Constructing formal normative concepts from natural language using machine learning

Data Systems Project, TNO, Group D2

Atanas Yonkov (15170616), Efstathios Trigkas (14765667), Dimitrios Tsiamouras (14896184), Lucas Belderink (12151750), Tawfeek Alkanbar (15256480)
University of Amsterdam

ABSTRACT

In both societal and business contexts, people's behaviour is often guided by normative texts written in natural language. Examples include laws governing society and organizational policies. The challenge inherent in these normative documents lies in the interpretation process, which often results in multiple, and at times conflicting, understandings. Formalized constructs such as the Flint Ontology can be employed to make these interpretations explicit to enhance transparency.

However, the time-intensive nature of creating normative models and the scarcity of experts capable of interpreting these documents pose significant challenges, making the task difficult to scale.

To address this hurdle, our project aims to automate the creation of formalized constructs by leveraging the capabilities of transformer-based machine learning. Our system utilizes these models to automatically construct formalized representations from sentences of natural language. Furthermore, the system utilizes a feedback mechanism to iteratively retrain the models and enhance their adaptability to various types of normative text.

While primarily trained on game rule sentences, our models exhibit promising accuracy in generating formalized constructs based on game rule sentences, and the retraining mechanism demonstrates adaptability to other types of normative texts. However, due to the dearth of diverse training data, the quality of the resulting frames is highly dependent on the complexity of the input sequence and the data our model has previously been trained on. In such cases, user intervention is necessary to refine the results and provide valuable feedback that will be used to retrain the models.

GITHUB REPOSITORY

https://github.com/dsp-norm-extractor

1 INTRODUCTION

In the modern world, it is no secret that organizational policies, legal frameworks and governmental regulations are becoming increasingly commonplace. These texts are written in natural language and are meant to guide people's behaviour in different environments. However, these types of texts introduce many challenges.

The most fundamental difficulty that people face when trying to navigate the labyrinth of normative texts is the fact that they are difficult to interpret, resulting in disagreements and conflicts. To address this severe hurdle, there is a necessity for the formalization of their interpretation.

To improve transparency and operational oversight, such interpretations can be made explicit by using formalized constructs, to be possibly used for further computational automation. One such examples is the Flint ontology, a formalized construct designed to make these interpretations explicit.

Despite the fact that these constructs are so useful for norm understanding, there are currently no state-of-the-art mechanisms to produce automatically standardized normative models out of these complex text segments. Noteworthy among the existing platforms is TNO's Norm Editor [8]; however, this framework requires expert knowledge, limiting its scalability and accessibility due to the manual creation and filling of FLINT frames [9]. Another initiative, Flint Filler SRL [2], focuses on using semantic role labelling to automatically label tokens, but it is confined to actions and lacks coverage for facts and duties.

These platforms highlight a crucial research gap – the need for an automated and universal approach to norm extraction that not only reduces expert reliance but is also designed to cover a broader domain of normative texts.

In light of these challenges, the research questions that guide our investigation are: 1. To what extent can state-of-the-art machine learning tools be leveraged to automate the accurate interpretation of normative text into a formal structure, and how does this automation contribute to scalability in the extraction of formalized constructs? 2. Given the difficulty in finding a validated dataset suitable for training, to what extent can we rely on a feedback loop of retraining to improve the accuracy and the generalization of models?

To answer these questions, due to resource and expertise limitations, we focused our work on normative texts specific to board games. Nevertheless, the significance of this paper extends beyond the boundaries of games, aiming to reach into broader domains, and can be applied to any field where norms are prevalent.

To address our research questions, we designed a system that utilizes transformer-based machine learning to predict formalized structures from natural language input. The formalized structures that the system produces are based on the Flint ontology's frames. For this purpose, we fine-tuned the last layers of a BERT model to perform 1) sentence classification based on the semantic meaning of the sentence and 2) token classification for individual words based on their semantic role in the sentence. The models' predictions are utilized to build and fill out the contents of their corresponding Flint frames.

An innovative feature that naturally arises from our approach is the ability to obtain and utilize user feedback. After the frames are automatically produced through our aforementioned pipeline, they are displayed and can be manually edited. At a surface level, this intervention allows for the creation of more complete and accurate

frames. However, this is not the only use of expert feedback. The edited frames are also being stored in our expanding database and then, through our ever-growing mechanism, can serve to expand our initially limited domain of applicability.

To ensure that our choice of direction was aligned with experts' needs and expectations, we conducted interviews and feedback sessions. Through these, we obtained invaluable feedback, a better understanding of the problems they face, and tailored our results to those.

2 BACKGROUND AND RELATED WORK

Several studies, most notably that of Semantic Role Labelling (SRL) for Dutch Law Texts [3] by TNO, have explored computational modeling for formalizing normative or rule-based data. Just like in this study the paper at hand and the application behind it involve a systematic application of the Flint Knowledge Representation (KR) language to create a formalized representation of normative text. The study contains a comparative analysis of a rule-based method and a transformer based method. The rule-based method uses established NLP based methods, part-of-speech tagging and chunk tagging, and allocates the results to the slots in a Flint frame using a designed set of rules. On the other hand the transformer based method consists of fine-tuning a transformer with the data found for the study and using it for a classification task similar to SRL. This is possible because the slots to be filled in a Flint frame are similar to semantic roles. The model used in this research is a specifically fine-tuned BERT [5] model, A Deep Bidirectional Transformer-based model specifically pre-trained for Language Understanding. This model can be fine-tuned with just a single additional output layer to create effective models for a wide range of language representation-based tasks.

The study has specifically narrowed down their domain to that of Dutch law texts and uses publicly available data from `wetten.nl`. Input text is split up in individual sentences for both approaches. No further preprocessing is needed for the rule-based method but the transformer-based method requires further preprocessing in the form of annotations for the sentences. To create an annotated dataset a group of human annotators were gathered and given instructions on how to annotate the data to specifically cater to the requirements of effectively training a bert model for the transformer task. Among these annotators were experts in the field of norm engineering including some of the authors of the original paper, assisting in creating a more reliable dataset. To prevent possible disagreement on the labeling between different annotators additional validation was performed on the annotated data using Fleiss' kappa [4].

One of the main differences between this study and that of the SRL of Dutch law texts is that the latter focuses on the predictive labeling of act frames only. The study shows promising results, especially for the use of a transformer-based method to automatically formalize text into Flint Act-frames, but doesn't generalize further to other Flint frames. This generalization is one of the problems this research attempts to engage with. Also the approach of the SRL paper remains an attempt of formalization in just a single domain. This approach has proven to be effective, especially as a starting point, but the question remains if creating a model that can perform formalization on more general normative data is possible. Finally, the result of the study is a static model to be used for the specific task of dutch law text labeling. One of the main limitations is a further lack of annotated data, highlighting the possibility of an approach that could possibly improve by retraining when given more annotated data. This also introduces the question if such a model could generalize, possibly dynamically, to other types of normative data as it is given different types of normative data than just law texts.

In the context of normative representation, the FLINT ontology, introduced by Acosta et al. [1], served a pivotal role in our work, offering a high-level description of normative systems. Its objective is to provide a structured and formalized representation of normative texts, essential for enabling machine interpretable versions of rules and regulations. The FLINT ontology achieves this by focusing on the dynamic perspective of norms, representing them in terms of normative acts and their associated pre- and postconditions.

The structure of FLINT frames plays a crucial role in this formalization process. As part of the ontology, Act and Fact frames are employed to categorize normative elements. Act frames describe the actions that agents may undertake, influencing the normative system's state. Fact frames, on the other hand, capture the aspects that characterize the system's state. This conceptual distinction allows for a clear and organized representation of normative data. Duties appear as a special kind of Fact, and are used to encode the type of behavior that is considered expected according to the norms. Similarly to Facts, Duties are created and terminated by acts, emphasizing their dynamic nature. It is imperative that each Duty have at least one creating and one terminating act. Furthermore, Duties inherently imply a duty holder and a claimant, establishing a clear Hohfeldian structure [7] within normative systems.

Another study, by Mills (2013) [6], has approached the same problem in a different domain of normative data, namely that of text data of board game rules. In this study an approach of computational modeling for formalizing normative or rule-based data is explored. More specifically, Mills investigated learning board game rules from instruction manuals, also leveraging both rule-based and machine-learning approaches. The dissertation dissected the problem into relevant components, one being Named Entity Recognition (NER) for identifying entities present in games. This work highlights the fact that, while supervised learning is commonly employed for NER purposes, the absence of a reasonably large labeled dataset complicates the process. We also encountered this constraint in our own work, necessitating manual data annotation, inadvertently introducing, however, due to the aforementioned limitation of resources, an element of human error, given the absence of an expert on our team. This further emphasizes the importance of the need for feedback from a human expert in the process of attaining annotated data.

As demonstrated in Mills' work, identifying relationships between entities is a crucial process. A notable distinction, however, between Mills' work and ours, lies in the methodology employed. More specifically, our work is aligned with normative representation using FLINT ontology, whereas Mills utilizes a novel design

for structuring game rules, comparing a rule-based and machine-learning-based approach. At a preliminary level, the former performed more accurately, but despite the improvements, Precision and F-measure metrics remained relatively low, revealing the fundamental challenges of the task and necessitating future work. The study does however show that machine learning-based, specifically transformer-based, approaches are effective in domains other than just that of law texts. This shows a possible direction of research for the further generalization of a formalizing model.

In summary, our research draws inspiration from these previous attempts of formalizing the interpretation of normative texts into standardized structures. Our contributions, however, lie in expanding these works in 2 main axes. First off, our work is designed to process normative sentences in English and categorize them into one of the 3 corresponding Flint frames, instead of just the Act one. More importantly, the retraining mechanism allows our platform to be utilized by experts, while simultaneously learning from them. To the best of our knowledge, this key functionality is unique to our work and is crucial to our platform's accuracy and ability to serve its initial purpose - the automatic interpretation of formalized structures in an unprecedented scale.

3 MATERIALS AND METHODS

In this section, we outline our research methodology, including but not limited to data reduction coding and analysis, as well as data collection strategies, which were integral to the development of our proof-of-concept system for interpreting normative game rules.

3.1 Data Collection, Purposive Sampling and Dataset Augmentation

Given the apparent lack of annotated data, we initiated our comprehensive analysis with the collection of data. For this purpose, we employed purposive sampling, specifically targeting board game rules as our primary source of normative text. This deliberate choice aligns with the proximity of board game rules to our research topic, enabling a focused and in-depth exploration of this unique normative domain. Our convenience sampling strategy was primarily centered around [https://playingcarddecks.com/pages/card-game-rules], supporting the collection of diverse rule sets for subsequent analysis.

However, it is essential to acknowledge the limitations our sampling strategy inadvertently introduced. The sample size, while comprehensive, does not fully adhere to the principle of saturation, with repetition being noted due to shared characteristics among game rules, especially in the areas of setup and objectives. Initial attempts to generalize our findings to all games encountered challenges, primarily stemming from substantial differences in the structure of normative sentences across various games. This complexity posed difficulties in both annotating sufficient data and achieving satisfactory accuracy across all categories. Despite these challenges, our focused data collection approach provides valuable insights into the distinctive features of normative texts in the context of board games.

Additionally, we acknowledge a limitation in our dataset related to the absence of postconditions. In Flint Ontology postconditions represent the desired outcomes or states expected after the execution of certain actions or events. Our data sampling approach, did not yield sufficient instances of post-conditions and, given the scarcity of relevant examples, we made a deliberate choice not to attempt to teach our models to recognize them. While this limitation restricts our model's ability to fully capture the entirety of normative language, it also underscores the importance of refining our sampling strategies in future endeavors to capture a broader range of scenarios and conditions.

To address some of the limitations in dataset size and diversity, we employed dataset augmentation. This involved perturbing existing examples by slightly changing words and their ordering while preserving the original meaning. This artificial creation of variations expanded our dataset, contributing to increased diversity and broader coverage of examples.

3.2 Data Reduction Coding and Analysis

Following the manual collection of our data, we transitioned to the data reduction phase — a pivotal element in our systematic approach to the analysis of normative text. Through meticulous examination of game rule texts, we identified underlying patterns in sentence structure, enabling the coding of sentences into FLINT frames: Acts, Facts, and Duties. This coding process extends to the subdivision of Acts into specific subcategories, distinguishing between Simple Acts (executable without conditions) and Conditional Acts (contingent on specific circumstances). Additional coding includes the further subdivision of Conditional Acts with regard to the creation and enforcement of Duties. Creating Acts initiates or establishes conditions, events, or rules, shaping the gameplay, while Enforcing Acts ensures compliance with established rules, often carrying penalties or obligations.

In parallel to the categorization of Acts, our analysis extended to the realm of Facts. There, the categories of Winning Condition, Setup, and Movement naturally emerged from our analysis. This distinction enables a more nuanced understanding of the semantic intricacies embedded in the rules. Our aspiration is that the rich human logic involved in interpreting game-related sentences is effectively translated into a computational framework, thereby bridging the gap between human understanding and machine learning capabilities. The emergence of these new Fact categories stands as a testament to the intricacy of normative texts, reflecting the depth and complexity inherent in the interpretation of game rules.

Having established a robust foundation through data collection and reduction coding analysis, we then refined our methodology by incorporating an annotation process and leveraging state-of-the-art machine learning techniques, particularly a BERT model that utilizes annotated knowledge for norm interpretation.

The results of our data analysis are presented in Table 1, Table 2, and Table 3. These tables provide valuable insights into the distribution of normative elements at both sentence and word levels.

Table 1 illustrates the distribution of examples for Acts, Facts and Duties at the sentence level. Tables 2 and 3 provide a more detailed breakdown of the distribution of word-level labels for Acts and Duties, respectively. These tables include information on the total number of examples and the corresponding percentage distribution for each label.

	Fact	Act	Duty
Examples	114	94	58
Percentage (%)	43.1	35.8	21.1

Table 1: Data for sentence level classification

	Total Examples	Percentage (%)
Actor	131	19.49
Action	76	11.31
Recipient	89	13.24
Object	102	15.18
O	173	25.74
Precondition	101	15.03
Total	**672**	**100%**

Table 2: Data for word level classification : Labels for Acts

	Total Examples	Percentage (%)
Duty Holder	100	23.39
Claimant	74	17.29
Enforcing Act	121	28.27
Creating Act	99	23.11
Action	76	17.74
Total	**427**	**100%**

Table 3: Data for word level classification : Labels for Duties

3.3 Data Preparation and Preprocessing

The dataset that resulted from the previous steps was stored in a CSV file. This file includes columns for the original sentence, a list of its words, and the label assigned to the entire sentence. Additionally, we created two distinct files, one for Acts and one for Duties. These files contained the complete sentences, the sentence tokenized in words, and a new column where each word is represented by a label based on its semantic role in the sentence. These files are then fed into our program and need to follow several steps in order to be utilized for training.

A crucial step was encoding labels for both sentences and tokens into numerical representations, since this is the format that BERT model is designed to process.

Subsequently, we employed DataCollatorWithPadding to dynamically pad sentences to the longest length within a batch during collation. This is done to optimize memory usage by padding only within batches, ensuring efficient processing without unnecessary memory consumption.

Another part of the preparation is splitting the data into train and test sets. This is accomplished using stratified K-Fold. This method maintains the distribution of normative elements across the training and testing sets, ensuring a representative sample for both model training and evaluation.

The processed data, now in a format compatible with BERT models, serves as the foundation for our subsequent steps, including model training, fine-tuning, and evaluation.

3.4 BERT Architecture and Experimental Insights

Our methodology leverages the power of BERT, a transformer-based machine learning approach, to interpret normative text within the FLINT ontology. The BERT model plays a central role in our framework, contributing to both sentence-level categorization and word-level semantic role assignment.

3.4.1 *Sentence-level Categorization.*

The first aspect involves training a single BERT model to predict the semantic label for the entire sentence. This model serves as a crucial decision point in our pipeline, determining whether the sentence falls into the categories of Act, Fact, or Duty. The success of subsequent stages heavily relies on the accuracy of this initial sentence-level prediction.

Notably, the dependency on correct sentence-level categorization introduces challenges. In cases where the model incorrectly predicts the sentence label, errors cascade into subsequent stages, impacting the overall accuracy of the norm interpretation. Recognizing this, we conducted experiments to assess the impact of incorrect sentence-level predictions on downstream tasks.

3.4.2 *Word-level Semantic Role Assignment.*

Following the sentence-level categorization, the pipeline branches into distinct models for Act and Duty categorization at the word level. Each word within the sentence is assigned a semantic role based on the predicted category, creating a fine-grained understanding of the normative elements.

Experimental results revealed a noteworthy observation. When the sentence-level prediction was bypassed by providing the correct label, and the models were used solely for word-level labelling, the accuracy of word labels increased significantly. This experiment demonstrated that the higher error rate observed in sentence-level predictions directly influenced the accuracy of word-level assignments.

The observed increase in word-labelling accuracy from 70% to 90% underlines the pivotal role of correct sentence-level categorization in ensuring precise semantic role assignment for individual words. This insight led us to emphasize the importance of refining the sentence-level prediction model for enhanced overall accuracy.

	Duty	Act	Fact
Presicion	0.63	0.80	0.92
Recall	0.67	0.80	0.85
F1-Score	0.65	0.83	0.93

Table 4: Training Validation Metrics

3.4.3 *Hyperparameter Optimization.*

For the BERT models, we conducted experiments to optimize their hyperparameters. We present the specific hyperparameters we fine-tuned on the best value that we obtained in Table 5. Fine-tuning these hyperparameters significantly contributed to the overall performance and effectiveness of our models.

3.4.4 *Data and Code Availability.*

For transparency and reproducibility, we provide access to our dataset and code used in the experiments. The dataset can be found

Hyperparameter	Value
num_train_epochs	5
learning_rate	2e-4
per_device_train_batch_size	16
weight_decay	0.01

Table 5: Hyperparameters and their corresponding values

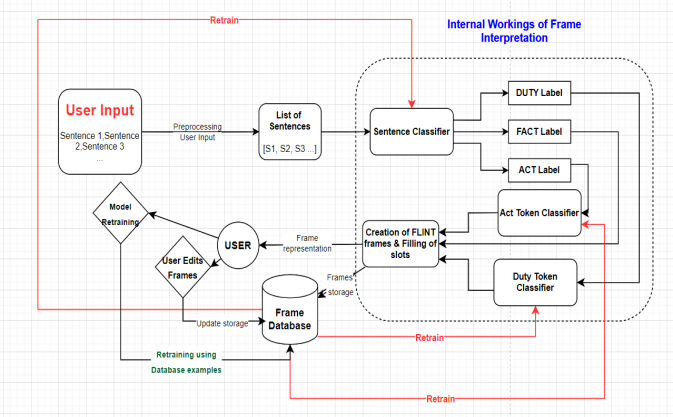

Figure 1: Architecture Pipeline

here and the code repository here. Researchers interested in replicating our experiments or extending our work are welcome to refer to and utilize these resources.

3.5 Prediction Pipeline

In this segment, we will break down the pipeline of our methodology, as depicted in 1, tying all the steps together to provide a better understanding of the prediction pipeline.

Initially, users input a segment of normative text, potentially consisting of multiple sentences. The text is processed, resulting in the creation of a list of sentences and a nested list of words for each sentence. Subsequently, our Sentence Classifier BERT model is employed to classify each individual sentence into an Act, Fact or Duty. This prediction is used to fork towards either an Act or a Duty Classifier that will attribute a semantic role label to each individual word.

Following this norm interpretation is the representation. For this purpose, the predicted label for each sentence is used to construct the corresponding Flint frame, its contents then being filled using the labels attributed to the words. The list of frames is, finally, presented to the expert who has the ability to edit them before submission to our database. Any modifications made by the expert are saved in our frame database along with previously seen data. These are all utilized when the user chooses to retrain.

Retraining involves training all models anew, starting from scratch and utilizing the entire available dataset. The newly trained models replace the previous ones, offering enhanced predictive accuracy and taking advantage of more diverse validated data.

3.6 Stakeholder Involvement and Expert Collaboration

In alignment with our commitment to accuracy and domain expertise, we actively involved professionals from the normative fields throughout the annotation and model development processes. Collaborative efforts included stakeholder interviews, discussions with researchers, and iterative feedback loops to enhance the alignment of our annotations with domain-specific nuances. This collaborative approach not only validated our methodology but also contributed to a more nuanced and accurate interpretation of normative game rules.

To validate the usability of our system, we collaborated with normative experts from our organization, TNO, to conduct usability tests using the think-aloud protocol. We invited an expert to interact with our system, allowing it to generate normative models automatically. Throughout the experiment, we encouraged the expert to vocalize their thoughts, enabling us to extract meaningful feedback. This iterative process played a crucial role in refining the system architecture.

3.7 Model Evaluation

3.7.1 *Sentence Predictive Model Accuracy.*

For the final sentence classification model, the F1-score for facts is 90%, for acts is 80%, and for duties is 67%. The challenges in duties' low accuracy are attributed to the lack of examples in games and the presence of implied modal verbs.

3.7.2 *Token Predictive Model Accuracy.*

To further assess the performance of our models, we conducted a detailed analysis of token-level predictive accuracy. In this evaluation, we considered the accuracy only in cases where the predicted sentence label was correct in the first place, excluding instances where the sentence was predicted incorrectly.

Duty Model Token Predictive Accuracy: The token predictive accuracy for the Duty model reveals impressive results. Specifically, for action, duty holder, object, and claimant, the accuracy is 90% or higher. The most challenging aspects are predicting the Creating Act (CA) and Enforcing Act (EA), both scoring an F1-score of 0.86. This emphasizes that, while the model generally performs well, there are complexities in accurately distinguishing between actions associated with creating conditions or events and actions related to enforcing rules, penalties, or obligations.

Act Model Token Predictive Accuracy: For the Act model, the token predictive accuracy is exceptional, consistently exceeding 95%. This high accuracy indicates the model's proficiency in capturing various actions within normative sentences.

3.7.3 *Train-Test Split Methodology.*

Our train-test split underwent an iterative process to optimize performance. We experimented with different approaches, including the standard manual split and the train_test_split function with various random states, and ultimately settled on the Stratified K-fold Cross-validation. This latter method ensures that each class has the same number of examples in the split, providing a more balanced representation. The choice of this methodology was motivated by the aim to mitigate potential bias and improve generalizability.

3.7.4 *Challenges and Insights.*

During our extensive evaluation, certain challenges and insights emerged. When visualizing frames with more complex sentences, it became evident that the precondition in an Act frame sometimes contains implicit information. In our tests, the creation of the related

frame is present in other frames, requiring human intervention to fill in the gaps. This observation aligns with our earlier strategy of using IDs to connect frames, underlining the importance of human oversight in handling intricate linguistic nuances.

The Duty model, while achieving high accuracy in several aspects, exhibits nuances in capturing semantic differences between actions associated with creating conditions or events and those related to enforcing rules, penalties, or obligations. The F1-score of 0.86 for the Creating Act (CA) and Enforcing Act (EA) suggests that the model is proficient but not perfect in these specific distinctions.

3.8 Conclusion

In conclusion, the evaluation of our final sentence classification model and token predictive models demonstrates a solid foundation with notable achievements in interpreting normative game rules. While achieving high accuracy in many aspects, we acknowledge the identified challenges and complexities, particularly in duties with implied modal verbs.

Moving forward, our focus will be on refining the Duty model to address specific challenges, enhancing the model's ability to accurately distinguish between creating conditions or events and enforcing rules or obligations. Additionally, continuous efforts in dataset expansion and refinement remain crucial for achieving a more comprehensive understanding and interpretation of varied linguistic patterns in diverse normative domains.

4 RESULTS

4.1 Generalization to other game domains

The model, originally trained on a dataset exclusively focused on card games, exhibited commendable performance during the generalization evaluation. Exhibiting an impressive precision of 89% in categorizing Act frames in unforeseen game domains, the model showcased a high level of accuracy in identifying activities regardless of the specific game context. Additionally, the model's weighted average precision and recall of 71% underline its ability to maintain a balanced performance across different classes.

However, this preliminary evaluation brought to light certain limitations. The model faced challenges in consistently recognizing expressions capturing the aspects that characterize the system's state (Fact frame) and duty-related statements (Duty) in contexts beyond its training data. Evident in the lower precision and recall scores for these classes, the confusion matrix highlights difficulty discerning between the two. These findings emphasize the need for further refinement to enhance the model's proficiency in understanding the nuanced features that define a system's state and assigned responsibilities in diverse gaming scenarios.

	Fact	Act	Duty
Presicion	93.00	68.00	44.00
Recall	20.00	77.00	55.00
F1-Score	33.04	72.28	48.89

Table 6: Metrics for Generalization to New Domains

4.2 Generalization to other normative domains

When extending the generalization evaluation to normative sentences outside the domain of games, including legal texts and government regulations, the model's weaknesses became even more apparent. Starting from the positives, The model showcased an overall accuracy of 58.83%, suggesting a solid foundation with a lot of room for improvement. This aligns with its previous success in game contexts, suggesting a certain level of versatility in understanding and categorizing diverse sentence structures.

For the "Act" label, the model maintained a balanced precision and recall of 68% and 77%, respectively, echoing its consistent performance in identifying activities across varied contexts. The model proved capable of correctly predicting Facts with a precision of 93%. However, its lower recall and f1-scores of 20% and 33% respectively highlight the inability to comprehensively capture all factual elements. These results, in conjunction with the Duty metrics being low, are consistent with those mentioned in the Methodology segment.

The challenges observed in recognizing expressions capturing the nuanced features of a system's state and duty-related statements remained, as evidenced by lower precision and recall scores for these classes.

4.3 In-depth investigation of the model's learning capacity

Upon closer examination of the LIME representations, it becomes evident that the model relies heavily on specific linguistic patterns to categorize normative text into their corresponding Flint frames. In particular:

Duty Frames: Modal verbs such as "must" and "should" play a crucial role in signaling duty-related statements, and the model assigns high weights to these words, indicating their significance in identifying obligations and responsibilities. This aligns with our human understanding of duties, where responsibilities often come with explicit directives. However, when these verbs are absent or implied, the model encounters difficulties in correctly identifying the duty. In such cases, it tends to default to categorizing the statement as a fact, showcasing the model's reliance on explicit linguistic cues. This observation underscores the need for the model to develop a more nuanced understanding of duty-related language, especially in scenarios where modal verbs are not explicitly stated.

Moreover, the Duty Holder, representing the entity responsible for the duty, consistently emerges as a salient feature. However, its importance diminishes when modal verbs are absent, contributing to misclassifications. The model's tendency to prioritize modal verbs over the Duty Holder in predicting duties suggests that the presence of explicit directives holds more weight in duty identification. This further emphasizes the importance of refining the model's understanding of duty-related language beyond the presence of modal verbs, ensuring accurate classification even in their absence.

Act Frames: In LIME representations, the operative verb, signifying the core action in an activity, stands out prominently. This underscores the model's emphasis on verb semantics when categorizing actions within a sentence. The model recognizes the pivotal

role of operative verbs in shaping the nature of activities, aligning with linguistic expectations.

Furthermore, surrounding elements, including the actor, object, and recipient, also carry substantial weight in the LIME representations. The model demonstrates an understanding of the contextual roles these entities play in defining an activity. This holistic consideration of actor, object, and recipient showcases the model's ability to grasp the interconnected components that collectively contribute to the characterization of diverse activities within different contexts.

4.4 Impact of Retraining on Accuracy metrics

	Precision	Recall	F1-Score	Accuracy
Initial Training	0.63	0.62	0.62	0.62
Expert Feedback	0.71	0.71	0.67	0.69

Table 7: Comparison of Results

As we have discussed above, a core pillar of our approach and something that distinguishes our work from others' is the retraining mechanism. Throughout our development, after obtaining expert feedback and establishing ground truth results, we performed a retraining using previous and new examples. This retraining had the intended results, as can be seen in Table 7, showcasing our approach's ability to effectively utilize feedback and suggesting a solid foundation that does not need rebuilding, but rather more diverse data to learn from.

4.5 Concluding Remarks

The examination of the model's performance in both game and normative contexts, coupled with insights into its learning capacity, underscores the continuous need for refinement and dataset expansion. While the model demonstrates a solid foundation, the observed inconsistencies may stem from our current knowledge limitations in the field.

The priority lies in enhancing the model's proficiency, particularly in discerning intricate linguistic features that distinguish Fact and Duty frames. Further fine-tuning, focusing on nuanced language structures, is essential for achieving a more robust classification across diverse domains. To address the challenges, diversifying duty examples, especially those with implicit modal verbs, is crucial. By doing so, the model should learn how to categorize duties more effectively, gaining a model robust understanding of the logic that separates them from factual statements. Due to the limitations of knowledge and time, exploring ways to improve such predictions remains a key avenue for future development.

Moreover, it's crucial to note that the reported results and accuracy metrics are based on the analysis of simple, self-contained sentences. In our interviews and feedback sessions with experts, we found that the accuracy for token classification and the resulting frame's contents significantly suffer when sentences lack completeness or contain implicit information. Recognizing this limitation, there is a clear need for further expansion of the complexity of training data. It's important to acknowledge that we intentionally refrained from attempting predictions for highly complex sentences given our inherent limitations and the associated prediction ceiling. This acknowledgment paves the way for future exploration into handling more intricate normative contexts.

5 DISCUSSION

In this section, we reflect on the implications and limitations of our design decisions. We additionally intend to suggest applications and highlight the potential offered by our platform.

5.1 Reflection on Design Decisions:

Our methodology, rooted in Machine Learning and leveraging pretrained BERT models, proved effective in categorizing normative text into meaningful frames. The retraining mechanism shows that the models have the ability to adjust to different types of normative texts.

The accuracy metrics presented in the Result section Table 1 and Table 3 prove that this direction is meaningful, and with enough training data and feedback from experts, this research could potentially help in the automation of the process of building formalized constructs.

The incorporation of stakeholder involvement and expert collaboration during annotation and model development phases validated our approach, ensuring alignment with domain-specific nuances.

The decision to use dataset augmentation and ontology extension played a pivotal role in addressing limitations in dataset size and diversity. The inclusion of Creating Acts (CA), Enforcing Acts (EA) and Preconditions (PR) in the FLINT ontology added granularity to the categorization of normative actions.

Nevertheless, it's crucial to note that our reliance on predicting the sentence label first introduces a cascading effect on the subsequent prediction of individual tokens. In cases where the sentence label is incorrectly predicted, errors propagate through the model, impacting the accuracy of word-level semantic role assignments. This insight underscores the importance of refining the sentence-level prediction model to enhance the overall precision of our framework.

Furthermore, our current approach treats each sentence individually, leading to a loss of contextual information. This limitation results in incomplete frames that necessitate manual intervention by experts to fill the gaps. While the model demonstrates proficiency in isolating normative sentences, the context provided by neighbouring sentences can be instrumental in achieving a more comprehensive understanding. Future iterations of our platform may explore strategies to incorporate contextual dependencies, potentially enhancing the completeness of extracted frames.

These considerations highlight the delicate balance between sentence-level predictions and maintaining context, emphasizing the ongoing refinement required to optimize the accuracy and completeness of normative text interpretation.

5.2 Real World Impact

A Tool that could semi-automate the process of normative model building could really benefit normative experts in their task. With this kind of help, the process could be scaled up and eventually reach a point where business and organizations lay down their rules not only in natural language but in their computational counterparts.

Our system, even despite its inherent limitations, is a small step towards that direction. Our solution is not a panacea, and it could not fully automate the process of normative model building. An expert supervising the system will always be necessary. But it could potentially save some time in order to allow the scale-up needed to address this challenge.

5.3 Ethical considerations

As we extend the generalization evaluation of our model to normative sentences outside the gaming domain, including legal texts and government regulations, it becomes imperative to address ethical considerations associated with the model's broader applicability.

The versatility demonstrated by our model in interpreting normative statements beyond its original training domain introduces potential implications. Given the sensitivity and impact of normative language, there are ethical considerations related to the potential misinterpretation of statements, especially when transitioning from gaming-specific to broader normative contexts.

5.3.1 ***Bias and Fairness.*** The model's training data, initially focused on card game rules, may introduce biases that could inadvertently affect its performance in broader normative domains. It is crucial to assess and mitigate biases to ensure fair and unbiased interpretations, particularly in applications where decisions based on normative analysis may have real-world consequences.

5.3.2 ***Interpretation Accuracy.*** The model's commendable adaptability raises the question of its accuracy and reliability in diverse normative contexts. It is essential to consider potential risks associated with misinterpretations, especially in legal and regulatory applications, where precision and correctness are paramount.

5.3.3 ***Transparency and Accountability.*** As our model is designed for broader normative applications, ensuring transparency in its decision-making processes becomes essential. Users and stakeholders should be informed about the model's capabilities, limitations, and potential areas of uncertainty to promote responsible and informed utilization.

5.3.4 ***Dataset Expansion.*** To address potential biases and enhance the model's proficiency in diverse normative domains, continuous efforts in dataset expansion are necessary. Diversifying training examples and including a wider array of normative texts can contribute to a more comprehensive understanding and interpretation of varied linguistic patterns.

While our current evaluation highlights the model's adaptability, these ethical considerations underscore the responsibility associated with the development and deployment of models that interpret normative language across different domains. Ongoing efforts in refining the model, addressing biases, and ensuring transparency are crucial steps towards responsible AI applications.

6 FUTURE WORK

Our research has established a foundation for interpreting normative game rules using machine learning techniques. Moving forward, there are a plethora of different directions that we would like to explore, the most important being presented in 6.1 and 6.2:

6.1 Continuous Annotating and Expert Feedback

Our primary goal is to continue annotating examples while opening up our platform to experts, enabling them to utilize the application and provide valuable feedback. This iterative process serves a dual purpose—expanding our dataset with diverse examples from various normative domains and, more importantly, utilizing expert insights to refine and retrain the baseline of our models. We remain committed to our initial goal of providing experts with a framework that assists them in creating formalized structures. By engaging experts from different fields, we aim to continually improve the accuracy and generalization of our system. This collaborative effort contributes to the advancement of norm interpretation and ensures our platform's practical utility in real-world scenarios.

Furthermore, as we progress in our research, we recognize the importance of gradually increasing the complexity of examples our models are trained on. Since our models are currently trained on simpler sentences from different domains, it is crucial that they are introduced to more complex examples with implicit information. However, at this stage, we propose to maintain the models' focus on simpler sentences to ensure a solid foundation before gradually transitioning to more challenging examples. This approach will allow us and the experts to effectively assess the models' performance, iteratively retrain them, and ensure steady progress in handling complex linguistic structures.

6.2 Incorporating Contextual Information and Multi-Sentence Integration

Simultaneously with expanding our domain and improving the accuracy of automatic norm interpretation, one should explore the integration of contextual information and the incorporation of multi-sentence context. In order to achieve this, we have envisioned an extension of our current architecture that will enable us to process full segments of normative text, rather than individual sentences (eg. by means of BERT or LLMs). This addition will be vital in our model's capacity to correctly fill out implicit fields of the frames, previously unavailable due to our aforementioned limitations. This change will be parallel to the utilization of frame IDs and they will both act synergistically to cover each others' weak spots, ultimately addressing our current weaknesses. As mentioned above, Frame IDs will be particularly useful in connecting frames retrospectively, after they have exited our automatic pipeline, and in order to manually add information that was either misclassified or entirely overlooked. By implementing these enhancements on a broad enough scale, the model will be able to acquire complex logic, often inherent in human understanding but challenging to effectively transfer to even the most advanced computational modules. This evolution is crucial for a more sophisticated and context-aware interpretation of normative language.

We believe that the 2 steps described above are vital for overcoming our current limitations, and if it weren't for lack of time and resources, we would have made significant progress towards these directions. Below, we present some additional ways in which the foundations we set could be utilized:

6.3 Expanding Platform Functionality to Game Execution

Beyond norm interpretation and leveraging the fact that our models were trained and evaluated on game rules, our work could be extended to not only creating structured norm interpretations but also facilitating the execution of games. By incorporating extended rules provided by users and then representing them in a formalized, universal manner, our platform could be extended to virtually facilitate the execution of a game. Users would have the capability to prompt actions, and the platform would verify whether the actions align with the defined norms. This expansion transforms our platform into a dynamic tool for not only understanding normative language but also practically applying and testing normative rules in various gaming scenarios.

6.4 Dynamic Adaptation to Evolving Norms

An additional avenue for future exploration involves devising strategies for the dynamic adaptation of the model to evolving norms. Recognizing that normative texts change over time, it is imperative that our approach seamlessly adjusts to these variations. Training the model with adaptability in mind ensures its continued relevance and applicability in dynamic environments. This task requires the development of mechanisms to update the model with new norms while preserving its existing knowledge.

In summary, ongoing efforts revolve around expert engagement and the integration of contextual information. We believe that the former is vital for leveraging our work's solid foundation, elevating our solution to a level that alleviates experts from the time-intensive task of manually creating normative models. The latter, on the other hand, will ensure that the contents of the resulting frames contain previously unavailable contextual information, minimizing the need for expert intervention. Working on these pillars should lead to a more robust, adaptable and well-rounded platform, which can then be utilized for more ambitious endeavours such as the facilitation of game-play or the dynamic updating of norms.

7 CONCLUSION

In this section, we would like to reflect on our initial objectives, obstacles that we found along the way, and design choices that were made to overcome them.

Our primary goal when we embarked on this journey was to build a tool that would automate the extraction of formalized constructs from diverse normative domains. However, the scarcity of annotated data and the complexities of normative sentences from different domains led us to narrow down to board game sentences. This allowed us to fine-tune our pipeline for the specific domain, without significantly hindering generalizability, as showcased by our extensions to diverse normative domains.

The next challenge that lay ahead was sentences containing implicit or incomplete information. This became even more apparent through our talks with experts, showcasing our model's difficulty in acquiring contextual information. The reason for this was that, even though the fact that we processed sentences individually allowed us to categorize them into their corresponding frame, the model no longer had access to surrounding sentences. Our solution was simplifying our example sentences, aiding our model to grasp complete semantic relationships. We additionally set up the infrastructure for incorporating contextual information using frame IDs, but at the time, we have not entirely harvested this potential.

Last but not least, bounded by our time and knowledge limitations, we were aware that the model's predictions were only as good as their training data. For that reason, we not only ensured the correctness of our predictions with experts, but we innovated by implementing a retraining mechanism. This mechanism allows our platform to benefit from expert feedback by storing it and utilizing it to improve predictions iteratively.

Looking ahead, on potential applications of our platform, it is our intention to continually improve our application's ability to facilitate experts in their work of creating formalized structures from normative texts. Their engagement in the annotation and feedback processes will expand our domain of use.

Simultaneously, we aim to address the limitations of contextual information by exploring strategies for multi-sentence integration, both by leveraging frame IDs and by expanding our BERT architecture beyond individual sentences.

In summary, we believe that our objectives were met and our research questions were adequately answered. While acknowledging its imperfections, we sincerely believe our work serves as a stepping stone towards bridging the apparent research gap towards automatic norm interpretation and subsequent creation of universal, standardized structures from a plethora of normative domains.

REFERENCES

[1] M Acosta et al. 2023. The FLINT Ontology: An Actor-Based Model of Legal Relations. In *Knowledge Graphs: Semantics, Machine Learning, and Languages: Proceedings of the 19th International Conference on Semantic Systems, 20-22 September 2023, Leipzig, Germany*, Vol. 56. IOS Press, 227.

[2] Roos Bakker, Romy van Drie, Daan Vos, and Maaike de Boer. 2020. Flint-Filler Semantic Role Labeling. https://gitlab.com/normativesystems/flintfillers/flintfiller-srl. (2020).

[3] Roos Bakker, Romy AN van Drie, Maaike de Boer, Robert van Doesburg, and Tom van Engers. 2022. Semantic role labelling for dutch law texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 448–457.

[4] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[5] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[6] Chad Mills. 2013. *Learning Board Game Rules from an Instruction Manual.* Ph.D. Dissertation.

[7] Joseph William Singer. 1982. The legal rights debate in analytical jurisprudence from Bentham to Hohfeld. *Wis. L. Rev.* (1982), 975.

[8] Ioannis Tolios, Erik Boertjes, Mike Wilmer, and Robert van Doesburg. 2022. Interpretation editor. https://gitlab.com/normativesystems/ui/interpretation-editor. (2022).

[9] Robert Van Doesburg, Tijs Van Der Storm, and Tom Van Engers. 2016. Calculemus: towards a formal language for the interpretation of normative systems. *AI4J Artif Intell Justice* 1 (2016), 73.