documents to only use the first 512 Word Pieces. More principled ways to deal with long documents in Transformers models will be explored in the next steps.

References

[1] R. Keeling, N. Huber-Fliflet, J. Zhang, and R. P. Chhatwal, "Separating the Privileged Wheat from the Chaff – Using Text Analytics and Machine Learning to Protect Attorney-Client Privilege," *Richmond Journal of Law and Technology*, 2019, Accessed: Apr. 08, 2019. [Online]. Available: https://jolt.richmond.edu/separating-the-privileged-wheat-from-the-chaff-using-text-analytics-and-machine-learning-to-protect-attorney-client-privilege/.

[2] N. M. Pace and L. Zakaras, *Where the money goes: Understanding litigant expenditures for producing electronic discovery*. RAND Corporation, 2012.

[3] G. V. Cormack and M. R. Grossman, "Evaluation of Machine-learning Protocols for Technology-assisted Review in Electronic Discovery," in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, New York, NY, USA, 2014, pp. 153–162, doi: 10.1145/2600428.2609601.

[4] R. P. Chhatwal, N. Huber-Fliflet, R. Keeling, J. Zhang, and H. Zhao, "Empirical evaluations of preprocessing parameters' impact on predictive coding's effectiveness," in *Big Data (Big Data), 2016 IEEE International Conference* on, 2016, pp. 1394–1401.

[5] S. Ruder, M. Peters, S. Swayamdipta, and T. Wolf, "Transfer Learning in Natural Language Processing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 2019, pp. 15--18, [Online]. Available: https://aclanthology.org/N19-5004.pdf.

[6] M. Gabriel, C. Paskach, and D. Sharpe, "The challenge and promise of predictive coding for privilege," in *Proceedings of the DESI V Workshop at ICAIL 2013*, 2013.

[7] P. Gronvall, N. Huber-Fliflet, J. Zhang, R. Keeling, R. Neary, and H. Zhao, "An Empirical Study of the Application of Machine Learning and Keyword Terms Methodologies to Privilege-Document Review Projects in Legal Matters," in *2018 IEEE International Conference on Big Data (Big Data)*, Dec. 2018, pp. 3282–3291, doi: 10.1109/BigData.2018.8621945.

[8] R. Chhatwal, R. Keeling, P. Gronvall, N. Huber-Fliflet, J. Zhang, and H. Zhao, "CNN Application in Detection of Privileged Documents in Legal Document Review," in *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, Dec. 2020, pp. 1485–1492, doi: 10.1109/BigData50022.2020.9378182.

[9] D. W. Oard, F. Sebastiani, and J. K. Vinjumur, "Jointly Minimizing the Expected Costs of Review for Responsiveness and Privilege in E-Discovery," *ACM Trans. Inf. Syst.*, vol. 37, no. 1, p. 11:1–11:35, Nov. 2018, doi: 10.1145/3268928.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *arXiv:1411.1792 [cs]*, Nov. 2014, Accessed: Nov. 19, 2021. [Online]. Available: http://arxiv.org/abs/1411.1792.

[11] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, Accessed: Nov. 20, 2021. [Online]. Available: http://arxiv.org/abs/1409.1556.

[12] N. Huber-fliflet, F. Wei, H. Zhao, H. Qin, S. Ye, and A. Tsang, "Image Analytics for Legal Document Review : A Transfer Learning Approach," in *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 4325–4328, doi: 10.1109/BigData47090.2019.9006551.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs], Oct. 2018, Accessed: Dec. 01, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805.

[14] A. Rogers, O. Kovaleva, and A. Rumshisky, "A Primer in BERTology: What we know about how BERT works," *arXiv:2002.12327 [cs]*, Nov. 2020, Accessed: Nov. 20, 2021. [Online]. Available: http://arxiv.org/abs/2002.12327.

[15] I. Mokrii, L. Boytsov, and P. Braslavski, "A Systematic Evaluation of Transfer Learning and Pseudo-labeling with BERT-based Ranking Models," *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2081–2085, Jul. 2021, doi: 10.1145/3404835.3463093.

[16] I. Chalkidis, M. Fergadiotis, S. Kotitsas, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels," *arXiv:2010.01653 [cs]*, Oct. 2020, Accessed: Oct. 15, 2021. [Online]. Available: http://arxiv.org/abs/2010.01653.

[17] Vinjumur, Jyothi K., et al. "Assessing the Reliability and Reusability of an E-Discovery Privilege Test Collection." *Proceedings of the 37th International ACM SIGIR Conference on Research \&\#38; Development in Information Retrieval*, 2014, pp. 1047--1050.

[18] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," presented at the *ICML 2014*, 2014.