License SPDX Classification

Gisele Aydi Saja Moussa

Supervised by: Dr. Montassar ben Messaoud

Tunis Business School - Univeristy of Tunis

Introduction: Motivation & Context

Definition

Open-source software (OSS) denotes software distributed with licenses that explicitly define usage, modification, and redistribution rights.

Context

- Modern software systems rely on large ***OSS dependency graphs***.
- Each dependency introduces legal constraints via its license.
- Automated license identification is required for:
  - ***Legal compliance***
  - ***Software supply chain security***
  - ***CI/CD governance***

eg, a single application may transitively depend on hundreds of licensed components.

Introduction: Problem Statement

Problem

- Manual license inspection is non-scalable and error-prone.
- License texts frequently appear with formatting variations or minor modifications.
- Canonical licenses typically have ***one reference text per family***.

Limitations of Existing Approaches

- Pattern-based tools rely on handcrafted rules and lack adaptability.
- Supervised classifiers assume sufficient labeled data per class.
- Explainability is often limited or absent.

Core Constraint

Extreme ***label sparsity*** transforms license identification into a ***one-shot learning problem***.

Introduction: Research Questions

Research Scope

This study evaluates computational paradigms for ***canonical OSS license identification*** under sparse supervision.

Primary Research Questions

1. Can ***supervised machine learning classifiers*** trained on semantic embeddings generalize under one-shot conditions?
2. Do ***similarity-based NLP methods*** provide superior performance for license family detection?

Evaluation Criteria

- License family accuracy
- Robustness to sparse labels
- Explainability of predictions

Choice of Dataset & Constraints: Dataset Selection

Objective: This study requires a dataset that provides ***canonical license texts*** with unambiguous legal semantics and standardized identifiers.

Evaluation Basis: *Coverage, structure, and suitability for controlled experimentation.*

Selected Dataset

The ***SPDX License List*** is an industry-maintained standard containing canonical reference texts and identifiers for open-source software licenses.

- 680 distinct licenses
- Standardized ***SPDX identifiers***
- Canonical, human-reviewed license texts

=> **Reproducible evaluation of license identification methods.**

Choice of Dataset & Constraints: Dataset Constraints

Selected Dataset: suitable for controlled benchmarking, introduces inherent constraints.

Structural Constraints

- Each license family is represented by ***one canonical reference text***.
- Intra-family variation is effectively absent.
- Modified or embedded license texts are not represented.

These constraints transform the task into a ***one-shot identification problem***, where traditional supervised learning assumptions do not hold.

Data Cleaning & Pre-processing

Objective: Remove formatting noise, preserve legally relevant semantics and ensure consistency across license texts.

Scope: *Uniform preprocessing* shared across all methods, and method-specific transformations, *eg, vectorization, chunking, embeddings*

Raw Data Representation: Design Principle

Only fields required for semantic license identification are retained:

- ***SPDX identifier***
- ***License name***
- ***Canonical license text***

Data Cleaning & Pre-processing: Text Normalization

Text Normalization

- Lowercasing
- Line break normalization
- Whitespace collapsing

Preservation Constraint

No stopword removal, *eg, shall, may.*

The resulting corpus is legally intact and compatible with all downstream methods.

Implemented Methods: Overview

Evaluate complementary computational paradigms for ***canonical OSS license identification***.

Implemented methods categories:

- ***Similarity-based retrieval*** methods
- ***Supervised machine learning*** classifiers

Design Goal

Assess the suitability of each paradigm under ***extreme label sparsity*** and ***one-shot identification*** constraints.

Method	**Paradigm**	**Training**	**Inference Type**
TF-IDF/SBERT Similarity	Similarity-based	None	Nearest Neighbor
Hybrid TF-IDF + SBERT	Similarity-based	Optional	Two-stage Retrieval
Supervised Classifiers	Classification	Required	Direct Prediction
ScanCode	Pattern-based	Manual	Rule Matching

Implemented Methods: Similarity-Based Retrieval – TF-IDF

Term Frequency–Inverse Document Frequency (TF-IDF) represents documents as weighted vectors reflecting term importance within a corpus.

Method

- Character-level n-grams ($n = 3$–5) are used to handle formatting variations.
- License texts are embedded into a high-dimensional sparse vector space.
- Similarity is computed using ***cosine similarity***.

Inference

- Perform nearest-neighbor search against reference licenses.
- Top-1 match determines the predicted ***license family***.

Implemented Methods: Similarity-Based Retrieval – SBERT

Sentence-BERT (SBERT) encodes license texts into dense semantic embeddings for similarity-based retrieval.

Method

- Paragraph-level chunking to respect context limits
- Independent encoding via transformer encoder
- ***Max pooling*** aggregation to preserve decisive clauses

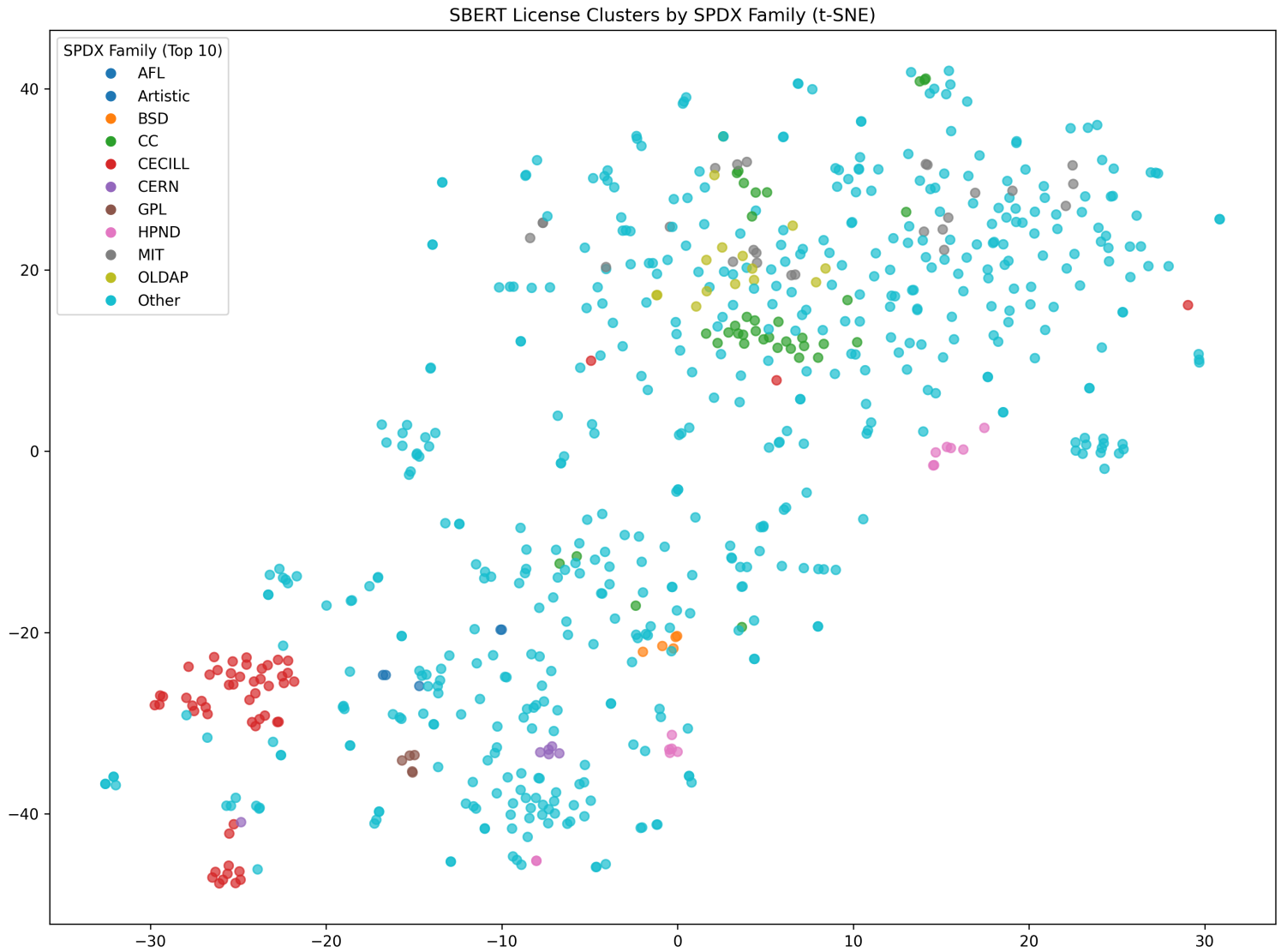
Fine-Tuning

- Contrastive learning using license family labels
- Improves intra-family similarity

Practical Constraint

One canonical text per license limits fine-tuning gains.

Implemented Methods: Similarity-Based Retrieval – SBERT



Implemented Methods: Similarity-Based Retrieval – Fine-Tuned SBERT

Implemented Methods: Hybrid TF-IDF + SBERT

The hybrid approach combines ***lexical filtering*** and ***semantic reranking*** in a two-stage retrieval pipeline.

Method

1. **Stage 1 — TF-IDF Pruning:** retrieve top-K candidates using lexical similarity.
2. **Stage 2 — SBERT Reranking:** compute semantic similarity on the reduced candidate set.

The final prediction corresponds to the highest-ranked license after SBERT reranking.

Implemented Methods: Hybrid TF-IDF + SBERT

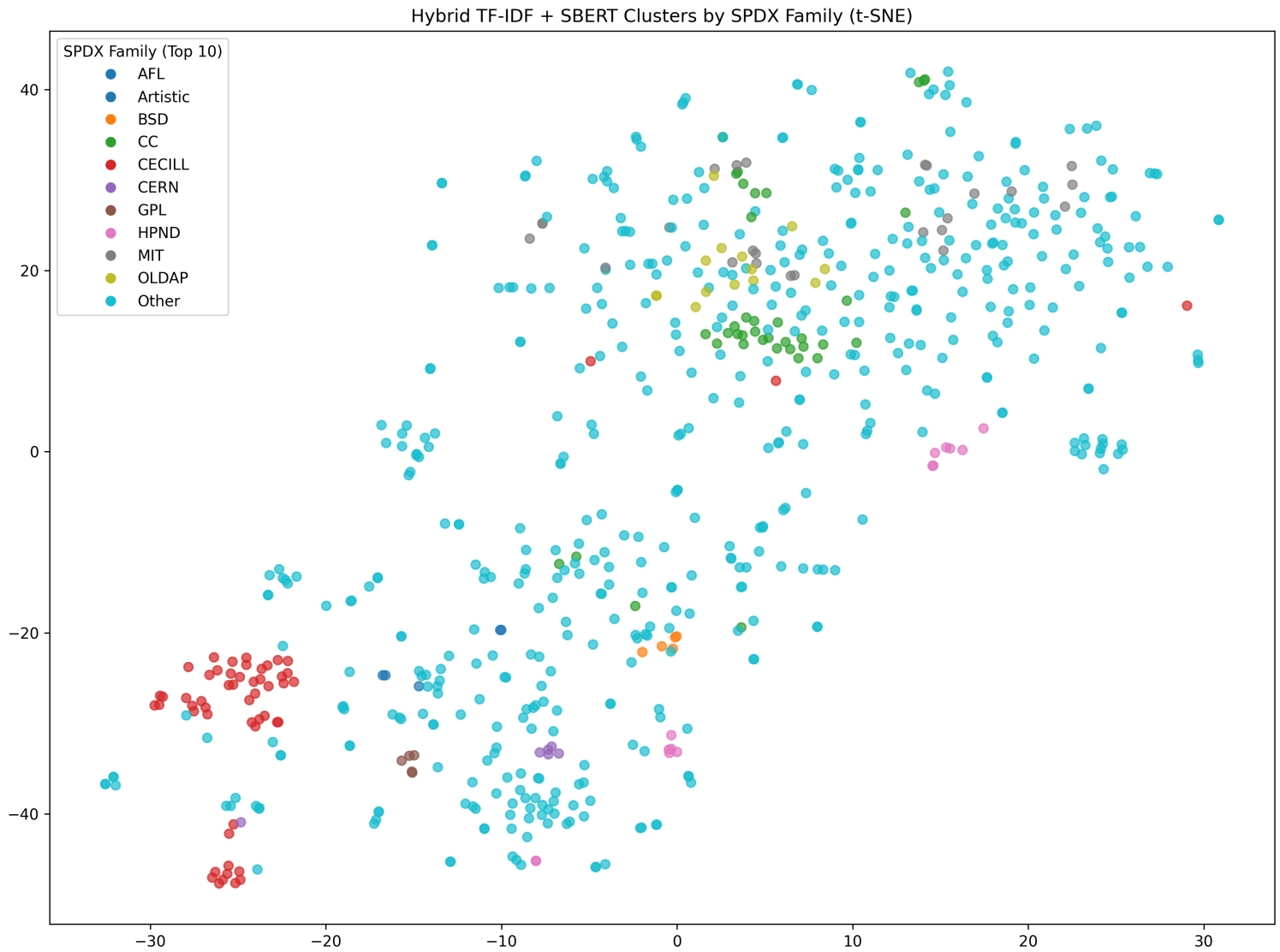


Figure 3: 2D t-SNE projection of the Hybrid method's reranked vectors

Implemented Methods: Supervised Classification Methods

In contrast to similarity-based retrieval, supervised learning formulates license identification as a ***multi-class classification*** problem.

Formulation

- Input: dense ***SBERT embeddings***
- Output: predicted ***license family***

All classifiers are trained using identical embeddings and evaluated under the same data split.

Key Assumption

Supervised classifiers assume sufficient labeled examples per class to learn generalizable decision boundaries.

Implemented Methods: Supervised Classification Methods

Models Evaluated

- ***Support Vector Machine*** — maximum-margin classifier
- ***Random Forest*** — ensemble of non-linear decision trees
- ***Logistic Regression*** — linear probabilistic baseline

Experimental Results & Benchmarking: Evaluation Setup

Evaluation Protocol

- Dataset split: 80% training / 20% testing
- Evaluation at ***license family level***
- One-shot setting: at most one canonical reference per family

Metrics

- Accuracy
- Detection rate
- Precision, recall, F1-score

Experimental Results & Benchmarking: Results – Similarity-Based Methods

Method	Family Accuracy	Detection Rate
TF-IDF Similarity	44.3%	100%
SBERT Similarity	77.2%	100%
SBERT (Fine-Tuned)	77.2%	100%
Hybrid TF-IDF + SBERT	**91.9%**	100%

Semantic similarity substantially outperforms lexical matching, while the hybrid method achieves the highest accuracy.

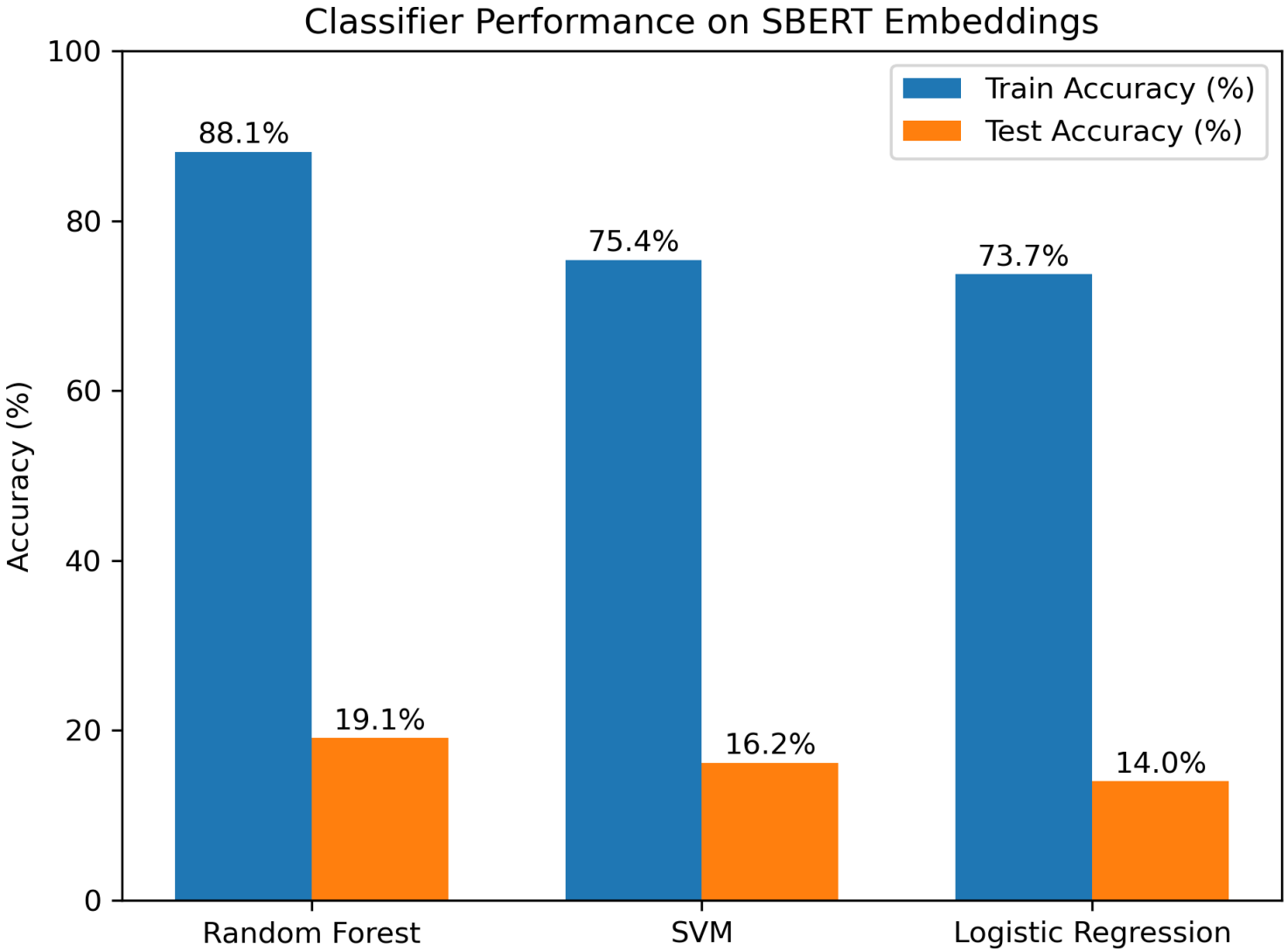
Experimental Results & Benchmarking: Results – Supervised Classification

Classifier	Train Acc.	Test Acc.	Precision	Recall
Random Forest	89.3%	22.8%	9.6%	9.7%
SVM	75.2%	16.2%	–	–
Logistic Regression	73.5%	14.0%	–	–

Observation

All supervised classifiers exhibit severe overfitting and near-random generalization performance.

Experimental Results & Benchmarking: Results – Supervised Classification



Experimental Results & Benchmarking: Results – Root Cause Analysis

Observed Failure Mode Supervised classifiers achieve high training accuracy but fail to generalize to unseen licenses.

Root Cause

Extreme ***label sparsity***: most license families are represented by a single canonical text.

Under this condition:

- Decision boundaries cannot be learned
- Models memorize individual embeddings
- Generalization is structurally impossible

Experimental Results & Benchmarking: ScanCode

Method	**Family Accuracy**	**Detection Rate**
ScanCode	94.9%	99.3%
Hybrid TF-IDF + SBERT	91.9%	100%

ScanCode achieves the highest accuracy through curated rule-based matching, while the hybrid method offers competitive performance with greater adaptability.

Experimental Results & Benchmarking: ScanCode

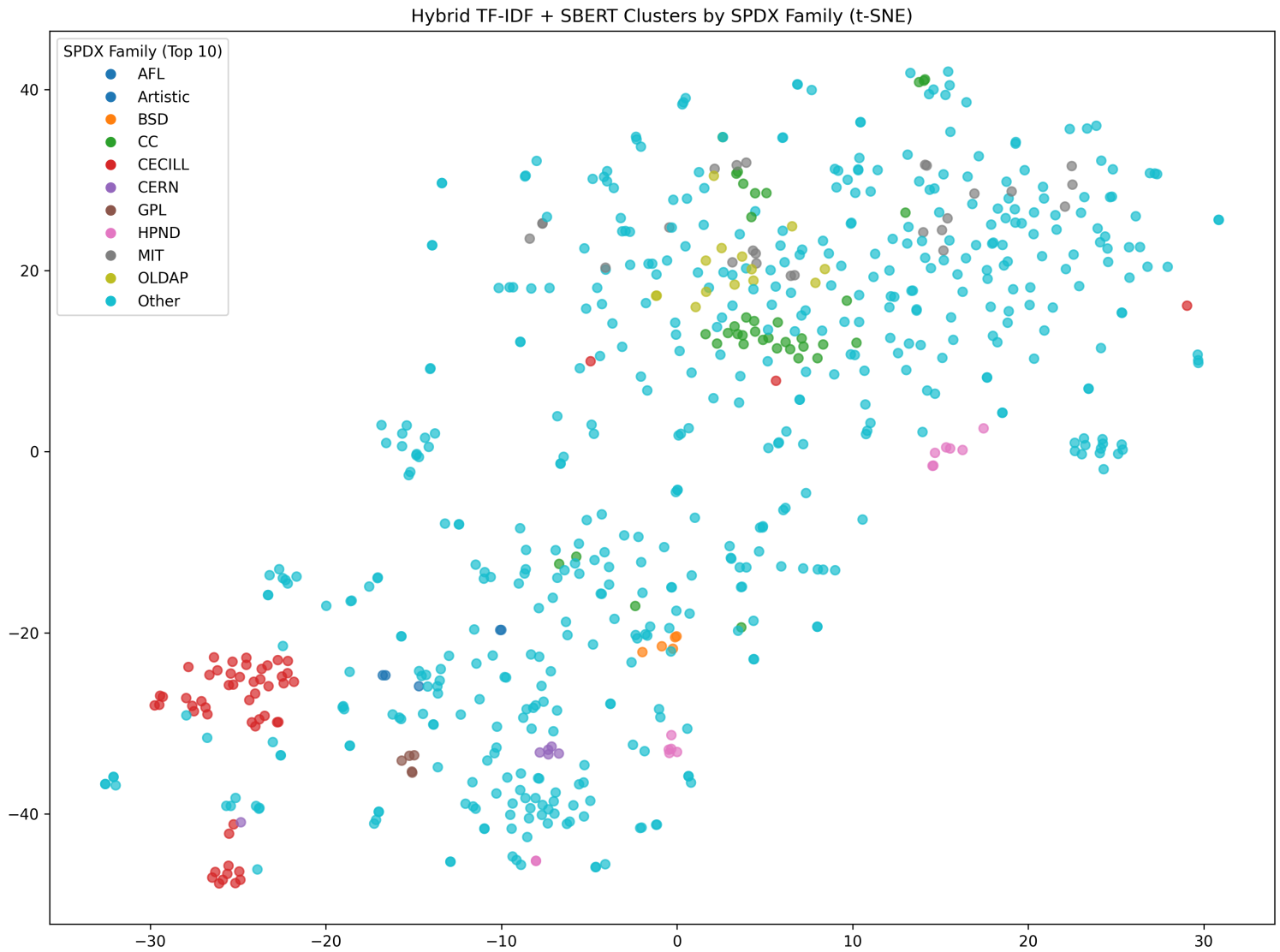
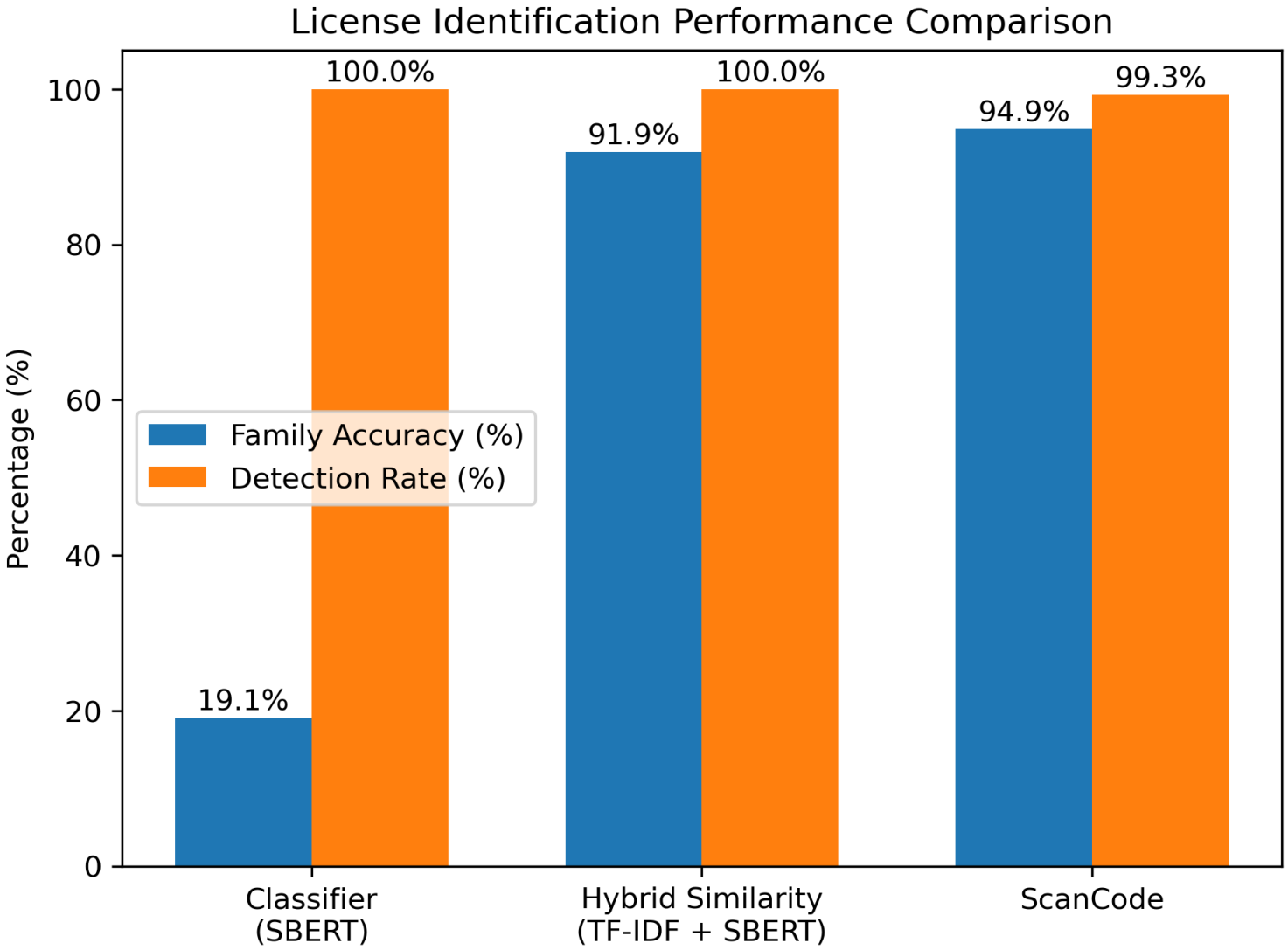


Figure 5: Bar plot comparing the hybrid similarity method to Scancode.

Experimental Results & Benchmarking: Overall Comparison

Method	**Accuracy**	**Explainability**	**Adaptability**
TF-IDF Similarity	Low	High	High
SBERT Similarity	Medium	Medium	High
Hybrid TF-IDF + SBERT	High	High	High
Supervised Classifiers	Very Low	–	Low
ScanCode	Very High	Low	Low

Experimental Results & Benchmarking: Overall Comparison



Key Analysis & Explainability: Paradigm Comparison

Similarity-Based Retrieval

- Distance-based matching to reference licenses
- Naturally supports ***one-shot identification***

Supervised Classification

- Learns global decision boundaries
- Requires multiple samples per class

Key Insight

Canonical license identification is a ***retrieval problem***, not a classification problem.

Key Analysis & Explainability: Similarity-Based Methods

Similarity-based methods provide ***intrinsic explainability*** through distance-based reasoning. **Explainability Signals**

- Similarity scores between query and reference licenses
- Ranked nearest neighbors
- Explicit comparison to known canonical texts

These signals allow practitioners to:

- Inspect why a license was matched
- Identify borderline or ambiguous cases
- Support human-in-the-loop review

Key Analysis & Explainability: Hybrid Method Explainability

- License identification under canonical data is inherently ***retrieval-based***.
- Similarity-based NLP methods outperform supervised classifiers under label sparsity.
- Hybrid TF-IDF + SBERT achieves strong accuracy with interpretable decisions.

Key Analysis & Explainability: Limitations

Despite their advantages, similarity-based methods have inherent interpretability limits.

Limitations

- Embedding dimensions lack direct human semantics
- Close similarity does not imply legal equivalence
- Permissive licenses may remain hard to distinguish

As a result, similarity scores should be interpreted as ***decision support***, not legal proof.

Advantages and Limitations of Implemented Methods
Key Takeaways: Advantages

Method	**Accuracy**	**Explainability**	**Adaptability**
TF-IDF Similarity	Low	High	High
SBERT Similarity	Medium	Medium	High
Hybrid TF-IDF + SBERT	High	High	High
Supervised Classifiers	Low	Low	Low
ScanCode	Very High	Low	Low

Similarity-based methods provide strong adaptability and transparent decision support, while the hybrid approach achieves the best balance between accuracy and interpretability.

Advantages and Limitations of Implemented Methods
Key Takeaways: Limitations

Methodological Limitations

- Canonical dataset lacks real-world license variations.
- Similarity does not guarantee legal equivalence.
- Semantic embeddings require non-trivial computational resources.

Model-Specific Constraints

- TF-IDF fails to capture deep legal semantics.
- SBERT may confuse closely related permissive licenses.
- Supervised classifiers are unsuitable under one-shot conditions.

Advantages and Limitations of Implemented Methods Key Takeaways: Key Takeaways

- License identification from canonical texts is fundamentally a ***retrieval problem***.
- Similarity-based NLP methods outperform supervised classifiers under extreme label sparsity.
- The hybrid TF-IDF + SBERT approach offers near–state-of-the-art accuracy with strong explainability.
- Pattern-based tools remain effective but lack adaptability and transparency.

These findings motivate retrieval-centric architectures for legal and compliance-oriented NLP tasks.

Conclusion and Future Directions

- This work demonstrates that similarity-based retrieval is the appropriate paradigm for ***one-shot open-source license identification***.
- The hybrid TF-IDF + SBERT approach balances accuracy, efficiency, and explainability, while supervised classifiers fail structurally due to data sparsity.

Thank You

Thank you!