Abstract

The increase in regulations of artificial intelligence (AI) worldwide, including the European Union AI Act, requires technology companies to evaluate compliance with regulations in their patent portfolio in a scalable manner. These dense, technical documents are not scalable to review manually, and they are likely to contain errors. Furthermore, conventional automation solutions fail to capture the contextual difference of the risks of AI.

This project details the design and the test of a supervised learning pipeline to solve this problem. An important methodological contribution is that a data preparation strategy that mitigates the extreme data sparsity by combining 33 risk labels into four categories. This was followed by the fine-tuning of a domain-specific Legal-BERT model on an human annotated data and employing a weighted loss function and a multi-stage optimization procedure.

The performance of this fine-tuned model was systematically evaluated against several established baseline approaches, where it demonstrated significantly improved efficacy in detecting multifaceted, context-specific risk patterns. The dissertation concludes with a thorough quantitative and qualitative investigation of the behavior of the final model and confirms the viability of this domain-specific methodology. The study provides a viable, reproducible process that enables the scalable screening of patent literature to provide a v practical pathway to support regulatory oversight in the field of AI.

Contents

Nomenclature

AI	Artificial Intelligence
BCS	British Computer Society
BERT	Bidirectional Encoder Representations from Transformers
CSV	Comma-Separated Values
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
IET	Institute of Engineering and Technology
IoT	Internet of Things
JSONL	JSON Lines
GDPR	General Data Protection Regulation
LLM	Large Language Model
NLP	Natural Language Processing
PII	Personally Identifiable Information
SBERT	Sentence-BERT
TF-IDF	Term Frequency-Inverse Document Frequency

List of Figures

List of Tables

1 Introduction

1.1 Overview

This project outlines the design, development, and evaluation of a supervised learning model for *Artificial Intelligence (AI)* risk automated classification of patent text. To address the demand to comply with complex regulatory regimes, such as the European Union's AI Act, this project employs cutting-edge *Natural Language Processing (NLP)* techniques to develop a domain-specific methodology. One of the main contributions of this work is a complete workflow for data processing that fully integrates unstructured patent texts with goal-oriented, human-annotated risk snippet. Such a strategy addresses the key issues of extreme data sparsity and class imbalance through strategic label aggregation and weighted loss function to yield a highly optimized and effective classification model.

One of the central concerns of this research is coping with extreme sparsity and class imbalance in data, difficulties that disqualify standard classification methods. The approach is therefore characterized by two main strategic interventions: first, the condensing of multiple sparse risk labels into a small number of robust, statistically valid aggregated categories; and secondly, the use of a weighted loss function that encourages the model to learn from infrequent, risk-positive instances. The third approach involved fine-tuning a domain language model, and then this was further optimized following a hyperparameter search process. This report provides a detailed account of the entire process, from the initial data collection and preparation through to the ultimate test of the model, which achieved robust predictive performance, particularly on top-risk categories.

1.2 Motivation and Significance

The global introduction of comprehensive regulations of AI, as in the European Union's risk-based AI Act, has created a critical implementation gap for technology companies. For tech firms with massive intellectual property portfolios, compliance with regulations necessitates thousands of patents to be examined systematically. The task is fundamentally challenging since these documents are characterized by complex legal and domain-specific language, and human examination is a laborious, error-prone, and ultimately unscalable process.

This project is expressly motivated by such limitations. It addresses the limitation of simple keyword-matching systems by developing a sophisticated tool based on supervised machine learning. By learning from a dataset of human-annotated content, the system learns to detect the subtle, context-dependent linguistic cues to risk. The approach enables scalable, consistent, and early-stage screening of patent portfolios, thereby enabling organizations to effectively manage prospective regulatory problems.

1.3 Research Aims and Objectives

This project establishes a supervised learning setting for automatic AI risk classification from patent documents. Transcending the limitations of traditional classification approaches in highly unbalanced and domain-specific data, it utilizes a fine-tuned language model specific to the law and technology contexts. The system takes patent abstracts and claims as input and produces multi-label classification of four merged risk classes.

The workflow envisioned generates two main outputs: (1) a classifier able to identify subtle patterns of risk-oriented language; and (2) a replicable process merging automated hyperparameter tuning with per-category threshold adjustment. To apply the European Union's AI Act's risk typology actionable in patent data sets, we evaluate appropriate metrics and employ weighted-average F1 score as the primary metric, with class-level diagnostics in addition. The model is trained and tested on a dedicated dataset that has been obtained from lens.org [1] and manually labeled. Additional preprocessing mechanisms, including contextually driven construction of dependent claims, are introduced to further improve input fidelity. This work aims to contribute to the field of legal informatics in offering an organized method to project technical patent text into novel regulation categories.

Therefore, our main goals and contributions are outlined as follows:

- We propose a complete workflow for AI risk classification of patents encompassing data collection, preprocessing, model building, and final optimization stage.
- We investigate and implement a targeted label aggregation technique in order to address data sparseness in the initial set of 33 risk categories.
- We describe approaches to improving the representational quality of legal text, with a focus on constructing context-aware inputs for dependent claims.
- We deploy and compare a weighted loss function to mitigate the effect of extreme class imbalance under multi-label constraints.
- We contrast a domain-adapted, fine-tuned model to classic and semantic-based baselines, with ensuing performance gains.
- We introduce a post-training threshold calibration procedure with the goal of improving per-category F1 scores and stabilizing classifier usefulness.

1.4 Outline of the Dissertation

The dissertation is organized as follows:

- Chapter 2 (Background and Related Work): The chapter provides a literature review, the contemporary approach to legal text analysis and imbalanced classification, and the key technologies that form the foundation of this project.
- Chapter 3 (Objectives, Specification and Design): The chapter restates the project objectives and the technical and functional specifications of the system. It then describes the design solution adopted, other designs that were being considered, and provides a justification for the selected design.
- Chapter 4 (Methodology and Implementation): This chapter provides a sequential, detailed description of the implementation of the project. It covers all phases of the workflow beginning from data preparation and collection to training and optimization steps for the end model.
- Chapter 5 (Results, Analysis and Evaluation): The chapter explains observed results of the experiments. It provides the comparison analysis between baseline models and the ultimate proposed solution and followed by a quantitative and qualitative analysis of the performance of the final model in detail.
- Chapter 6 (Legal, Social, Ethical and Professional Issues): This chapter provides a rational justification for the legal, social, ethical, and professional issues in constructing and

deploying an automated AI risk classification system.

- Chapter 7 (Conclusion and Future Work): The final chapter of the present project concludes by summarizing the major findings and outlining the major contributions. It also goes further to discuss the limitations of the current study and outlines areas of future research.

2 Background and Related Work

2.1 Overview

To provide context for the methodological choices in this project, the following chapter introduces the required background and previous research. This chapter is organized into four key areas, beginning with the broader regulatory context and progressing to specific technical challenges and their established solutions. Firstly, we outline the AI regulation and risk frameworks that offer the real-world rationale for this study. Second, the chapter examines the challenges of legal and privacy-sensitive text, combining an introduction to NLP application in the legal sector with a privacy review. Third, we present the core NLP technologies for text understanding, defining the two primary paradigms evaluated in this project: semantic similarity with *Sentence-BERT (SBERT)* and supervised fine-tuning. Finally, to tackle the significant issue of data imbalance, we review several techniques for imbalanced classification.

2.2 AI Regulation and Risk Frameworks

2.3 The Challenge of Legal and Privacy-Sensitive Text

2.4 NLP Technologies for Text Understanding

2.4.1 Semantic Representation with *Sentence-BERT (SBERT)*

2.4.2 Supervised Fine-tuning with Domain-Specific Models

2.5 Techniques for Imbalanced Classification

3 Objectives, Specification and Design

3.1 Overview

In this chapter, we describe the project objectives and the rationale of the system design. The chapter begins by restating the underlying problem as well as the specific objectives of this research. The chapter then describes the technical and functional requirements which the system was designed to meet. Lastly, it provides a general description of the final system design, including a description of alternative approaches being considered and an evidence-based justification of the adopted methodology.

3.2 Problem Overview and Objectives

The core problem addressed by this project is the automatic multi-label classification of patent texts into four pre-defined aggregated categories of AI risk: Unacceptable risk, High risk, Transparency risk, and Human right. This is an activity that is characterized by severe class imbalanced dataset where risk-positive examples are exceptionally rare compared to non-risk examples. The underlying objective is therefore to build a model which, in addition to performing well, as measured in terms of the weighted F1 score, also provides a reproducible and transparent pipeline to detect these identified risks. The system must be capable of processing the complex, domain-related vocabulary of patents to create reliable classifications that can support regulatory compliance and responsible innovative initiatives.

3.3 Requirements and Specifications

The present section spells out the specifications and requirements of the project in question. The requirements part points the exact functionality that the model should achieve, and the following technical specifications subpart defines hardware, software, and dataset requirements.

3.3.1 Functional Requirements

The system was designed to meet a range of key functional requirements. The system needs to accept a patent text string (abstract or claim) as input and, after processing, provide an independent classification decision for each of the four pre-defined risk aggregated groups. The system is also required to be capable of handling advanced legal and technical vocabulary found in patent texts, enabling it to learn from contextual cues rather than relying solely on simple keyword matching.

3.3.2 Technical Specifications

For the experiments, the technical specifications are described as follows:

- **Hardware**: All experiments and hyperparameter tuning were conducted using Google Colaboratory [13], with a T4 GPU for calculation.
- **Software**: The experiment was implemented in Python 3, with the primary use of the PyTorch [14], Transformers [15], [16] and Scikit-learn [17] libraries (more detail on supplemental files).

- **Dataset**: The primary dataset was a collection of manually labeled CSV documents and a larger .jsonl document of patent data.

3.4 Design

In this subsection, we outline thoroughly our system design. We start with the design paradigm that we adopted, a comprehensive supervised learning process based on a fine-tuned, domain-specific language model. We then outline the alternative designs that were considered and subsequently dismissed, i.e., a semantic similarity-based and a machine learning baseline, and present experimental results highlighting their limitations. Lastly, we provide a full explanation of why we selected the fine-tuning method as our final design.

3.4.1 Design Approach

The selected design is an entire supervised learning pipeline that begins with a preprocessing stage which involves cleaning the raw data, merging sources, and constructing contextual text features. This is followed by a scheduled data aggregation step to create four stable target risk groups. The centerpiece of the design is the fine-tuning of a domain-specialized Legal-BERT [18] model, supplemented with a custom training process that incorporates a weighted loss function to address class imbalance. The training process is also improved through an automated hyperparameter search. The process is concluded by a crucial post-processing step where the prediction threshold of each risk class is individually optimized to achieve the maximum possible F1 score.

3.4.2 Alternative Designs

To validate the selected design, two primary alternative solutions were implemented and evaluated as baselines. The first, a semantic similarity solution, employs a Sentence-Transformer to determine the semantic relevance between risk definitions and patent documents. The second, a traditional machine learning solution, employs *Term Frequency-Inverse Document Frequency (TF-IDF)* vectors in combination with a logistic regression classifier for prediction. The results of these baselines are presented in Table 1, providing the comparative context for the final model's evaluation.

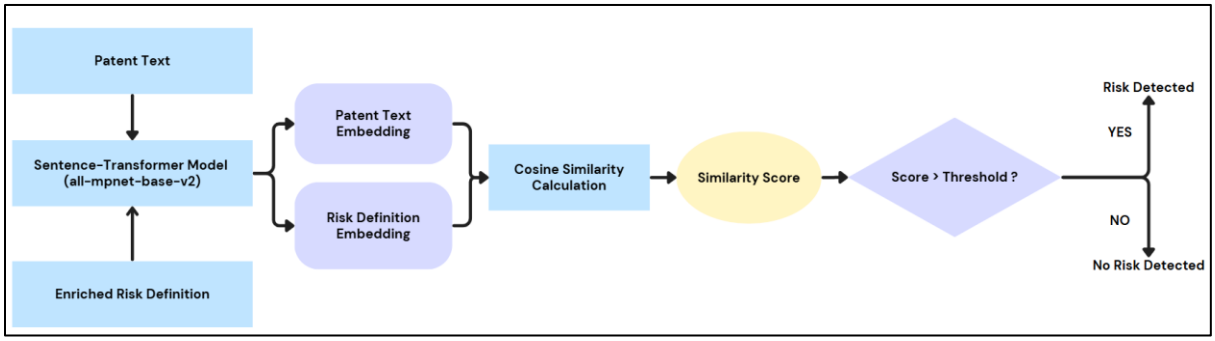

Figure 1**: Process flow of the semantic similarity design process.** A patent text and a risk definition augmented with the additional information are encoded independently by a pre-trained Sentence-Transformer model (all-mpnet-base-v2) [19] to produce high-dimensional vector representations. The cosine similarity between the two vectors is then calculated, producing a score between -1 and 1. If the similarity score exceeds an optimized threshold, the text is classified as 'Risk Detected'; otherwise, it is classified as 'No Risk Detected'.

Alternative method 1: Semantic similarity using Sentence-Transformers

The process of this approach, illustrated in Figure 1, is semantic matching rather than supervised training. The method involved using a pre-trained Sentence-Transformer model (all-mpnet-base-v2) [19] to generate high-dimensional vector embeddings of both the patent documents and enriched, accurate definitions of each of the four risk macro-categories. The cosine similarity between the vector of a patent document and every vector of the risk definitions was then computed. A patent was marked for a specific risk if its similarity score exceeded a globally optimized threshold.

Observed Failure: As evident from Table 1**,** although this approach was computationally cheap, its performance (Weighted Avg F1 score: 0.19) was low. The model was effective at identifying topical relevance such as marking up any text that included "biometrics" or "scoring". However, it consistently failed to distinguish between a benign mention of technology and a description of a risky use of that technology. The general semantic understanding of the model was lacking concrete; explicit knowledge of legal and situational risk required for this task and generated a large number of false positives and a low F1 score.

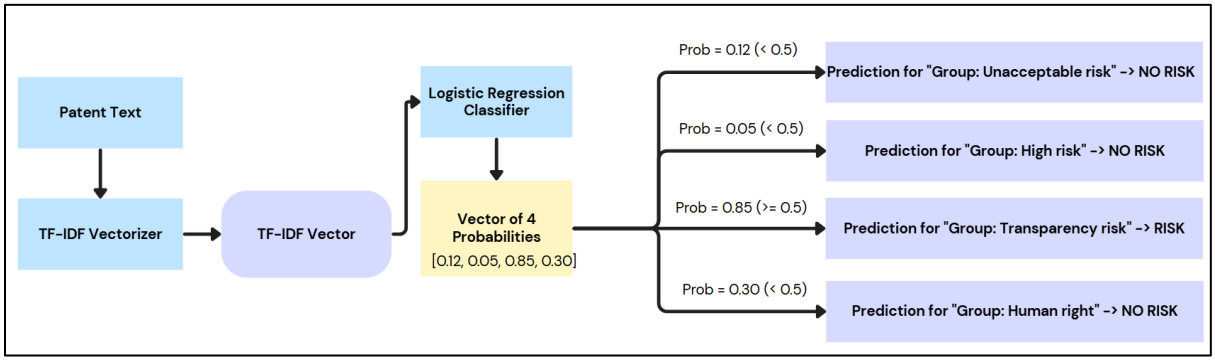

Figure 2**: Process workflow of the TF-IDF and logistic regression design approach.** A patent document is processed as input initially by a TfidfVectorizer, which converts the document into a high-dimensional numerical vector based on the word frequencies. This TF-IDF vector is applied to an already trained logistic regression classifier. The classifier gives a vector of four distinct probabilities, one for each risk aggregated category. As shown, each probability is then associated with a decision boundary (e.g., 0.5) to make a final binary classification for that category.

Alternative method 2: Classical Machine Learning with TF-IDF

This approach, illustrated in Figure 2, served as a traditional NLP baseline. It involved converting patent texts into numerical vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm, which considered both single words (unigrams) and two-word phrases (bigrams). A OneVsRestClassifier using a logistic regression base model was subsequently trained on these TF-IDF vectors.

Observed Failure: The initial run with this design, without alteration of the class imbalance, experienced overall model failure (F1 score: 0.00), as the classifier learned to predict only the majority class ("no risk"). This outcome perfectly showcased the risks in applying standard algorithms to highly imbalanced data. However, after incorporating the class_weight='balanced' parameter, performance improved substantially, yielding a respectable baseline F1 score of 0.55. Nevertheless, this result was significantly lower than that achieved by the fine-tuned model, further confirming that TF-IDF's feature space is less powerful than deep contextual embeddings for this task.

Model	weighted average F1 score
Semantic similarity (Enhanced)	0.19
TF-IDF + Logistic Regression (Standard)	0.00
TF-IDF + Logistic Regression (with class weight= balanced)	0.55

Table 1**: The alternative design baselines' performance on the four macro-categories.** This table provides the final weighted average F1 scores that were achieved by the two competing approaches on the validation set.

3.4.3 Justification for Selected Design

There are three main reasons why the fine-tuning direction was used instead of the other architectures. The former is that fine-tuning enables the model to explicitly acquire the specific patterns and intricacies of risk in our labelled data, leading to much stronger performance than the more generally applicable semantic similarity and TF-IDF correlations. Second, the existence of a weighted loss function in the training process proposes an edge clear and powerful means of addressing the implicit issue of the imbalance of classes, neither brought easily nor adapted in the other architectures. Third, the high quality of the selected fine-tuning approach can also be seen in the final weighted average F1 score, 0.73 which is a result significantly higher than the performance level of the alternative structures offered as a comparison (Table 6). The efficacy of the model is also supported by the remarkable F1 score of 0.88 that was obtained on the crucial class of high risk.

4 Methodology and Implementation

4.1 Overview

This project attempts to produce an autonomous risk classification system of AI in patent documentation. To achieve this aim, the following two main tasks in the methodology include (1) preparation and consolidation of the risk classes, and (2) supervised fine-tuning of the classification model. These components are explained further in the following discussion, including the process of selecting and preparing datasets, consolidation of risk classes, selection of model architecture, the objective function to be used to overcome the class imbalance problem, and the whole training and optimization pipeline.

4.2 Datasets

4.2.1 Data Selection

The patent filing information utilized in this project was collected from the lens.org [1] database. It was restricted by the following, chosen to identify applicable AI technology in terms of security-based keywords.

- **Search pattern**:
- **Date range**: From January 4, 2018, to January 4, 2023.
- **Document type**: Granted patent
- **Jurisdiction**: US

4.2.2 Data Collection

4.2.3 Data Preparation

Data Integration and Initial Corpus Construction.

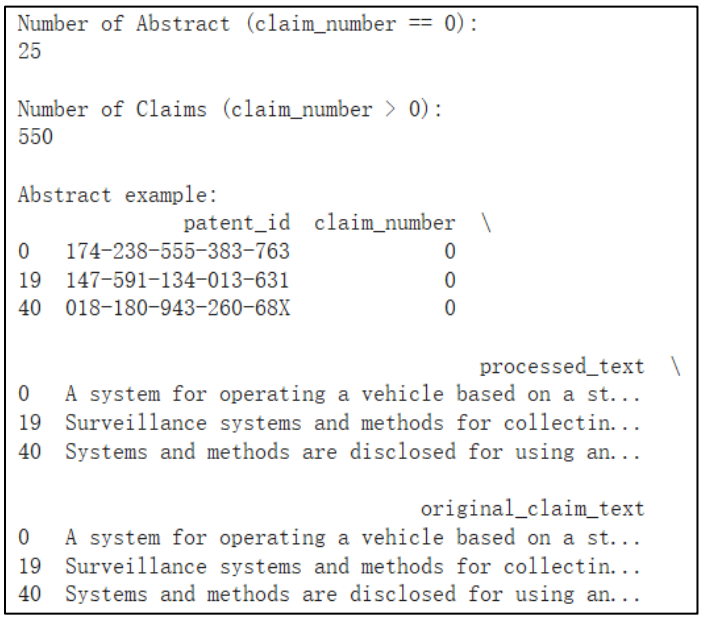

Figure 3**: Structure of the processed patent text.** A processed text database was constructed by programmatic parsing of the master .jsonl file. Output ensures successful extraction of key fields patent id, original claim text, and processed text. Abstracts are assigned a claim number of 0 for organizational reasons to differentiate from sequentially numbered claims. Total counts indicate a starting corpus of 25 abstracts, and 550 claims was successfully created to proceed with the next label matching phase.

Contextual Text Building.

Relaxed Label Matching and Hybrid Dataset Construction.

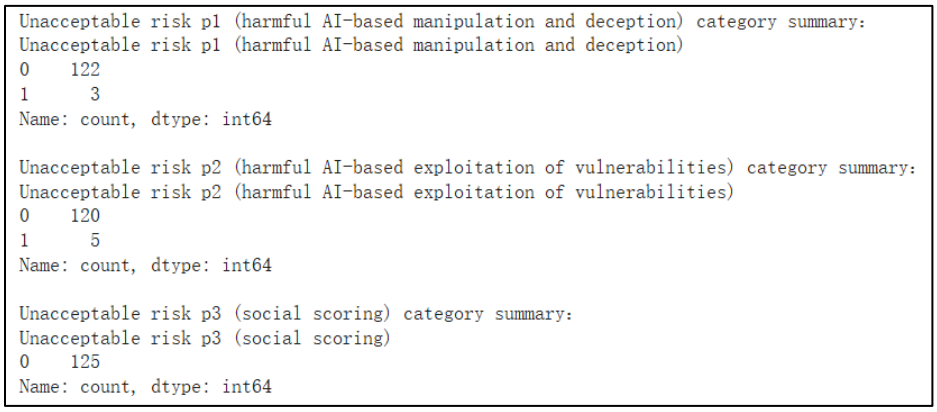

Figure 4**: Distribution of positive and negative samples in original risk categories.** This figure plots a part of summary of the label distribution for a sample of the 33 original risk categories in the master annotation dataset. As indicated in the value counts for categories such as 'Unacceptable risk p1' (3 positive vs. 122 negative), 'p2' (5 positive vs. 120 negative) and 'p3' (no positive), the dataset is exhibited class imbalance. This finding was a key motivator for the risk category aggregation methodology in Section 4.4 and the incorporation of the weighted loss function in Section 4.5.

4.3 Model Architecture

The architecture of the model is a determining factor in its performance across domain-specific

tasks. While extremely powerful general-purpose models like the initial BERT are incredibly strong, training them on massive, generic datasets renders them non-specialized in being used with the particular dictionary and syntax utilized in technical and legal writing. Domain-specific models fill this gap. To provide for this, the nlpaueb/legal-bert-base-uncased model [20] was selected as the base architecture for the supervised fine-tuning approach.

Legal-BERT is not a new architecture but a basic BERT model that has been undergone continuous pre-training over a huge set of legal and administrative texts. The domain adaptation process fine-tunes the internal representations of the model such that the model becomes extremely proficient in understanding legal language, complex sentence construction, and the implied meaning words have in a legal environment. This specific specialization means that it is a much stronger foundation for any NLP legal task than a universal model.

In the multi-label classification task of this project, a classification "head" was merely added on top of the pre-trained Legal-BERT model [20]. This was achieved using the AutoModelForSequenceClassification class of the Hugging Face Transformers [15] library. This class automatically adds on top of the base BERT architecture a randomly initialized, fully connected neural network layer. This head receives the last hidden-state vector of the special [CLS] token, which encodes the overall meaning aggregated in the entire input sequence and projects it to a vector of the same dimensionality as the number of target labels to be utilized such as four aggregated risk groups. This vector of uncalibrated scores, known as logits, is subsequently fed through a Sigmoid activation function to give independent probabilities for each risk category such that a single text may have more than one risk.

4.4 Objective Function

One critical issue that was identified early on in the project was that the 33 risk labels were extremely sparse, with a majority of them having very few positive examples for effective model training. To address this issue, a strategic decision was taken: the 33 labels were aggregated under four broader parent categories: Group: Unacceptable risk, Group: High risk, Group: Transparency risk, and Group: Human right. This was accomplished by introducing a new column for each group and flagging an entry as 1 if any of its underlying sub-categories were present in that patent text. This grouping effectively increased the number of positive instances for all the target classes, converting the problem from an unsolvable sparse classification problem to one which could be solved and providing a more concrete, balanced training input to the model. The resulting positive label distribution for the merged categories is shown in Table 2.

Risk category group	Number of positive samples
Group: Unacceptable risk	25
Group: High risk	90
Group: Transparency risk	44
Group: Human right	35

Table 2**: Distribution of the positive samples across the four aggregated risk groups.** Even after aggregation, the risk labels in the data remained sparse and imbalanced. To handle this, a weighted loss function (BCEWithLogitsLoss with a pos_weight parameter) was employed. It penalizes misclassifications of the minority (positive) class more heavily than those of the majority class, so the model will be more concentrated on them while training and can't simply predict "no risk" for all instances. It is a supervised method that trains the model in our specific data to specialize the model for this task. The model is trained on the patterns in the data and, for any given text, calculates the probability that it will be in each of the four risk categories.

4.5 Training and Optimization

The training and the hyperparameter optimization were done in three sequential phases, managed by the Hugging Face Trainer API [21]. The first phase was an automated hyperparameter search to find an optimal set of them. This was followed by a final, full training run with the found hyperparameters. The last phase was a post-processing step to adjust the prediction thresholds for each risk category.

4.5.1 Hyperparameter Search

To move beyond manual trial and error, a hyperparameter search was automated using the Optuna [22] library, which fits in seamlessly together with the Trainer [21]. The objective of this search was to find the optimal combination of the identified hyperparameters that would produce the optimal performance of the model on the validation set. The search iterated through the pre-determined space for the *learning rate* (log-uniform over 1e-5 to 5e-5) and the per device train *batch size* (categorical choice between 4 and 8). 10 runs were executed, each training a new model for fewer epochs to be cost-effective. The optimisation target was set as maximising evaluation F1 weighted metric. The search stopped with the identification of the best-performing combination (*learning rate*: 3.59e-05, *batch size*: 4) as the final, full-scale training was then conducted using the combination.

4.5.2 Final Training

With the optimal hyperparameters discovered via the search, the final model was then trained for a complete 8 epochs. The model performance was tracked on the validation set at the completion of each epoch using the weighted average F1 score. The weighted average F1 score is a single robust, consensus metric which averages precision against recall across all four risk categories equally, with preference to categories containing high support. The model checkpoint which achieved the best performance according to this process of evaluation was selected as the final model for any future analysis. This process is a powerful form of early stopping to prevent overfitting and ensures that the resulting model represents the point of peak performance during training.

4.5.3 Threshold Optimization

For each of the four risk classes, its output from the optimized model is a logit that is converted to a 0-to-1 probability using a sigmoid function. A default threshold value of 0.5 is often not optimal for imbalanced datasets. Therefore, a critical post-processing operation was performed to find the optimal prediction threshold for each of the four risk classes individually. For each category, the precision recall curve function in Scikit-learn [17] was used to compute the precision and recall values for all possible thresholds on the prediction probabilities of the validation set. Then, the F1 score was computed for all thresholds from these curves. The threshold that resulted in the maximum F1 score for a particular category was then found and stored. This process resulted in a dictionary of four optimal thresholds, which were used in generating the final classification report and follow-on inference tasks.

To demonstrate the practical application of the final model, the whole prediction pipeline was encapsulated within an interactive graphical user interface. The complete design and implementation of this interface are presented in Appendix A.1.

5 Results, Analysis and Evaluation

5.1 Overview

In this section, the findings and general analysis of the experiments conducted within the project are highlighted. The primary purpose is to validate the performance of various methodologies employed to address the problem of automated AI risk categorization in patent documents. The chapter traces the methodological development of the project. It begins with an accurate outline of the fundamental experimental setup, such as measurement statistics and data division, which are to be equally used in all the subsequent experiments. The focus is redirected towards a comparative study, focusing on the early attempt in 33 sparse risk classes classification. These failures show the limitations of direct classification and justify the necessity for developing a data aggregation strategy. Then, the chapter presents an extensive comparison of the three major baseline models- semantic similarity, TF-IDF with logistic regression, and the manually fine-tuned supervised fine-tuning approach on the four aggregated risk categories. Finally, the chapter finishes with an extensive evaluation of the final, automatically fine-tuned Legal-BERT model.

5.2 Base Experimental Setup

To have consistency and comparability across all experiments, a default experimental setup was created.

Dataset Split:

Evaluation Metrics:

Implementation Details: The experimental pipeline was set up and run in Google Colaboratory cloud [13], using a T4 GPU to train and infer faster. Programmatic implementation was done using Python 3 and was based on a set of libraries: PyTorch (2.1.0) [14] to perform fundamental deep-learning tasks; Transformers (4.40) [15], [16] to perform language-model-related operations; Pandas to delete and manipulate data; and Scikit-learn [17] to divide data and calculate evaluation metrics.

5.3 Ablation Study / Comparative Analysis

This part includes a comparative review of various methodologies investigated in this project. The objective is to outline the progression of the method, from early experiments that revealed the major problems of the task, and towards the final successful model. This review is intended to validate the most important methodological choices made, especially the need for risk category aggregation.

5.3.1 Experiments on 33 Sparse Risk Categories

The first series of experiments were performed on the dataset using its original 33 detailed risk labels. The intention was to determine if current NLP models can cope with this very sparse, multi-label classification problem. Three models were compared.

Model (33 risk labels)	Weighted average F1 score
Semantic similarity (Baseline)	0.03
Semantic similarity (Enhanced)	0.13

Supervised Fine-tuning (Legal-BERT)	0.33

Table 3**: Comparison of model performances on 33 original sparse risk categories.** The table represent the weighted average F1 score of the three models: Semantic similarity (Baseline), Semantic similarity (Enhanced), and Supervised Fine-tuning (Legal-BERT).

The results, as presented in Table 3, were markedly low in all approaches. The supervised fine-tuning of Legal-BERT, despite being the most sophisticated approach, registered a weighted average F1 score of 0.33. Detailed analysis of the classification report revealed that the model had adopted a simplistic strategy of predicting the negative class (0) for most samples in all 33 classes. This was because the support for nearly all the classes in the validation set was zero or one, as seen in Figure 5. Since data was highly sparse, the model was unable to learn any helpful patterns in the positive classes and defaulted to the majority class in order to decrease its loss.

The semantic similarity experiments performed were slightly above zero but remained unsatisfactory. While they did find some topically related documents, the inability to identify subtle risk meant their F1 scores were extremely low (below 0.15). These experiments did have one significant lesson: the extreme sparsity of the original 33 labels rendered the classification task difficult. This insight led to the strategic pivot towards label aggregation, which is discussed in the next section.

claim_text	Unacceptable risk p1 (harmful AI-based manipulation and deception)	Unacceptable risk p2 (harmful AI-based exploitation of vulnerabilities)	Unacceptable risk p3 (social scoring)	Unacceptable risk p4 (individual criminal offence risk assessment or prediction)	Unacceptable risk p5 (untargeted scraping of the internet or CCTV material to create or expand facial recognition databases)	Unacceptable risk p6 (emotion recognition in workplaces and education institutions)	...	Human right to respect for private and family life	Human right to freedom of thought, conscience and religion	Human right to freedom of expression
for operating e based on a st...	0	0	0	0	0	0	...	0	0	0
A system for ing a vehicle based on a...	0	0	0	0	0	0	...	0	0	0
tem of claim , wherein the sensory...	0	0	0	0	0	0	...	0	0	0

Figure 5**: Distribution of Positive Samples Over a Subset of the 33 Initial Risk Categories.** The below image illustrates the severe class imbalance in the raw dataset. It is a visualization of counts of a few of the 33 fine-grained risk classes, which illustrates the phenomenal sparsity of positive samples (label 1) over negative samples (label 0).

5.3.2 Experiments on 4 Risk Groups

After establishing that the initial label structure was not practical, the 33 sparse labels were grouped into four meaningful risk groups as explained in Section 4.4. This provided a new dataset with a far more achievable distribution of positive samples per class. On this dataset, four different modeling methods were applied and contrasted.

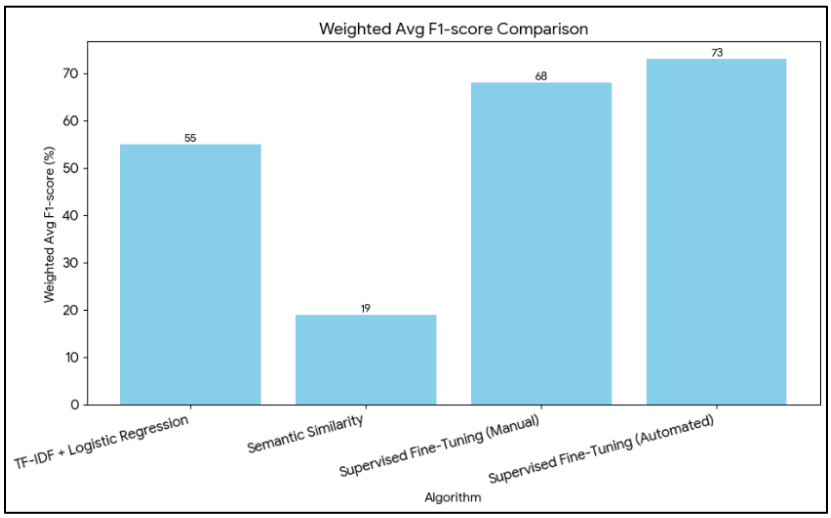

Figure 6**: Overall performance comparison of models on the four aggregated risk groups.** The results illustrate a clear performance, with both supervised fine-tuning approaches well above the baselines of TF-IDF and semantic similarity. Notably, automated hyperparameter search yielded the best-performing model, with a final F1 score of 73%.

The Figure 6 presents an overview of the four models' overall end performance on the validation set. The results clearly show that supervised fine-tuning approaches far outperform the baseline approaches. The end model that employed automated hyperparameter search achieved the best overall score. A detailed comparison of the results above is presented in Table 4**.**

Model	Precision	Recall	F1 score (weighted average)	F1 score (macro average)
TF-IDF + logistic regression	0.57	0.60	0.55	0.50
Semantic similarity	0.11	0.63	0.19	0.15
Supervised fine-tuning (manual)	0.62	0.83	0.68	0.59
Supervised fine-tuning (automated)	0.70	0.79	0.73	0.64

Table 4**: Model performance in the four risk aggregation groups compared.** The table provides precise assessment of the recent performance metrics of the four major forms of modeling on the validation set.

Various interesting results can be revealed after analyzing Table 4. The TF-IDF + logistic regression model with class_weight='balanced applied gives a reasonable baseline (weighted average F1 score of 0.55) which shows that even a model based on word-frequency is capable of discovering some risk patterns without needing such deep semantic understanding. In contrast, the Semantic similarity model yields a rather poor F1 score (F1 score: 0.19), indicating that its non-specialized linguistic knowledge is unusable in the specific task.

The most convincing results are attained by the Supervised Fine-tuning strategies. The last model having an automated hyperparameter search generating a weighted average F1 score of 0.73 (see Figure 6 and Figure 7), is also a significant improvement over all other methods and confirms the philosophy of design, which was followed in fine-tuning a domain-knowledge model on a preprocessed dataset strategically. The following sections will give a critical analysis of this last model.

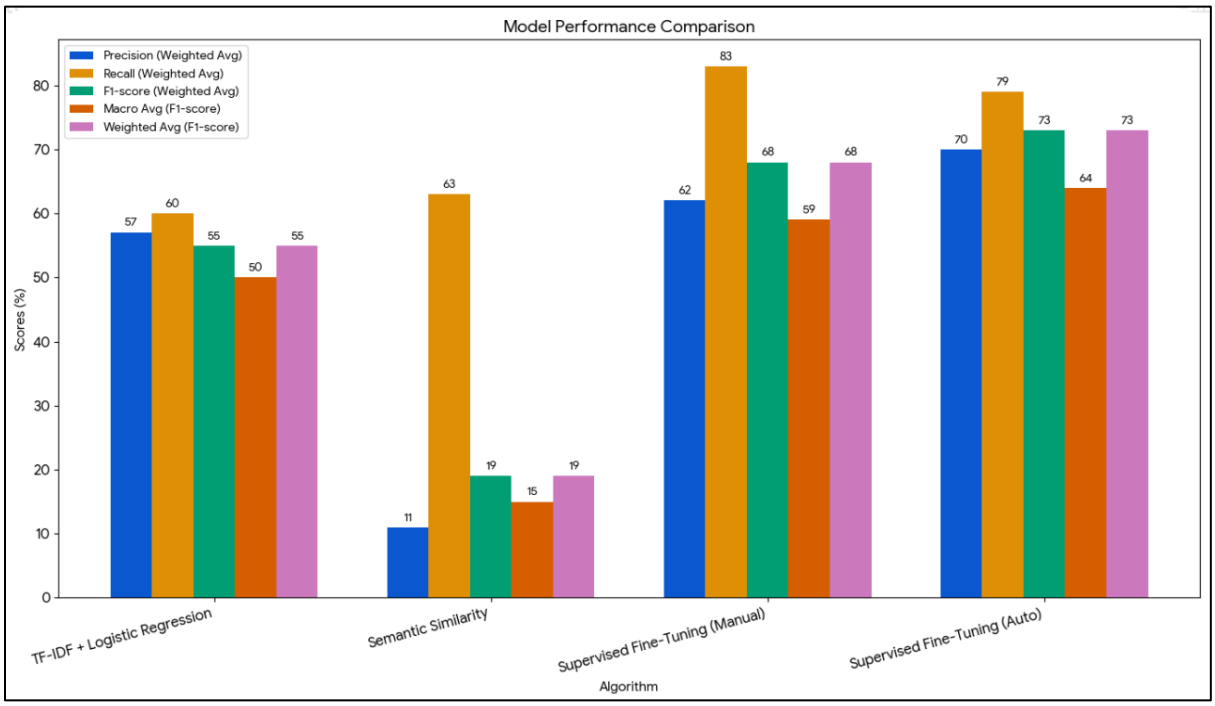

Figure 7**: Comparison of per-category F1 score for all models evaluated.** This plot provides a close-up comparison of the final F1 scores for all four major modeling approaches across each of the four macro-categories of combined risk. The results readily illustrate the performance ordering between the approaches. As with all categories, both supervised fine-tuned approaches have steadily high performance compared to the variable but low-performing TF-IDF and semantic similarity baselines. The third and final method of automated fine-tuning exhibits the highest efficacy.

5.4 Results and Evaluation of Final Model

This subsection presents the final performance results of the Legal-BERT model. The model being tested here is the one trained using the best hyperparameters (*learning rate*: 3.59e-05, *batch size*: 4) that were discovered by the automatic search in Section 4.5.1. The evaluation is on the 131-sample validation set, to which the model has never been exposed during training. The evaluation is given in two sections: a quantitative analysis of the statistical performance measures, and a qualitative analysis with the use of Explainable AI for understanding the decision-making process of the model.

5.4.1 Quantitative Results and Analysis

Quantitative assessment developed out of a methodical exploration of the training dynamics and reached a final analysis of overall accuracy of classification. It follows that the training log (Table 5) will document the performance of the model at the point in time after every one of the 8 training epochs.

Epoch	Training Loss	Validation Loss	F1 Weighted
1	1.427000	1.165676	0.642767
2	1.001600	1.194590	0.646214
3	0.906100	1.275145	0.666587
4	0.635300	1.220247	0.647925
5	0.391000	1.460789	0.650178

6	0.563000	1.437216	0.673420
7	0.251900	1.251168	0.674817
8	0.476300	1.318607	0.677564

Table 5**: Training and validation performance log over 8 epochs.** This table provides a summary of the model's performance on the training set and validation set after each training epoch.

A graph of the training log provides valuable insight into learning dynamics for the model. Whereas training loss shows an overall downward, albeit unstable, trend, the Validation loss behaves otherwise. It is minimized during the first epoch and then trends upward, diverging from the training loss. This divergence between training and validation loss is a classic indicator of overfitting, i.e., while the model was optimizing its performance on the training set, its performance in the aspect of generalizing to new unseen validation data did not improve from the initial epoch. Although there was such a trend in the loss, the primary measurement criterion, weighted average F1 score, was still showing marginal improvements and achieved its best value of 0.678 at the final epoch. The training configuration was to select the best F1-scoring model checkpoint at the end, thereby ensuring the chosen model for testing is this peak, and not the state near the end of a training session that could be overfitted.

The final evaluation of this best-performing model was completed in a two-stage process to demonstrate the impact of threshold optimization. First, a baseline classification report was generated using a default, uniform prediction threshold of 0.5 for all risk classes. Then, the optimal, per-class thresholds were calculated as described in Section 4.5.3. These empirically optimal thresholds, which maximize each per-group F1 score, are presented in Table 6.

Risk category group	Number of positive samples	F1 score	Optimal threshold
Group: Unacceptable risk	25	0.50	0.85
Group: High risk	90	0.88	0.92
Group: Transparency risk	44	0.56	0.37
Group: Human right	35	0.64	0.75

Table 6**: Category optimal thresholds and corresponding potential F1 scores.** This table shows the empirically established best prediction threshold by risk category, calculated on the validation set to maximize the F1 score.

Table 6 also highlights ideal thresholds with greatly high values for some classes, Group: High risk (0.926) and Group: Unacceptable risk (0.851). The higher thresholds indicate that the model was designed to generate highly certain and discriminative predictions for these specific classes. The model's prediction probabilities on the validation set for the risks were highly divided: on positive samples it correctly classified, its predicted probabilities were always highly in the extremes, and on negative samples, the probabilities were highly in the extremes with very few in the middle. Due to this, the algorithm for maximizing the F1 score picked a high threshold as optimal at classifying the two classes cleanly without giving out recall.

The final performance comparison is presented in Figure 8 where the baseline report (with cutoff value 0.5) is being compared with the end of optimization report.

```
============================================================
--- Report 1: use【(general) 0.5 thresholds】eval report ---
============================================================
                          precision    recall  f1-score   support

Group: Unacceptable risk       0.33      0.60      0.43         5
        Group: High risk       0.77      0.88      0.82        26
Group: Transparency risk       0.44      0.67      0.53        12
      Group: Human right       0.44      0.89      0.59         9

               micro avg       0.56      0.81      0.66        52
               macro avg       0.50      0.76      0.59        52
            weighted avg       0.59      0.81      0.68        52
             samples avg       0.14      0.19      0.15        52


============================================================
--- Report 2: use【best threshold(each group)】eval report (final result) ---
============================================================
                          precision    recall  f1-score   support

Group: Unacceptable risk       0.43      0.60      0.50         5
        Group: High risk       0.92      0.85      0.88        26
Group: Transparency risk       0.45      0.75      0.56        12
      Group: Human right       0.54      0.78      0.64         9

               micro avg       0.64      0.79      0.71        52
               macro avg       0.58      0.74      0.64        52
            weighted avg       0.70      0.79      0.73        52
             samples avg       0.13      0.18      0.15        52
```

Figure 8**: Comparison of overall performance with baseline and optimized thresholds.** The table contrasts the classification results with a hard 0.5 threshold with those with the per-category optimal thresholds from Table 6.

The utilization of threshold classification levels yields an overall weighted average F1 score of 0.73. The model performs best in the Group: High risk category, yielding an F1 score of 0.88, along with precision of 0.92 and recall of 0.85. These outcomes indicate a well-balanced and high capability of accurately classifying patents in the most critical risk group, also being the most frequently classified category found in the data set.

The model performs well not only within one category, but also in all categories of risk groups. It achieved 0.64 in Human right and 0.56 in Transparency risks, which means that it can perform well under a wide range of semantic patterns. Even with extremely limited examples, like the Unacceptable risk group, it still performed with a score of 0.50. In general, the results show that the model can be consistent with a range of language environments and limited data situations.

In order to obtain a more detailed evaluation of the performance of the classifier as compared to the four risk aggregated groups, confusion matrices per risk category are created, as shown in Figure 5.4. These matrices graphically decompose the true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) and thus clarify the exact character of the mistakes the model made.

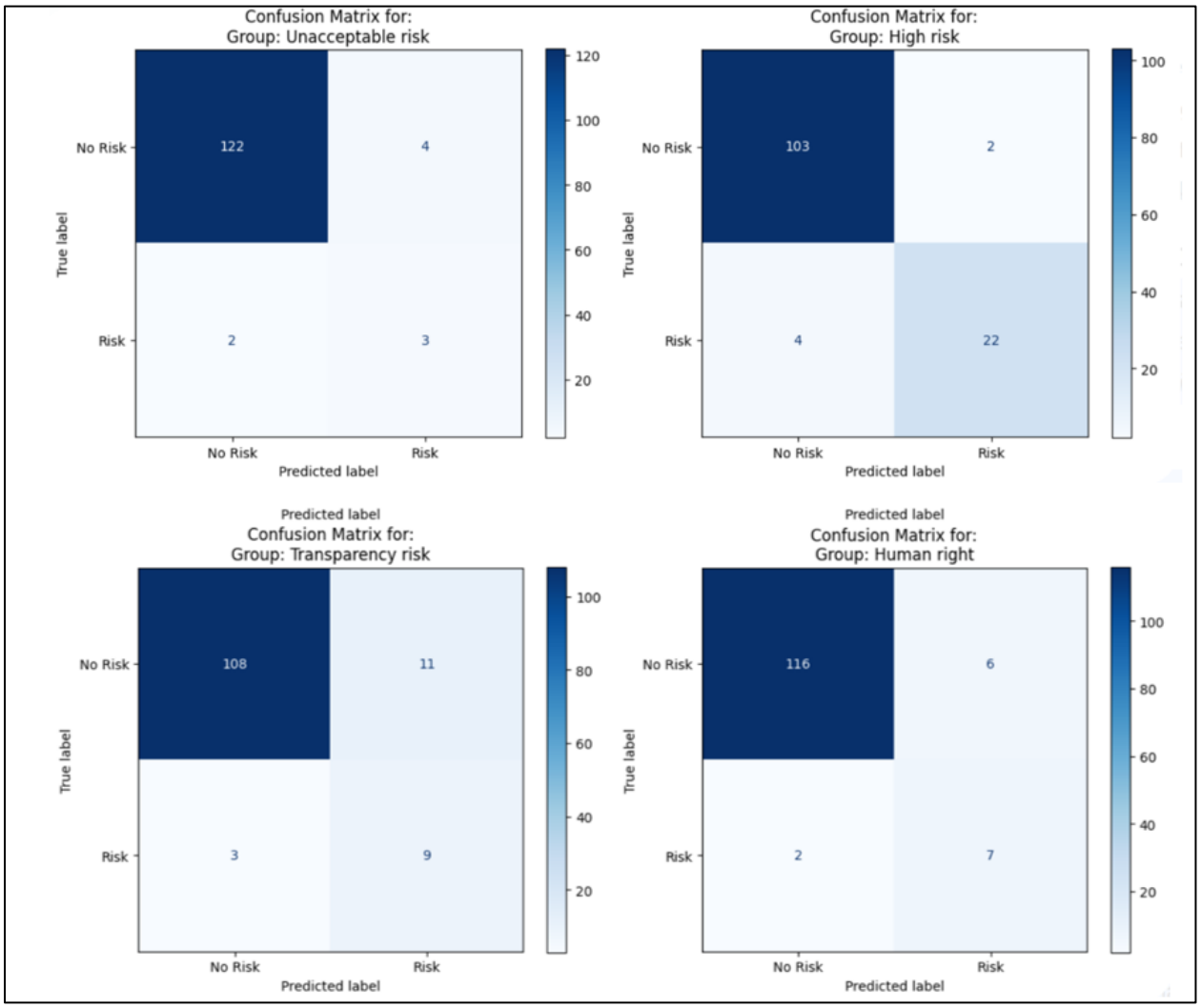

Figure 9**: Per-category confusion matrices for the final optimized model.**

The Figure 9 displays the confusion matrix for each of the four risk categories on the validation set. Each matrix shows the number of correct ("No Risk" to "No Risk" or "Risk" to "Risk") and incorrect ("No Risk" to "Risk" or "Risk" to "No Risk") predictions.

The results displayed in the previous classification report (Figure 8) are further clarified by the confusion matrices. The matrix that relates to the Group: High risk category reveals an effective classifier: 22 TP against 4 FN and with a low FP count of 2. The low number of FPs is reflected in high precision. The classifier is (0.92) for this type and that when the classifier gives an outcome of high risk, the likelihood of it being correct is very high.

In the rest of the categories, the matrices indicate trade-offs in performance. A matrix shows a balanced performance. In the Group: Human right matrix which also correctly predicted 7 out of 9 positive cases and incorrectly classified 2 as negative (FN=2). In contrast, the Group: Transparency risk matrix has a rate of FP greater than the rate of FN (FP=11 > FN=3), which indicates the model is potential to over-classifying for this risk class.

The most important and least frequently occurring category, Group: Unacceptable risk, the matrix shows that 3 of the 5 total number of risk occurrences were accurately correctly identified (TP=3). The two false negatives (FN=2) represent a significant error rate, thus necessitating the optimization of the error rate. However, the outcomes show the ability of the model to distinguish between different types of risks, and this is the model's foundational ability to identify this risk type even with limited data. Taken together, the examined matrices provide a

detailed description of the model behavior highlighting both its strong performance on well-represented classes and its specific error patterns on classes with limited data.

5.4.2 Qualitative Analysis

token	Attribution score	token	Attribution score
change	0.031	a	0.011
the	0.017	responsive	0.010
of	0.016	indicates	0.006
to	0.012	oarameters	0.005
indication	0.011	operating	0.005

Table 7**: Top 10 contributing tokens for a 'High risk' prediction with gradient attribution.** The table above shows the top 10 tokens in the input with ranking based on their attribution score computed through the Layer Integrated Gradients algorithm [22]; the higher the score, the more positive the contribution to the final probability of Group: High risk.

5.5 Discussion

This section examines the implications of the experimental results more broadly, in terms of the model's limitations as determined from an analysis of its failure cases, and its scalability within a real-world application context.

5.5.1 Failure Cases

5.5.2 Scalability

6 Legal, Social, Ethical and Professional Issues

6.1 Overview

The current chapter forms a methodical analysis of the legal, social, ethical and professional aspects of building and implementing an automatic AI risk-classification architecture of patent literature. It locates the wider implications of such an undertaking, especially, potential future implications on society, challenges to the reliability of software, intellectual property regimes, and inevitable issues of data bias within a scheme of professional responsibility.

6.2 *British Computer Society (BCS)* Code of Conduct & Code of Good Practice

The development of the project and its execution were influenced by the ethical codes of the *British Computer Society (BCS)* [23], supplemented by suitably relevant *Institute of Engineering and Technology (IET)* [24] considerations. Significant aspects such as the project's orientation towards the public interest and its open documentation demonstrate conformity with BCS values [23] such as "Public Interest," "Professional Competence," and "Duty to Relevant Authority." These values led the project in its efforts to identify risks in patent literature and monitor regulatory reforms for responsible AI.

Under these codes, the system has been constructed to enable expert judgment subject to the limitation of automatic classification. Rather than superseding human judgment, it encourages knowledgeable management. The IET [24] values, particularly professional responsibility and public good, also determined the system's intended use and user interface to assure ongoing ethical use in technical environments.

6.3 Issues of the Work

7 Conclusion and Future Work

7.1 Conclusion

This study suggested a systematic process for the identification of AI-related risks within patent literature using NLP techniques applied for domain-specific terminology. Data regrouping and algorithmic weighting were employed to mitigate the effect of class imbalance. The initial 33 sparse risk tags were categorized into four general groups in order to simplify model learning. Despite limitations related to class imbalance and data sparsity, this work confirms the feasibility of risk classification from patent texts through domain-adapted language models. The training strategy, with label grouping, algorithm weighting, and model-specific hyperparameter tuning, demonstrated successful performance across all the risk levels. Threshold tuning also helped make the framework more flexible, particularly for high-risk detection. The conclusions can be used as the foundation for future research aimed at enhancing regulatory oversight, cross-jurisdictional risk labeling, and bulk compliance analysis.

7.2 Limitation

Though this project achieved its overall objectives, this study has several limitations that frame the context of its findings.

The primary limitation is the shortage of manually annotated risk-positive samples is the most significant limitation. Despite the fact that the final dataset comprises 654 text instances taken from 487 patents, those instances labeled with a specific risk are remains low. The direct consequence of this limited positive signal is best observed in the Group: Unacceptable risk category, the one with the fewest positive examples and thus the lowest F1 score. The model's generalization capability is consequently constrained by both diversity and quantity of risk-positive examples the model was exposed to during training.

Secondly, there was an attempt to add a generative explanation module but failed as intended due to a number of reasons. Initial attempts, the intention was to introduce this feature utilizing *LLMs (Large Language Model)* for deployment in constraint-limited environments but was hindered by software compatibility concerns. Other attempts to overcome such constraints were marred by project timeliness and resource availability. This illustration points to the gap between theoretical availability of LLMs and empirical challenges in using reliably integrating them into interconnected risk analysis systems.

Finally, model generalization fine-tuning is an inherent limitation. The model performed very well on the validation set, which was drawn from the same distribution as the training set. However, its performance on patent texts that employ very novel linguistic structures or describe technology far outside the scope of what was used to train it has yet to be demonstrated. As discussed in the failure case analysis (Section 5.5.1), the model's knowledge is eventually limited by what it has been trained on and can stall on out-of-distribution instances.

7.3 Future Work

From the results and constraints of the project, several potential directions for future research can be identified:

Expansion of the Annotated Dataset: The most straightforward path to improved performance, particularly for the low-data classes, is to expansion of the manually annotated dataset. Future

work would include more emphasis on creating a larger and more varied set of manually annotated examples. That would provide a more robust foundation for all post modeling methodologies and is expected to improve the model's generalization performance across all risk classes.

Implementation of a Hybrid Explanation System: The attempts at building a local generative explanation module clarified the key technical challenges. A successful extension would be to formally implement the two-stage system conceived in this paper with a solid, API-based Generative AI service. The fine-tuned Legal-BERT model would make a fast and efficient first-pass filter, and for any risk it identifies, a good API can be called to programmatically generate the detailed, human-readable analytic report, providing the essential "why" for the prediction.

Enhancing Model Transparency using Visualization: While the Captum [22] analysis was revealing, this can be developed into a more integrated feature. One potential avenue of future work is designing an interactive dashboard wherein users are not only capable of seeing the risk prediction for any input text but also visualize the corresponding gradient attributions. This would allow professionals to immediately see what words and phrases contributed most to the model's decision, therefore towards a more transparent and trustworthy review process.

References

[1] The Lens, “The Lens,” [Online]. Available: https://www.lens.org/. Accessed: Aug. 6, 2025.

[2] M. Petrolini, S. Cagnoni, and M. Mordonini, “Automatic detection of sensitive data using transformer-based classifiers,” *Future Internet, vol. 14, no. 8*, 2022.

[3] R. Vasanthi, H. Raj DK, H. Kumar, and D. Kumar, “Automated terms and condition analyser using natural language processing,” in *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, 2024.

[4] B. Breve, G. Cimino, and V. Deufemia, “Identifying security and privacy violation rules in trigger-action IoT platforms with NLP models,” *IEEE Internet of Things Journal, vol. 10, no. 6*, 2022.

[5] T. Perera and T. Perera, “Barrister-processing and summarization of terms & conditions/privacy policies,” in *2021 6th International Conference for Convergence in Technology (I2CT)*, 2021.

[6] P. Silva, C. Gonçalves, C. Godinho, N. Antunes, and M. Curado, “Using NLP and machine learning to detect data privacy violations,” in *IEEE INFOCOM Workshops*, 2020.

[7] A. García-Pablos, N. Perez, and M. Cuadros, “Sensitive data detection and classification in Spanish clinical text: Experiments with BERT,” *arXiv preprint arXiv:2003.03106*, 2020.

[8] A. J.-P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, “Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports,” *Automation in Construction, vol. 62*, 2016.

[9] B. Rosadini, A. Ferrari, G. Gori, A. Fantechi, S. Gnesi, I. Trotta, and S. Bacherini, “Using NLP to detect requirements defects: An industrial experience in the railway domain,” in *International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ)*, 2017.

[10] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” *arXiv preprint arXiv:1908.10084*, 2019.

[11] P. G. Bizzaro, E. Della Valentina, M. Napolitano, N. Mana, and M. Zancanaro, “Annotation and classification of relevant clauses in terms-and-conditions contracts,” *arXiv preprint arXiv:2402.14457*, 2024.

[12] D. Braun and F. Matthes, “NLP for consumer protection: Battling illegal clauses in German terms and conditions in online shopping,” in *Proc. 1st Workshop on NLP for Positive Impact*, 2021.

[13] Google, “Google Colaboratory,” [Online]. Available: https://colab.research.google.com/. Accessed: Aug. 6, 2025.

[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, “PyTorch: An imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems, 2019.

[15] Hugging Face, “Transformers,” [Online]. Available: https://huggingface.co/docs/transformers/en/index. Accessed: Aug. 6, 2025.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf,

M. Funtowicz, and J. Davison, “Transformers: State-of-the-art natural language processing,” in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, “Scikit-learn: Machine learning in *Python,” J. Mach. Learn. Res.*, 2011.

[18] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The muppets straight out of law school,” arXiv preprint arXiv:2010.02559, 2020.

[19] Hugging Face, “sentence-transformers/all-mpnet-base-v2,” [Online]. Available: https://huggingface.co/sentence-transformers/all-mpnet-base-v2. Accessed: Aug. 6, 2025.

[20] Hugging Face, “nlpaueb/legal-bert-base-uncased,” [Online]. Available: https://huggingface.co/nlpaueb/legal-bert-base-uncased. Accessed: Aug. 6, 2025.

[21] Hugging Face, “Transformers: Trainer class,” [Online]. Available: https://huggingface.co/docs/transformers/en/main_classes/trainer. Accessed: Aug. 6, 2025.

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-Generation Hyperparameter Optimization Framework,” GitHub Repository. [Online]. Available: https://github.com/optuna/optuna. Accessed: Aug. 6, 2025.

[23] British Computer Society, “BCS Code of Conduct,” [Online]. Available: https://www.bcs.org/membership/become-a-member/bcs-code-of-conduct/. Accessed: Aug. 6, 2025.

[24] Institute of Engineering and Technology, “Rules of Conduct,” [Online]. Available: https://www.theiet.org/about/governance/rules-of-conduct/. Accessed: Aug. 6, 2025.

[25] R. Frissen, K. J. Adebayo, and R. Nanda, “A machine learning approach to recognize bias and discrimination in *job advertisements,” AI & Society*, 2023.

[26] A. Abid, A. Abdalla, A. Abid, D. Khan, A. Alfozan, and J. Zou, “Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild,” *arXiv preprint arXiv:1906.02569*, 2019.

Appendices

A.1 Interactive Risk Analysis Tool

The performance of a final classification model was implemented practically by creating an interactive user interface with the help of the Gradio Python [26] library.

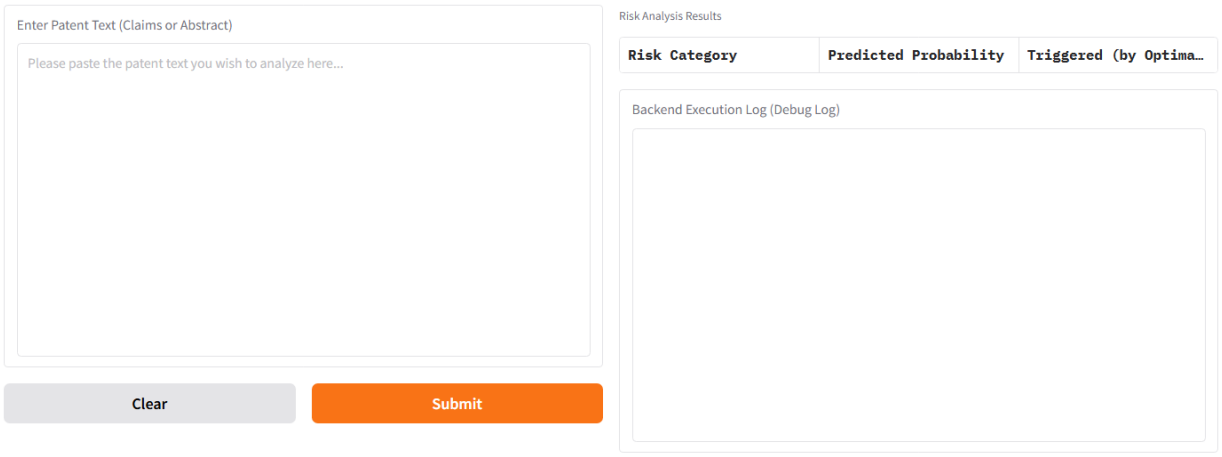

Figure 10**: AI Risk Analysis Tool interactive user interface.**

The figure shows a left panel with the input text area and the right panel with output components, including a classification report and a backend execution log.

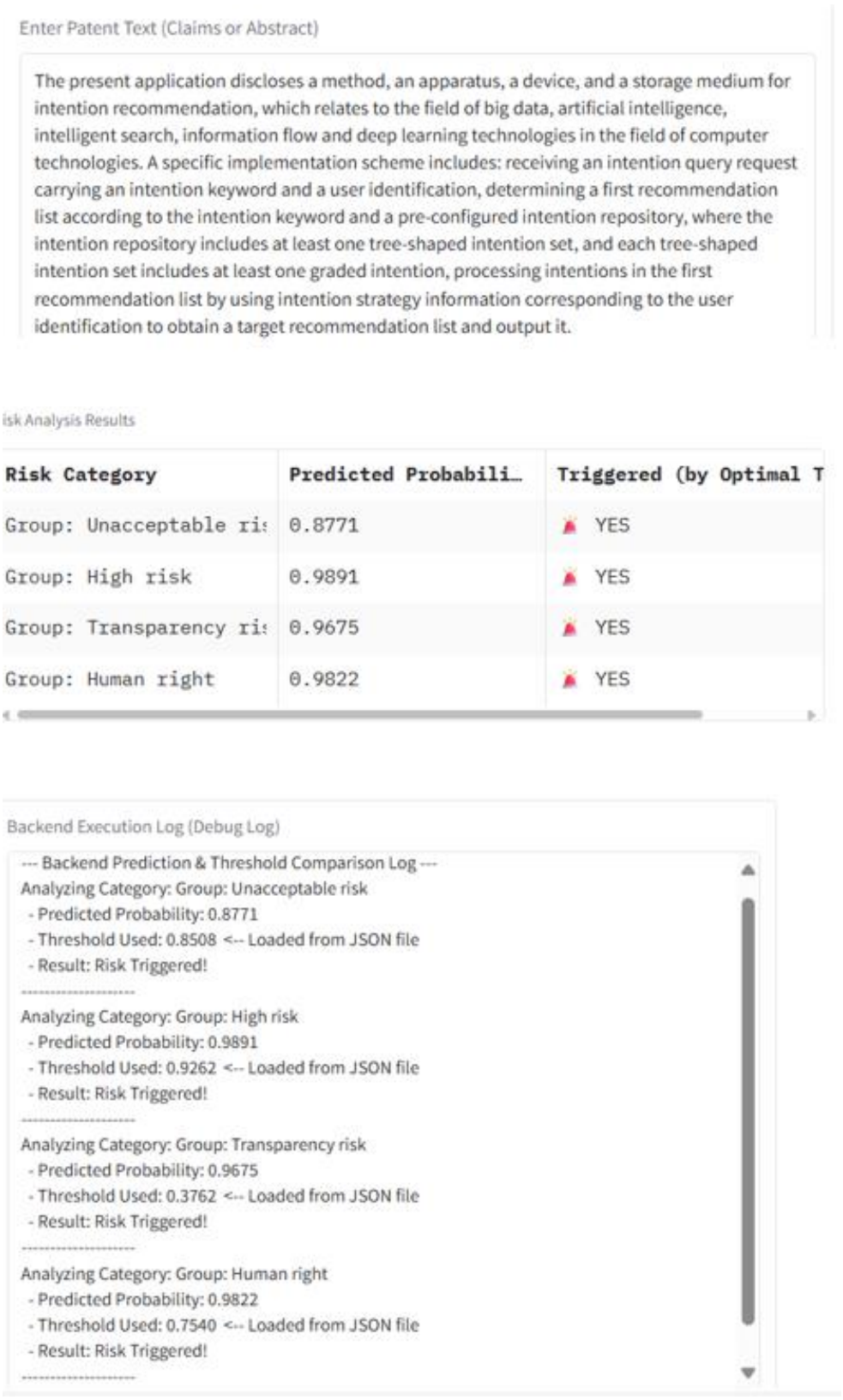

Figure 11: **An example of a Gradio result of analysis.**

The output is shown in the figure in the form of output panels filled with a classification decision and subsequent backend log as a result of the processing of a sample text.