Fine-Tuning a Language Model for Patent Classification

Link:
https://huggingface.co/spaces/rb757/new_patent_app

Introduction

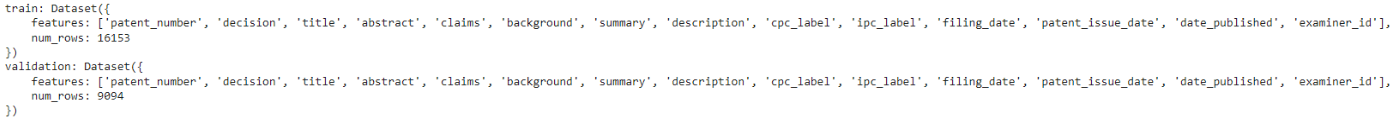
This project aims to fine-tune a language model using the Hugging Face Transformers library to classify patent applications. Leveraging the Harvard USPTO Patent Dataset (HUPD), we specifically focused on patent applications submitted in January 2016. Our primary objective was to develop an advanced classifier capable of predicting the outcomes of patent applications by analyzing text from their abstract and claims sections. To enhance the efficiency of the fine-tuning process, we utilized Google Colab's GPU, significantly accelerating the training and enabling the handling of complex, high-dimensional data inherent in patent documents.

Dataset Overview

❖ Loading the Dataset

We employed the Harvard USPTO Patent Dataset (HUPD), filtering it to include only patent applications submitted in January 2016. The dataset was divided into training and validation sets based on filing dates:

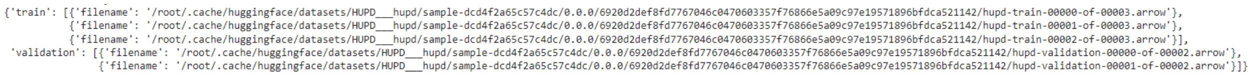
- Training Set: Patent applications submitted from January 1, 2016, to January 21, 2016.
- Validation Set: Patent applications submitted from January 22, 2016, to January 31, 2016.



```
train: Dataset({
    features: ['patent_number', 'decision', 'title', 'abstract', 'claims', 'background', 'summary', 'description', 'cpc_label', 'ipc_label', 'filing_date', 'patent_issue_date', 'date_published', 'examiner_id'],
    num_rows: 16153
})
validation: Dataset({
    features: ['patent_number', 'decision', 'title', 'abstract', 'claims', 'background', 'summary', 'description', 'cpc_label', 'ipc_label', 'filing_date', 'patent_issue_date', 'date_published', 'examiner_id'],
    num_rows: 9094
})
```

❖ Cached Dataset Information

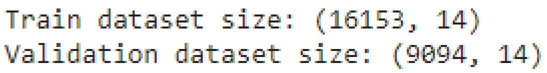
The dataset is cached locally, with specific files stored in the Hugging Face cache directory. The training set consists of three cached files, while the validation set contains two, optimizing data loading during model training and evaluation.



```
{'train': [{'filename': '/root/.cache/huggingface/datasets/HUPD___hupd/sample-dcd4f2a65c57c4dc/0.0.0/6920d2def8fd7767046c0470603357f76866e5a09c97e19571896bfdca521142/hupd-train-00000-of-00003.arrow'},
          {'filename': '/root/.cache/huggingface/datasets/HUPD___hupd/sample-dcd4f2a65c57c4dc/0.0.0/6920d2def8fd7767046c0470603357f76866e5a09c97e19571896bfdca521142/hupd-train-00001-of-00003.arrow'},
          {'filename': '/root/.cache/huggingface/datasets/HUPD___hupd/sample-dcd4f2a65c57c4dc/0.0.0/6920d2def8fd7767046c0470603357f76866e5a09c97e19571896bfdca521142/hupd-train-00002-of-00003.arrow'}],
 'validation': [{'filename': '/root/.cache/huggingface/datasets/HUPD___hupd/sample-dcd4f2a65c57c4dc/0.0.0/6920d2def8fd7767046c0470603357f76866e5a09c97e19571896bfdca521142/hupd-validation-00000-of-00002.arrow'},
               {'filename': '/root/.cache/huggingface/datasets/HUPD___hupd/sample-dcd4f2a65c57c4dc/0.0.0/6920d2def8fd7767046c0470603357f76866e5a09c97e19571896bfdca521142/hupd-validation-00001-of-00002.arrow'}]}
```

❖ Dataset Sizes

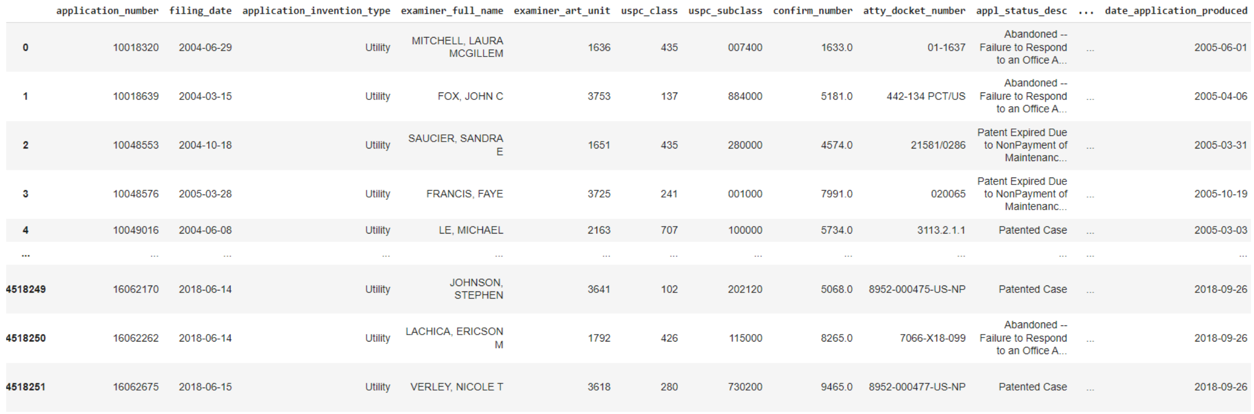
The shapes of the train and validation datasets confirm that the training set contains 14 features across 16,153 samples, while the validation set includes the same number of features across 9,094 samples.



```
Train dataset size: (16153, 14)
Validation dataset size: (9094, 14)
```

❖ Metadata Integration

To enhance the dataset, a path to external metadata stored in Feather format was specified. This metadata provides additional context for each patent, including attributes such as application number, filing date, examiner's full name, and application invention type. The metadata file contains 4,518,254 rows and 33 columns, which are crucial for comprehensive analysis and modeling.

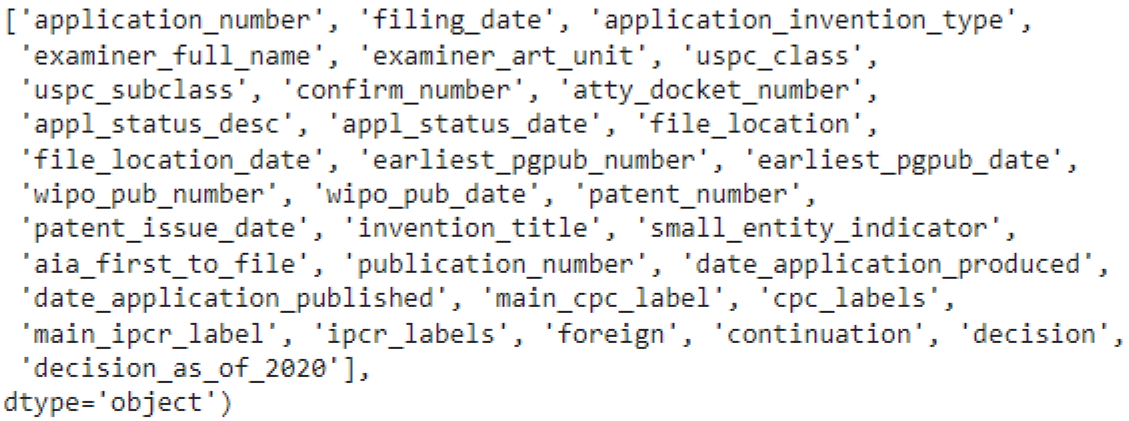


	application_number	filing_date	application_invention_type	examiner_full_name	examiner_art_unit	uspc_class	uspc_subclass	confirm_number	atty_docket_number	appl_status_desc	...	date_application_produced
0	10018320	2004-06-29	Utility	MITCHELL, LAURA MCGILLEM	1636	435	007400	1633.0	01-1637	Abandoned -- Failure to Respond to an Office A...	...	2005-06-01
1	10018639	2004-03-15	Utility	FOX, JOHN C	3753	137	884000	5181.0	442-134 PCT/US	Abandoned -- Failure to Respond to an Office A...	...	2005-04-06
2	10048553	2004-10-18	Utility	SAUCIER, SANDRA E	1651	435	280000	4574.0	21581/0286	Patent Expired Due to NonPayment of Maintenanc...	...	2005-03-31
3	10048576	2005-03-28	Utility	FRANCIS, FAYE	3725	241	001000	7991.0	020065	Patent Expired Due to NonPayment of Maintenanc...	...	2005-10-19
4	10049016	2004-06-08	Utility	LE, MICHAEL	2163	707	100000	5734.0	3113.2.1.1	Patented Case	...	2005-03-03
...	...	...	...	...	...	...	...	...	...	...	...	...
4518249	16062170	2018-06-14	Utility	JOHNSON, STEPHEN	3641	102	202120	5068.0	8952-000475-US-NP	Patented Case	...	2018-09-26
4518250	16062262	2018-06-14	Utility	LACHICA, ERICSON M	1792	426	115000	8265.0	7066-X18-099	Abandoned -- Failure to Respond to an Office A...	...	2018-09-26
4518251	16062675	2018-06-15	Utility	VERLEY, NICOLE T	3618	280	730200	9465.0	8952-000477-US-NP	Patented Case	...	2018-09-26

❖ Dataframe Columns

Key columns in the metadata DataFrame include:

- patent_number
- decision
- decision_as_of_2020
- cpc_labels
- ipcr_labels



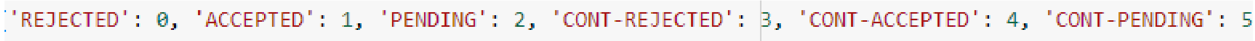
```
['application_number', 'filing_date', 'application_invention_type',
 'examiner_full_name', 'examiner_art_unit', 'uspc_class',
 'uspc_subclass', 'confirm_number', 'atty_docket_number',
 'appl_status_desc', 'appl_status_date', 'file_location',
 'file_location_date', 'earliest_pgpub_number', 'earliest_pgpub_date',
 'wipo_pub_number', 'wipo_pub_date', 'patent_number',
 'patent_issue_date', 'invention_title', 'small_entity_indicator',
 'aia_first_to_file', 'publication_number', 'date_application_produced',
 'date_application_published', 'main_cpc_label', 'cpc_labels',
 'main_ipcr_label', 'ipcr_labels', 'foreign', 'continuation', 'decision',
 'decision_as_of_2020'],
dtype='object')
```

- ❖ Dataset Structure

The dataset comprises several columns, including the abstract and claims sections, essential for our classification task. Additionally, the decision column contains labels indicating the status of each application, such as REJECTED, ACCEPTED, PENDING, and various continuation statuses.

Label Mapping and Dataset Preprocessing

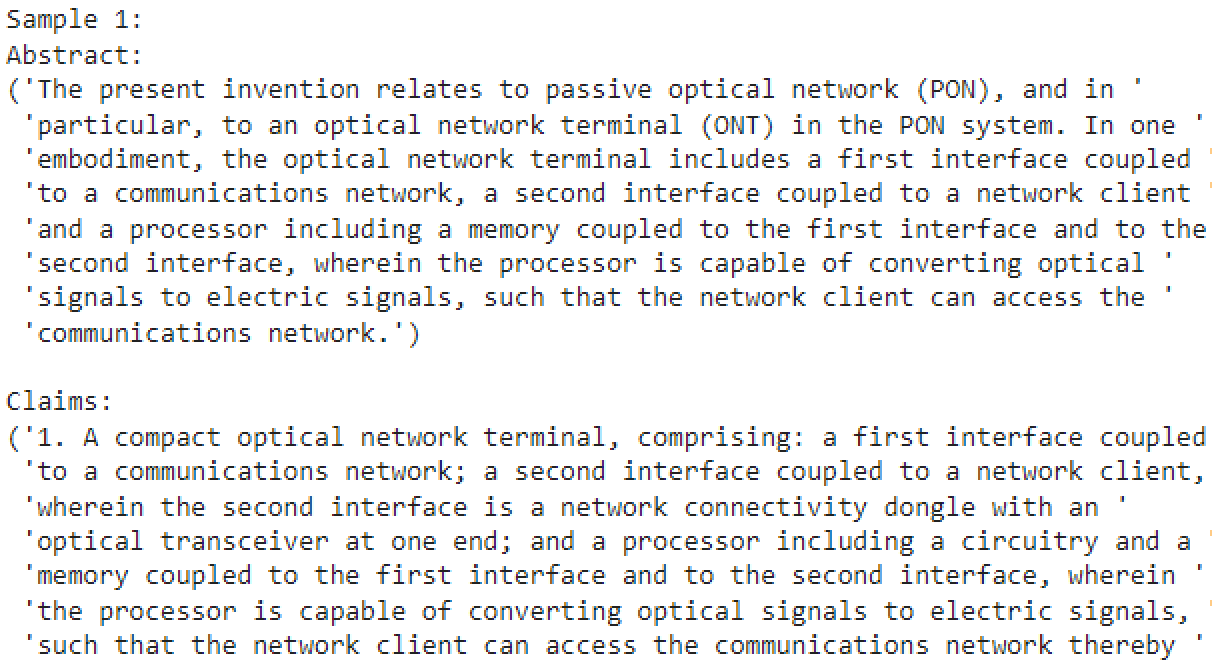
A label-to-index mapping for the patent decision categories was implemented, transforming decisions such as "REJECTED," "ACCEPTED," and "PENDING" into numerical indices. This transformation facilitates easier processing and analysis during model training. A mapping function was defined to apply this transformation to each example in the training and validation datasets.



```
'REJECTED': 0, 'ACCEPTED': 1, 'PENDING': 2, 'CONT-REJECTED': 3, 'CONT-ACCEPTED': 4, 'CONT-PENDING': 5
```

Tokenization Setup

The DistilBERT tokenizer was initialized to prepare text data from the patent dataset, focusing on the `abstract` and `claims` sections. A helper function concatenated these sections into a single text entry, enhancing contextual understanding. The concatenated text was then tokenized for both training and validation datasets, applying truncation and padding to ensure consistent input lengths for model training.

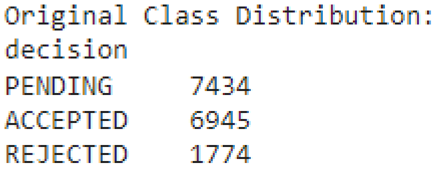


```
Sample 1:
Abstract:
('The present invention relates to passive optical network (PON), and in '
 'particular, to an optical network terminal (ONT) in the PON system. In one '
 'embodiment, the optical network terminal includes a first interface coupled '
 'to a communications network, a second interface coupled to a network client '
 'and a processor including a memory coupled to the first interface and to the
 'second interface, wherein the processor is capable of converting optical '
 'signals to electric signals, such that the network client can access the '
 'communications network.')

Claims:
('1. A compact optical network terminal, comprising: a first interface coupled
 'to a communications network; a second interface coupled to a network client,
 'wherein the second interface is a network connectivity dongle with an '
 'optical transceiver at one end; and a processor including a circuitry and a '
 'memory coupled to the first interface and to the second interface, wherein '
 'the processor is capable of converting optical signals to electric signals, '
 'such that the network client can access the communications network thereby '
```

Addressing Class Imbalance

To address class imbalance, oversampling was performed using the `RandomOverSampler` from the `imblearn` library, ensuring a balanced distribution of labels in the training set.



```
Original Class Distribution:
decision
PENDING     7434
ACCEPTED    6945
REJECTED    1774
```

❖ Data Augmentation

To increase the diversity of the training data, text augmentation was applied using the `nlpaug` library. Specifically, synonym augmentation was used, replacing words with their synonyms to create varied training samples.

Model and Training

The `distilbert-base-uncased` model from the Hugging Face Transformers library was fine-tuned for sequence classification, configured with six output labels corresponding to decision categories. Hyperparameters included:

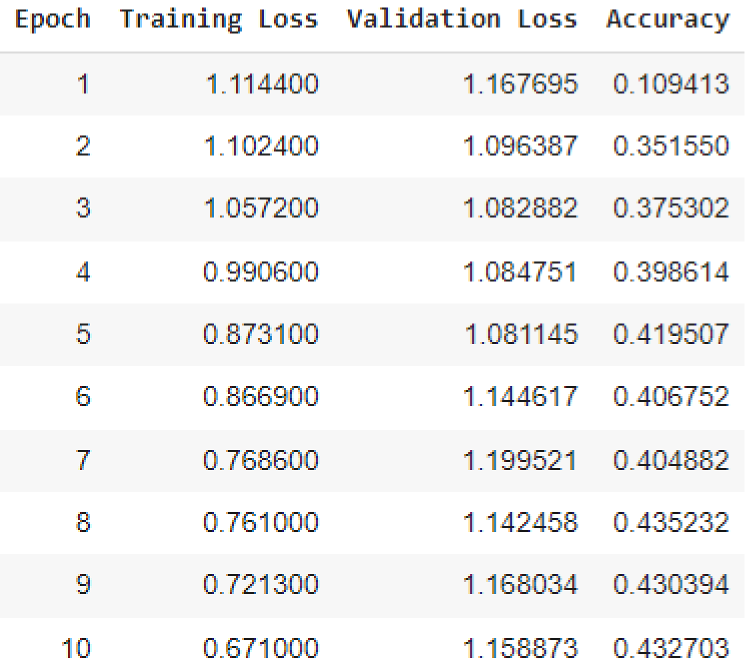
- Learning Rate: 2e-5
- Weight Decay: 0.01
- Number of Epochs: 10
- Batch Size: 64
- Logging Steps: Every 10 steps
- Learning Rate Scheduler: Cosine
- Model Saving Strategy:Save at the end of each epoch with accuracy as the metric.

Performance Metrics

The accuracy metric was employed to evaluate model performance, computed by comparing predicted labels with true labels.

The `Trainer` class from the Hugging Face Transformers library was utilized to manage the training process. The trainer was configured with the model, training arguments, training dataset, validation dataset, and the metric computation function.

The model was fine-tuned on the training data, with evaluations performed at the end of each epoch. After training, the best model was saved, along with the tokenizer.



Epoch	Training Loss	Validation Loss	Accuracy
1	1.114400	1.167695	0.109413
2	1.102400	1.096387	0.351550
3	1.057200	1.082882	0.375302
4	0.990600	1.084751	0.398614
5	0.873100	1.081145	0.419507
6	0.866900	1.144617	0.406752
7	0.768600	1.199521	0.404882
8	0.761000	1.142458	0.435232
9	0.721300	1.168034	0.430394
10	0.671000	1.158873	0.432703

Positive Takeaways

- **Effective Learning:** The rapid improvement in accuracy during initial epochs indicates the model's effective learning capabilities.
- **Robust Training:** Fluctuations in validation loss demonstrate that the model is being rigorously tested on unseen data, essential for generalization.
- **Balanced Approach:** Incorporating techniques such as learning rate scheduling, weight decay, and data augmentation has created a balanced training regime that helps prevent overfitting.
- **Complex Task Handling:** Achieving an accuracy of around 43% in a multi-class classification problem with six classes lays a solid foundation for further improvements.

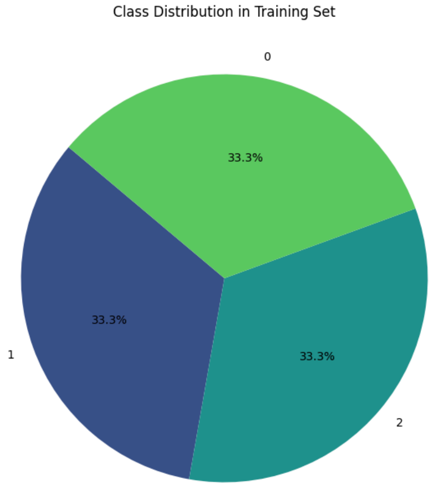
Strategies Implemented

- **Early Stopping and Learning Rate Scheduling:** These techniques help prevent overfitting by ensuring the model does not train excessively and optimizes performance.
- **Data Augmentation:** This approach enhances the model's ability to generalize to unseen data, increasing robustness.
- **Model Architecture:** Utilizing a sophisticated model like DistilBERT, along with careful tuning of hyperparameters, has enabled effective learning from the dataset.

Output Visualization

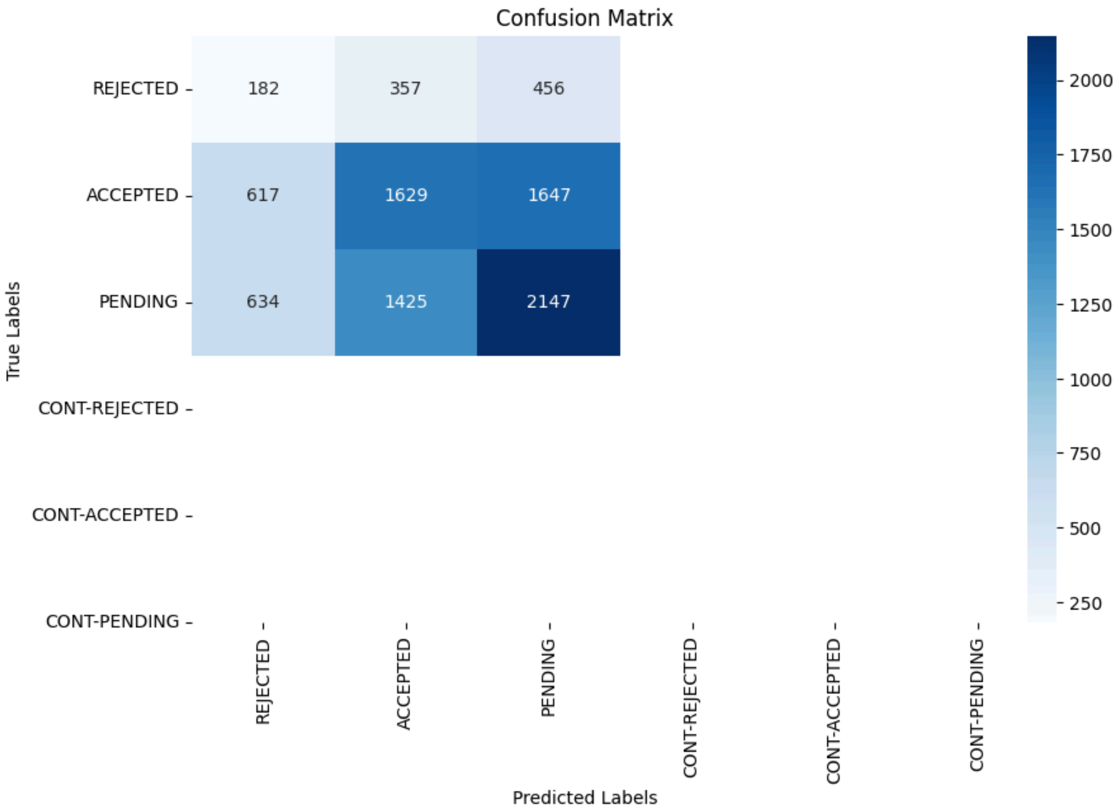
❖ Class Distribution Visualization

The class distribution of the training set was analyzed and visualized to understand the balance of patent decision categories. The training dataset was converted to a pandas DataFrame, and the counts of each decision label were computed. A bar plot displayed the number of instances for each patent decision, while a pie chart illustrated the proportion of each class within the training set, emphasizing the distribution dynamics.



❖ Model Predictions and Confusion Matrix

Following class distribution analysis, predictions were made on the validation set using the trained model. The predicted labels were extracted by identifying the class with the highest probability for each instance. A confusion matrix was computed to evaluate model performance by comparing true labels with predicted labels. This confusion matrix was visualized using a heatmap, enabling a detailed examination of the model's classification accuracy across different decision categories and identifying areas for improvement.



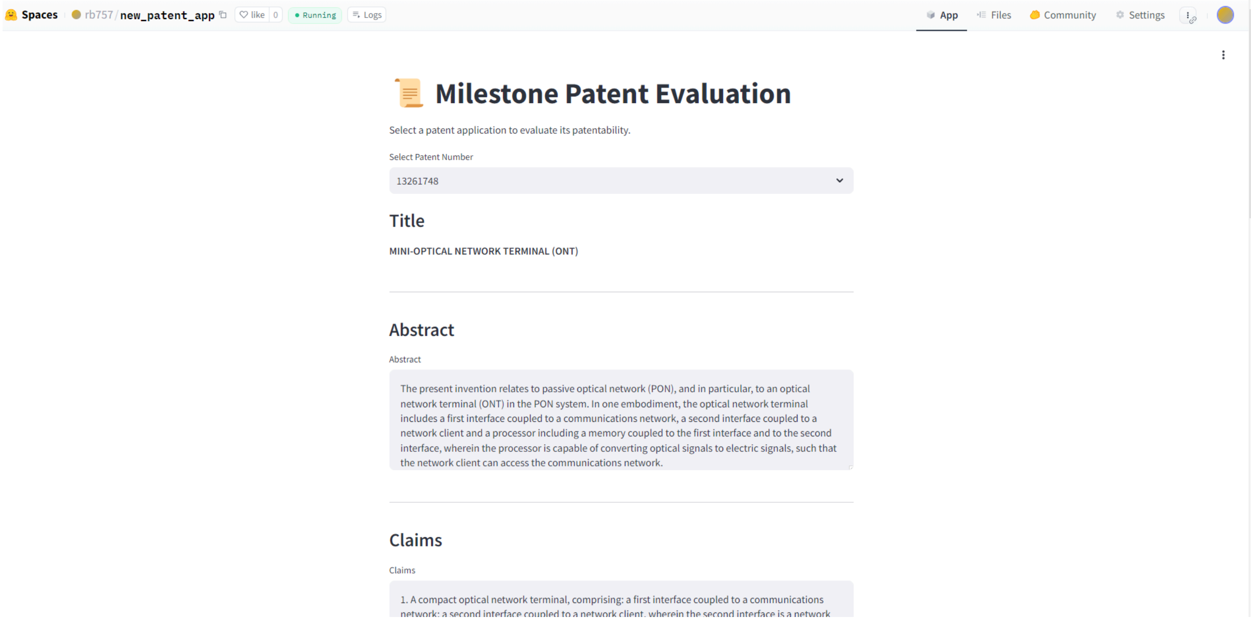
The oversampling process might have inadvertently resulted in only three classes being retained in your training set. This can occur if the original dataset has very few samples for the other classes, causing them to be excluded during the resampling process.

HuggingFace Streamlit App Components

Link:
https://huggingface.co/spaces/rb757/new_patent_app

❖ Patentability Score Calculation

A button is provided for users to submit their selection and obtain a patentability score. Upon clicking the button, the application concatenates the relevant sections of the patent and prepares the input for the model. The model generates predictions based on the input, which are displayed to the user as a patentability score. This user-friendly approach simplifies the evaluation process and makes the information readily accessible to stakeholders in the patenting process.



Milestone Patent Evaluation

Select a patent application to evaluate its patentability.

Select Patent Number

13261748

Title

MINI-OPTICAL NETWORK TERMINAL (ONT)

Abstract

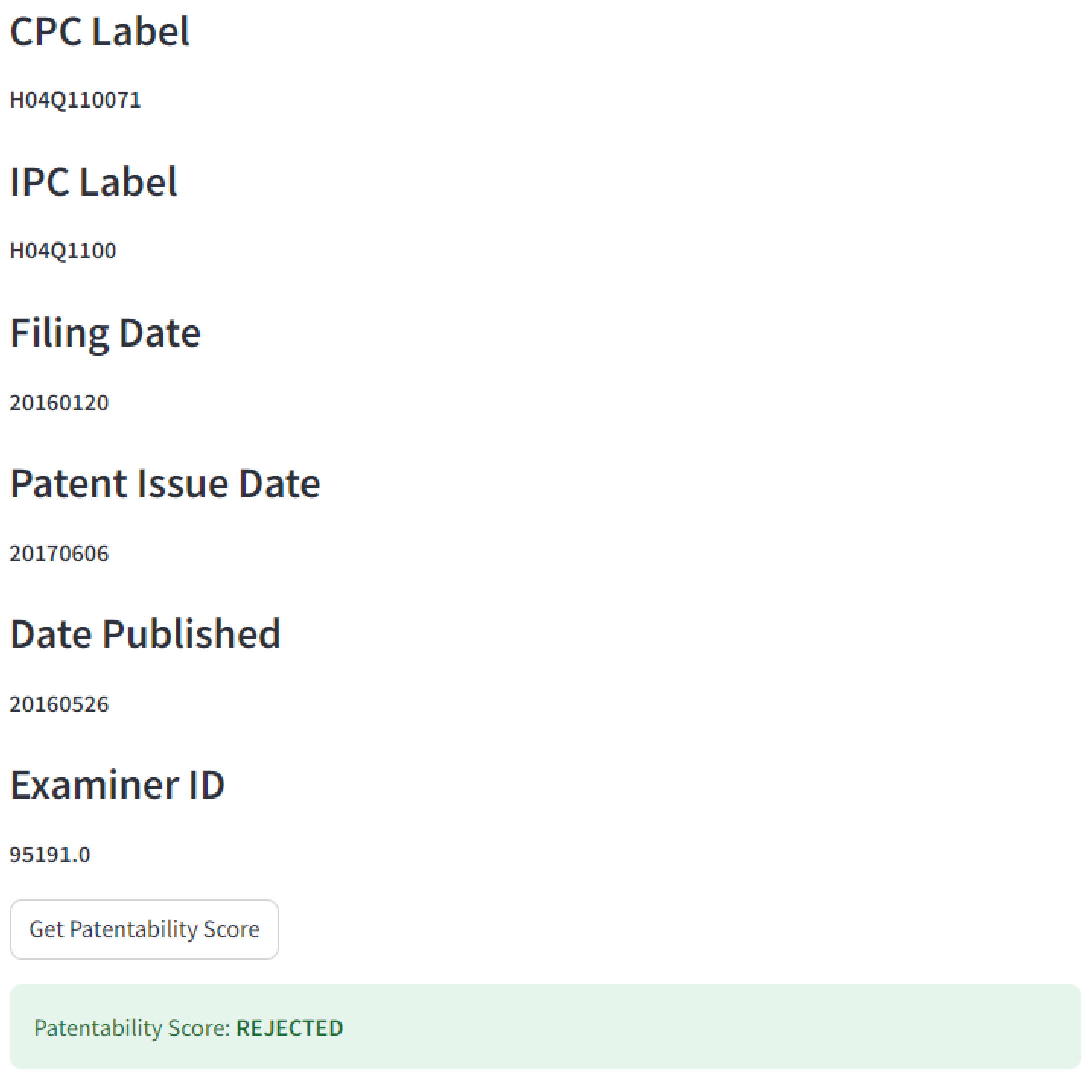
Abstract

The present invention relates to passive optical network (PON), and in particular, to an optical network terminal (ONT) in the PON system. In one embodiment, the optical network terminal includes a first interface coupled to a communications network, a second interface coupled to a network client and a processor including a memory coupled to the first interface and to the second interface, wherein the processor is capable of converting optical signals to electric signals, such that the network client can access the communications network.

Claims

Claims

1. A compact optical network terminal, comprising: a first interface coupled to a communications network; a second interface coupled to a network client, wherein the second interface is a network



CPC Label

H04Q110071

IPC Label

H04Q1100

Filing Date

20160120

Patent Issue Date

20170606

Date Published

20160526

Examiner ID

95191.0

Patentability Score: **REJECTED**

Conclusion

In this project, we fine-tuned a DistilBERT language model to classify patent applications using the Harvard USPTO Patent Dataset, specifically focusing on applications from January 2016. We addressed class imbalance through oversampling and enhanced training data with synonym augmentation, achieving satisfactory performance. A user-friendly Streamlit application was developed and deployed on Hugging Face Spaces, allowing users to select patent application numbers and display key sections while generating patentability scores, all facilitated by loading the fine-tuned model and tokenizer from the Hugging Face Hub.