Pairwise Contrastive Fine-Tuning for Patent Classification

By Mridul Jain and Lynne Wang
July 2025 - NLP Course (MIDS 266), UC Berkeley
https://github.com/jain-mridul/w266_final_project

Abstract

The pairwise contrastive fine-tuning patent classification architecture proposed herein achieves a micro F1 score of 0.81 and an instance-average F1 score of 0.84 at the section level, outperforming state-of-the-art models, including PatentSBERTa (2024), Shajalal et al. (2023), and PatentBERT (2020). The architecture includes two stages. In the first stage, a sentence embedding model (e5-base-v2) is fine-tuned using contrastive learning on balanced positive and negative patent pairs sampled by the CPC section, enhancing semantic separability in the embedding space. In the second stage, the resulting embeddings serve as input to multiple classifiers, where the Mixture of Experts (MoE) ensemble—comprising logistic regression, KNN, and SVM—demonstrates superior classification performance. The methodology is extendable to deeper levels of the CPC taxonomy and offers a generalizable framework for improving hierarchical multi-label classification in patents.

1 Introduction

The objective of this project is to improve hierarchical patent classification by leveraging contrastive embedding fine-tuning to address semantic ambiguity in the CPC taxonomy.

Patents filed with the U.S. and European Patent Offices are classified using the CPC system through a primarily manual process. The CPC system is a deeply hierarchical taxonomy including nine top-level sections (A–H, Y), each covering a broad technological domain.

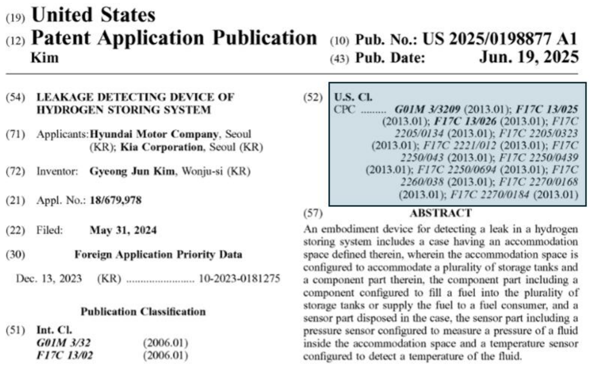
Below is a list of the nine top-level CPC Sections:

- A – Human Necessities (e.g., agriculture, food, health, personal items)
- B – Performing Operations; Transporting (e.g., manufacturing, vehicles, handling materials)
- C – Chemistry; Metallurgy (e.g., inorganic/organic chemistry, metal treatment)
- D – Textiles; Paper (e.g., spinning, weaving, paper production)
- E – Fixed Constructions (e.g., buildings, roads, water supply, mining)
- F – Mechanical Engineering; Lighting; Heating; Weapons (e.g., engines, machines, refrigeration)
- G – Physics (e.g., measuring, optics, computing)
- H – Electricity (e.g., basic electric elements, communication)
- Y – General Tagging of New Technological Developments (e.g., cross-sectional technologies like climate change, smart grids, nanotech)

The nine top-level sections further expand into over 250,000 subgroups through successive levels: classes (~650), subclasses (~1,300), main groups (>25,000), and subgroups. This structure enables detailed technical distinctions but poses major challenges for automation due to its granularity and semantic complexity.

A single patent may span multiple sections and subclasses—for example, US20250198877A1 is classified under both Physics (G01M 3/3209 for leak testing) and Mechanical Engineering (F17C categories for gas vessels and acoustic sensors). Errors at lower levels can cascade through the hierarchy, and the label distribution is highly imbalanced. These factors make CPC classification a demanding task, requiring models with fine-grained semantic understanding and hierarchical reasoning.

The cover page of the example patent (reproduced below) lists its CPC classification (highlighted in blue), and the complete taxonomy tree for this patent is provided in Appendix A.


(19) United States
(12) Patent Application Publication (10) Pub. No.: US 2025/0198877 A1
Kim (43) Pub. Date: Jun. 19, 2025

(54) LEAKAGE DETECTING DEVICE OF HYDROGEN STORING SYSTEM

(71) Applicants: Hyundai Motor Company, Seoul (KR); Kia Corporation, Seoul (KR)

(72) Inventor: Gyeong Jun Kim, Wonju-si (KR)

(21) Appl. No.: 18/679,978

(22) Filed: May 31, 2024

(30) Foreign Application Priority Data

Dec. 13, 2023 (KR) 10-2023-0181275

Publication Classification

(51) Int. Cl.
G01M 3/32 (2006.01)
F17C 13/02 (2006.01)

(52) U.S. Cl.
CPC ***G01M 3/3209*** (2013.01); ***F17C 13/025*** (2013.01); ***F17C 13/026*** (2013.01); *F17C 2205/0134* (2013.01); *F17C 2205/0323* (2013.01); *F17C 2221/012* (2013.01); *F17C 2250/043* (2013.01); *F17C 2250/0439* (2013.01); *F17C 2250/0694* (2013.01); *F17C 2260/038* (2013.01); *F17C 2270/0168* (2013.01); *F17C 2270/0184* (2013.01)

(57) ABSTRACT

An embodiment device for detecting a leak in a hydrogen storing system includes a case having an accommodation space defined therein, wherein the accommodation space is configured to accommodate a plurality of storage tanks and a component part therein, the component part including a component configured to fill a fuel into the plurality of storage tanks or supply the fuel to a fuel consumer, and a sensor part disposed in the case, the sensor part including a pressure sensor configured to measure a pressure of a fluid inside the accommodation space and a temperature sensor configured to detect a temperature of the fluid.

2 Background

Recent efforts in automatic patent classification have explored a range of deep learning and embedding-based techniques.

- PatentBERT (Lee & Hsiang, 2020) fine-tuned BERT-Base on over 3 million U.S. patents using only patent claims. It evaluated both section-level (9 CPC sections) and subclass-level (656 labels) classification. The model achieved an instance-average F1 score of 80.98% at the section level and 66.83% at the subclass level.
- Shajalal et al. (2023) used FastText embeddings and Bi-LSTM/CNN architectures and achieved a micro F1 score of 0.78 (section level) on the BigPatent dataset.
- PatentSBERTa (Bekamiri et al., 2024) fine-tuned a domain-specific Sentence-BERT model using CPC supervision and evaluated it across three classification levels: The model achieved an instance-average F1 score of 0.82 at the section level.

Unlike the approaches described above, we introduce a two-stage architecture that integrates contrastive learning with a Mixture of Experts (MoE) classifier to address semantic ambiguity and class imbalance within the CPC hierarchy. Our model surpasses previously reported state-of-the-art results at the section level, achieving a micro F1 score of 0.81 and an instance-average F1 score of 0.84. With adequate training data, this framework can be effectively extended to class, subclass, and deeper levels of the CPC taxonomy.

3 Dataset and Preprocessing

To support the hierarchical classification of patents into CPC categories, we downloaded about 20K patents from the USPTO XML Bulk Data published in July 2025. The data consisted of structured XML-format patent documents containing both metadata and full-text content. We implemented a parsing pipeline to process each XML document and extract relevant patent metadata and classification fields. Specifically, we extracted the following CPC hierarchy levels:

- Section (e.g., A, B, C, D, E, F, G, H, Y)
- Class (e.g., A61, G06)
- Subclass (e.g., A61B, G06F)
- Main Group (e.g., A61B 5)
- Subgroup (e.g., A61B 5/020)

In parallel, the textual content of each patent was extracted from the following fields: Title, Abstract, and claims. These fields were concatenated into a single input text per patent, following basic preprocessing steps such as whitespace normalization and HTML tag removal. The result of this preprocessing pipeline was saved as a structured CSV file: metadata.csv. A flow diagram illustrating the pipeline is included in Appendix B.

The CPC classification space exhibits a highly imbalanced distribution. Dominant sections such as G (Physics) and H (Electricity) appear in thousands of patents, while rare sections like D (Textiles) or Y (General) occur far less frequently. Similar skewed distributions are observed at the class level within each section.

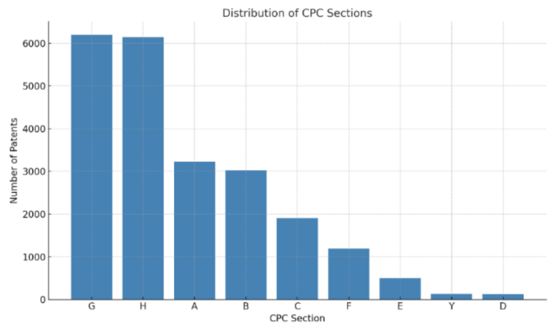


FIG. 1

To address this imbalance, we later apply pairwise sampling strategies that generate balanced positive and negative pairs across sections and classes for contrastive learning. This shift from sample-level to pair-level balancing helps ensure that both common and rare classes are meaningfully represented during model training.

4 Methodology – Contrastive Learning for Embedding Finetuning

A two-stage architecture was implemented for patent classification (illustrated in FIG. 2): (1) contrastive fine-tuning of a SentenceTransformer model (e.g., e5-base-v2), followed by (2) a section-level classification model. This pipeline is designed to enhance semantic separability in the embedding space, address class imbalance, and support scalable multi-label classification across levels of the CPC hierarchy.

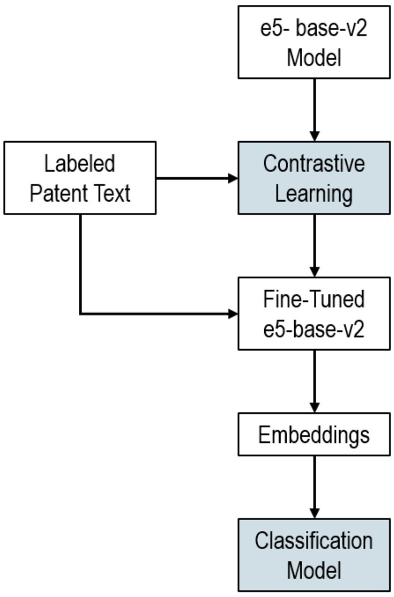


FIG. 2

4.1 Baseline Model

We evaluated several pre-trained embedding models, including all-MiniLM-L6-v2 and e5-base-v2. For each patent, we generated text embeddings using these models and applied standard classifiers—logistic regression, SVM, XGB—for section-level classification. Among the models, e5-base-v2 yielded the best performance. Depending on the classifier used, the best-performing setup reached F1 scores between 0.5 and 0.7.

4.2 Contrastive Learning

As shown in FIG. 3 below, to improve performance, we apply contrastive learning to fine-tune e5-base-v2 embedding model, encouraging the model to cluster semantically similar patents (same section) while pushing dissimilar ones apart.

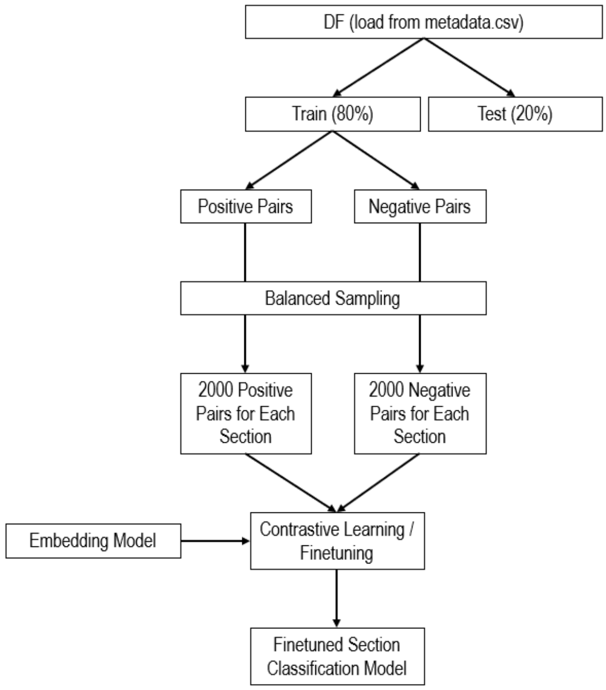


FIG. 3

We employed an 80/20 train-test split to partition the dataset for model training and evaluation. For contrastive learning, we constructed a dataset of patent pairs labeled as either positive or negative based on their CPC sections. All pairs were generated exclusively from the training set to prevent data leakage during evaluation.

Positive pairs were defined as those consisting of patents that shared the same CPC section, while negative pairs were sampled from patents belonging to different sections. This pairwise sampling approach was designed to capture both semantic similarity and dissimilarity across CPC sections.

To address the significant class imbalance inherent in the CPC taxonomy, we implemented a pair balancing strategy by generating 2,000 positive and 2,000 negative pairs for each section and class. This ensured that rare categories were adequately represented during training and prevented the loss function from being dominated by more frequent classes.

Using this balanced dataset, we fine-tuned the e5-base-v2 sentence embedding model with a contrastive loss function based on cosine similarity. The fine-tuned model produced embeddings that were both section-aware and class-sensitive, and served as the input representations for all downstream classification tasks.

4.3 Section Level Classification with Mixture of Experts

Following contrastive finetuning, we perform classification using LSTM, logistic regression, KNN, SVM, ensemble, and a Mixture of Experts (MoE) framework. FIG. 4 illustrates the steps of training the classification model. Both the train and test sets are converted into embeddings using the fine-tuned section embedding model. The train set embeddings are further split into train' and val embedding sets, which are used to train the classification model. The test embedding set is a holdout set, and will only be used to test the trained models.

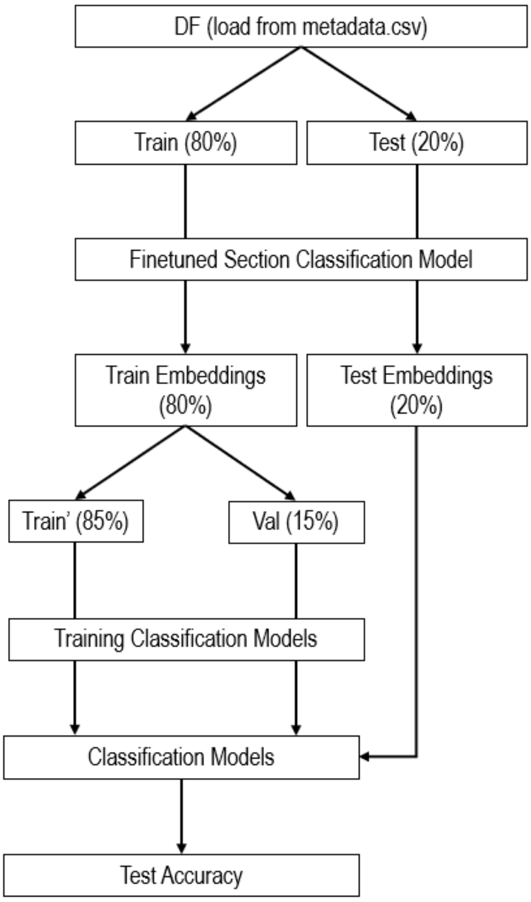


FIG. 4

The best performance is achieved by our Mixture of Experts (MoE) model, which combines logistic regression, K-nearest neighbors (KNN), and support vector machine (SVM) as its expert components. This ensemble approach yields a micro-averaged F1 score of 0.81 and an instance-average F1 score of 0.84, outperforming PatentSBERTa (2024), Shajalal et al. (2023), and PatentBERT (2020) models. Detailed per-class and overall evaluation metrics for MoE are presented in Table 1 below.

4.4 Section Level Classification Based on Taxonomy-Aware Embeddings

We also implemented a taxonomy-aware approach by generating section-level embeddings derived from CPC class descriptions for each section, which were then used to measure semantic similarity between patent embeddings and CPC taxonomy nodes.

Cosine similarity between each patent embedding and each of the section-level taxonomy vectors was computed, resulting in a 9-dimensional feature vector per patent representing its closeness to each CPC section. A One-vs-Rest Logistic Regression classifier was trained on these similarity vectors to predict CPC sections. The classifier achieved a Micro F1 of 0.8 and an instance-average F1 of 0.83, also outperforming PatentSBERTa (2024), Shajalal et al. (2023), and PatentBERT (2020) models. Additional details about the generation of section-level embeddings are included in Appendix C.

5 Results and Discussion

We evaluated our contrastively fine-tuned embedding model and Mixture of Experts (MoE) classifier and taxonomy-aware classifier on the holdout test set. Our models achieved state-of-the-art performance, outperforming existing benchmarks including PatentSBERTa (2024), Shajalal et al. (2023), and PatentBERT (2020).

5.1 Section-Level Performance

As shown in Table 1 below, our MoE ensemble, combining logistic regression, K-nearest neighbors, and SVM, achieved a micro F1 score of 0.81 and an instance-average F1 of 0.84, improving upon prior bests by 2–4 percentage points. The taxonomy-aware classifier using cosine similarity also performed competitively (Micro F1 = 0.80, Instance F1 = 0.83).

Section	Our Model (MoE)	Our Model (Taxonomy Aware)	PatentSBERTa (2024)	Shajalal et al. (2023)
A	**0.87**	**0.85**	N/A	0.85
B	**0.75**	**0.72**	0.76	0.70
C	**0.80**	**0.76**	0.86	0.81
D	**0.37**	**0.16**	0.64	0.73
E	**0.64**	**0.64**	0.74	0.67

Section	Our Model (MoE)	Our Model (Taxonomy Aware)	PatentSBERTa (2024)	Shajalal et al. (2023)
F	**0.72**	**0.68**	0.78	0.70
G	**0.83**	**0.83**	0.85	0.82
H	**0.84**	**0.84**	0.86	0.82
Y	**0.07**	**0.00**	0.56	0.41
Micro Avg.	**0.81**	**0.80**	0.80	0.78
Macro Avg.	**0.65**	**0.61**	0.80	N/A
Instance Avg.	**0.84**	**0.83**	0.82	N/A

Table 1

Notably, performance varied across CPC sections. The model excelled in high-frequency sections such as H (Electricity, F1 = 0.84) and G (Physics, F1 = 0.83), showing strong semantic consistency. However, sections D (Textiles, F1 = 0.37) and Y (General/Interdisciplinary, F1 = 0.07) remained difficult to classify due to limited training data.

To address these limitations, future work will include expanding the training dataset, particularly for underrepresented sections like D and Y. Given that millions of labeled patents are publicly available through sources such as the USPTO and Google Patents, targeted data acquisition is both feasible and scalable. By increasing the number of labeled examples in these minority sections, we aim to reduce class imbalance and improve the model's ability to generalize.

5.2 Comparison to Baseline

Compared to the original e5-base-v2 model, our contrastively fine-tuned e5-base-v2 model, when paired with a Mixture of Experts (MoE) classifier, achieved a significantly higher micro F1 of 0.81 and instance-average F1 of 0.84. This represents an improvement of approximately 14–30% over the baseline, underscoring the impact of domain-specific contrastive finetuning.

5.3 Visualization of Embedding Spaces

FIGS. 5A and 5B are visualizations of the embeddings generated by the original e5-base-v2 and the fine-tuned section classification model. After fine-tuning, the embedding space exhibits significantly improved class separation, illustrating the effectiveness of domain-specific contrastive learning in producing more semantically coherent representations.

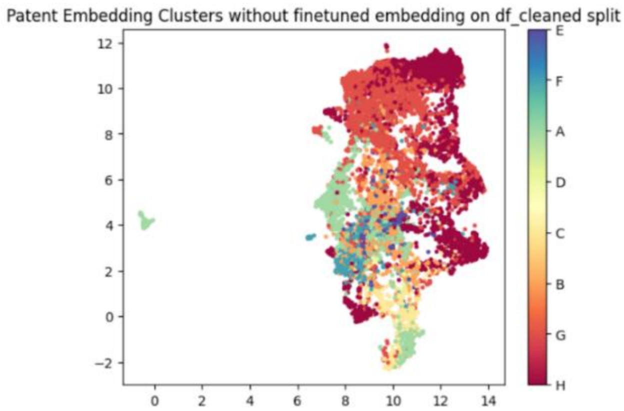


FIG. 5A

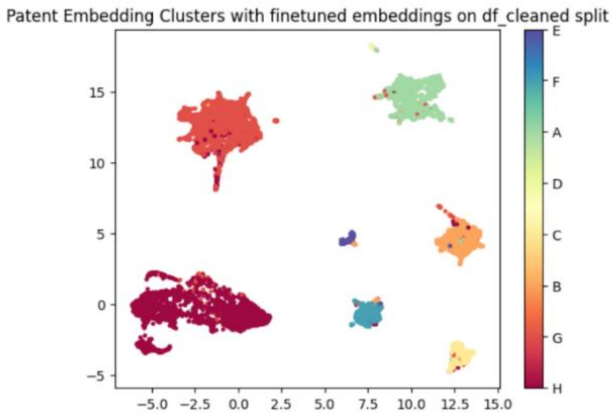


FIG. 5B

6 Conclusion

We presented a contrastive learning-based approach to hierarchical patent classification using fine-tuned sentence embeddings, a Mixture of Experts (MoE) classifier, and a taxonomy-guided model. Our method achieved strong performance at the CPC section level, outperforming PatentSBERTa (2024) and Shajalal et al. (2023), achieving a micro F1 of 0.81 and instance-average F1 of 0.84. While performance on underrepresented sections like Textiles (D) and General (Y) was lower due to limited training data, our findings highlight the value of contrastive embedding finetuning and provide a strong foundation for future improvements with more data.

The same general approach is also applicable to the class, subclass, and deeper levels of the CPC hierarchy.

We also conducted class-level classification using this limited dataset, with details and results provided in Appendix D.

7 Next Steps

While our model performs well at the CPC section and class levels, several extensions could further improve performance. A key next step is expanding training data, especially for underrepresented sections like D (Textiles) and Y (General), where limited data hampers classification. Public patent datasets offer ample opportunity for targeted collection to address this imbalance.

We also aim to extend the classification hierarchy to include finer CPC levels such as subclass, main group, and subgroup. These are essential for detailed patent analysis but pose challenges due to label sparsity. Our contrastive learning pipeline is well-suited to adapt.

Lastly, we plan to integrate taxonomy-aware embeddings by generating CPC group vectors from official descriptions and comparing them with patent embeddings. This may enhance model interpretability and support zero- or few-shot classification in long-tail categories.

References

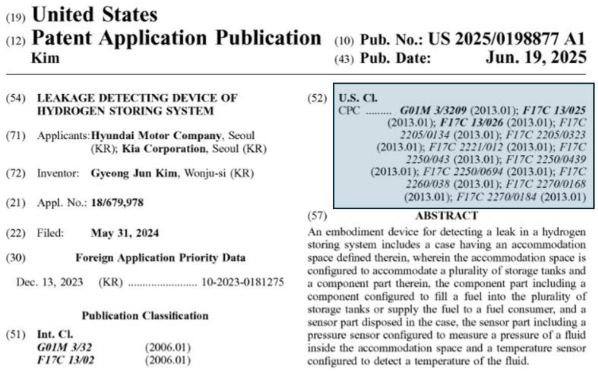
[1] Lee, J.-S., & Hsiang, J. (2020), *Patent classification by fine-tuning BERT language model.* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://arxiv.org/pdf/1906.02124

[2] Shajalal, M., Denef, S., Karim, M. R., Boden, A., & Stevens, G. (2023), *Unveiling Black-boxes: Explainable Deep Learning Models for Patent Classification.* arXiv preprint arXiv:2310.20478. https://arxiv.org/abs/2310.20478

[3] Bekamiri, H., Hain, D. S., & Jurowetzki, R. (2024), *PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT.* Technological Forecasting and Social Change, 206, 123536. https://doi.org/10.1016/j.techfore.2024.123536

[4] Zou, T., Yu, L., Ye, J., Sun, L., Du, B., & Wang, D. (2024), *Adaptive Taxonomy Learning and Historical Patterns Modelling for Patent Classification.* Journal of the ACM (J. ACM), 37(4), Article 111. https://arxiv.org/abs/2308.05385

Appendix A


(19) **United States**
(12) **Patent Application Publication**
Kim

(10) **Pub. No.: US 2025/0198877 A1**
(43) **Pub. Date: Jun. 19, 2025**

(54) **LEAKAGE DETECTING DEVICE OF HYDROGEN STORING SYSTEM**

(71) Applicants: **Hyundai Motor Company**, Seoul (KR); **Kia Corporation**, Seoul (KR)

(72) Inventor: **Gyeong Jun Kim**, Wonju-si (KR)

(21) Appl. No.: **18/679,978**

(22) Filed: **May 31, 2024**

(30) **Foreign Application Priority Data**

Dec. 13, 2023 (KR) 10-2023-0181275

Publication Classification

(51) **Int. Cl.**
G01M 3/32 (2006.01)
F17C 13/02 (2006.01)

(52) **U.S. Cl.**
CPC ***G01M 3/3209*** (2013.01); ***F17C 13/025*** (2013.01); ***F17C 13/026*** (2013.01); *F17C 2205/0134* (2013.01); *F17C 2205/0323* (2013.01); *F17C 2221/012* (2013.01); *F17C 2250/043* (2013.01); *F17C 2250/0439* (2013.01); *F17C 2250/0694* (2013.01); *F17C 2260/038* (2013.01); *F17C 2270/0168* (2013.01); *F17C 2270/0184* (2013.01)

(57) **ABSTRACT**

An embodiment device for detecting a leak in a hydrogen storing system includes a case having an accommodation space defined therein, wherein the accommodation space is configured to accommodate a plurality of storage tanks and a component part therein, the component part including a component configured to fill a fuel into the plurality of storage tanks or supply the fuel to a fuel consumer, and a sensor part disposed in the case, the sensor part including a pressure sensor configured to measure a pressure of a fluid inside the accommodation space and a temperature sensor configured to detect a temperature of the fluid.

G - Physics
└── G01 - Measuring; Testing
  └── G01M - Testing static or dynamic balance of machines or structures; Testing of structures or apparatus, not otherwise provided for
    └── G01M 3/3209 - Leak testing using fluid detection, etc.

F - Mechanical Engineering; Lighting; Heating; Weapons; Blasting
└── F17 - Storing or distributing gases or liquids
  └── F17C - Vessels for storing or distributing compressed, liquefied or solidified gases
    ├── F17C 13/025 - Arrangements for detecting or preventing leakage
    ├── F17C 13/026 - Arrangements for preventing corrosion
    ├── F17C 2205/0134 - Type of vessel: Rigid vessel with outer jacket
    ├── F17C 2205/0323 - Material: Metal only (e.g., aluminum, steel)
    ├── F17C 2221/012 - Insulating means: Vacuum insulation
    ├── F17C 2250/043 - Leak detection using pressure or vacuum change
    ├── F17C 2250/0439 - Leak detection by means of acoustic sensing
    ├── F17C 2250/0694 - Protective devices or arrangements (e.g., relief valves)
    ├── F17C 2260/038 - Use or application: Cryogenic liquefied gases (e.g., LNG, liquid nitrogen)
    ├── F17C 2270/0168 - Features related to maintenance: Monitoring of physical parameters
    └── F17C 2270/0184 - Features related to maintenance: Data processing or control arrangements

Appendix B

FIG. 6 illustrates the preprocessing pipeline for extracting and organizing patent metadata from USPTO XML files. Title, abstract, and claims are parsed into a combined text field, while CPC hierarchy levels (sections, classes, etc.) are extracted and saved in metadata.csv for downstream classification.

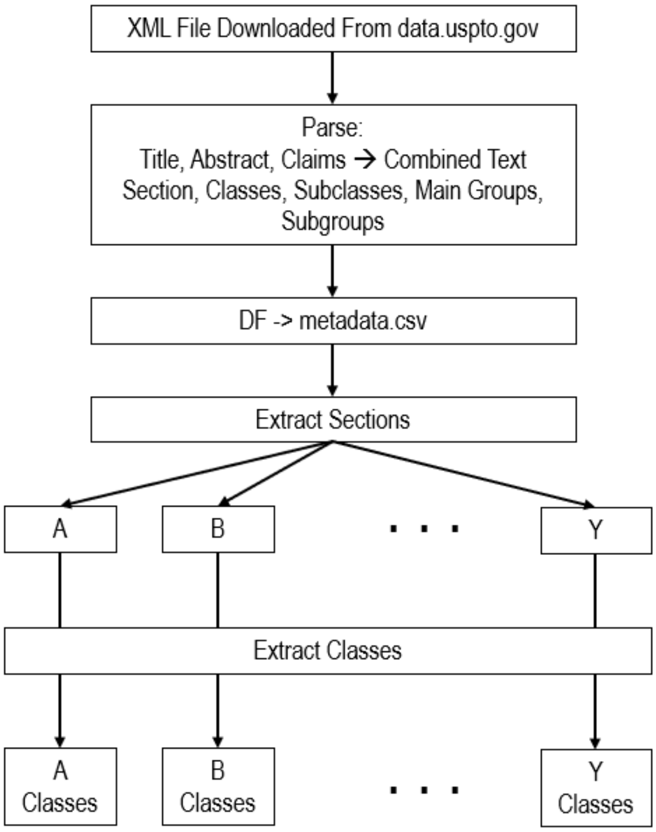


FIG. 6

Appendix C

Below are the CPC class descriptions, which were grouped by section (e.g., A–H, Y) to form aggregated descriptions for each section. These aggregated descriptions were then embedded using the fine-tuned SentenceTransformer model to generate reference vectors representing each CPC section in the taxonomy.

```
cpc_class_list = [
    ["A01", "Agriculture; Forestry; Animal Husbandry; Hunting; Trapping; Fishing"],
    ["A21", "Baking; Equipment for making or processing doughs; Doughs for baking"],
    ["A22", "Butchering; Meat treatment; Processing poultry or fish"],
    ["A23", "Foods or foodstuffs; Their treatment, not covered by other classes"],
    ["A24", "Tobacco; Cigars; Cigarettes; Smokers' requisites"],
    ["A41", "Wearing apparel"],
    ["A42", "Headwear"],
    ["A43", "Footwear"],
    ["A44", "Haberdashery; Jewelry"],
    ["A45", "Hand or travelling articles"],
    ["A46", "Brushware"],
    ["A47", "Furniture; Domestic articles or appliances"],
    ["A61", "Medical or veterinary science; Hygiene"],
    ["A62", "Life-saving; Fire-fighting"],
    ["A63", "Sports; Games; Amusements"],
    ["B01", "General physical or chemical methods or apparatus"],
    ["B02", "Crushing, pulverising, or disintegrating; Preparatory treatment of grain"],
    ["B03", "Separation of solid materials using liquids or using pneumatic tables or jigs"],
    ["B04", "Centrifugal apparatus or machines for carrying-out physical or chemical processes"],
    ["B05", "Spraying or atomising in general; Applying liquids or other fluent materials to surfaces"],
    ["B06", "Generating or transmitting mechanical vibrations in general"],
    ["B07", "Separating solids from solids; Sorting"],
    ["B08", "Cleaning"],
    ["B09", "Waste disposal"],
    ["B21", "Mechanical metal-working without essentially removing material"],
    ["B22", "Casting; Powder metallurgy"],
    ["B23", "Machine tools; Metal-working not otherwise provided for"],
    ["B24", "Grinding; Polishing"],
    ["B25", "Hand tools; Portable power-driven tools; Handles for hand implements"],
    ["B26", "Hand cutting tools; Cutting; Severing"],
    ["B27", "Working or preserving wood or similar material"],
    ["B28", "Working cement, clay, or stone"],
    ["B29", "Working of plastics; Working of substances in a plastic state in general"],
    ["B30", "Presses"],
    ["B31", "Making paper articles or working paper"],
    ["B32", "Layered products"],
    ["B33", "Additive manufacturing technology"],
    ["B41", "Printing; Lining machines; Typewriters"],
    ["B42", "Bookbinding; Albums; Filing appliances"],
    ["B43", "Writing or drawing implements; Bureau accessories"],
    ["B44", "Decorative arts"],
    ["B60", "Vehicles in general"],
    ["B61", "Railways"],
    ["B62", "Land vehicles for travelling otherwise than on rails"],
    ["B63", "Ships or other waterborne vessels"],
    ["B64", "Aircraft; Aviation; Cosmonautics"],
    ["B65", "Conveying; Packing; Storing goods"],
    ["B66", "Hoisting; Lifting; Haulage"],
    ["B67", "Opening; Closing; Emptying; Refilling; Dispensing"],
    ["B68", "Saddlery; Upholstery"],
    ["B81", "Micro-structural technology; Micro-structural devices"],
    ["B82", "Nanotechnology"],
    ["C01", "Inorganic chemistry"],
    ["C02", "Treatment of water, waste water, sewage, or sludge"],
    ["C03", "Glass; Mineral or slag wool"],
    ["C04", "Cements; Concrete; Artificial stone; Ceramics"],
    ["C05", "Fertilizers; Manufacture thereof"],
    ["C06", "Explosives; Matches"],
    ["C07", "Organic chemistry"],
    ["C08", "Organic macromolecular compounds; their preparation or chemical working-up"],
    ["C09", "Dyes; Paints; Polishes; Adhesives; Compositions not otherwise provided for"],
    ["C10", "Petroleum, gas or coke industries; technical gases"],
    ["C11", "Animal or vegetable oils, fats, fatty substances"],
    ["C12", "Biochemistry; Beer; Spirits; Wine; Vinegar; Microbiology; Enzymology"],
    ["C13", "Sugar industry"],
    ["C14", "Skins; Hides; Pelts; Leather"],
    ["C21", "Metallurgy of iron"],
    ["C22", "Metallurgy; Ferrous or non-ferrous alloys; Treatment of alloys or metals"],
    ["C23", "Coating metallic material; Coating material with metallic material; Surface treatment"],
    ["C25", "Electrolytic or electrophoretic processes"],
    ["C30", "Crystal growth"],
    ["C40", "Combinatorial chemistry; Libraries thereof"],
    ["C99", "Subject matter not otherwise provided for in this section"],
    ["D01", "Natural or artificial threads or fibres; Spinning"],
    ["D02", "Yarns; Mechanical finishing of yarns or ropes"],
    ["D03", "Weaving"],
```

```
        ["D04", "Braiding; Lace-making; Knitting; Netting"],
        ["D05", "Sewing; Embroidering; Tufting"],
        ["D06", "Treatment of textiles or the like"],
        ["D07", "Ropes; Cables"],
        ["D10", "Paper-making; Production of cellulose"],
        ["D21", "Paper-making; Production of cellulose"],
        ["E01", "Construction of roads, railways, or bridges"],
        ["E02", "Hydraulic engineering; Foundations; Soil-shifting"],
        ["E03", "Water supply; Sewerage"],
        ["E04", "Building"],
        ["E05", "Locks; Keys; Window or door fittings"],
        ["E06", "Doors, windows, shutters, or roller blinds"],
        ["E21", "Earth or rock drilling; Mining"],
        ["F01", "Machines or engines in general; Engine plants in general"],
        ["F02", "Combustion engines"],
        ["F03", "Machines or engines for liquids"],
        ["F04", "Positive displacement machines for liquids; Pumps"],
        ["F15", "Fluid-pressure actuators; Hydraulic or pneumatic systems"],
        ["F16", "Engineering elements or units"],
        ["F17", "Storing or distributing gases or liquids"],
        ["F21", "Lighting"],
        ["F22", "Steam generation"],
        ["F23", "Combustion apparatus; Combustion processes"],
        ["F24", "Heating; Range; Ventilation"],
        ["F25", "Refrigeration or cooling"],
        ["F26", "Drying"],
        ["F27", "Furnaces; Kilns; Ovens; Retorts"],
        ["F28", "Heat-exchange apparatus"],
        ["F41", "Weapons"],
        ["F42", "Ammunition; Blasting"],
        ["G01", "Measuring; Testing"],
        ["G02", "Optics"],
        ["G03", "Photography; Cinematography; Apparatus or processes"],
        ["G04", "Horology"],
        ["G05", "Controlling; Regulating"],
        ["G06", "Computing; Calculating; Counting"],
        ["G07", "Checking-devices"],
        ["G08", "Signalling"],
        ["G09", "Educating; Cryptography; Display; Advertising; Seals"],
        ["G10", "Musical instruments; Acoustics"],
        ["G11", "Information storage"],
        ["G16", "Information and communication technology specially adapted for specific applications"],
        ["G21", "Nuclear physics; Nuclear engineering"],
        ["H01", "Basic electric elements"],
        ["H02", "Generation, conversion, or distribution of electric power"],
        ["H03", "Basic electronic circuitry"],
        ["H04", "Electric communication technique"],
        ["H05", "Electric techniques not otherwise provided for"],
        ["H99", "Subject matter not otherwise provided for in this section"],
        ["Y02", "Technologies or applications for mitigation or adaptation against climate change"],
        ["Y02A", "Technologies for adaptation to climate change"],
        ["Y02B", "Climate change mitigation technologies related to buildings (e.g., housing, appliances)"],
        ["Y02C", "Capture, storage, sequestration or disposal of greenhouse gases (GHG)"],
        ["Y02D", "Climate change mitigation technologies in ICT (aimed at reducing their own energy use)"],
        ["Y02E", "Reduction of GHG emissions related to energy generation, transmission or distribution"],
        ["Y02P", "Climate change mitigation technologies in the production or processing of goods"],
        ["Y02T", "Climate change mitigation technologies related to transportation"],
        ["Y02W", "Climate change mitigation technologies related to wastewater treatment or waste management"],
        ["Y04", "Information or communication technologies having an impact on other technology areas"],
        ["Y04S", "Systems integrating power network operations, communication, or IT for smart grids"],
        ["Y10", "Technical subjects covered by former USPC cross reference art collections"],
        ["Y10S", "Technical subjects covered by former USPC cross reference art collections (XRACs) and digests"],
        ["Y10T", "Technical subjects covered by former US (USPC) classification (post 2015)"]
    ]
```

Appendix D

D1. Further Fine-Tuning of the Embedding Model for Class-Level Classification

For class level classifications, we applied contrastive learning again to fine-tune the previously fine-tuned section embedding model further. FIG. 7 illustrates the steps we performed to create positive and negative pairs. For each section A-H and Y, we generated positive and negative pairs within the section. Note, our pairs are only within the section, i.e., there is no pair between a class in Section A and a class in Section B. Similar to the pair sampling for the section level pairs, to balance the minority classes, we sampled 400 positive pairs and 400 negative pairs for each class.

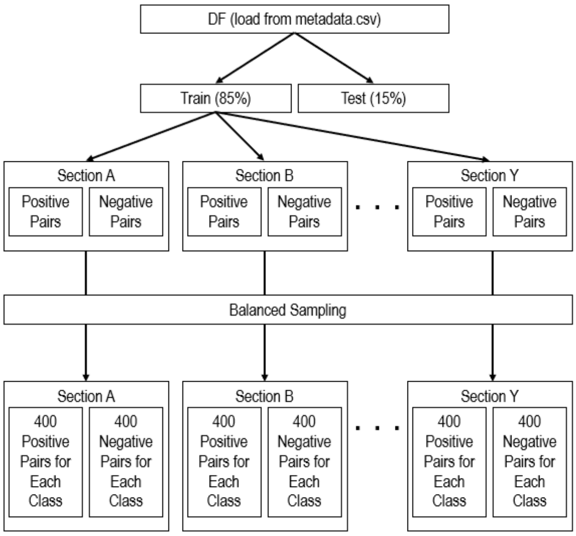


FIG 7

We then perform fine-tuning iteratively based on the pairs generated for each section. As illustrated in Figs. 8A-8C, section A pairs are used to fine-tune the previously fine-tuned section embedding model, the output of which is a fine-tuned section A embedding model. This model is then fine-tuned based on section B pairs, the output of which is a fine-tuned section AB embedding model. This process repeats until all the pairs (including section Y pairs) are used to fine-tune and output a fine-tuned section ABCDEFGH/Y embedding model.

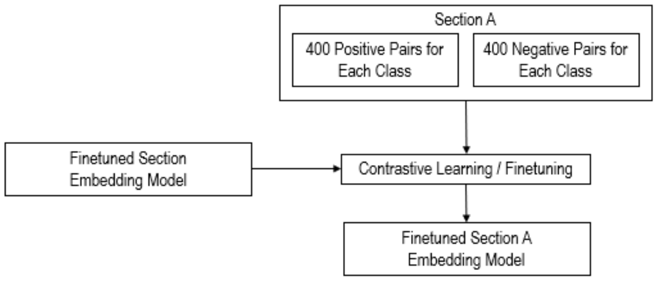


FIG. 8A

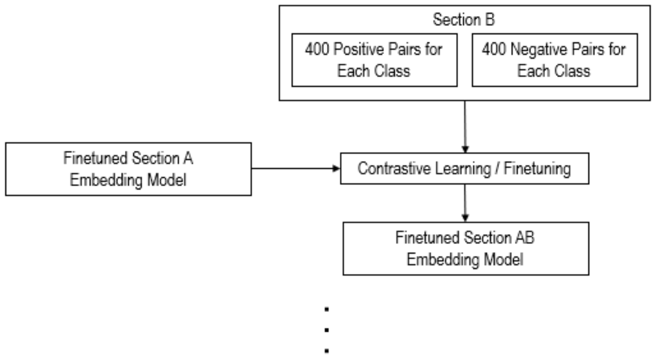


FIG. 8B

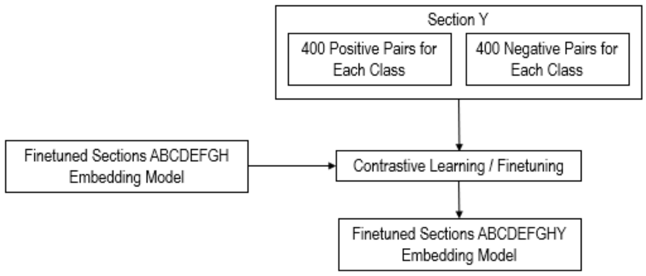


FIG. 8C

As illustrated in FIG. 9 below, the fine-tuned sections ABCDEFGHY model is then applied to both the train and test sets to generate embeddings. For each section, the train embeddings are divided into 85/15 train' and val embedding sets, which are then used to train a classification model for classes within the section. For instance, Section A includes Classes A01, A21–A24, A41–A47, and A61–A63, and a model is trained to classify among these. This process is repeated for Sections B through H and Y.

The dataset we used in this project only contains approximately 18k. Given the large number of classes, many have only a few examples, resulting in limited training data per class. This scarcity significantly impacts the performance of class-level classification.

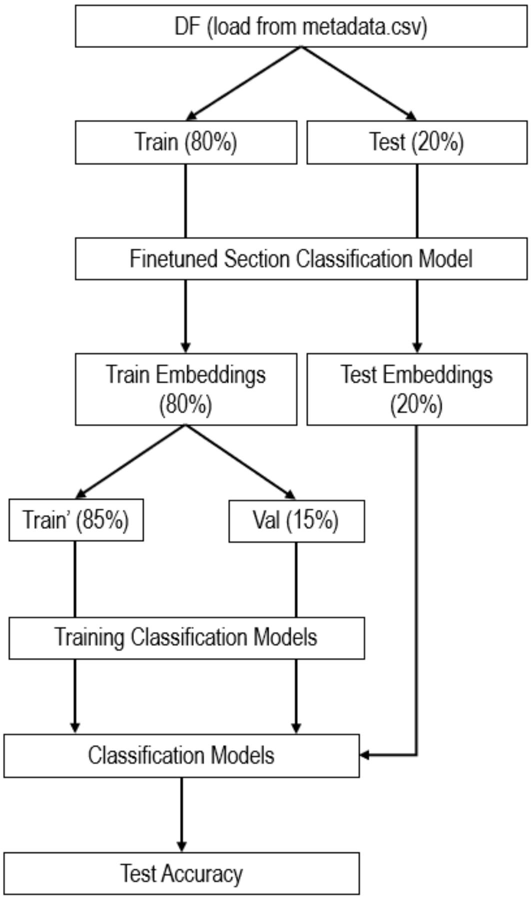


FIG. 9

D2. Class-Level Results

Table 2 below shows the metrics of our class-level models. The results demonstrate that our methodology is effective for classes with sufficient training samples. For example, consistently strong performance was observed in high-frequency classes such as A61, G06, and H04. However, many classes had very limited support—sometimes fewer than 10 examples—which significantly impacted performance in those areas. This limitation stems not from flaws in the model design, but from data sparsity across the long tail of the CPC hierarchy. Encouragingly, this experiment serves as a proof of concept: when enough labeled examples are available, the model performs well. Obtaining additional data is highly feasible—over 10,000 new patents are published daily by the USPTO, and millions of labeled examples are accessible through publicly available datasets. With broader data coverage, especially for underrepresented classes, this approach has strong potential to scale effectively across the full CPC taxonomy.

Section	Class	Precision	Recall	F1-Score	Support
A	A01	0.65	0.42	0.51	36
A	A21	0	0	0	1
A	A22	0	0	0	0
A	A23	0	0	0	8
A	A24	1	0.33	0.5	3
A	A41	0	0	0	3
A	A42	0	0	0	1
A	A43	0	0	0	3
A	A44	0	0	0	3
A	A45	0	0	0	9
A	A46	0	0	0	0
A	A47	0.75	0.35	0.47	26
A	A61	0.91	0.91	0.91	223
A	A62	0	0	0	2
A	A63	0.5	0.44	0.47	18
B	B01	0.78	0.39	0.52	36
B	B60	0.73	0.75	0.74	81
C	C01	0.33	0.13	0.19	15
C	C07	0.85	0.62	0.71	65
D	D01	1	0.33	0.5	6
D	D05	1	1	1	1
E	E02	1	0.17	0.29	6
E	E04	0.8	0.5	0.62	16
F	F02	0.8	0.47	0.59	17
F	F03	1	0.58	0.74	12
G	G01	0.6	0.47	0.53	121
G	G06	0.83	0.79	0.81	346
H	H01	0.78	0.7	0.74	194
H	H04	0.95	0.93	0.94	285
Y	Y02	0.69	1	0.82	9
Y	Y10	1	0.33	0.5	9

Table 2