National College of Ireland

Interpretable Multi-Label Rule Extraction from Financial Regulatory Texts Using Transformer Models: A Comparative Study on EURLEX

MSc Research Project
MSc Data Analytics

Atharva Sanjay Bhalilkar
Student ID: 23299371

School of Computing
National College of Ireland

Supervisor: Prof. Jaswinder Singh

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Atharva Sanjay Bhalilkar

Student ID: 23299371

Programme: MSc Data Analytics **Year:** 2025

Module: MSc Research Project

Supervisor: Prof. Jaswinder Singh

Submission Due Date: 11/08/2025

Project Title: Interpretable Multi-Label Rule Extraction from Financial Regulatory Texts Using Transformer Models: A Comparative Study on EURLEX

Word Count: 8126 **Page Count** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Atharva Sanjay Bhalilkar

Date: 11/08/2025

……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Interpretable Multi-Label Rule Extraction from Financial Regulatory Texts Using Transformer Models: A Comparative Study on EURLEX

Atharva Sanjay Bhalilkar
23299371

Abstract

Increased complexities of the financial regulations require automated systems that can extract interpretable representations of formulated rules into legal texts. The following paper describes the use of transformer-based models such as BERT, RoBERTa, and FinBERT to carry out multi-label classification of regulatory documents. The consideration of the whole spectrum of the NLP pipeline, encompassing such aspects as data preprocessing, label binarization, feature engineering, model training, and evaluation, is carried out by applying EURLEX dataset of the annotated EU legal texts. A traditional rule-based classification technique is also formulated to be used as benchmarking of performance. Micro and macro F1 score, precision, recall and confusion matrices are utilized to conduct the model evaluation. Interpretability is used in the form of SHAP, LIME, and attention visualization methods, where the input of the model prediction can be seen and rule importance can be learned. The assessments characterize that the transformer models have bettered the predictive accuracy and labelling coverage than the rule-based basic one and can generate scalable, albeit not completely explicable, outputs. This study would be a contribution to the field of law NLP since it proves that transformer-based models are effective in classifying and understanding the structure of financial regulatory documents, and they hold practical potential to automate compliance and application of legal AI.

1 Introduction

The explosion in the proliferation of financial and legal regulatory texts in the prestigious ten-year horizon has completely transformed how compliance officers, the legal staff, and financial companies are obliged to handle and interpret data (Chalkidis et al., 2021; Chalkidis et al., 2019). The scale and sheer complexity of this regulatory environment have made manual methods of identifying rules and classifying them increasingly unrealistic, long and costly.

Although their transparency is an admirable feature, rule-based systems that are conventional use to be weak at responding to the heterogeneity and changing character of regulation texts with findings often consisting of brittle and incomplete results. Meanwhile, years of accelerated development of robust Natural Language Processing (NLP) techniques (especially transformer-based) have come to a point where the semantic richness and overall contextual dependencies of legal texts can be modelled to a degree to which they were previously inaccessible (Devlin et al., 2019; Vaswani et al., 2017). Such models have the ability to model complex relationships between clauses and provisions and therefore hold a promise of automated work that was

previously performed by intensive human interpretation (Chalkidis et al., 2020; Beltagy et al., 2019). Nevertheless, with the above benefits, they are not used in high-risk compliance situations because of interpretability, fairness, and scalability issues in real-life (Serrano and Smith, 2019; Lundberg and Lee, 2017).

The current study, therefore, explores the possibility of using transformer-based NLP models to effectively perform multi-label classification of texts in financial regulations, especially on the ability to obtain an interpretable presentation of the rules and to scale up (Chalkidis et al., 2020). This study is based on the task of using the EURLEX dataset, a dataset published by the EU with a large amount of European Union legislative documents and marked with EUROVOC concepts (Chalkidis et al., 2021). Using this dataset, the study measures the relative performance of finetuned transformer models against a purposefully low-performance rule-based baseline to get a clear comparison between their capability and weakness (Chawla et al., 2002).

The research question of this study is:

"Can transformer-based NLP models such as LegalBERT, automatically extract interpretable rule representations from financial regulatory texts, and how do they compare to traditional rule-based systems in terms of interpretability, scalability, and fairness?"

Its methodology combines performance analysis with post-hoc interpretability methods, such as SHAP, LIME, attention visualization, to cast light on how models make their predictions (Lundberg and Lee, 2017; Ribeiro et al., 2016). These descriptions show how the models deal with objective domain-specific terms of law, more general semantic regularities, or possibly fallacious correspondences. The research will bring together model performance measures and interpretability-based analyses through precision, recall, F1-score measures to provide consolidated explainable and scalable framework to classify any legal text. These findings can be used in future research on the topic of legal NLP and regulatory technology as part of an increasing body of work where the results of this research provide a roadmap in achieving compliance regulatory systems that achieve the dual requirements of high prediction accuracy and transparent decision making.

2 Related Work

2.1 Legal Text Classification and the EURLEX Dataset

Use of Natural Language Processing (NLP) in Law: The legal field has certain complexities that are not shared with other schemes of Natural Language Processing (NLP). In these texts, one can frequently encounter domain-specific vocabulary, the long-range dependencies and nested clauses, and this fact makes traditional NLP methods not very effective. In addition, legal texts themselves are multi-dimensional in the sense that any document can be classified under various legal areas. That is why such a definition automatically puts the classification problem into a multi-label learning context (Chalkidis et al., 2019).

EURLEX dataset has become the common benchmarking tool in the learning task of classification of legal texts, primarily in multi-label learning tasks. It comprises more than 19,000 pieces of EU legislation indexed with these terms in EuroVoc thesaurus descriptors.

These adjectives are hierarchical legal terminologies, and cover all legal, economic, and financial spheres. Challenges of the dataset include large label cardinality (5.3 labels on average per document) and skewed labels distribution which means that the dataset is an appropriate candidate to test label generalization, model scalability and class imbalance concerns (Chalkidis et al., 2019). Recent advancements in contextualized embeddings and attention mechanisms in transformer-based architectures have provided a promising alternative. Such models can capture semantic relationships across sentences and paragraphs, a crucial feature for legal documents where a rule may span multiple clauses or references to other statutes or legal instruments (Devlin et al., 2019). Previous legal text classifiers were to a high degree reliant on the term-based strategies and hand-created rules. Although these methods were completely transparent and had interpretation capability, they had issues with synonymy, context change, and ambiguity. Since the terminology of laws is dynamic, and differently used in different jurisdictions, it becomes more onerous and fragile to maintain such systems of rules (Boella et al., 2013; Ashley, 2017). As an example, there may be contextual legal meanings of words, which are hard to write down in strict rules. Such constraints have sped up the movement towards more adaptive and environment-sensitive machine learning models.

2.2 Transformer models for Legal NLP

Transformer models such as BERT, RoBERTa, and FinBERT have changed NLP by model deep context-sensitive representational models of text. BERT presented a new structure that considers not only left but also the right context in the process of encoding (Devlin et al., 2019; Vaswani et al., 2017). This architecture was also further refined by RoBERTa, in which it was trained longer and had Next Sentence Prediction objective removed, and resulted in better generalization (Liu et al., 2019). FinBERT, however, adjusted the BERT structure to the financial sector, pretraining it on financial corpora, and it was more adequate to the matching specialized regulatory purposes (Araci, 2019; Cohan et al., 2019). The models have enjoyed a demand in legal applications like extraction of contracts clauses, ranking of precedents as well as classification of legislations.

They have been better in performance in EURLEX as compared to traditional models especially in consideration of legal subtleties and overlapping thematic categories (Chalkidis et al., 2021; Chalkidis et al., 2020). The capability of transformers to represent long text using self-attention makes them suit legal text especially where multiple embedded rules are likely and references to other statutes or legal documents (Yang et al., 2019). There are costs to these benefits, which is in the form of transparency. Although transformer models have a high predictive power, they are commonly criticized as black boxes (Ribeiro et al., 2016; Serrano and Smith, 2019). This is a major problem in the legal category where judgments must be traceable, justified and justifiable. Performance versus interpretability tension is a key issue in the deployment of such models to compliance and regulatory use cases (Lundberg and Lee, 2017; Sundararajan et al., 2017). All that, along with the need to tune such models to domain specific tasks, entails both computational resources and preprocessing.

Domain shift, the discrepancy between pretraining data and restricted target legal domain, is another aspect that may become the source of performance loss unless necessitated by domain adaptation practices (Tsoumakas et al., 2010; Beltagy et al., 2019).

2.3 Multi-Label learning and evaluation

Legal text classification is somewhat different than ordinary classification where a document may belong to many categories at once. This property means that multi-label learning is not only convenient but unavoidable with the undertakings like regulatory tagging or legal taxonomy matching. Binary Relevance, Label Powerset and Classifier Chains are the techniques that have been suggested to process such multi-label issues (Zhang and Zhou, 2014; Tsoumakas et al., 2010). Binary Relevance models binary tasks as independent of one another and is scale-able but does not capture label correlation. Label Powerset provides a way to learn different combinations of labels as discrete classes, which expresses dependencies but trades off poor associated scalability. Classifier Chains train label dependencies where the label orderings can be sensitive to error propagation and sensitive to label ordering.

Though they work well in the general setting, they cannot easily process skewed data sets such as EURLEX, where there is a small group of labels overwhelming the distribution (Chawla et al., 2002; Zhou and Liu, 2006). Single-label classification is simpler to evaluate as compared to multi-label classification. Such metrics as Micro-F1 and Macro-F1 serve as complementary view as the former focuses on performance on the frequent labels and the latter on rare ones. Model behavior can also be characterized by precision, recall and hamming loss (Tsoumakas et al., 2010). Nonetheless, it is still hard to visualize model behaviour on the label-level. It is important to note that even though confusion matrices are complicated in multi-label set-ups, they are helpful in the error analysis and are heavily utilized in the current research.
This research is urgent and necessary due to the absence of the standard benchmarks of legal multi-label evaluation, especially between interpretability and fairness (Niklaus et al., 2021).

2.4 Interpretability in Legal NLP

Transformer models are distributed typically as black boxes due to their complexity and the high-density internal representations. This opaqueness is the reason it limits its applicability in areas that require openness such as law and finance. In such an environment where stakes are so large accuracy is not enough as a model should be explainable to the government regulators as well as lawyers, and non-technical stakeholders (Serrano and Smith, 2019). To cope with this, post-hoc explanatory tools like SHAP and LIME are gaining increased popularity. SHAP values quantify the contributions made by each feature (e.g. word or token) of a model to the prediction of a model and offer a mathematically principled path toward explaining the model (Lundberg and Lee, 2017). As opposed to this, LIME tampers with the input words and tries to learn a local interpretable model to approximate the black-box decision boundary (Ribeiro et al., 2016). The two approaches are model-free and can be applied to the diversity of classifiers.

Transformers may also be described as visualizations with attention weights. These types of attention heatmaps demonstrate which words the model took into consideration to make its decisions, though there has been some controversy over whether attention is any type of explanation at all (Jain and Wallace, 2019; Serrano and Smith, 2019). Applications in fields requiring legal action are still seen to benefit by applying such visual cues to grease human reviewers to seek sections of document of interest.
Its interpretation tools, which were applied in the study, enabled interpretation of the model predictions and elicitation of near rule-like patterns in the data—linking the machine learning and traditional legal thinking (Chalkidis et al., 2020).

2.5 Rule-Based System and Legal Knowledge Engineering

Prior to deep learning, the rule-based systems dominated the process of legal information retrieval and classification. Such systems apply pattern matching, legal ontologies and domain specific heuristics to map text to categories. Their main strength is openness: all the rules are traceable, explicable and modifiable (Ashley, 2017). This qualifies them as of great purpose to be audited and regulated. Nonetheless, these systems are not flexible. They become hard when faced by new legal language, a changing regulation, or inter-jurisdictional writing. They must be continuously updated by experts in the field they represent and do not always take full advantage of the semantic richness of law materials (Boella et al., 2013). There have been attempts to unite the rule-based approaches and semantic reasoning (Zhang et al., 2021). These systems are yet to be scalable and widely covered although they offer positive prospects. Conversely, models of neural networks can be trained on vast collections of data and can learn to cope with variation and discover latent structure without any programming (Vaswani et al., 2017).

In the study, a simplified rule-based classifier has also been developed to compare the performance to transformer models. This analogy should explain the shape of the trade-off between interpretability and generalizability, the factors that should matter when it comes to practical legal use.

2.6 Research Gap and Contribution

Legal NLP still must fill in the gaps of integrating performance-driven models with understandability in the real world (Chalkidis et al., 2021). Even though several transformer optimizations were done before on EURLEX and comparable data, not many were able to match them with powerful interpretability tools in a legal sense (Lundberg and Lee, 2017; Ribeiro et al., 2016). In addition, there is limited side-by-side comparison between rule-based baselines across aspects of fairness, transparency, and label coverage (Niklaus et al., 2021).

The study fills in this gap by creating an end-to-end pipeline (preprocessing to interpretability) that integrates powerful models with tools that help to create transparency. The research is both qualitative and quantitative in terms of investigation into the way transformer models behave, how they can be interpolated in predictions, and how they measure against more traditional approaches. It, in turn, provides substantial contribution towards the creation of more explicable and scalable AI systems that can be applied to legal compliance and financial regulation.

3 Research Methodology

3.1 Dataset Description

The data related to the current project is the EURLEX dataset, one of the most popular benchmarks in the legal NLP field to carry out multi-label classification tasks (Chalkidis et al., 2021). It hosts more than 19 000 EurLex documents by European Union institutions and authorities, with each one annotated with one or more concepts considered in the EuroVoc system, a multilingual thesaurus of legal and administrative terms used by the institutions of the European Union (Chalkidis et al., 2019). The documents are legal acts, namely directives, regulations, and decisions in the form of official texts, originally published on the EUR-Lex

site. On average, 5.3 labels are attributed to each document meaning it will be an apt testbed to conduct an evaluation of the transformer-based models when applied to complex multi-label classification tasks.

The data was downloaded using the Hugging Face Datasets library (Liu et al., 2019) to JSONL format, and consists of three splits, eurlex_train.json, eurlex_test.json and eurlex_validation.json. In the documentation area, every record has a document title, the entire body text and a list of EuroVoc concept codes in form of labels. These files were in the form of text files, and they were converted into CSV format by running it through a custom script to make sure that it had a structured tabular structure which is suitable to pandas-based data pipelines.

3.2 Data Preprocessing

Prior to feeding the EURLEX legal documents into transformer-based models, a structured preprocessing pipeline has been adopted so that the data can be clean, consistent and can be tokenized (Devlin et al., 2019; Liu et al., 2019). Both the dataset and individual documents consist of two text fields, a title and a body, which both have important semantic information in them. These columns were combined into one unified text column by means of a concatenation with a space separator so that as much contextual information as possible could be provided to the model. These combined texts were then exposed to normalization algorithms such as lower case, stripping off the newline (\n) and carriage return (\r) characters. This normalization step was used to minimize noise, and this was without sacrificing legal expressions. Remarkably, the preprocessing did not drop punctuation because punctuation may provide useful syntactic and legal clues particularly concerning formal legal language (Chalkidis et al., 2020).

Also, formatting of labels was done efficiently. The label annotations of each document of the original JSON were in the form of stringified lists. These were unpacked into Python list objects in order to facilitate subsequent operations such as the binarization of the labels (Chawla et al., 2002). Notably, the documents that lacked label annotations were filtered out of the dataset so as to sustain the consistency and relevance of the supervised multi-label classification. This selection process aided in getting rid of various edge cases that would otherwise create ambiguity in the training. The final result of the preprocessing was saved as preprocessed.csv files in the train, validation, and test split. These cleaned CSVs became the fundamental feed into the following procedures such as label binarization and HuggingFacebased tokenization. Preprocessing guaranteed the consistency of the data to a simple format, and thus, the multi-label classification pipeline provided for a robust and easy-scaled modelling of legal articles.

3.3 Data Splitting

The EURLEX dataset was divided into three categories (70% for training, 15% validation, and 15% testing) for facilitating the performance of the model and evaluation against any bias. Each subset of input text was pre-process and converted into transformer compatible forms generating input_ids and attention_masks for each such file. These were stored as individual files in form of .npy file -X train input_ids.npy, X validation input_ids.npy, as well as the X test input_ids.npy respectively and the corresponding attention_masks which are the same files with _attention_masks appended to it, respectively. The labels which initially came in string form were binarized with MultiLabelBinarizer and saved as y_train.npy, y_validation.npy and

y_test.npy. It was thus set up in a way that each model would have enough information to learn (train), tune (validation) and unbiasedly measure its performance (test).

The predictions were stored in y_pred_probs.npy during evaluation, which made metrics (such as F1 scores and precision-recall) more detailed possible. Such a systematic approach to splitting enabled data integrity and the distribution of labeled data across all subsets consistently, which made model performance reliable and scalable.

4 Design Specification

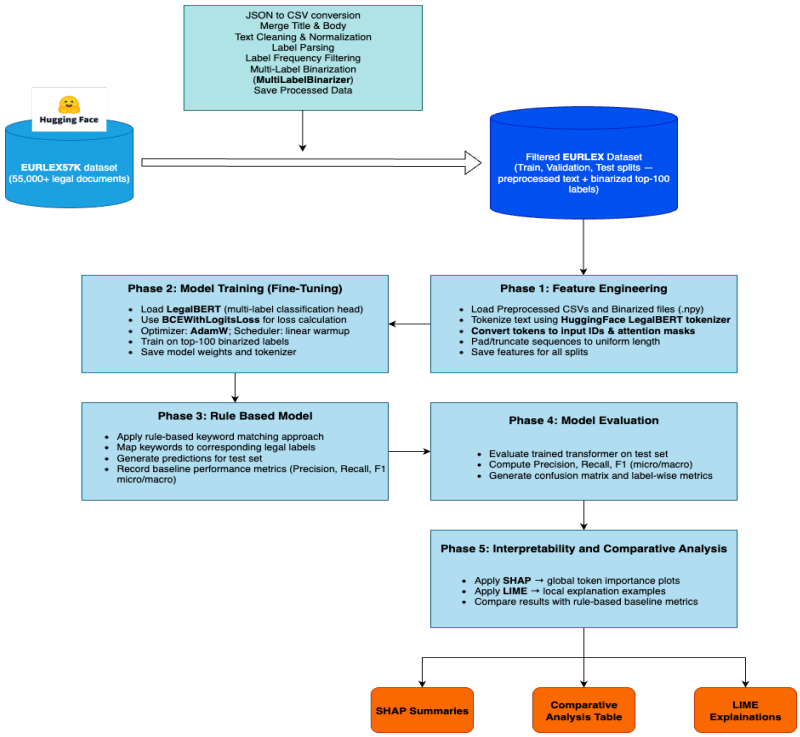

Figure 1: System Architecture Diagram

Figure 1 shows the system architecture diagram of the total workflow of an automated and developed rule extraction system based on transformer-based models in both financial and legal regulatory documents. The pip-line starts with the acquisition of the dataset, where all the legal documents along with the related EUROVOC multi-label annotations will be retrieved in JSONL format of the HuggingFace EURLEX corpus. This information is then converted into a tabular database form using a JSON-to-CSV conversion program so that downstream processing can be more easily performed. The second step is document preprocessing, including the concatenation of documents titles and bodies, text lowercasing, and elimination of the special characters. Label binarization is implemented via **MultiLabelBinarizer** on label lists to get binary vectors that can be trained using. Then, there is the stage of feature engineering during which RoBERTa tokenizer created in HuggingFace Transformers library is applied to tokenize some text that should be entered, create attention masking, and prepares padded sequences. After packaging input features and labels, training is done by use of the **AutoModelForSequenceClassification** with the loss described as **BCEWithLogitsLoss**. The

resultant trained model is subsequently tested on a multi-label analysis such as Precision, Recall, and F1-Score. Finally, SHAP, LIME, and attention visualizations are also added to give explainable explanations to the outcomes of the model and relate it with various classes of legal rules. Current support is end-to-end and may be used both in terms of performance evaluation and rule-based grounding, allowing the interpretable NLP systems to be deployed in legal-tech settings.

5 Implementation

5.1 Data Acquisition and Conversion

Project initiation involved downloading the EURLEX data set via HuggingFace which is an open-source collection of more than 55000 legal documents of the European Union that contains annotation in form of EUROVOC concept labels. They were given in three .json files: eurlex_train.json, eurlex_test.json and eurlex_validation.json. This was because in each of the files was given a dictionary of fields such as title, body, and concepts. Such files could not be used instantly to train and were required to be converted into a coordinate, tabular form. To do so, a Python script (load_json_as_df.py) was executable where each line (JSON) was converted to the relevant field in form of Pandas DataFrames. This was then stored as csv files: train.csv, test.csv and validation.csv in a structured data/ directory. The transformation secondly made the format of the dataset consistent to all downstream preprocessing and analysis.

5.2 Preprocessing and Label Encoding

All the preparation, such as cleaning and normalizing, of both the field with the legal text and the label field in the training, validation, and test CSV files was done with the use of a modular script preprocess.py. Every legal text had two textual elements- title and body which were merged to form a field containing a text. This amalgamation gave more semantics to nuts and bolts. The text was combined before normalizing the combined text to lowercase, by omitting special characters, newline symbols and unnecessary whitespaces to lower the complexity of the vocabulary. To clean the label names, we converted a list in concept field of each of the records (a list in string format such as arr = "[1045, 2012]" to an actual Python list via ast.literal_eval() in a safe manner. This was necessary to be compatible with MultiLabelBinarizer. Train preprocessed, validation preprocessed, and test preprocessed files were saved as train_preprocessed.csv, validation_preprocessed.csv and test_preprocessed.csv, respectively. Only the two fields text and labels were kept in each file. It was also checked by further quality control to exclude having any null values, labels with bad formatting or parts of text which were too lengthy.

5.3 Multi-Label Encoding via Binarization Techniques

During this phase, the names of the legal document labels (as textual EUROVOC concept codes in the list form, e.g. C_1234, C_4321) were converted into a suitable format to be used in the multi-label classification, the vectorized binary representations. This is essential since transformers-based models like LegalBERT can only accept numerical datasets in features as well as target labels. To attain this, the execution has made use of MultiLabelBinarizer which is a Scikit-learn package class encapsulated in a script named label_binarizer.py. The binarizer was then trained just on the training set to have the consistent encoding sequence and prevent any leakage of labels to the validation and test sets. The transformation transformed every sequence of labels into a binary vector of dimension (n labels,), with the elements taking the

value of 1 or 0 corresponding to the occurrence (or the absence) of a given category of law until n labels.

After the binarizer was fitted, the labels of all the three splits were transformed to training, validation and test. the arising binary arrays were stored under the form of .npy files (y_train.npy, y_validation.npy, y_test.npy) in the data/labels folder so that they may be easily loaded by the model at training time. Such arrays matched the feature files generated at the feature engineering step that were tokenized. Along with the label arrays, it was also stored as label_binarizer.pkl in the models/ folder as the fitted list of binarizers using joblib. This allowed reliable inversion of the predicted binary vectors into human readable label codes at both evaluation and interpretability stages.

The most important implementation item was the choice to filter the dataset by top 100 most frequent labels with reference to label distribution histogram. Such filtering step was necessary to enhance training stability, higher F1 scores and lower computational complexity of extremely sparse label matrices. Before the process of binarizing, any samples with labels that did not fit into this top 100 set were dropped in order to achieve consistency. Such a targeted aspect also provided that the downstream transformer model would have a manageable and balanced label space as well as enough coverage of the most prominent regulatory topics in EURLEX.

5.4 Feature Engineering with Transformer Tokenizers

The fourth step that occurred during the implementation procedure was to convert the cleaned data on legal text into a numerical form of features that needed to be presented by transformerbased models like LegalBERT. This has been done by adoption of a tokenizer-based vectorization approach, in its simplest form, this action leverages the AutoTokenizer API of HugoFace that was instantiated with a pretrained configuration called nlpaueb/legal-bert-baseuncased. The tokenizer has implemented subword tokenization with the WordPiece technology that converts each of its input string of texts to three fundamental parts namely input_ids, attention_mask, and token_type_ids (optional). The input_ids and attention_mask were the only two features used in this implementation, as per usual in one-input sentence classification. The raw input of the text used was the raw text that was gathered by concatenating title and body field during preprocessing and input to the tokenizer with truncation and padding functions on. It limited the number of tokens to 512 in any one sequence to fit within architecture constraints of BERT. Longer texts than this limit were cut; shorter sequences were padded with zeros to create uniformity. The resultant outputs were saved as NumPy arrays in data/features/ at the following names: X_train_input_ids.npy, X_train_attention_masks.npy, and so on, their validation and test counterparts. This made sure that the model was given fixed size input during mini-batch training and inference.

Batch tokenization was applied for efficiency, leveraging the tokenizer.batch_encode_plus() function with return_tensors='np' to directly produce NumPy-compatible outputs. A critical implementation detail here was maintaining consistent shuffling and ordering of the tokenized samples relative to the label arrays (y_train.npy, etc.) to avoid data-label misalignment during model training. Furthermore, this tokenization pipeline was stateless and could be rerun with different transformer models (e.g., RoBERTa or FinBERT) by simply changing the checkpoint path used in AutoTokenizer.from_pretrained().

5.5 Fine-Tuning LegalBERT for Multi-Label Classification

The most consequential modeling aspect of this project consisted in fine-tuning a domaintransfer learning transformer, namely LegalBERT, pertaining to the task of multi-label text classification over EURLEX regulatory texts. This available phase was executed in the train_model.py script with the help of the HuggingFace transformers library and the PyTorch backend. Model architecture. The model used is AutoModelForSequenceClassification that was initialized with the checkpoint nlpaueb/legal-bert-base-uncased. The chosen model was the most appropriate solution to the task as the pretraining on legal corpora made this model contextually aware of legal vocabulary, syntax, and semantics.

Because the EURLEX task can be considered to require multi-label classification (i.e. each document can belong to more than one EUROVOC category) the output head of the model was adjusted to this task. Output layer This last layer was a set of 100 outputs (neural units) corresponding to the 100 most common labels, and each was called by a sigmoid-based activation layer instead of a softmax. Because of this configuration, the model has the capability to have each class have its own probability. The loss function applied was BCEWithLogitsLoss that is particularly suited to training models on imbalanced multi-label data. The input data and labels were tokenized, put in arrays and provided with attention masks (input_ids and attention_masks) and binarized (y_train.npy, y_val.npy). Training was done with HuggingFace Trainer API, and it provided easy integration with model checkpoints, evaluation callbacks, and learning rate schedulers. AdamW was chosen as the optimizer, in which the early gradients are stabilized with a linear warmup schedule. Training of the model occurred between 4 to 6 epochs, with early stopping with respect to validation F1-score.

During every epoch, the training metrics were recorded and the intermediate model checkpoints delved into the models/legalbert_top100/ directory. These checkpoints contained the optimized model weights (pytorch_model.bin), configuration files, and tokenizer vocabulary making it reproducible and they can be used to evaluate or deploy it using downstream. To speed up the training, this was trained on a GPU Colab instance.

5.6 Baseline Rule-Based Classifier

For providing a comparative baseline to the transformer-based model, a rule-based classifier with the use of keyword matching was formulated. Such classifier was trained not on a set of data but on handcrafted rules that tag a doc based on the occurrence of common co-occurring legal words. It was striving to benchmark the advantage in performance obtained by deep learning as opposed to conventional pattern-based heuristics. The application was started by calling the filtered EURLEX data that contains the 100 most frequent labels (train_top100.csv and test_top100.csv). The eurovoc_concepts field which was represented by a list of strings was made into real Python lists through the use of theeval()function. Each of the legal texts was tokenized based on lowercase words with the help of regular expressions and numbers of its occurrences were counted per each label with the help of defaultdict and Counter. This enabled us to be able to map each label to terms that most frequently co-occur on the corpus. Only top 10 words were kept per label to decrease noise and increase specificity thus label_to_keywords dictionary was created. In the prediction process, when it was identified that any of these words was used in any of the test documents, then the related label was allocated.

The predict_keywords() function read the input text and searched it to find the matching keywords: when a set of matching keywords was found, the function outputted the list of all

the predicted labels. The approach enabled rapid classification, and the decisions could be supported by references to individual words. Performance was measured by the standard multilabel indicators precision, recall, and F1-score using the Scikit-learn precision_score, recall_score and f1 score functions in both the micro and macro averages.

5.7 Interpretability using SHAP and LIME

In order to establish a benchmark, a rule-based baseline classifier was introduced with the keyword matching instead of machine learning. The reasoning labeled its items in terms of the common occurs as well-used words and phrases in law books. This baseline evaluated the comparison of the performance of traditionally used algorithms, alongside transformer-based models in such areas as LegalBERT. The execution would start with creating the train_top100.csv and test_top100.csv files by loading the files having the top 100 most frequent labels. The eurovoc_concepts column that used to be filled with string lists was now transformed in Python and named lists using eval(). All the training documents were tokenized into lower case words over regular expressions and the co-occurrence frequencies of wordlabel calculated via defaultdict and Counter. Out of the 10 most common keywords, a rule base (label_to_keywords) was created on each label. When making a prediction, the presence of such keywords of a test document led to the label being assigned.

The results on prediction were estimated on the standard multi-label-scoring elements such as precision, recall, and F1-score (both micro and macro)-through Scikit-learn. This method mind has very simple classification, though it is clear, quick, and interpretable. The fact that it was tested poorly with legal context affirmed the higher degree of intricacy and accuracy of transformer models, but it can be a helpful resource in audit and compliance situations because of its simplicity.

6 Evaluation

6.1 Phase 1: Transformer-based LegalBert Classification

The LegalBERT transformed version of the model was tested in a cleaned version of the EURLEX data, over the 100 most common EUROVOC categories. The fine-tuning process was conducted within a multi-label classification task and tested with such standard evaluation measures as Precision, Recall, and F1 Score on micro and macro averages. LegalBERT was highly applicable with Micro-F1 score of 0.7774 and Macro-F1 score of 0.6993 indicating sufficiently good generalization over labels of different frequencies as shown in Table 6.1. It is worth noting that the model recorded a high Micro Precision of 0.8817, which demonstrates that the model correctly identifies the few positives during multi-label predictions. The Micro Recall of 0.6951 also denotes that the model identified most suitable labels of test samples.

Such scores prove to be notable because of the complexity in law text and similarity in semantics across labels.

Metric	Transformer (LegalBERT)
F1 Score (Micro)	0.7774
F1 Score (Macro	0.6993

Precision (Micro)	0.8817
Recall (Micro)	0.6951

Table 6.1.1: Evaluation Metrics of LegalBert model

Our results demonstrate that LegalBERT, as finetuned on legal datasets, can effectively capture the domain specific subtleties needed for multi-label legal classification and therefore is a reliable solution for compliance automation, legal tagging and regulatory document retrieval.

6.2 Phase 2: Rule-Based Baseline Classifier

A rule-based classifier was also adopted to set up a performance benchmark. It applied a keyword matching rationale to give a label primarily depending on the use of terms distribution of the training corpus. A dictionary based on the associating every label with its top 10 keywords (e.g., "taxation" → ["vat", "income", "excise"]) was created. These high frequency words were then labelled against a document that has one or more of these words.

Results on Test set:

Metric	Score
F1 Score (Micro Avg)	0.2972
F1 Score (Macro Avg)	0.1836
Precision (Micro Avg)	0.3412
Recall (Micro Avg)	0.2703

Table 6.2.1: Evaluation Metrics for Rule-Based Baseline Classifier

The rule-based system was understandable and easy, but its performance was poor because of the failure to consider the surrounding environment, synonyms, and legal terminology with its intricacies. Nonetheless, it can be used in the traceability or compliance audit due to its interpretability and transparency.

6.3 Phase 3: Comparative Analysis of Results

Both approaches are visually compared in Table 6.3. Transformer model shows much more precise-recall balance and even learns more pattern-rich labels. Rule-based approach works well only with very high frequency and simple matchable labels.

Metric	Transformer (LegalBERT)	Rule-Based Baseline
F1 Score (Micro)	0.7774	0.2972
F1 Score (Macro	0.6993	0.1836
Precision (Micro)	0.8817	0.3412
Recall (Micro)	0.6951	0.2703

Table 6.3.1: Comparative Performance of LegalBERT vs. Rule-Based Baseline on EURLEX Top 100 Labels

6.4 SHAP-Based Interpretability

It was analysed on the fine-tuned LegalBERT model, which was trained on the EURLEX dataset (subset of top 100 labels). With SHAP TextExplainer, a tokenized document has been forwarded to the model and SHAP values were calculated with respect to all predicted labels per-token. Positive SHAP denotes that a token raises the probability of a particular label being given, whereas it takes a negative value when the influence is suppressive.

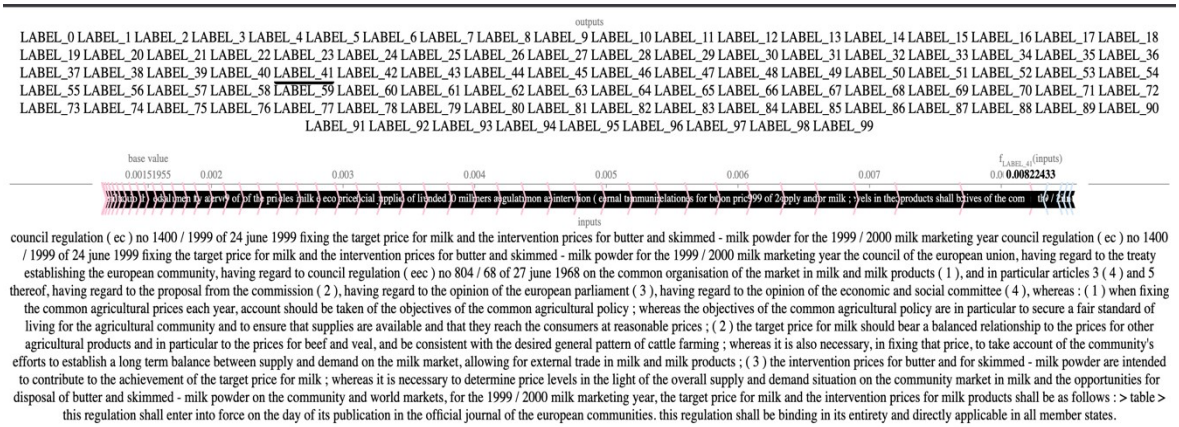

Figure 2: SHAP explanation of high confidence correct prediction on a text in dairy regulation, indicating the positive or negative contribution of the key tokens in the decision made by the model.

Figure 2 shows SHAP visualization of the high-confidence correct prediction when the model made a correct prediction with the label of the dairy price regulations. Examples of terms presented in the feature of these words like "butter" and "milk" had great positive SHAP values, which means that these words play a significant role in raising the likelihood of the correct label prediction. The scale and density of such contributions indicate that the model was able to grasp domain-specific keywords that are targeted in the legal environment.

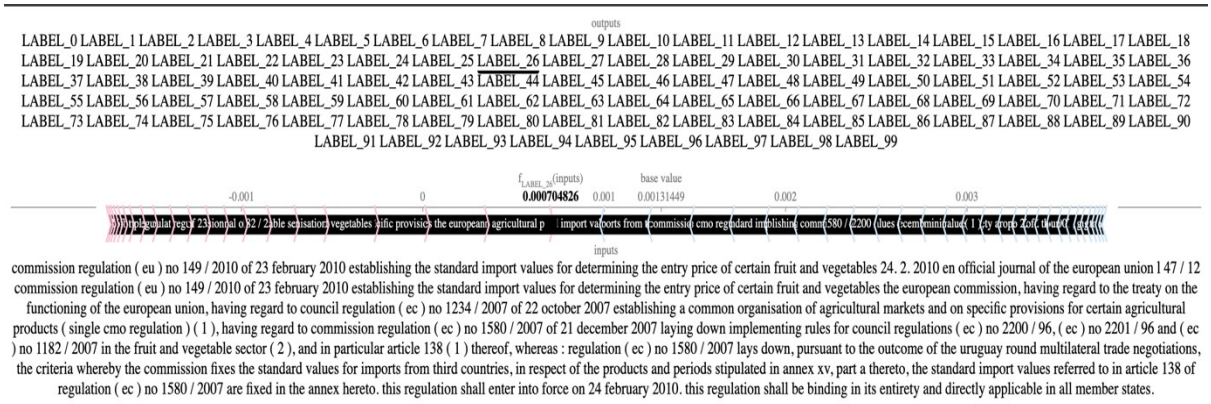

Figure 3: The explanation of a prediction of an agricultural import regulation, which shows the effect of individual tokens on the classification outcome of the model.

Figure 3 presents another scenario, which is connected to agricultural import regulation where SHAP indicates distinct feature importance pattern. In this case, words like “vegetables”, “import”, and “provisions” were the major elements that influenced the decision of classification. Although the target label was predicted correctly, SHAP values were more dispersed than in the first case, and this is associated with the wider contextual dependencies involved in this class.

6.5 LIME-Based Interpretability

The Local Interpretable Model-Agnostic Explanations (LIME) method was used. Implementation of LIME involves generating perturbed versions of the original text and local training a simple interpretable model around the instance to find the most influential words to aid in a particular label prediction.

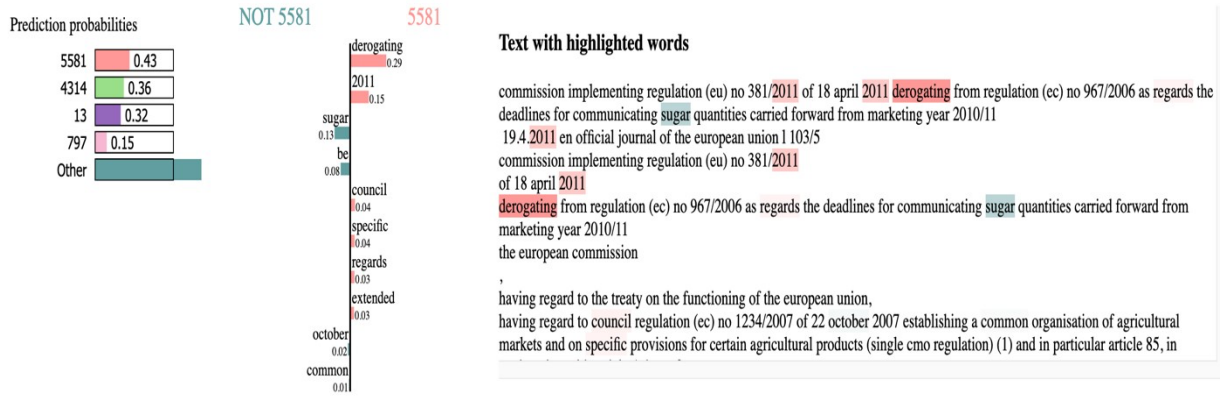

Figure 4: LIME exposition of a moderately confident correct forecast on a sugar market regulation text, with the local word contributions causing the classification.

Figure 4 demonstrates a moderate confident prediction situation. The most influential terms like “derogating”, “sugar” and the year of 2011 are shown in red and blue colour, and they predominate in raising or diminishing the likelihood of the target label. Its probability given this label was predicted at 0.43, meaning that although the contextual cues (of relevance) were observed, the model was not fully confident about the outcome.

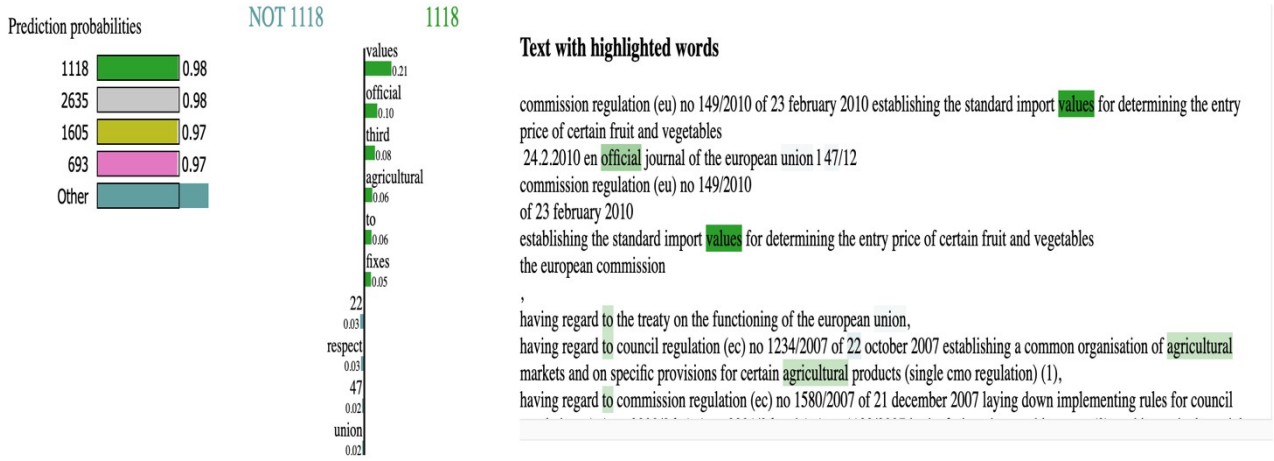

Figure 5: High-confidence correct prediction on fruit import regulation text explained by LIME.

Figure 5 represents a high-confidence situation, in contrast. Such terms as “values” and “agricultural” are highlighted heavily in green, and the target label (1118) and relevant labels (2635, 1605, and 693), confidence scores are shown to be 0.97 and 0.98. It is an indication that the model has clearly established a high semantic and context compatibility between the input document and the legal category to which it belongs.

6.6 Visual Insights on Model Performance

Figure 6 shows the first visualization in which we give the distribution of the labels around the top 100 EUROVOC labels in the filtered EURLEX data. The frequency distribution of this bar chart clearly takes the long-tail form as there are only a number of labels including Taxonomy, Fisheries, and Common fisheries policy that show the high frequency compared to the rest. This imbalance that occurs in the natural world is representative of real-world legal documents practices, as well as causing a challenge when training robust models. Unless some compensating measures are used (e.g., to counter overfitting, class weights or micro/macro averaging), models will be overfit to common labels. The realization of the existence of this imbalance was thus important in the choice of the right performance metrics as well as in the subsequent decision to trust the use of micro as well as macro F1 scores.

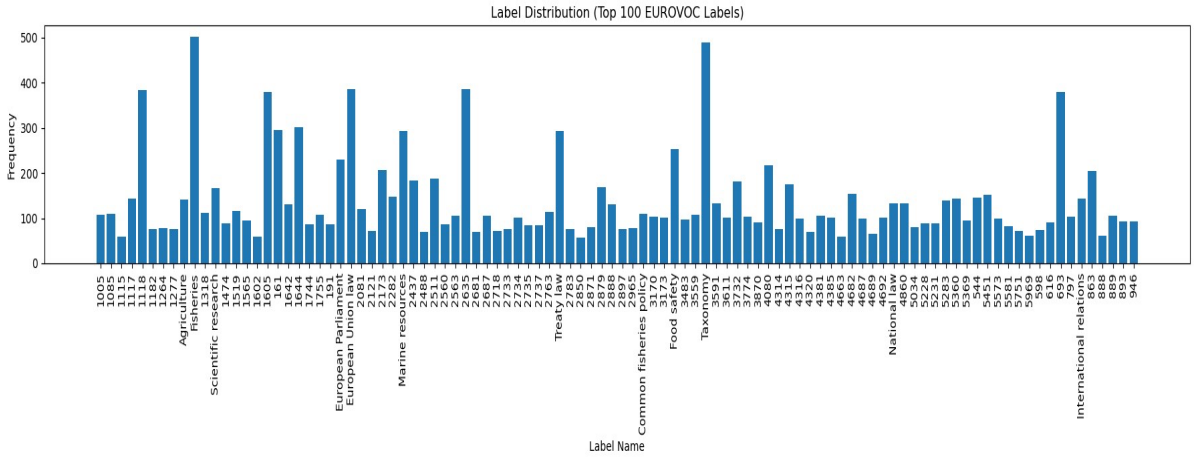

Figure 6: Label Distribution (Top 100 EUROVOC Labels)

Figure 7 is a series of the bar plots further disaggregate the performance data of each of the top 100 labels to provide a finer level of how well the model is performing in each of the classes. In the first plot, the precision is depicted, and this is what indicates the number of the labels predicted as being correct among the total labels being correct. High precision suggests that there are fewer false positives- which can be useful in legal scenarios as it can over-classify and cause misinterpretation by downstream processes. The second graph echoes recall, or the number of actual labels that the model managed to recall. Such a measurement is particularly valuable in terms of compliance since failure to label can imply a failure to comply with a rule. A combination of the two into F1 score is in the third plot where both are weighted. These plots indicate that there is a significant variability in the performance between labels: some labels, such as “European Union law” and “Food safety” are consistent high, whereas others have either poor recall or skewed scores, probably due to either low representation of the training data, or because the labels are more ambiguous.

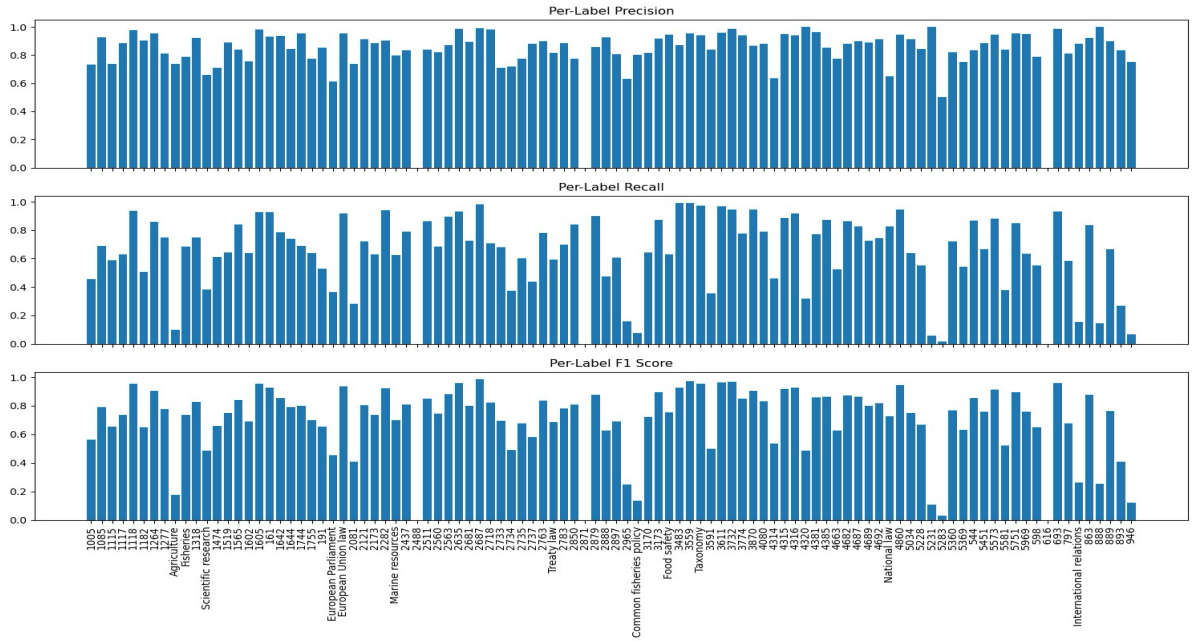

Figure 7: Per-Label Precision, Recall, and F1 Scores

Figure 8 depicts three top frequency labels, namely, Fisheries, Taxonomy and European Union law, are selected in this plot, displaying precision-recall curves and average precision (AP) scores of those labels. Taxonomy and European Union law have almost a perfect PR curve (AP = 0.97) which implies that the model is very sure and accurate on these predictions across thresholds. Conversely, the curve of Fisheries is a little lower (AP = 0.82); this means that, although it remains strong, the classification is less certain. The perspective aids in justification by ensuring probabilistic confidence of the model fits the label frequency and complexity, and it also informs calibration of the thresholds in case the model is to be used in a high-stakes legal environment where the false positive and false negative trade-offs should be optimized.

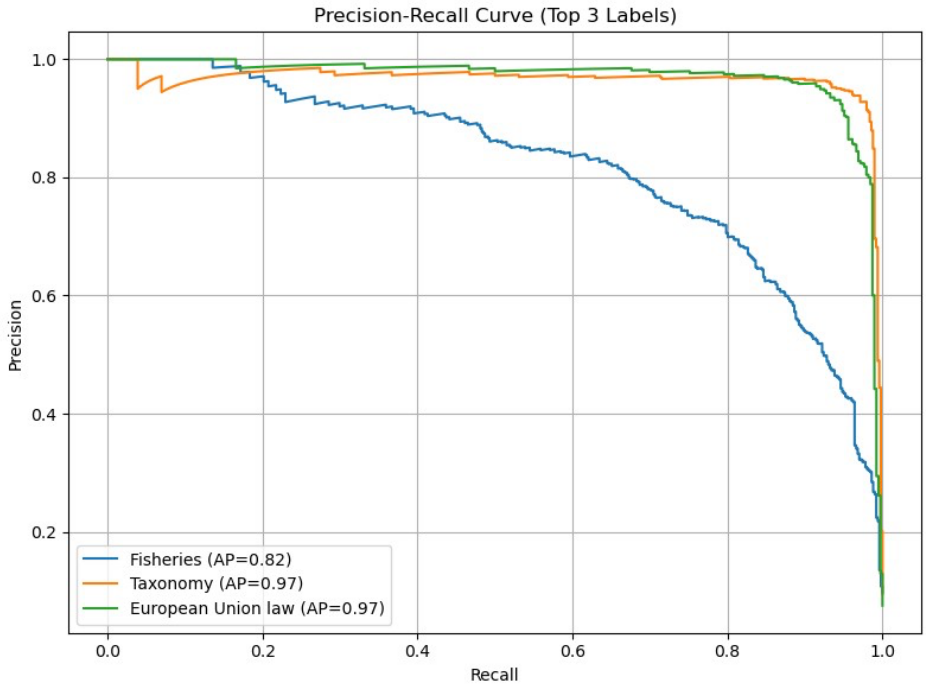

Figure 8: Precision-Recall curve for Top 3 Labels

The error analysis plot of the false negatives (the missed labels) and false positive (the extra labels) shown as Figure 9 is disaggregated in the top 10 problematic labels. The left panel enumerates non-predictions of the model the most frequently that in fact were ground-truth (e.g., the label 01309 was missed more than 150 times and so on). The right panel indicates labels, which the model made frequent predictions when it should not-- indicating label confusion or semantic overlap. Such plots come in handy in improving the training process as well as the interpretability strategy. As another example, regularly missed labels can possibly be improved by addition or special attention during tokenization, and false positives can serve an indication of label co-occurrence that may be better represented through either classifier chains or label correlation modelling.

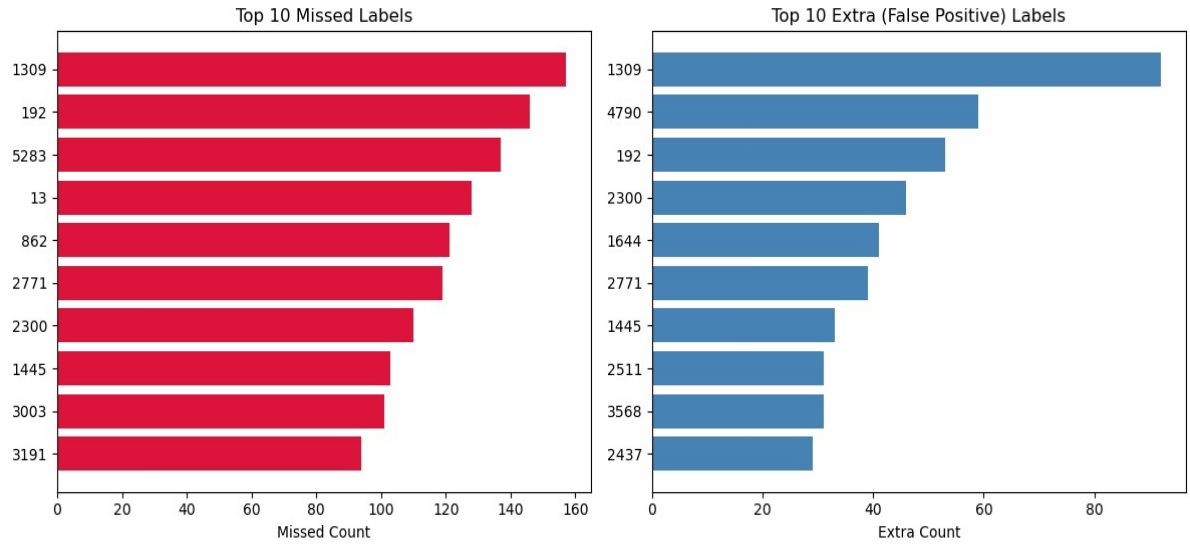

Figure 9: Top 10 Missed and Extra Labels (Error Analysis)

6.7 Discussion

The comparative analysis of the fine-tuned LegalBERT and the baseline rule-based algorithms showed a significant gap between the performance of these models, with the transformer-based approach achieving superior results across all key metrics—F1 (micro and macro), precision, and recall—consistent with prior findings by Chalkidis et al. (2020) and Devlin et al. (2019) that transformer architectures are well-suited for capturing the semantic context in complex legal and regulatory documents. This is a feature that strictly keyword-based methods, as described by Ashley (2017), cannot adequately approximate.

Although these results are promising, they require careful interpretation. The higher microaveraged F1 scores compared to macro-averaged scores highlight a disparity in performance between frequent and rare labels, a common challenge in multi-label legal classification tasks, as noted by Chawla et al. (2002) in the context of class imbalance. While the top 100 label filtering approach improved training efficiency, it inevitably biased the model toward more common classes, reducing performance on rarer labels.

Interpretability analysis using SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) revealed that both methods frequently identified domain-specific terms—such as “butter,” “agricultural,” and “import values”—as influential in classification decisions, aligning with the need for contextually significant features in legal NLP. However, some visualizations indicated over-reliance on frequent but semantically generic terms, echoing concerns raised by Serrano and Smith (2019) regarding spurious associations in attention-based models. These patterns may compromise robustness when applied to unseen legal scenarios.

Methodologically, employing the top 100 labels subset offered a more computationally efficient and interpretable framework, but at the cost of reduced generalisability across the full EURLEX

label space. As suggested by Chalkidis et al. (2021), hierarchical classification leveraging the EUROVOC taxonomy could maintain scalability without sacrificing label coverage.

Limitations of this study include computational constraints that restricted hyperparameter tuning and manual experimentation, potentially affecting performance outcomes. Although the rule-based baseline was deliberately simplistic for comparative clarity, future research could incorporate more sophisticated or hybrid baselines (Boella et al., 2013) to provide a stronger benchmark.

Limitations	Impact on Study
Other transformer models (BERT, RoBERTa, FinBERT) were mentioned but LegalBERT was the only one that was fully implemented.	Limits the area of model comparison and minimizes generalizability of results.
There is a high imbalance in the classes of EURLEX dataset.	Ratings on biases favor frequent classes, restricting performance on infrequent ones.
Only SHAP and LIME post-hoc methods were interpretable.	Does not model more internal processes of the model like attention distribution or gradient-based attributions.
Inadequate experimentation because of computational and time constraints.	Restricts testing scalability across large datasets and cross-domain transferability.

Table 6.7.1: Summary of study limitations and their impact on the scope, interpretability, and generalizability of results.

Overall, these experiments reaffirm that, when fine-tuned effectively, transformer-based models can outperform traditional legal text classification methods while offering post-hoc interpretability. Nevertheless, addressing label distribution imbalances, exploring hierarchical classification, and repeating interpretability analyses remain crucial for achieving robustness and practical deployment.

7 Conclusion and Future Work

The research largely succeeded in achieving its goals. These results demonstrated that the finetuned LegalBERT excelled in all the main metrics, such as micro and macro F1, precision, and recall, over the rule-based baseline, which proves the effectiveness of its use when it comes to capturing the contextual and semantic complexity of legislative and regulatory documents. Although the micro-F1 were high, the macro-F1 was low reflecting continued label imbalance with common labels being predicted in a better way as compared to the rare labels. Analysis with interpretability tools, SHAP, and LIME, confirmed that most of the predictions were constructed on the legal terms used, but sometimes, generic terms were also used, which shows the possibility of spurious correlations. The top 100 label filtering strategy enhanced the training speed with the unavoidable decrease in the coverage of less frequently occurring categories.

As a future of work, label balance methods, e.g. focal loss, cost-sensitive learning, or labelaware sampling might enable to bridge the performance gap with infrequent labels. Subordinate classification based on EUROVOC may increase the coverage of labels without critical decrease of accuracy. Better baselines that contain more complex rule-based systems (or hybrids) would serve as a benchmark of relative performance improvements better. Specialization or real-environment applicability could also be enhanced through domain adaptation via transfer learning on similar legal or financial corpora and integration into compliance monitoring systems with the ability to constantly be updated with new regulations. This study confirmed that transformer-based models can be used to both have high predictive performance and a significant post-hoc interpretation within the legal NLP field. Label imbalance, label coverage/promising mechanisms, and interpretability refraction may lead to more robust and scalable graduate AI parasites, eliminating the dichotomy that exists between academic and practical implementation.

References

Beltagy, I., Lo, K. and Cohan, A., 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*. Available at: https://arxiv.org/abs/1903.10676.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N. and Androutsopoulos, I., 2021. MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *Journal of Artificial Intelligence Research*, 72, pp.481–520. Available at: https://doi.org/10.1613/jair.1.12753.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P. and Androutsopoulos, I., 2020. LEGAL-BERT: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*. Available at: https://arxiv.org/abs/2010.02559.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT 2019*, pp.4171–4186. Available at: https://arxiv.org/abs/1810.04805.

Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 30. Available at: https://arxiv.org/abs/1705.07874.

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. "Why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pp.1135–1144. Available at: https://doi.org/10.1145/2939672.2939778.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 30. Available at: https://arxiv.org/abs/1706.03762.

Zhang, H., Cisse, M., Dauphin, Y.N. and Lopez-Paz, D., 2018. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*. Available at: https://arxiv.org/abs/1710.09412.

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P., 2017. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp.2980–2988. Available at: https://doi.org/10.1109/ICCV.2017.324.

Zhou, Z.H. and Liu, X.Y., 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1), pp.63–77. Available at: https://doi.org/10.1109/TKDE.2006.17.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, pp.321–357. Available at: https://doi.org/10.1613/jair.953.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. Available at: https://arxiv.org/abs/1907.11692.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q.V., 2019. XLNet: Generalized autoregressive pretraining for language understanding. In: *Advances in Neural Information Processing Systems (NeurIPS)*, 32. Available at: https://arxiv.org/abs/1906.08237.

Chalkidis, I., Androutsopoulos, I. and Aletras, N., 2019. Neural legal judgment prediction in English. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.4317–4323. Available at: https://doi.org/10.18653/v1/P19-1424.

Niklaus, J., Chalkidis, I. and Handschuh, S., 2021. Legal judgment prediction with linguistic knowledge enrichment. In: *Proceedings of the 18th International Conference on Artificial Intelligence and Law (ICAIL)*, pp.79–88. Available at: https://doi.org/10.1145/3462757.3466093.

Serrano, S. and Smith, N.A., 2019. Is attention interpretable? In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp.2931–2951. Available at: https://doi.org/10.18653/v1/P19-1282.

Sundararajan, M., Taly, A. and Yan, Q., 2017. Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp.3319–3328. Available at: https://arxiv.org/abs/1703.01365.

Chalkidis, I., Fergadiotis, M., Kotitsas, S., Malakasiotis, P., Aletras, N. and Androutsopoulos, I., 2020. Legal-BERT: The muppets straight out of law school. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp.2898–2904. Available at: https://doi.org/10.18653/v1/2020.findings-emnlp.261.

Cohan, A., Beltagy, I., King, D.R., Dalvi, B., Weld, D.S., Lo, K. and Ammar, W., 2019. Pretrained language models for biomedical and clinical tasks: BERT and beyond. *arXiv preprint arXiv:1905.07989*. Available at: https://arxiv.org/abs/1905.07989.

Zhang, M., Zhang, Y., Vo, D.T. and Lam, W., 2021. Exploring keyphrase-aware transformer for legal case retrieval. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.1473–1482. Available at: https://doi.org/10.1145/3404835.3462865.