Configuration Manual for Enhancing Risk Assessment in Legal Documents Through Advanced Machine Learning

MSc Research Project
MSc Data Analytics

Sai Teja Pusarla
X23193638

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sai Teja Pusarla		
Student ID:	X23193638		
Programme:	MSc Data Analytics	**Year:**	2024
Module:	Configuration Manual		
Lecturer:	Jorge Basilio		
Submission Due Date:	12-12-2024		
Project Title:	Enhancing Risk Assessment in Legal Documents Through Advanced Machine Learning		
Word Count:	958 **Page Count:** 19		

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: Sai Teja Pusarla ………………………………………………………………………………………………………

Date: 12-12-2024 ………………………………………………………………………………………………………

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):	

Configuration Manual

Sai Teja Pusarla
Student ID: x23193638

1. Introduction

This manual outlines a step-by-step replication guide for Assessing Risk whether it's Risky or Not Risky and Anomaly detection using advanced machine learning models. The experiment compares several models, including BERT, Legal BERT, Isolation Forest whereby Legal BERT surpasses other in accuracy and error minimization. This document lists the packages and configurations of software for the experimental environment so that it can replicate similar results.

2. Deployment Environment

2.1 Hardware Specification

- **Processor:** Intel Core i5 or equivalent
- **RAM:** 8 GB or higher
- **GPU:** Min 2 cores or higher.

2.2 Software Specification

- **Operating System:** Windows 10/11, macOS, or Linux-based OS
- **Programming Language:** Python 3.11
- **IDE:** Jupyter Notebook or VS Code (with Python extension)

2.3 Core Libraries:

- os: For file and directory related work.
- torch: Used to implement BERT and Legal-BERT architectures.
- PyPDF2: For reading and extracting text from pdfs.
- pyrna: To deal with tabular data
- numpy: For numerical calculations.
- seaborn and matplotlib: For visualisation of data.

2.4 Machine Learning and NLP:

- torch. nn. functional: Used for activation functions and computing loss.
- torch. utils. data: This is used to create DataLoader and TensorDataset.
- transformers: For all things model and tokenizer, pre-trained on BERT, Legal-BERT, etc.
- scikit-learn: For model evaluation, PCA, train-test split and anomaly detection
- imbalanced-learn performing the oversampling (SMOTE) to deal with the class imbalance.

2.5 Additional Tools:

- huggingface-hub — to download pre-trained models e.g. BERT, Legal-BERT

2.6 Download the Required Files:

Place the following files in the project directory:

- master_clauses.csv (used in risk assessment)
- Legal_docs.csv (used in anomaly detection)

You need to install the following packages to execute this code

```
!pip install pandas
!pip install matplotlib
!pip install scikit-learn
!pip install numpy
!pip install PyPDF2
!pip install tqdm
!pip install transformers
!pip install torch
```

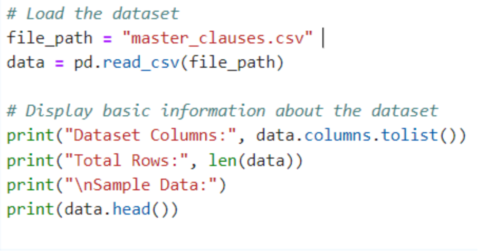
3. Configuration for Risk Assessment

3.1 **Open the Notebook**:

Open the file Risk Assessment.ipynb in Jupyter Notebook.

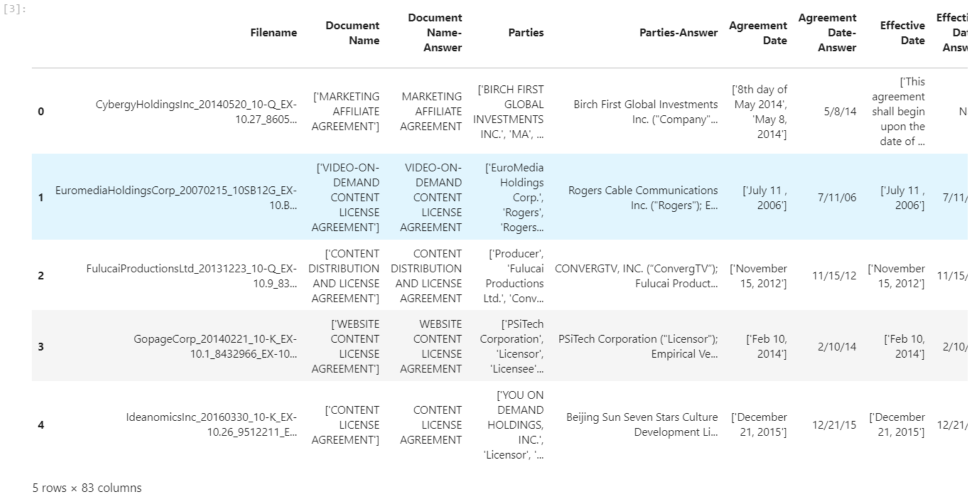
3.2 **Prepare the Dataset**:

- Ensure **master_clauses.csv** is in the same directory as the notebook.
- Update the file path in the code if needed:



```
# Load the dataset
file_path = "master_clauses.csv"
data = pd.read_csv(file_path)

# Display basic information about the dataset
print("Dataset Columns:", data.columns.tolist())
print("Total Rows:", len(data))
print("\nSample Data:")
print(data.head())
```

- To see the structure of dataset I’m printing the first few rows



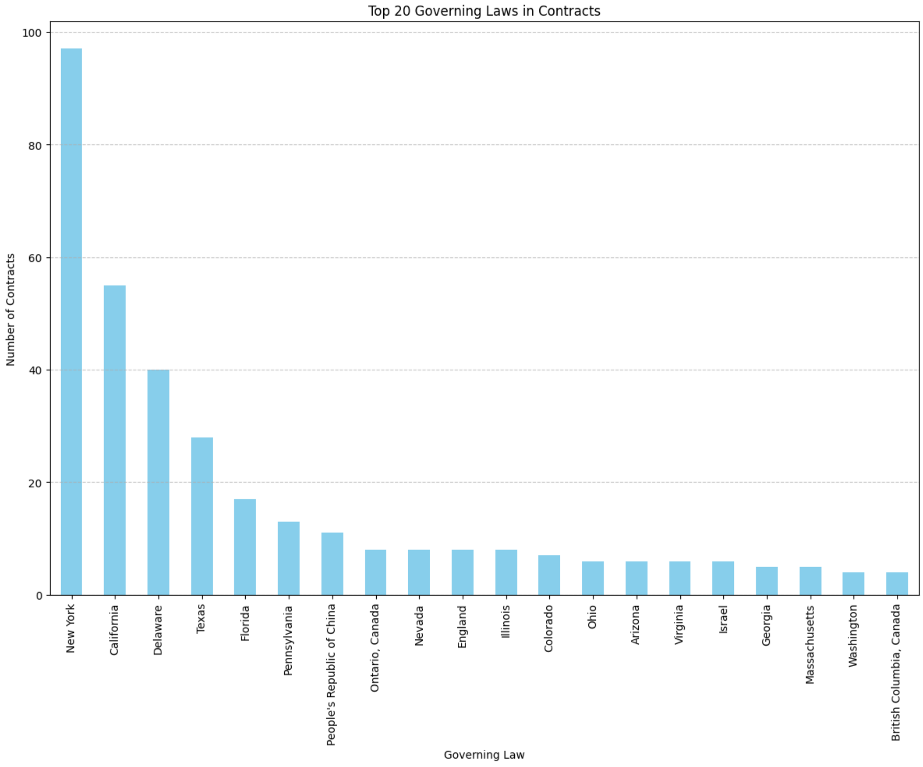
[3]:

	Filename	Document Name	Document Name-Answer	Parties	Parties-Answer	Agreement Date	Agreement Date-Answer	Effective Date	Effecti Da Answ
0	CybergyHoldingsInc_20140520_10-Q_EX-10.27_8605...	['MARKETING AFFILIATE AGREEMENT']	MARKETING AFFILIATE AGREEMENT	['BIRCH FIRST GLOBAL INVESTMENTS INC.', 'MA', ...	Birch First Global Investments Inc. ("Company"...	['8th day of May 2014', 'May 8, 2014']	5/8/14	['This agreement shall begin upon the date of ...	N
1	EuromediaHoldingsCorp_20070215_10SB12G_EX-10.B...	['VIDEO-ON-DEMAND CONTENT LICENSE AGREEMENT']	VIDEO-ON-DEMAND CONTENT LICENSE AGREEMENT	['EuroMedia Holdings Corp.', 'Rogers', 'Rogers...	Rogers Cable Communications Inc. ("Rogers"); E...	['July 11 , 2006']	7/11/06	['July 11 , 2006']	7/11/
2	FulucaiProductionsLtd_20131223_10-Q_EX-10.9_83...	['CONTENT DISTRIBUTION AND LICENSE AGREEMENT']	CONTENT DISTRIBUTION AND LICENSE AGREEMENT	['Producer', 'Fulucai Productions Ltd.', 'Conv...	CONVERGTV, INC. (“ConvergTV”); Fulucai Product...	['November 15, 2012']	11/15/12	['November 15, 2012']	11/15/
3	GopageCorp_20140221_10-K_EX-10.1_8432966_EX-10...	['WEBSITE CONTENT LICENSE AGREEMENT']	WEBSITE CONTENT LICENSE AGREEMENT	['PSiTech Corporation', 'Licensor', 'Licensee'...	PSiTech Corporation ("Licensor"); Empirical Ve...	['Feb 10, 2014']	2/10/14	['Feb 10, 2014']	2/10/
4	IdeanomicsInc_20160330_10-K_EX-10.26_9512211_E...	['CONTENT LICENSE AGREEMENT']	CONTENT LICENSE AGREEMENT	['YOU ON DEMAND HOLDINGS, INC.', 'Licensor', '...	Beijing Sun Seven Stars Culture Development Li...	['December 21, 2015']	12/21/15	['December 21, 2015']	12/21/

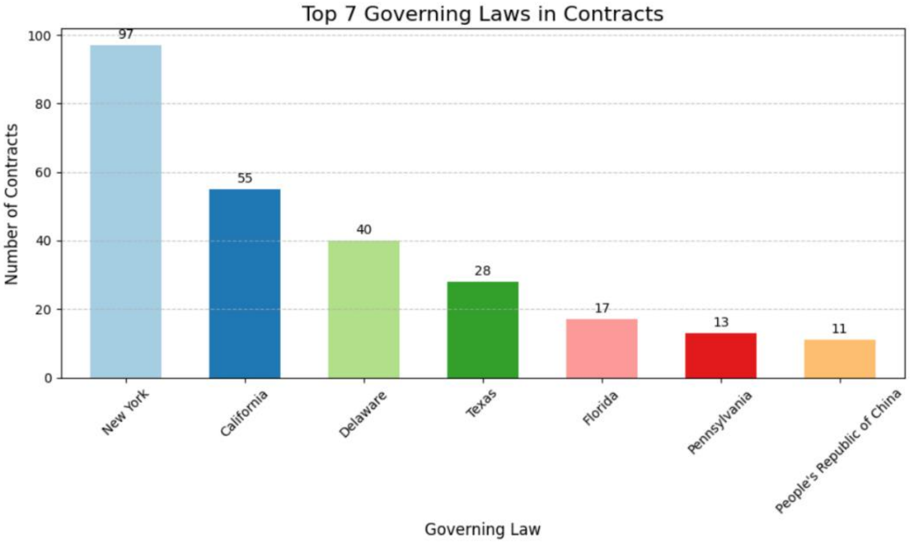
5 rows × 83 columns

3.3 Exploratory Data Analysis:

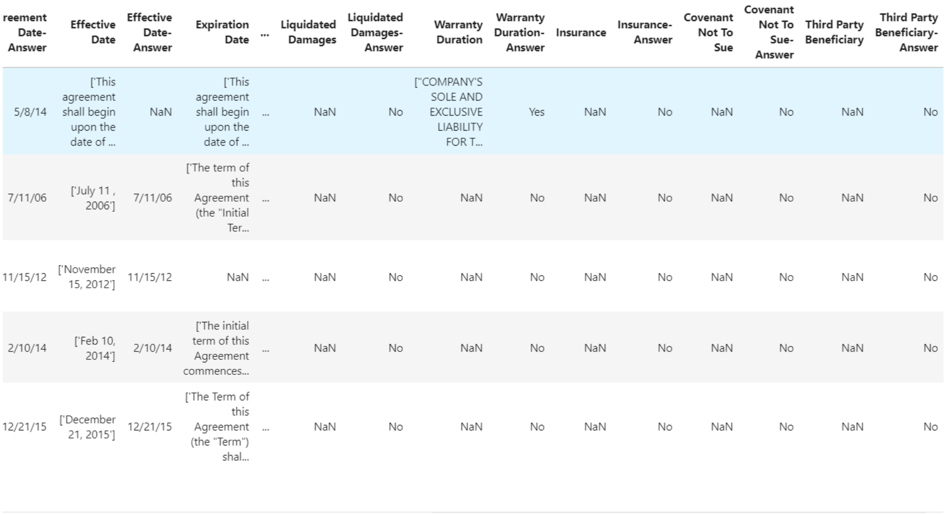
- **Top 20 governing laws in contracts**



- **Top 7 Governing laws in contracts**

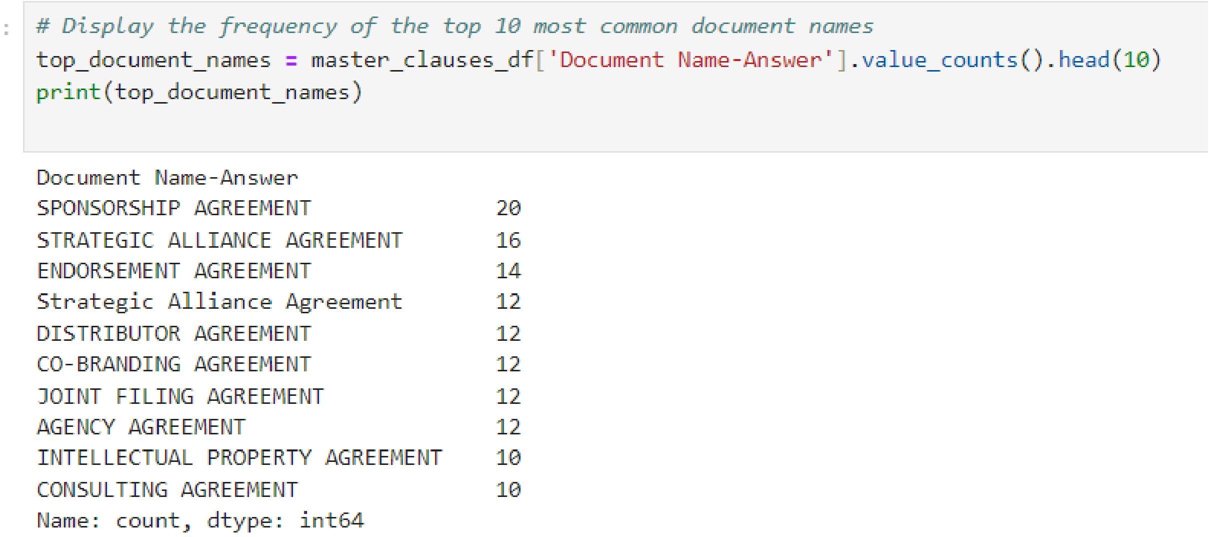


- Replace '[]' with NaN



reement Date-Answer	Effective Date	Effective Date-Answer	Expiration Date	...	Liquidated Damages	Liquidated Damages-Answer	Warranty Duration	Warranty Duration-Answer	Insurance	Insurance-Answer	Covenant Not To Sue	Covenant Not To Sue-Answer	Third Party Beneficiary	Third Party Beneficiary-Answer
5/8/14	['This agreement shall begin upon the date of ...	NaN	['This agreement shall begin upon the date of ...	...	NaN	No	["COMPANY'S SOLE AND EXCLUSIVE LIABILITY FOR T...	Yes	NaN	No	NaN	No	NaN	No
7/11/06	['July 11 , 2006']	7/11/06	['The term of this Agreement (the "Initial Ter...	...	NaN	No	NaN	No	NaN	No	NaN	No	NaN	No
11/15/12	['November 15, 2012']	11/15/12	NaN	...	NaN	No	NaN	No	NaN	No	NaN	No	NaN	No
2/10/14	['Feb 10, 2014']	2/10/14	['The initial term of this Agreement commences...	...	NaN	No	NaN	No	NaN	No	NaN	No	NaN	No
12/21/15	['December 21, 2015']	12/21/15	['The Term of this Agreement (the "Term") shal...	...	NaN	No	NaN	No	NaN	No	NaN	No	NaN	No

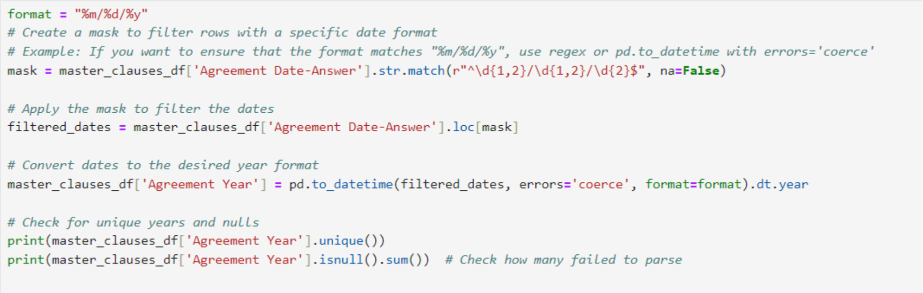
- **Displaying top 10 common documents**



```
# Display the frequency of the top 10 most common document names
top_document_names = master_clauses_df['Document Name-Answer'].value_counts().head(10)
print(top_document_names)
```

```
Document Name-Answer
SPONSORSHIP AGREEMENT                  20
STRATEGIC ALLIANCE AGREEMENT           16
ENDORSEMENT AGREEMENT                  14
Strategic Alliance Agreement           12
DISTRIBUTOR AGREEMENT                  12
CO-BRANDING AGREEMENT                  12
JOINT FILING AGREEMENT                 12
AGENCY AGREEMENT                       12
INTELLECTUAL PROPERTY AGREEMENT        10
CONSULTING AGREEMENT                   10
Name: count, dtype: int64
```

- ## Date Filtering and Year Extraction

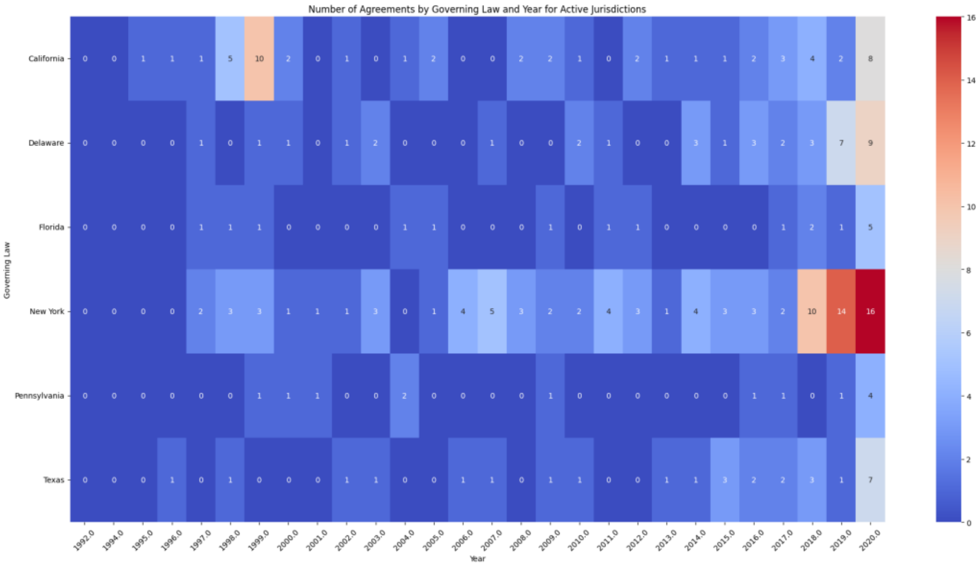


```
format = "%m/%d/%y"
# Create a mask to filter rows with a specific date format
# Example: If you want to ensure that the format matches "%m/%d/%y", use regex or pd.to_datetime with errors='coerce'
mask = master_clauses_df['Agreement Date-Answer'].str.match(r"^\d{1,2}/\d{1,2}/\d{2}$", na=False)

# Apply the mask to filter the dates
filtered_dates = master_clauses_df['Agreement Date-Answer'].loc[mask]

# Convert dates to the desired year format
master_clauses_df['Agreement Year'] = pd.to_datetime(filtered_dates, errors='coerce', format=format).dt.year

# Check for unique years and nulls
print(master_clauses_df['Agreement Year'].unique())
print(master_clauses_df['Agreement Year'].isnull().sum())  # Check how many failed to parse
```

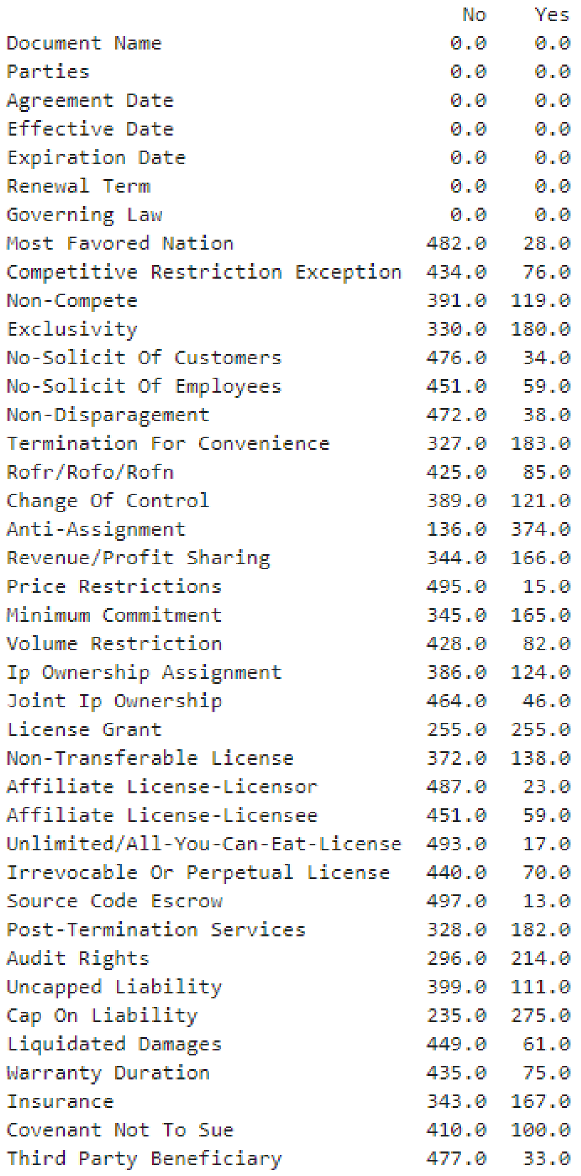
Output:

```
[2014. 2006. 2012. 2015. 1999. 2000. 2013. 2005. 2019. 2018. 2020. 2017.
 2008. 2016.   nan 2009. 2001. 2003. 2007. 2011. 2010. 2004. 1994. 1998.
 2002. 1997. 1992. 1995. 1996. 1990.]
19
```

- ## Heatmap:

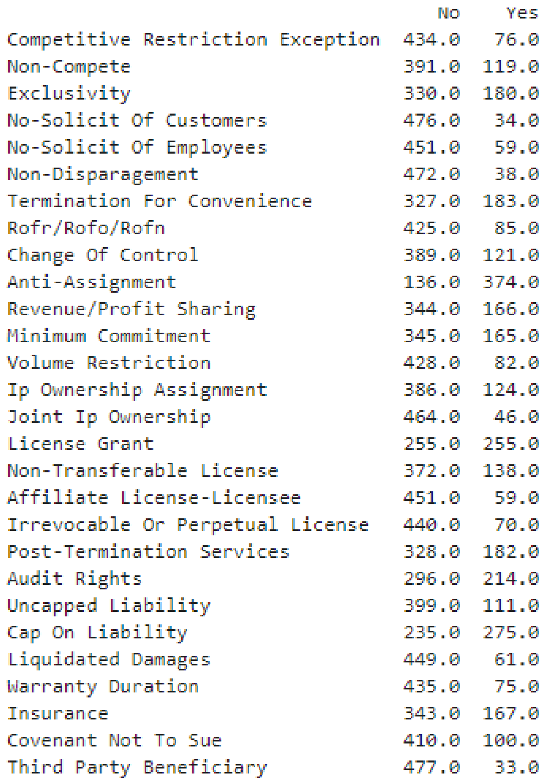


- ## Analysis and Export of Cleaned Clause Distribution Data



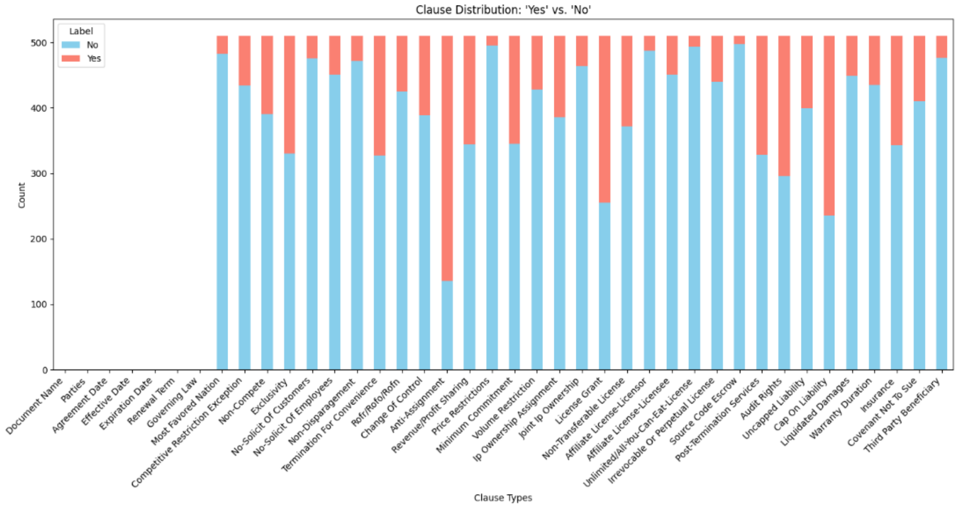
```
                                      No    Yes
Document Name                         0.0    0.0
Parties                               0.0    0.0
Agreement Date                        0.0    0.0
Effective Date                        0.0    0.0
Expiration Date                       0.0    0.0
Renewal Term                          0.0    0.0
Governing Law                         0.0    0.0
Most Favored Nation                 482.0   28.0
Competitive Restriction Exception   434.0   76.0
Non-Compete                         391.0  119.0
Exclusivity                         330.0  180.0
No-Solicit Of Customers             476.0   34.0
No-Solicit Of Employees             451.0   59.0
Non-Disparagement                   472.0   38.0
Termination For Convenience         327.0  183.0
Rofr/Rofo/Rofn                      425.0   85.0
Change Of Control                   389.0  121.0
Anti-Assignment                     136.0  374.0
Revenue/Profit Sharing              344.0  166.0
Price Restrictions                  495.0   15.0
Minimum Commitment                  345.0  165.0
Volume Restriction                  428.0   82.0
Ip Ownership Assignment             386.0  124.0
Joint Ip Ownership                  464.0   46.0
License Grant                       255.0  255.0
Non-Transferable License            372.0  138.0
Affiliate License-Licensor          487.0   23.0
Affiliate License-Licensee          451.0   59.0
Unlimited/All-You-Can-Eat-License   493.0   17.0
Irrevocable Or Perpetual License    440.0   70.0
Source Code Escrow                  497.0   13.0
Post-Termination Services           328.0  182.0
Audit Rights                        296.0  214.0
Uncapped Liability                  399.0  111.0
Cap On Liability                    235.0  275.0
Liquidated Damages                  449.0   61.0
Warranty Duration                   435.0   75.0
Insurance                           343.0  167.0
Covenant Not To Sue                 410.0  100.0
Third Party Beneficiary             477.0   33.0
```

- **Balancing Data by Filtering Based on Label Threshold**

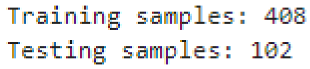


```
                                    No    Yes
Competitive Restriction Exception  434.0   76.0
Non-Compete                        391.0  119.0
Exclusivity                        330.0  180.0
No-Solicit Of Customers            476.0   34.0
No-Solicit Of Employees            451.0   59.0
Non-Disparagement                  472.0   38.0
Termination For Convenience        327.0  183.0
Rofr/Rofo/Rofn                     425.0   85.0
Change Of Control                  389.0  121.0
Anti-Assignment                    136.0  374.0
Revenue/Profit Sharing             344.0  166.0
Minimum Commitment                 345.0  165.0
Volume Restriction                 428.0   82.0
Ip Ownership Assignment            386.0  124.0
Joint Ip Ownership                 464.0   46.0
License Grant                      255.0  255.0
Non-Transferable License           372.0  138.0
Affiliate License-Licensee         451.0   59.0
Irrevocable Or Perpetual License   440.0   70.0
Post-Termination Services          328.0  182.0
Audit Rights                       296.0  214.0
Uncapped Liability                 399.0  111.0
Cap On Liability                   235.0  275.0
Liquidated Damages                 449.0   61.0
Warranty Duration                  435.0   75.0
Insurance                          343.0  167.0
Covenant Not To Sue                410.0  100.0
Third Party Beneficiary            477.0   33.0
```

- **"Yes" and "No" counts for all clauses**

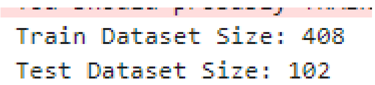


- **Preparation and Split of Dataset for Non-Compete Clause Classification**



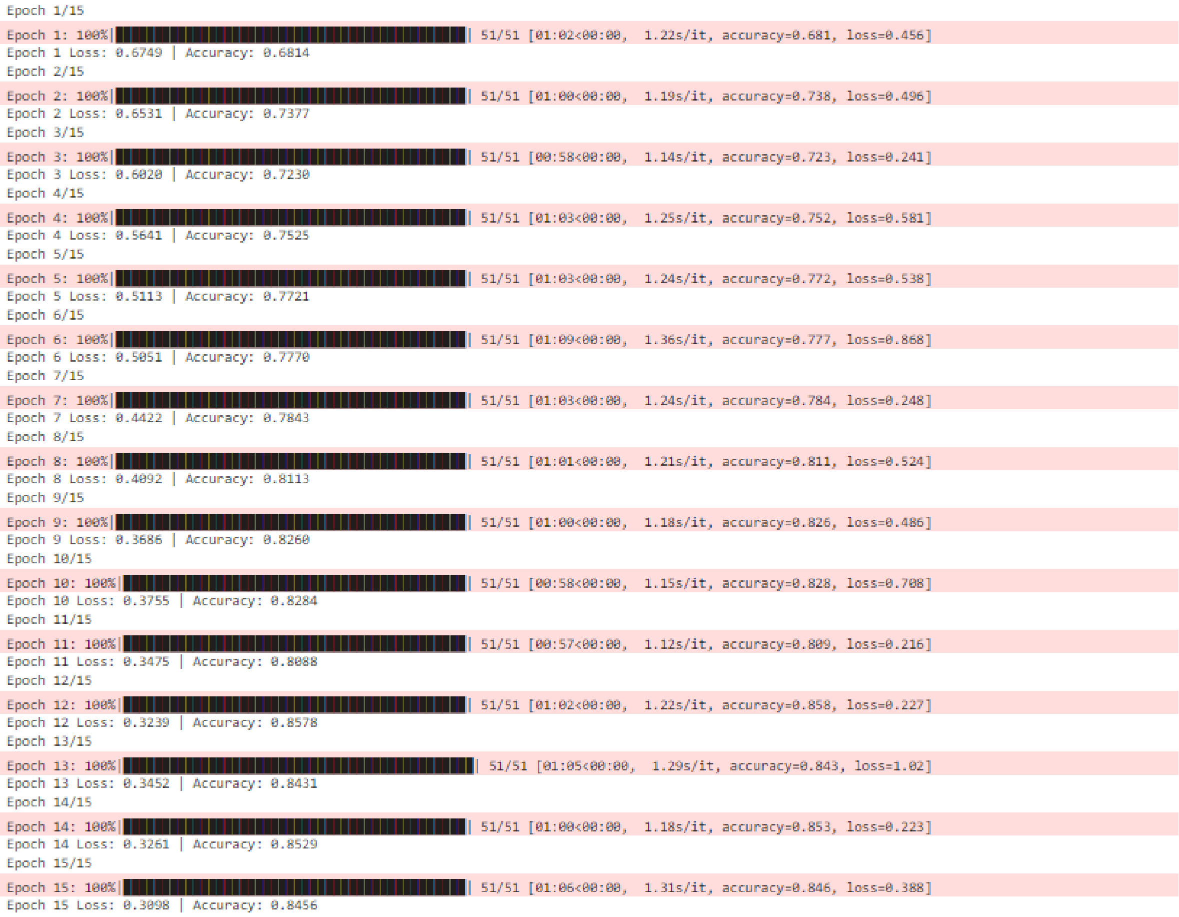
```
Training samples: 408
Testing samples: 102
```

- **BERT Model Setup and Data Preparation for Training**



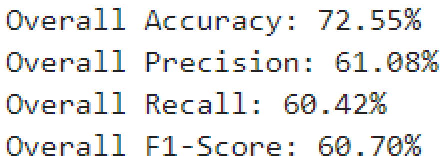
```
Train Dataset Size: 408
Test Dataset Size: 102
```

- **BERT Model Training with Class Weighting and Learning Rate Scheduling**



```
Epoch 1/15
Epoch 1: 100%|████████████████████████████████████| 51/51 [01:02<00:00,  1.22s/it, accuracy=0.681, loss=0.456]
Epoch 1 Loss: 0.6749 | Accuracy: 0.6814
Epoch 2/15
Epoch 2: 100%|████████████████████████████████████| 51/51 [01:00<00:00,  1.19s/it, accuracy=0.738, loss=0.496]
Epoch 2 Loss: 0.6531 | Accuracy: 0.7377
Epoch 3/15
Epoch 3: 100%|████████████████████████████████████| 51/51 [00:58<00:00,  1.14s/it, accuracy=0.723, loss=0.241]
Epoch 3 Loss: 0.6020 | Accuracy: 0.7230
Epoch 4/15
Epoch 4: 100%|████████████████████████████████████| 51/51 [01:03<00:00,  1.25s/it, accuracy=0.752, loss=0.581]
Epoch 4 Loss: 0.5641 | Accuracy: 0.7525
Epoch 5/15
Epoch 5: 100%|████████████████████████████████████| 51/51 [01:03<00:00,  1.24s/it, accuracy=0.772, loss=0.538]
Epoch 5 Loss: 0.5113 | Accuracy: 0.7721
Epoch 6/15
Epoch 6: 100%|████████████████████████████████████| 51/51 [01:09<00:00,  1.36s/it, accuracy=0.777, loss=0.868]
Epoch 6 Loss: 0.5051 | Accuracy: 0.7770
Epoch 7/15
Epoch 7: 100%|████████████████████████████████████| 51/51 [01:03<00:00,  1.24s/it, accuracy=0.784, loss=0.248]
Epoch 7 Loss: 0.4422 | Accuracy: 0.7843
Epoch 8/15
Epoch 8: 100%|████████████████████████████████████| 51/51 [01:01<00:00,  1.21s/it, accuracy=0.811, loss=0.524]
Epoch 8 Loss: 0.4092 | Accuracy: 0.8113
Epoch 9/15
Epoch 9: 100%|████████████████████████████████████| 51/51 [01:00<00:00,  1.18s/it, accuracy=0.826, loss=0.486]
Epoch 9 Loss: 0.3686 | Accuracy: 0.8260
Epoch 10/15
Epoch 10: 100%|███████████████████████████████████| 51/51 [00:58<00:00,  1.15s/it, accuracy=0.828, loss=0.708]
Epoch 10 Loss: 0.3755 | Accuracy: 0.8284
Epoch 11/15
Epoch 11: 100%|███████████████████████████████████| 51/51 [00:57<00:00,  1.12s/it, accuracy=0.809, loss=0.216]
Epoch 11 Loss: 0.3475 | Accuracy: 0.8088
Epoch 12/15
Epoch 12: 100%|███████████████████████████████████| 51/51 [01:02<00:00,  1.22s/it, accuracy=0.858, loss=0.227]
Epoch 12 Loss: 0.3239 | Accuracy: 0.8578
Epoch 13/15
Epoch 13: 100%|████████████████████████████████████| 51/51 [01:05<00:00,  1.29s/it, accuracy=0.843, loss=1.02]
Epoch 13 Loss: 0.3452 | Accuracy: 0.8431
Epoch 14/15
Epoch 14: 100%|███████████████████████████████████| 51/51 [01:00<00:00,  1.18s/it, accuracy=0.853, loss=0.223]
Epoch 14 Loss: 0.3261 | Accuracy: 0.8529
Epoch 15/15
Epoch 15: 100%|███████████████████████████████████| 51/51 [01:06<00:00,  1.31s/it, accuracy=0.846, loss=0.388]
Epoch 15 Loss: 0.3098 | Accuracy: 0.8456
```

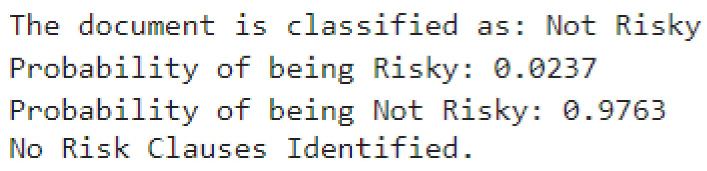
- **Evaluation Metrics**



```
Overall Accuracy: 72.55%
Overall Precision: 61.08%
Overall Recall: 60.42%
Overall F1-Score: 60.70%
```

- **Realtime Testing and Assessing Risk by Analysing Legal Possibilities:**

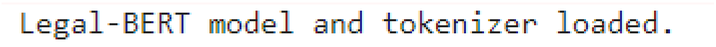
Once you run the cells with an actual legal contract, the output will provide a summary of the document risk assessment. In the first step, open the document, extract the text and process it. Using a pre-trained BERT model, the processed text is input to classify the degree of risk the document poses, with outputs indicating the document is "Risky" or "Not Risky," based on the content. Based on probabilities representing the likelihood of each risk category, as predicted by the model.



```
The document is classified as: Not Risky
Probability of being Risky: 0.0237
Probability of being Not Risky: 0.9763
No Risk Clauses Identified.
```

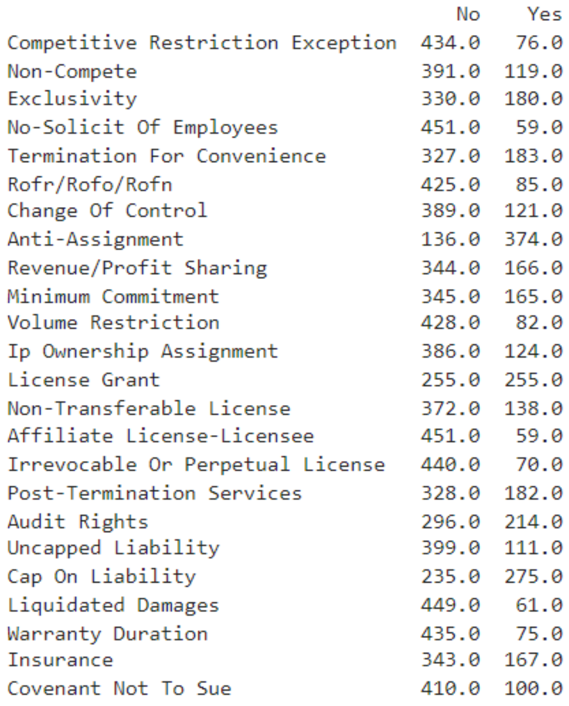
Legal BERT

After executing the Legal BERT code cell by cell you can see the following outputs



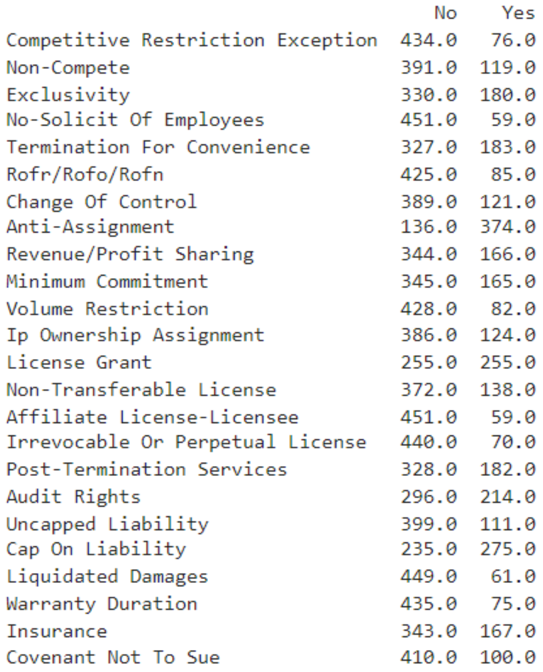
```
Legal-BERT model and tokenizer loaded.
```

- **Filtering and Balancing Clauses Based on Yes/No Distribution**



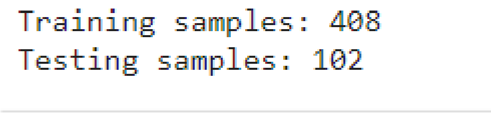
```
                                     No    Yes
Competitive Restriction Exception  434.0   76.0
Non-Compete                        391.0  119.0
Exclusivity                        330.0  180.0
No-Solicit Of Employees            451.0   59.0
Termination For Convenience        327.0  183.0
Rofr/Rofo/Rofn                     425.0   85.0
Change Of Control                  389.0  121.0
Anti-Assignment                    136.0  374.0
Revenue/Profit Sharing             344.0  166.0
Minimum Commitment                 345.0  165.0
Volume Restriction                 428.0   82.0
Ip Ownership Assignment            386.0  124.0
License Grant                      255.0  255.0
Non-Transferable License           372.0  138.0
Affiliate License-Licensee         451.0   59.0
Irrevocable Or Perpetual License   440.0   70.0
Post-Termination Services          328.0  182.0
Audit Rights                       296.0  214.0
Uncapped Liability                 399.0  111.0
Cap On Liability                   235.0  275.0
Liquidated Damages                 449.0   61.0
Warranty Duration                  435.0   75.0
Insurance                          343.0  167.0
Covenant Not To Sue                410.0  100.0
```

- **Clause Distribution Analysis and Balancing for Classification**



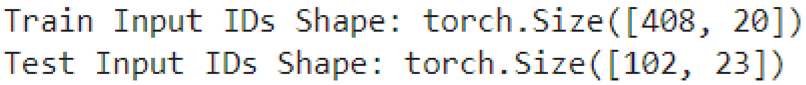
```
                                        No    Yes
Competitive Restriction Exception    434.0   76.0
Non-Compete                          391.0  119.0
Exclusivity                          330.0  180.0
No-Solicit Of Employees              451.0   59.0
Termination For Convenience          327.0  183.0
Rofr/Rofo/Rofn                       425.0   85.0
Change Of Control                    389.0  121.0
Anti-Assignment                      136.0  374.0
Revenue/Profit Sharing               344.0  166.0
Minimum Commitment                   345.0  165.0
Volume Restriction                   428.0   82.0
Ip Ownership Assignment              386.0  124.0
License Grant                        255.0  255.0
Non-Transferable License             372.0  138.0
Affiliate License-Licensee           451.0   59.0
Irrevocable Or Perpetual License     440.0   70.0
Post-Termination Services            328.0  182.0
Audit Rights                         296.0  214.0
Uncapped Liability                   399.0  111.0
Cap On Liability                     235.0  275.0
Liquidated Damages                   449.0   61.0
Warranty Duration                    435.0   75.0
Insurance                            343.0  167.0
Covenant Not To Sue                  410.0  100.0
```

- **Preparation and Split of Dataset for Specific Clause Classification**



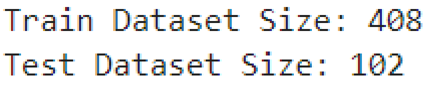
```
Training samples: 408
Testing samples: 102
```

4 Tokenize the training and testing data



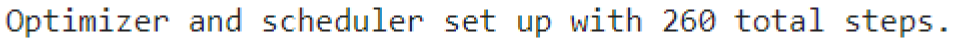
```
Train Input IDs Shape: torch.Size([408, 20])
Test Input IDs Shape: torch.Size([102, 23])
```

5 Dataset and DataLoader Configuration for Model Training



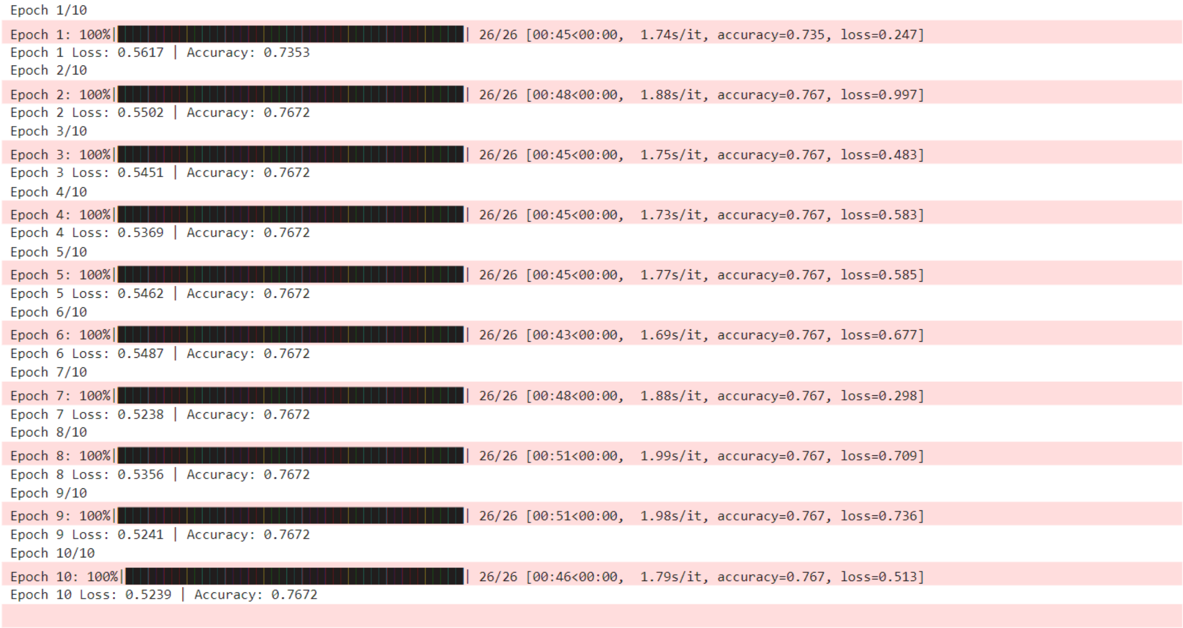
```
Train Dataset Size: 408
Test Dataset Size: 102
```

6 Optimizer and Learning Rate Scheduler Setup for Training



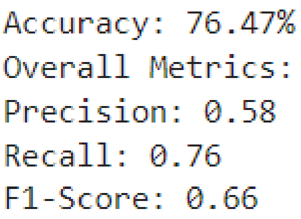
```
Optimizer and scheduler set up with 260 total steps.
```

7 Model Training Loop



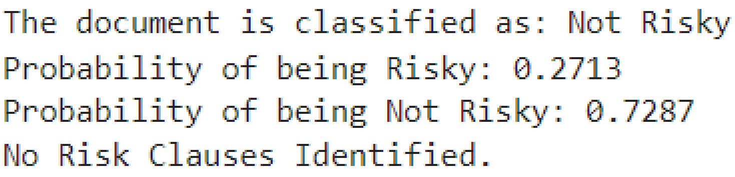
```
Epoch 1/10
Epoch 1: 100%|██████████████████████████████████████████████| 26/26 [00:45<00:00,  1.74s/it, accuracy=0.735, loss=0.247]
Epoch 1 Loss: 0.5617 | Accuracy: 0.7353
Epoch 2/10
Epoch 2: 100%|██████████████████████████████████████████████| 26/26 [00:48<00:00,  1.88s/it, accuracy=0.767, loss=0.997]
Epoch 2 Loss: 0.5502 | Accuracy: 0.7672
Epoch 3/10
Epoch 3: 100%|██████████████████████████████████████████████| 26/26 [00:45<00:00,  1.75s/it, accuracy=0.767, loss=0.483]
Epoch 3 Loss: 0.5451 | Accuracy: 0.7672
Epoch 4/10
Epoch 4: 100%|██████████████████████████████████████████████| 26/26 [00:45<00:00,  1.73s/it, accuracy=0.767, loss=0.583]
Epoch 4 Loss: 0.5369 | Accuracy: 0.7672
Epoch 5/10
Epoch 5: 100%|██████████████████████████████████████████████| 26/26 [00:45<00:00,  1.77s/it, accuracy=0.767, loss=0.585]
Epoch 5 Loss: 0.5462 | Accuracy: 0.7672
Epoch 6/10
Epoch 6: 100%|██████████████████████████████████████████████| 26/26 [00:43<00:00,  1.69s/it, accuracy=0.767, loss=0.677]
Epoch 6 Loss: 0.5487 | Accuracy: 0.7672
Epoch 7/10
Epoch 7: 100%|██████████████████████████████████████████████| 26/26 [00:48<00:00,  1.88s/it, accuracy=0.767, loss=0.298]
Epoch 7 Loss: 0.5238 | Accuracy: 0.7672
Epoch 8/10
Epoch 8: 100%|██████████████████████████████████████████████| 26/26 [00:51<00:00,  1.99s/it, accuracy=0.767, loss=0.709]
Epoch 8 Loss: 0.5356 | Accuracy: 0.7672
Epoch 9/10
Epoch 9: 100%|██████████████████████████████████████████████| 26/26 [00:51<00:00,  1.98s/it, accuracy=0.767, loss=0.736]
Epoch 9 Loss: 0.5241 | Accuracy: 0.7672
Epoch 10/10
Epoch 10: 100%|█████████████████████████████████████████████| 26/26 [00:46<00:00,  1.79s/it, accuracy=0.767, loss=0.513]
Epoch 10 Loss: 0.5239 | Accuracy: 0.7672
```

8 Evaluation Metrics



```
Accuracy: 76.47%
Overall Metrics:
Precision: 0.58
Recall: 0.76
F1-Score: 0.66
```

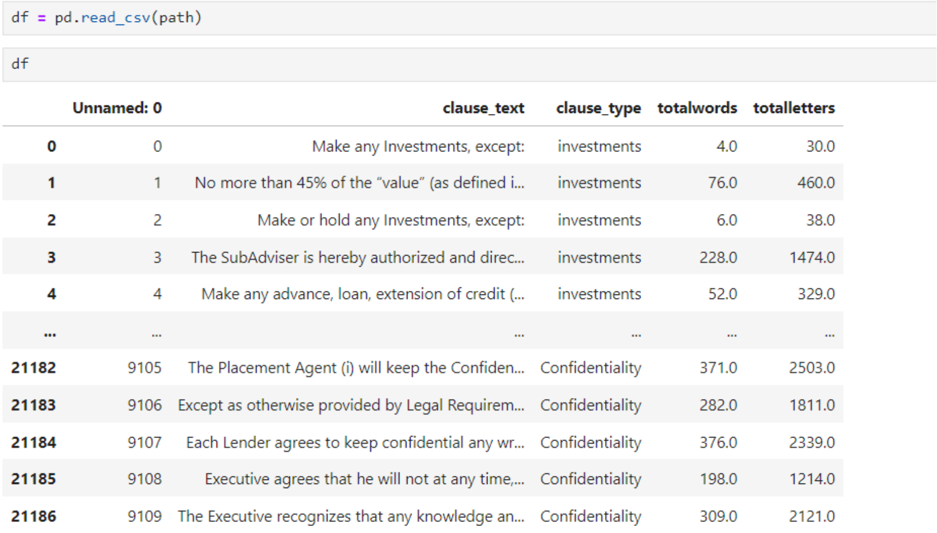
- **Realtime testing**



```
The document is classified as: Not Risky
Probability of being Risky: 0.2713
Probability of being Not Risky: 0.7287
No Risk Clauses Identified.
```

4. Configuration for Anomaly Detection

- **Structure of dataset**

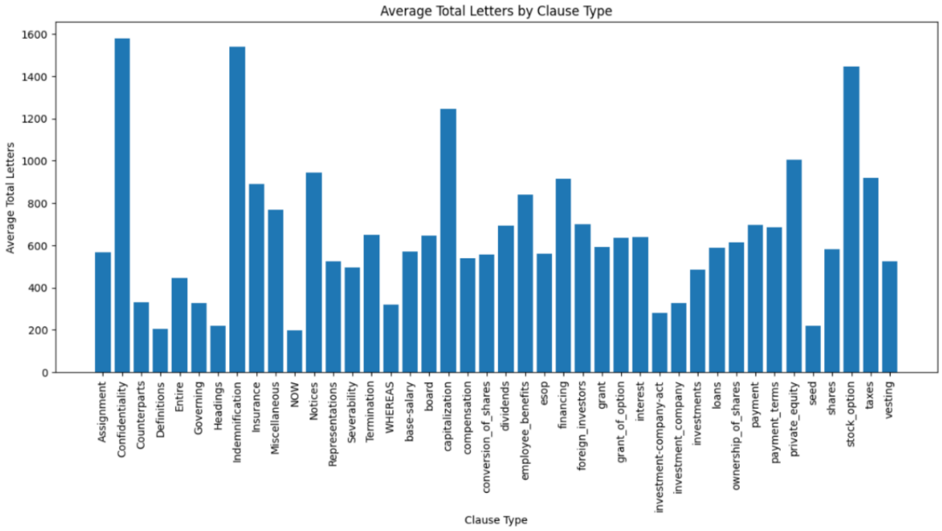


```
df = pd.read_csv(path)
```

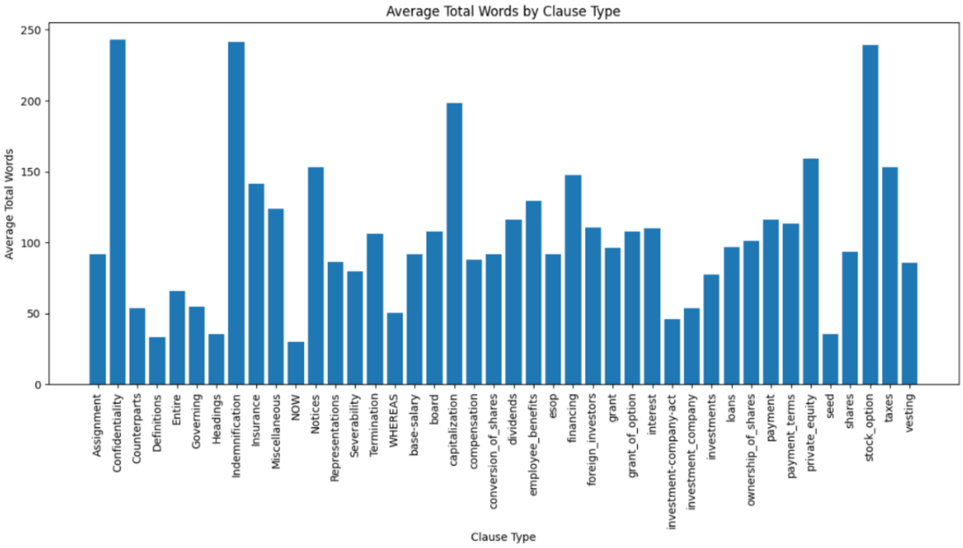
```
df
```

	Unnamed: 0	clause_text	clause_type	totalwords	totalletters
0	0	Make any Investments, except:	investments	4.0	30.0
1	1	No more than 45% of the "value" (as defined i...	investments	76.0	460.0
2	2	Make or hold any Investments, except:	investments	6.0	38.0
3	3	The SubAdviser is hereby authorized and direc...	investments	228.0	1474.0
4	4	Make any advance, loan, extension of credit (...	investments	52.0	329.0
...	...	...	...	...	...
21182	9105	The Placement Agent (i) will keep the Confiden...	Confidentiality	371.0	2503.0
21183	9106	Except as otherwise provided by Legal Requirem...	Confidentiality	282.0	1811.0
21184	9107	Each Lender agrees to keep confidential any wr...	Confidentiality	376.0	2339.0
21185	9108	Executive agrees that he will not at any time,...	Confidentiality	198.0	1214.0
21186	9109	The Executive recognizes that any knowledge an...	Confidentiality	309.0	2121.0

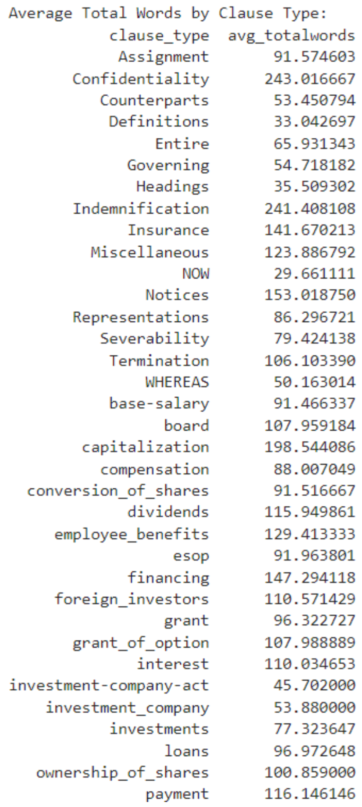
- ## Average Total Letters by Clause Type



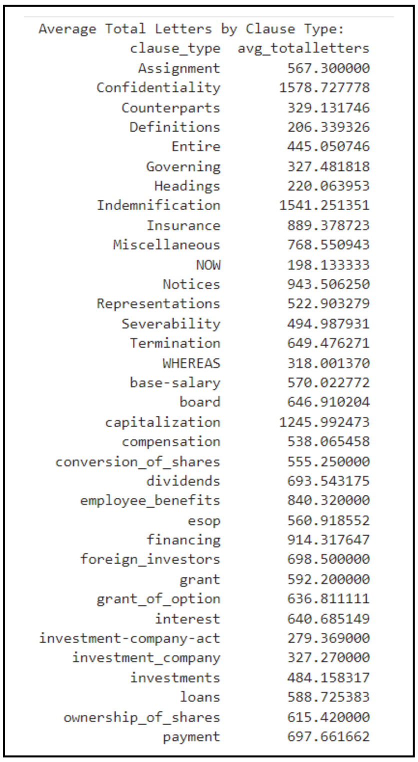
- ## Average Total Words by Clause Type



- ### Displaying Average Words and Letters

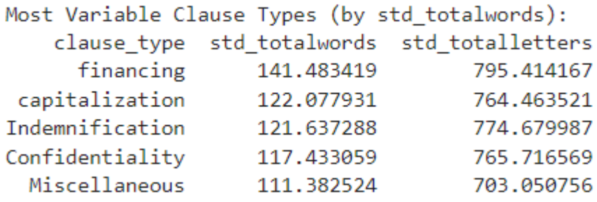


```
Average Total Words by Clause Type:
                 clause_type  avg_totalwords
                  Assignment       91.574603
             Confidentiality      243.016667
                Counterparts       53.450794
                 Definitions       33.042697
                      Entire       65.931343
                   Governing       54.718182
                    Headings       35.509302
             Indemnification      241.408108
                   Insurance      141.670213
               Miscellaneous      123.886792
                         NOW       29.661111
                     Notices      153.018750
             Representations       86.296721
                Severability       79.424138
                 Termination      106.103390
                     WHEREAS       50.163014
                 base-salary       91.466337
                       board      107.959184
              capitalization      198.544086
                compensation       88.007049
         conversion_of_shares       91.516667
                   dividends      115.949861
            employee_benefits      129.413333
                        esop       91.963801
                   financing      147.294118
            foreign_investors      110.571429
                       grant       96.322727
              grant_of_option      107.988889
                    interest      110.034653
       investment-company-act       45.702000
          investment_company       53.880000
                 investments       77.323647
                       loans       96.972648
           ownership_of_shares      100.859000
                     payment      116.146146
```

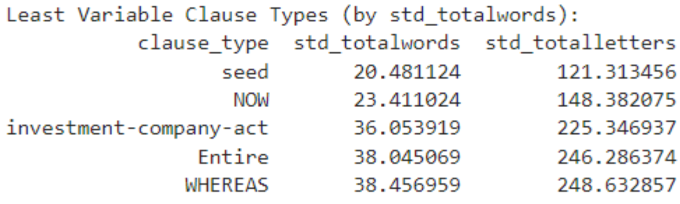


```
Average Total Letters by Clause Type:
                 clause_type  avg_totalletters
                  Assignment        567.300000
             Confidentiality       1578.727778
                Counterparts        329.131746
                 Definitions        206.339326
                      Entire        445.050746
                   Governing        327.481818
                    Headings        220.063953
             Indemnification       1541.251351
                   Insurance        889.378723
               Miscellaneous        768.550943
                         NOW        198.133333
                     Notices        943.506250
             Representations        522.903279
                Severability        494.987931
                 Termination        649.476271
                     WHEREAS        318.001370
                 base-salary        570.022772
                       board        646.910204
              capitalization       1245.992473
                compensation        538.065458
         conversion_of_shares       555.250000
                   dividends        693.543175
            employee_benefits       840.320000
                        esop        560.918552
                   financing        914.317647
            foreign_investors       698.500000
                       grant        592.200000
              grant_of_option       636.811111
                    interest        640.685149
       investment-company-act       279.369000
          investment_company        327.270000
                 investments        484.158317
                       loans        588.725383
           ownership_of_shares      615.420000
                     payment        697.661662
```

- ### Sorting and Displaying Variability

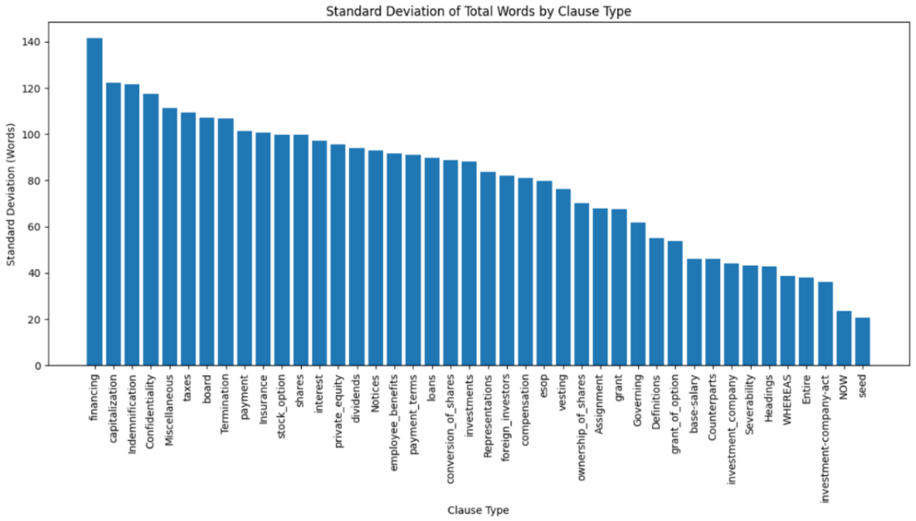


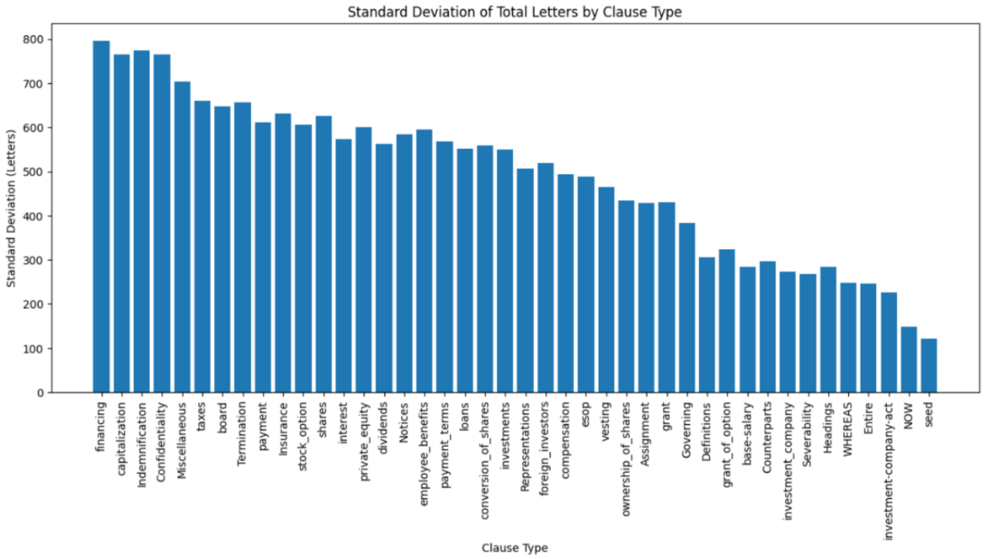
```
Most Variable Clause Types (by std_totalwords):
     clause_type  std_totalwords  std_totalletters
       financing      141.483419        795.414167
  capitalization      122.077931        764.463521
 Indemnification      121.637288        774.679987
 Confidentiality      117.433059        765.716569
   Miscellaneous      111.382524        703.050756
```



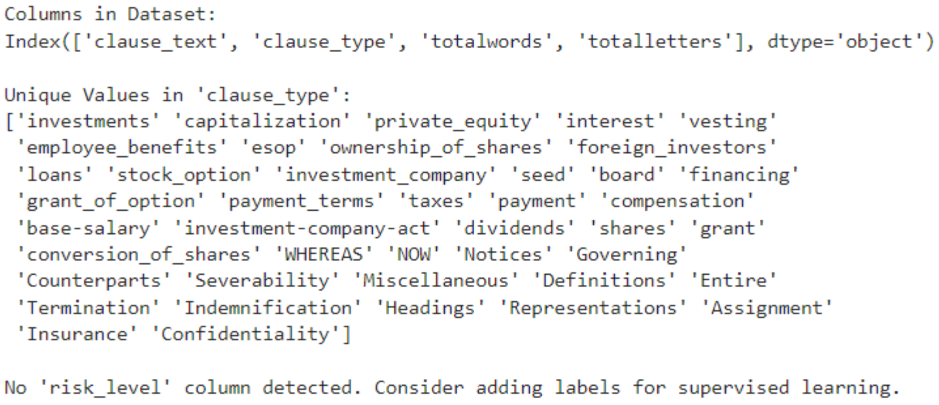
```
Least Variable Clause Types (by std_totalwords):
             clause_type  std_totalwords  std_totalletters
                    seed       20.481124        121.313456
                     NOW       23.411024        148.382075
  investment-company-act       36.053919        225.346937
                  Entire       38.045069        246.286374
                 WHEREAS       38.456959        248.632857
```

- ### Standard Deviations Plot



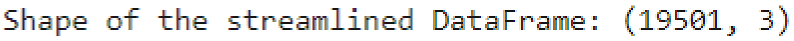


- **Data Inspection**



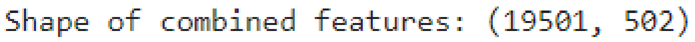
```
Columns in Dataset:
Index(['clause_text', 'clause_type', 'totalwords', 'totalletters'], dtype='object')

Unique Values in 'clause_type':
['investments' 'capitalization' 'private_equity' 'interest' 'vesting'
 'employee_benefits' 'esop' 'ownership_of_shares' 'foreign_investors'
 'loans' 'stock_option' 'investment_company' 'seed' 'board' 'financing'
 'grant_of_option' 'payment_terms' 'taxes' 'payment' 'compensation'
 'base-salary' 'investment-company-act' 'dividends' 'shares' 'grant'
 'conversion_of_shares' 'WHEREAS' 'NOW' 'Notices' 'Governing'
 'Counterparts' 'Severability' 'Miscellaneous' 'Definitions' 'Entire'
 'Termination' 'Indemnification' 'Headings' 'Representations' 'Assignment'
 'Insurance' 'Confidentiality']

No 'risk_level' column detected. Consider adding labels for supervised learning.
```

- **Data Cleaning and Outlier Removal**



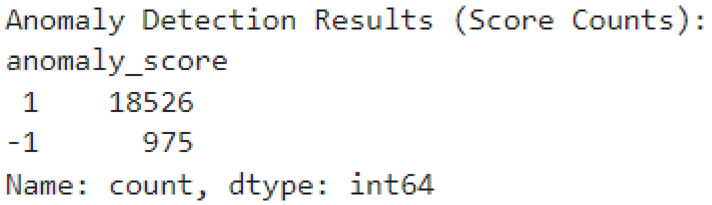
```
Shape of the streamlined DataFrame: (19501, 3)
```

- **Feature Engineering: TF-IDF and Scaling**



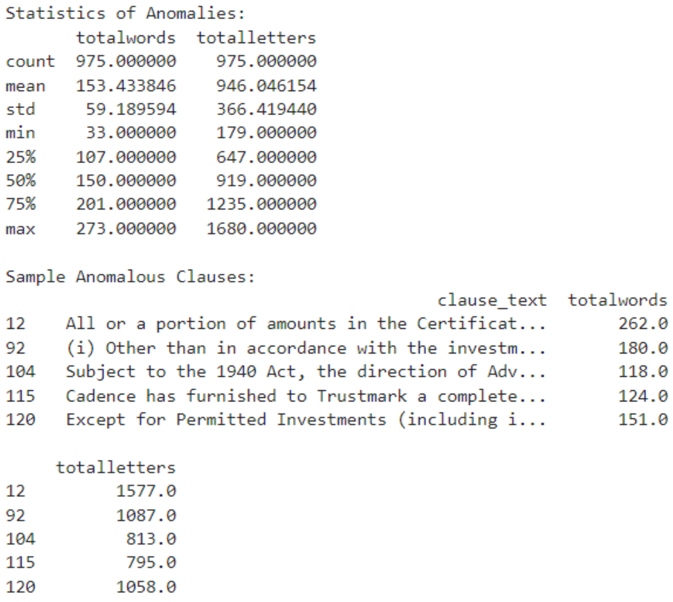
```
Shape of combined features: (19501, 502)
```

- **Anomaly Detection with Isolation Forest**



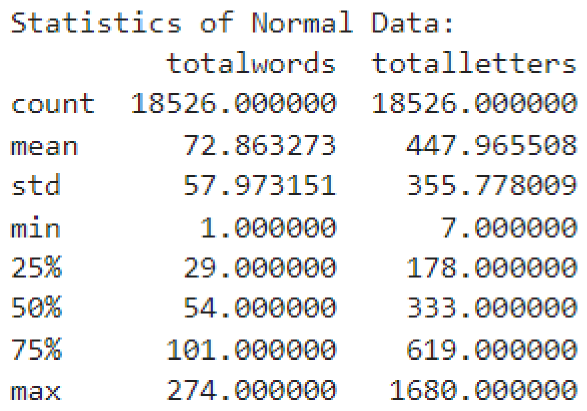
```
Anomaly Detection Results (Score Counts):
anomaly_score
 1     18526
-1       975
Name: count, dtype: int64
```

- **Analysis and Display of Detected Anomalies**



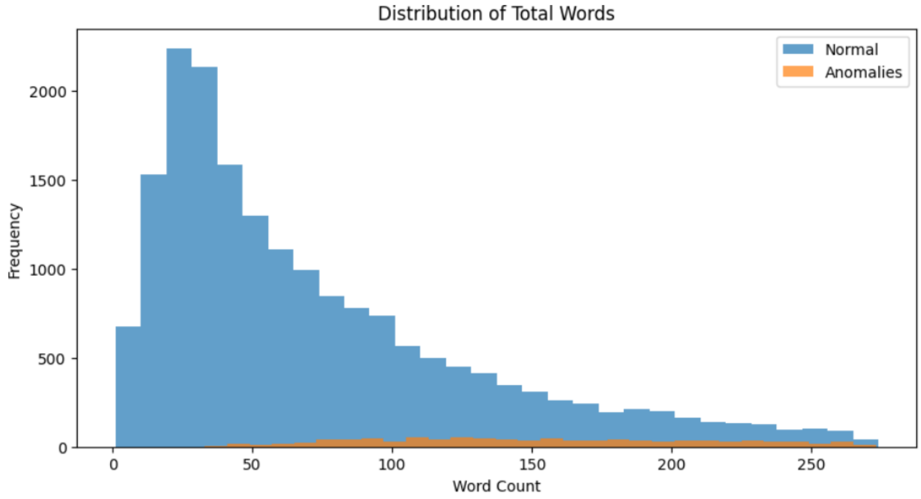
```
Statistics of Anomalies:
       totalwords  totalletters
count  975.000000    975.000000
mean   153.433846    946.046154
std     59.189594    366.419440
min     33.000000    179.000000
25%    107.000000    647.000000
50%    150.000000    919.000000
75%    201.000000   1235.000000
max    273.000000   1680.000000

Sample Anomalous Clauses:
                                            clause_text  totalwords  \
12    All or a portion of amounts in the Certificat...       262.0
92    (i) Other than in accordance with the investm...       180.0
104   Subject to the 1940 Act, the direction of Adv...       118.0
115   Cadence has furnished to Trustmark a complete...       124.0
120   Except for Permitted Investments (including i...       151.0

     totalletters
12         1577.0
92         1087.0
104         813.0
115         795.0
120        1058.0
```

9 Statistics for Normal data

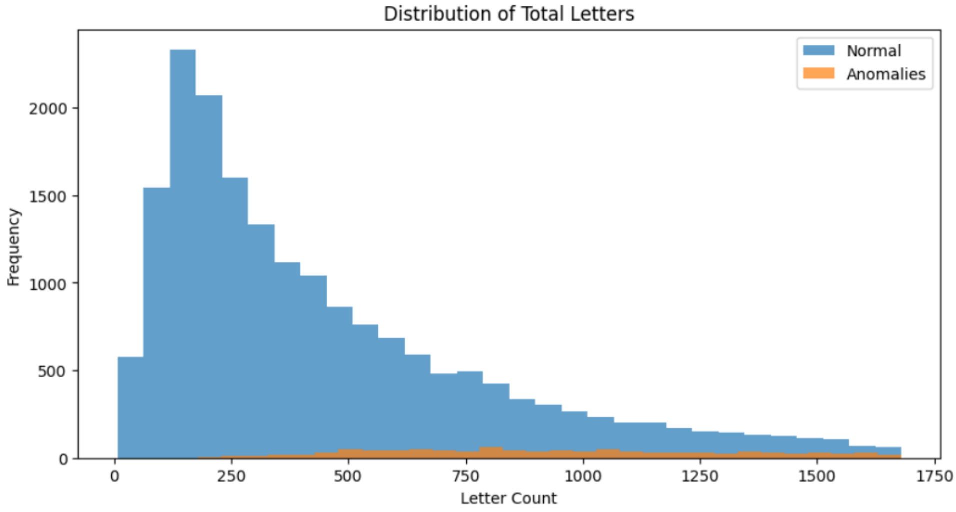


```
Statistics of Normal Data:
         totalwords  totalletters
count  18526.000000  18526.000000
mean      72.863273    447.965508
std       57.973151    355.778009
min        1.000000      7.000000
25%       29.000000    178.000000
50%       54.000000    333.000000
75%      101.000000    619.000000
max      274.000000   1680.000000
```

- **Distributions of Totalwords**



- **Distributions of Total Letters**



References

Python:_https://www.python.org

Dataset 1: CUAD Dataset https://www.atticusprojectai.org/cuad
Dataset 2: legal clauses dataset link:
https://www.kaggle.com/datasets/mohammedalrashidan/contracts-clauses-datasets