Legal Document Classification using BERT and Traditional ML Techniques on LexGLUE

Authors

Sara Samir Nasr – 22101068

shams abdelhalim 22101040

Abstract

Legal document classification is a crucial task in legal NLP, aiming to automatically assign category labels to complex legal texts. In this paper, we explore the effectiveness of transformer-based models (BERT) and traditional machine learning techniques (TF-IDF + Logistic Regression) on the LexGLUE benchmark using the LEDGAR dataset. Experimental results show that BERT significantly outperforms classical methods in terms of classification performance.

1 Introduction

The increasing volume of legal documents requires efficient classification tools for legal professionals. Manual classification is costly, time-consuming, and error-prone. We focus on automating this task using the LEDGAR dataset from the LexGLUE benchmark. We compare a transformer-based model (BERT) with a traditional approach (TF-IDF + Logistic Regression) to assess performance trade-offs and evaluate their applicability in legal contexts.

2 Related Work

Previous work includes keyword-based methods and topic modeling, which lack context awareness. Recently, deep learning models, especially transformers like BERT and LegalBERT, have been applied with significant success. The LexGLUE benchmark introduced by Chalkidis et al. provides standardized datasets and tasks for legal NLP evaluation, making it a key resource in this area.

3 Methodology:

Dataset:
We used the LEDGAR subset of the LexGLUE benchmark, which consists of clauses from legal contracts categorized into multiple classes. We used 1000 samples for training and 200 for testing.
Models:
TF-IDF + Logistic Regression:
Built using a scikit-learn pipeline with a TF-IDF vectorizer (max features = 3000) and Logistic Regression classifier (max iterations = 1000).
BERT (bert-base-uncased):
Fine-tuned using HuggingFace Transformers. Texts were tokenized and padded using AutoTokenizer and trained using the Trainer API.
GPT-2 Prompting (Baseline):
As a baseline, we used a text-generation approach with GPT-2, where prompts asked the model to generate the label for a given document.

4 Keywords

n this section, we describe the models and methods we implemented in our sentiment classification pipeline on Twitter data.

5 ModelBasics

In this project, we employed two distinct approaches for legal document classification:
BERT (Bidirectional Encoder Representations from Transformers):
BERT is a transformer-based pre-trained model that captures bidirectional context, making it highly effective for complex legal language. We used the bert-base-uncased model from HuggingFace Transformers and fine-tuned it on the LEDGAR dataset. Tokenization was handled using AutoTokenizer with padding and truncation to standardize input lengths.
TF-IDF + Logistic Regression:
This traditional machine learning pipeline combines:
TF-IDF (Term Frequency–Inverse Document Frequency): Converts text into numerical features by evaluating word importance.
Logistic Regression: A simple yet powerful linear classifier.
The model was implemented using Scikit-learn’s Pipeline, trained on 1000 samples, and evaluated using classification metrics.

6 Experimental Results

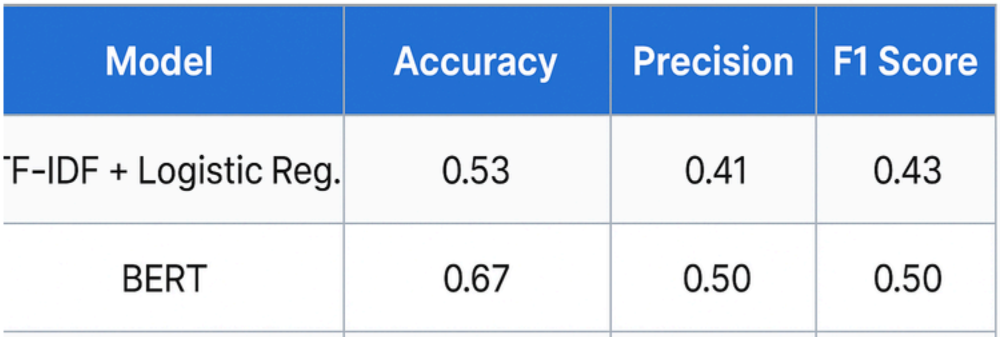
To evaluate the performance of the models on legal document classification, we conducted experiments using a subset of the LexGLUE - LEDGAR dataset. We selected 1000 samples for training and 200 samples for testing. The models were assessed using standard evaluation metrics: Accuracy, Precision, Recall, and F1 Score.

The results are summarized as follows:

The TF-IDF + Logistic Regression model achieved an accuracy of 53%, with relatively lower precision and F1 score. While simple and fast, this model struggles with the complex language and structure of legal texts.

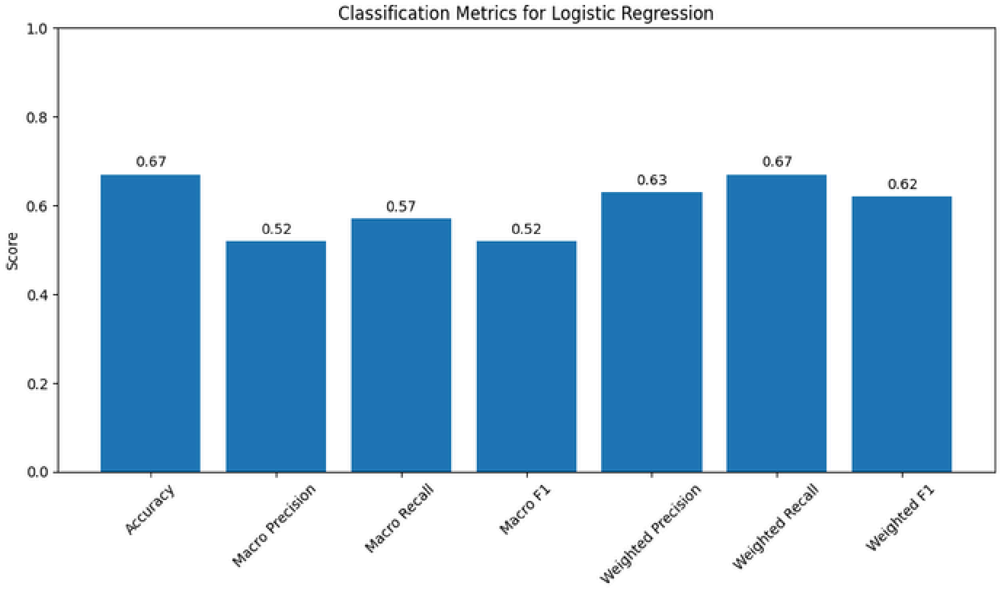
The BERT (bert-base-uncased) model significantly outperformed the traditional baseline, achieving 67% accuracy and stronger overall scores across all metrics. Its contextualized embeddings allowed better understanding of legal clause semantics.

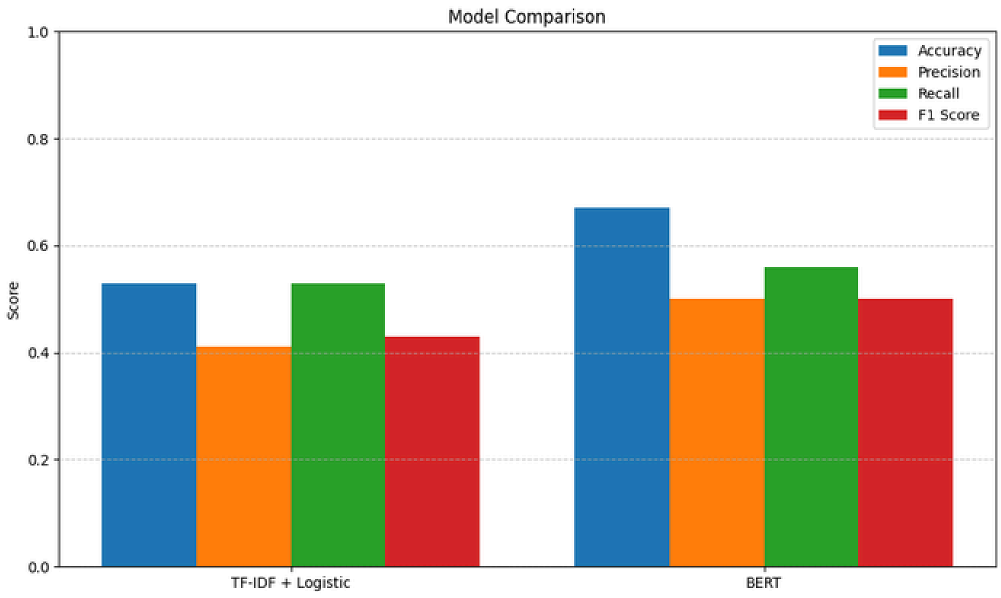
7 Experiments and Results



Model	Accuracy	Precision	F1 Score
TF-IDF + Logistic Reg.	0.53	0.41	0.43
BERT	0.67	0.50	0.50

Visualization:





8 Discussion:

BERT significantly outperformed the traditional machine learning model. Its ability to capture deep contextual information makes it more suitable for understanding legal texts. However, the TF-IDF + Logistic model is faster to train and requires fewer resources, making it a viable choice for baseline comparisons. A limitation of our work is the small dataset size (1000 samples), which may affect generalization.

9 Conclusion:

We presented a comparative study of BERT and traditional machine learning methods for legal document classification using the LexGLUE LEDGAR dataset. Our findings confirm the advantage of transformer-based models in legal NLP tasks, providing strong performance improvements over classical methods.

10 References:

1. Chalkidis, Ilias, et al. “LexGLUE: A Benchmark Dataset for Legal Language Understanding in English.” arXiv preprint arXiv:2110.00976, 2022.
2. Devlin, Jacob, et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL, 2019.