Automated Identification of Deontic Modalities in Software Engineering Contracts: A Domain Adaptation-based Generative Approach

Gokul Rejithkumar
TCS Research, India
gokul.rejithkumar@tcs.com

Preethu Rose Anish
TCS Research, India
preethu.rose@tcs.com

Smita Ghaisas
TCS Research, India
smita.ghaisas@tcs.com

***Abstract*—Contracts are legally binding agreements between parties that establish rights, obligations, and terms for their business relationship. They articulate the deontic modalities (Obligations, Permissions, Prohibitions, and Exclusions) that apply to those involved in the contractual agreement. Deontic modalities can be leveraged to effectively elicit Software Engineering (SE) requirements. In this paper, we propose a novel approach for identifying deontic modalities from contracts, employing a combination of text generation and domain adaptation techniques. Among the SOTA approaches we experimented with, the T5-large model yielded the best results, with an average precision and recall of 0.96 and 0.95, respectively. Comparing our approach with previous methods, the results show that our approach handles significant class imbalances in the training data and demonstrates good generalization ability on new datasets.**

Keywords—contracts, regulations, text generation, t5, domain adaptation, transfer learning, deep learning, natural language processing

I. INTRODUCTION

Contracts are an integral part of business and legal transactions. They outline expectations, protect rights and interests, allocate responsibilities and obligations, and provide a framework for resolving disputes and seeking legal remedies between the contractual signatories [1]. Deontic modalities are words or phrases that are often used in contracts to specify the rights and responsibilities of the parties involved. Contracts contain deontic modalities such as *Obligations*, *Prohibitions*, *Permissions*, *Exclusions*, *Facts,* and *Definitions* [10]. Deontic modalities can be leveraged to effectively elicit Software Engineering (SE) requirements. Modalities such as *Obligations* and *Permissions* can be used to discern the high-level functional and non-functional requirements [2]. Also, modalities such as *Prohibition* translate into business rules and organizational policies [10]. Thus, the identification of deontic modalities from contracts can complement the Requirements Engineering (RE) phase of the Software Development Life Cycle (SDLC).

Contracts are large documents and use legalese-like language, which can be intricate and ambiguous. Identification of deontic modalities from contracts is therefore a challenging task. For example, the term *"must"* can be utilized to convey both an *Obligation* and, in certain instances, even a *Prohibition* or *Exclusion*. Limited work has been reported on the automated identification of deontic modalities in contracts. Existing works are based on rules, transformer-based classification models, or a combination of both. In [3, 4] the authors proposed a rule-based method for the identification of different modal expressions. In [5], the authors proposed a syntax-based and logic-based method for the extraction of rules from contracts. In [6-8], the authors proposed machine learning and deep learning-based methods for the extraction of modal expressions. In [9], the authors showcased the application of transfer learning using the RoBERTa model on agreements to extract deontic modalities. Identification of deontic modalities in contracts presents an additional challenge stemming from the restricted availability of contract data, owing to its confidential nature. As a result, the application of domain adaptation from a related source becomes particularly significant in this context [10].

In our prior work [10], we employed a domain adaptation-based discriminative approach with the BERT model to automate the identification and classification of deontic modalities in SE contracts, encompassing *Obligations*, *Prohibitions,* and *Permissions.* We leveraged the fact that regulations and contracts share the same taxonomy, and both contain deontic modalities. Therefore, we utilized the publicly available regulations data as the source dataset and contracts data as the target dataset. We achieved an average precision and recall of 0.90 and 0.89 respectively for classification of deontic modalities. We posited in [10] that the taxonomy of regulations and contracts encompass *Obligations*, *Prohibitions*, *Permissions*, and *Exclusions*; which are vital from a RE perspective. In [10], we did not report the results of *Exclusions* due to the limited availability of training data. However, since *Exclusions* are crucial to comprehensively capture deontic modalities; in this work, we include the *Exclusions* modality as well, despite the limited availability of training data.

An *Exclusion* can be defined as a statement in the legal contract that specifies certain conditions and the absence of permission required for an actor to perform certain actions, as well as the absence of any explicit obligations or prohibitions imposed on the actor regarding that action. To illustrate, consider the following examples of *Exclusions* in regulations and contracts; In regulations, an instance of an *Exclusion* would be – *A health plan may not mandate a healthcare provider, who has already been assigned an NPI*

(National Provider Identifier), to acquire an additional NPI. In contracts, an example of an Exclusion would be – *Where enforcement is provided by the password protected screen saver, Application/XYZ enforcement is not required.*

In recent times, text generation has emerged as a prominent area of research and application in the field of Natural Language Processing (NLP). Based on our experiments, generative models such as T5, GPT-2, and PEGASUS demonstrate a better ability to capture the intricate nuances of contractual language, including syntax, semantics, and context when compared to discriminative models such as BERT. Building upon our previous work on the automated detection of deontic modalities in SE contracts [10], in this work we further enhance our prior approach using a text generation and domain adaptation-based approach for the automated identification of deontic modalities in SE contracts.

The contributions of our paper are three-fold: (1) We model the classification problem as a domain adaptation-based text generation problem and experiment with multiple text generation models such as T5, PEGASUS, and GPT-2. We apply this approach to both the source and target datasets as in our previous work [10] to facilitate a comparison with the previous study that utilized the BERT model.; (2) We automate the extraction of an additional modality – *Exclusions*. (3) We evaluate the text generation-based approach and prior approach on a completely new dataset that is larger and from different domains than the dataset used in our previous study [10], and thus demonstrate that the proposed text generation-based approach has more generalization ability.

The findings of our experiments indicate that the T5-large model, with the text-to-text generation approach, gave the best results and effectively addressed the significant class imbalances and limited training data. Furthermore, a text generation-based approach demonstrates more generalizability across a wide range of data variations compared to our previous discriminative approach using the BERT model.

The remainder of the paper is structured as follows: Section 2 discusses the automated identification of deontic modalities; Section 3 presents the results and Section 4 concludes the paper.

II. Automated Identification of Deontic Modalities

In this section, we present the details of our dataset and the experiments.

A. Dataset Description

Our source regulations dataset and target contracts datasets are derived from our previous work [10], with the difference being the inclusion of *Exclusions*. We refer to this target contracts dataset as Dataset-A. Additionally, in this paper, we also evaluate our approach on a completely new and larger target dataset of contracts from different domains to check the generalizability of the discriminative and generative approaches. We refer to this new dataset as Dataset-B. Fig. 1 illustrates the frequency distribution plot of

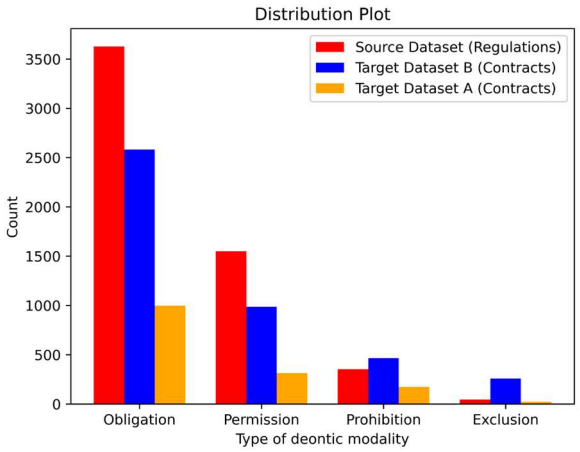

Fig. 1. Distribution plot - deontic modality types

the deontic modality types in source dataset, target domain contracts Dataset-A and target domain contracts Dataset-B.

The source domain dataset of regulations was prepared from the following regulations: Gramm Leach Bliley Act, Health Information Portability and Accountability Act, MiFID Directives, MiFID Regulation, Data Protection Act, General Data Protection Regulation, Anti-Money Laundering UK, Australia Prudential Regulation Authority, Banking Regulation Act, Banking Special Provision Act. The source domain dataset consists of 3628 *Obligation* sentences, 1549 *Permission* sentences, 353 *Prohibition* sentences, and 45 *Exclusion* sentences.

The target domain dataset was annotated by Subject Matter Experts (SMEs) in our organization. Both target domain dataset of contracts consists of real-life SE contractual clauses. The target domain contracts Dataset-A consists of SE contractual clauses from nine application domains, namely healthcare, automotive, finance, banking, pharmaceuticals, telecom, clothing, retail and supermarket. It consists of 998 *Obligation* sentences, 314 *Permission* sentences, 174 *Prohibition* sentences, and 23 *Exclusion* sentences. The target domain contracts Dataset-B consists of real-life SE contractual clauses from two new application domains, namely aerospace and logistics. It consists of 2581 *Obligation* sentences, 986 *Permission* sentences, 464 *Prohibition* sentences, and 258 *Exclusion* sentences.

B. Identification of Deontic Modalities

We formulate the classification task of deontic modalities as a text generation task. We explore multiple Pre-trained Language Models (PLMs) that have shown SOTA performance for text generation tasks, namely T5-large [11], PEGASUS-large [12], and GPT-2-large [13]. Fig. 2 depicts a basic schematic for identification of the deontic modalities.

T5 (Text-to-Text Transfer Transformer) is a transformer-based language model developed for multiple downstream tasks, including text summarization, natural language inference, translation, and more. The specific version of the model we used was *t5-large* [14], which has 770M parameters and was pre-trained on the C4 corpus.

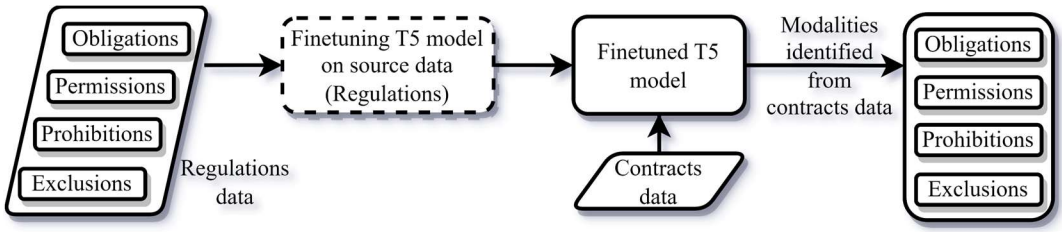

Fig. 2. Basic schematic for automated identification of deontic modalities in SE contracts

PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) is a transformer-based language model specifically designed for abstractive summarization. We utilized the *google/pegasus-large* [15] version of the model, which has 568M parameters and was pre-trained on C4 and the Huge News dataset. GPT-2 (Generative Pre-trained Transformer 2) is a transformer-based language model developed by OpenAI for text generation. It is pre-trained on a large corpus known as WebText. The specific version of the model we experimented with was *gpt2-large* [16].

We fine-tune the PLMs on the source domain dataset of regulations. We experiment with batch size ∈ {2, 4}, number of epochs ∈ {2, 3, 4, 5, 6, 7, 8}, and learning rate ∈ {3e-5, 4e-5}. The fine-tuning and inference of the PLMs were performed on an NVIDIA Tesla V100 GPU with 32 GiB of memory.

The T5-large model took about 454 minutes to fine-tune on the regulations dataset and performed better than PEGASUS-large and GPT-2-large models when evaluated on the contractual target datasets–Dataset-A and Dataset-B. TABLE I presents the hyperparameter values of the T5-large model. We discuss the results of the classification in Section 3.

TABLE I. HYPERPARAMETER VALUES OF T5-LARGE MODEL

Hyperparameters	Values
learning_rate	3e-5
weight_decay	0.01
train_batch_size	2
num_train_epochs	8
max_length	512

III. RESULTS AND DISCUSSION

Our experiments indicate that the T5-large model performed better than the PEGASUS-large and GPT-2-large models, in terms of achieving high F1-scores on both Dataset-A and Dataset-B. We report the results of the T5-large model in TABLE II, while also including the results of our previous work [10] that used the BERT model, to enable an easy comparison.

We obtained an average precision and recall (macro average) of 0.96 and 0.95, respectively, in the classification of the deontic modalities on target domain contracts Dataset-B. On the target domain contracts Dataset-A, we obtained an average precision and recall of 0.95 and 0.94 respectively. Inference on the T5-large model took about 0.8 seconds per statement on average. Fig. 3 depicts the combined classification confusion matrix on both Dataset-A and Dataset-B.

Based on our results in TABLE II, it is evident that the previous BERT-based approach does not generalize well to new datasets, as indicated by an average precision of 0.84 and an average recall of 0.75 on Dataset-B. In contrast, the new generative approach demonstrates good generalizability, achieving an average precision of 0.96 and an average recall of 0.95 on Dataset-B. Additionally, the presence of a significant class imbalance between the modalities in the training data and the limited amount of training data has no discernible effect on the obtained results.

It is also to be noted that the generative models occasionally generated output that was indecipherable. In the case of T5-large model, out of the total 5798 instances evaluated, such occurrences were observed in only about 19 instances (0.003%). Upon manual inspection, it was determined that most of these instances belonged to the modality of *Exclusion*. The exact cause of this issue cannot be explained; however, we surmise that the limited availability of training data in the *Exclusion* class contributed to this issue.

Lately, there has been a surge in new language models such as GPT-J, LLaMA, and other instruction fine-tuned models such as Dolly 2.0, Vicuna, Alpaca, and more. Because of the significant computational resources required

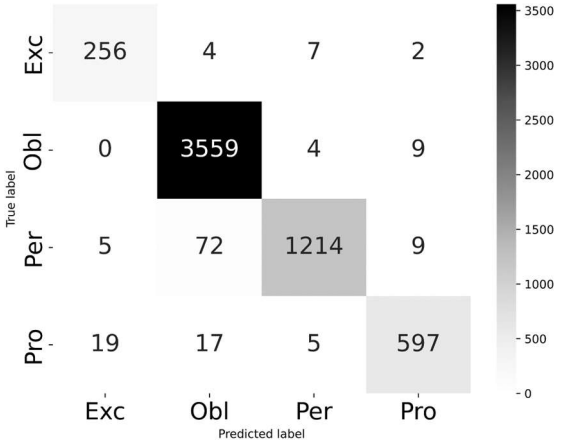

Fig. 3. Confusion matrix (Combined on Dataset-A and Dataset-B)

TABLE II. RESULTS OF MODALITY CLASSIFICATION

Dataset	Modality Type	Precision		Recall		F1-Score	
		T5-large	BERT	T5-large	BERT	T5-large	BERT
A	**Obligation**	0.96	0.94	0.99	0.97	**0.98**	0.95
	Permission	0.98	0.91	0.87	0.81	**0.92**	0.85
	Prohibition	0.96	0.85	0.95	0.91	**0.96**	0.87
	Exclusion	0.92	-	0.95	-	**0.94**	-
B	**Obligation**	0.98	0.93	0.99	0.98	**0.99**	0.95
	Permission	0.98	0.88	0.96	0.90	**0.97**	0.89
	Prohibition	0.97	0.82	0.93	0.77	**0.95**	0.79
	Exclusion	0.91	0.72	0.91	0.35	**0.91**	0.47

for fine-tuning these models, we couldn't perform full model fine-tuning of these models. However, we fine-tuned LLaMA-7B [18] and GPT-J-6B [19] using Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA) method [17]. Despite our efforts, the results obtained through LoRA-based fine-tuning were unsatisfactory.

IV. CONCLUSION

In this paper, we address the challenge of automating the identification and classification of deontic modalities in Software Engineering (SE) contracts. We enhance our prior approach by employing text generation and domain adaptation-based techniques. We experiment with different text generation models, including T5, PEGASUS, and GPT-2, and find that the T5-large model using the text-to-text generation approach yields the best results. We evaluate the text-generation-based approach and our prior approach on a new, larger dataset and find that the text-generation-based approach outperforms our previous approach in terms of both F1 scores and generalizability to a new dataset. Furthermore, we incorporate an additional modality that is important from an SE requirement perspective–*Exclusions*.

In future work, we plan to explore instruction fine-tuning on models such as LLaMA, GPT-J, and Falcon [20] to extract and provide more detailed insights into deontic modalities in SE contracts. Although our ideas in this direction are still in the nascent stages, we believe it holds great promise for further advancements in the field.

REFERENCES

[1] M. Curtotti and E. C. McCreath, "Corpus Based Classification of Text in Australian Contracts," Econometrics: Computer Programs & Software eJournal, 2010.

[2] A. Sainani, P. R. Anish, V. Joshi and S. Ghaisas, "Extracting and Classifying Requirements from Software Engineering Contracts,"2020 IEEE 28th International Requirements Engineering Conference (RE), Zurich, Switzerland, 2020, pp. 147-157, doi: 10.1109/RE48521.2020.00026.

[3] E. Ash, J. Jacobs, B. MacLeod, S. Naidu and D. Stammbach, "Unsupervised Extraction of Workplace Rights and Duties from Collective Bargaining Agreements," 2020 International Conference on Data Mining Workshops (ICDMW), Sorrento, Italy, 2020, pp. 766-774, doi: 10.1109/ICDMW51313.2020.00112.

[4] W. P. Peters and A. Wyner, "Legal Text Interpretation: Identifying Hohfeldian Relations from Text," In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia. European Language Resources Association (ELRA). pp. 379–384, 2016.

[5] M. Dragoni, S. Villata, W. Rizzi, and G. Governatori, "Combining natural language processing approaches for rule extraction from legal documents," Lecture Notes in Computer Science, pp. 287–300, 2018. doi:10.1007/978-3-030-00178-0_19

[6] J. O. Neill, P. Buitelaar, C. Robin, and L. O. Brien, "Classifying sentential modality in legal language," Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law, 2017. doi:10.1145/3086512.3086528

[7] I. Chalkidis, I. Androutsopoulos, and A. Michos, "Obligation and prohibition extraction using hierarchical rnns," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018. doi:10.18653/v1/p18-2041

[8] D. Bracewell, D. Hinote, and S. Monahan., "The Author Perspective Model for Classifying Deontic Modality in Events," In The Twenty-Seventh International Flairs Conference., 2014

[9] A. Sancheti, A. Garimella, B. Srinivasan, and R. Rudinger, "Agent-Specific Deontic Modality Detection in Legal Language," In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, UAE, pp. 11563–11579, 2022

[10] V. Joshi, P. R. Anish, and S. Ghaisas, "Domain adaptation for an automated classification of deontic modalities in software engineering contracts," Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021. doi:10.1145/3468264.3473921

[11] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, vol. 21, no. 140, Jan. 2020.

[12] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: pre-training with extracted gap-sentences for abstractive summarization," in Proceedings of the 37th International Conference on Machine Learning (ICML), vol. 119, JMLR.org, 2020, pp. 11328-11339.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019

[14] "T5-large · Hugging Face," t5-large · Hugging Face, https://huggingface.co/t5-large (accessed May 7, 2023).

[15] "Google/Pegasus-large · Hugging Face," google/pegasus-large · Hugging Face, https://huggingface.co/google/pegasus-large (accessed May 7, 2023).

[16] "GPT2-large · Hugging Face," gpt2-large · Hugging Face, https://huggingface.co/gpt2-large (accessed May 7, 2023).

[17] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685 [cs.CL], Jun. 2021.

[18] H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint arXiv:2302.13971, 2023.

[19] B. Wang and A. Komatsuzaki, "GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model," [Online]. Available: https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

[20] E. Almazrouei et al., "Falcon-40B: an open large language model with state-of-the-art performance," 2023.