Scientific Paper Categorizer: Methodology, Experiments, Results & Analysis

---

Introduction

This project presents a robust NLP-based system for automatic classification of scientific paper abstracts into multiple scientific domains. The system leverages both traditional (TF-IDF) and modern (SBERT) feature extraction methods, and supports multiple machine learning models. It is designed for high accuracy, interpretability, and production-readiness.

Methodology

Data Collection & Preprocessing

- **Data Source:** Abstracts are fetched from the arXiv API, covering categories such as Computer Science, Physics, Biology, Mathematics, etc.
- **Preprocessing**:
  - – Lowercasing, punctuation and digit removal
  - – Stopword removal (NLTK)
  - – Lemmatization (spaCy)
  - – Augmentation for underrepresented classes (synonym replacement)
- **Features Extraction :**
  - **TF-IDF:**
    - Bag-of-words with ngrams (1-2)
    - Max features: 10,000
    - Dimensionality reduction (SVD) for SVM models
  - **SBERT:**
    - Sentence-BERT embeddings (all-MiniLM-L6-v2)
    - 384-dimensional dense vectors
- **Model Architectures :**
  - **TF-IDF Models:**
    - Logistic Regression (OneVsRest)
    - Linear SVM (OneVsRest, with SVD + StandardScaler)
  - **SBERT Models:**
    - Logistic Regression (OneVsRest)
    - Multi-layer Perceptron (OneVsRest, 256 hidden units)

Training & Evaluation

- **Multi-label Binarization:** For multi-label classification
- **Cross-validation:** 5-fold for robust evaluation
- **Metrics:**
  - Hamming loss (lower is better)
  - Exact Match Accuracy
  - Macro / Micro Precision, Recall, F1
  - Per-class metrics

Experiments

Experimental Setup

- **Training Data:** Balanced across 9+ arXiv categories, with augmentation for minority classes.
- **Feature Methods:** Both TF-IDF and SBERT pipelines evaluated.
- **Model Selection:** Logistic Regression, SVM and MLP compared.
- **Evaluation:** Performed on held-out test set and via cross-validation.

Experimental Pipeline

- **Data Loading:** Fetch and preprocess abstracts.
- **Feature Extraction:** Generate TF-IDF or SBERT features.
- **Model Training:** Train selected model(s) with cross-validation.
- **Evaluation:** Compute all metrics and generate confusion metrices.
- **Explainability:** Generate SHAP explanations for prediction

---

Results

Experimental Setup

Model	Exact Match Accuracy	Macro F1	Hamming Loss
SBERT MLP	93.48%	95.61%	0.0096
TF-IDF SVM	82.38%	91.41%	-
TF-IDF Logistic Reg.	79.15%	89.92%	-

Category Highlights:

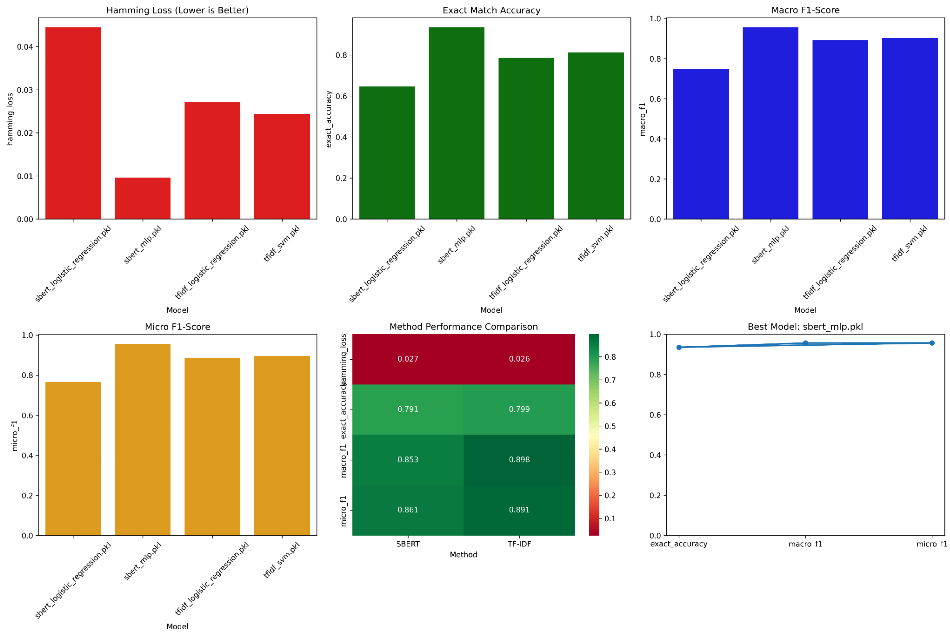
- **Signal Processing (eess.SP)**: 99.26% F1
- **Biology (q-bio.BM):** 97.81% F1
- **Physics (physics.gen-ph)**: 97.32% F1
- **Mathematics (math.ST)**: 95.11% F1

Model Comparison

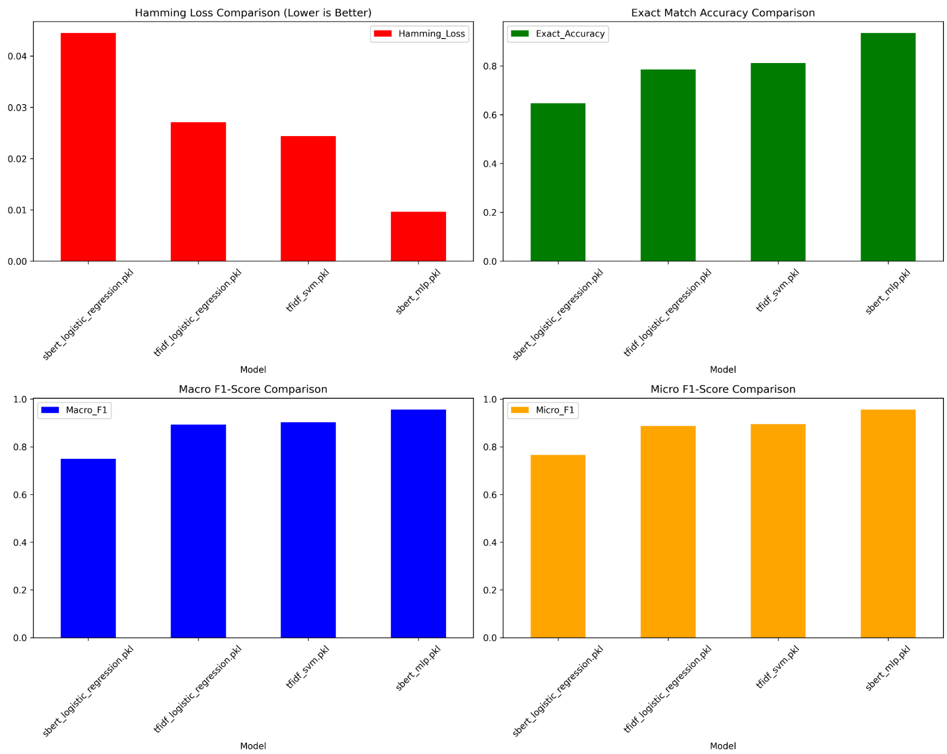
- **SBERT MLP** outperforms all other models in both accuracy and F1-score.
- **TF-IDF SVM** is a strong traditional baseline, especially for interpretable explanations.

Visualizations

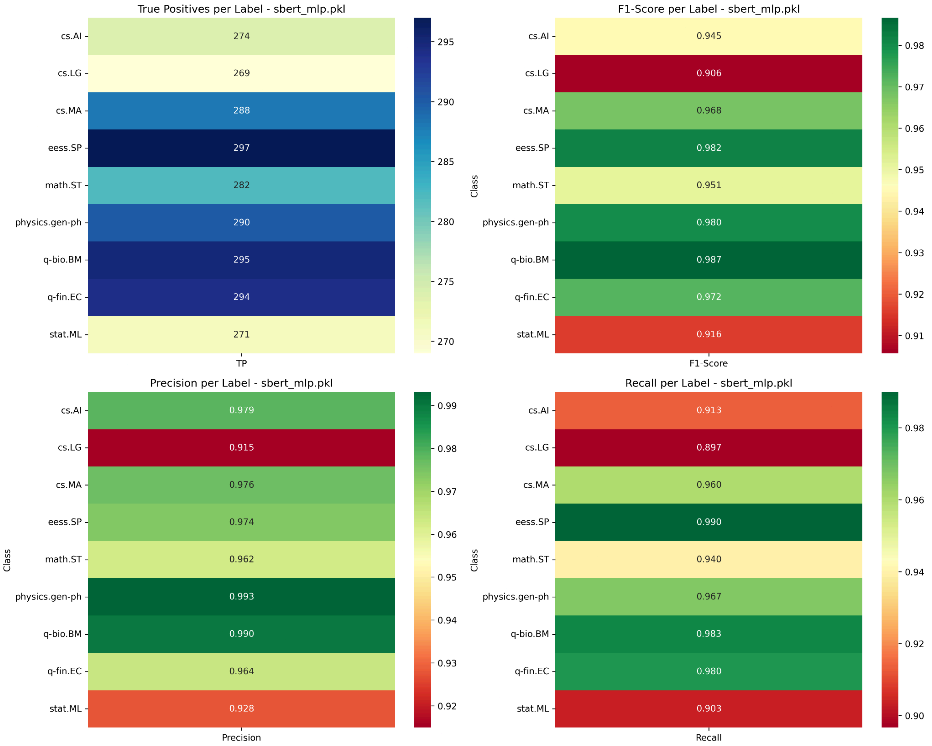
- **Performance Summary :**



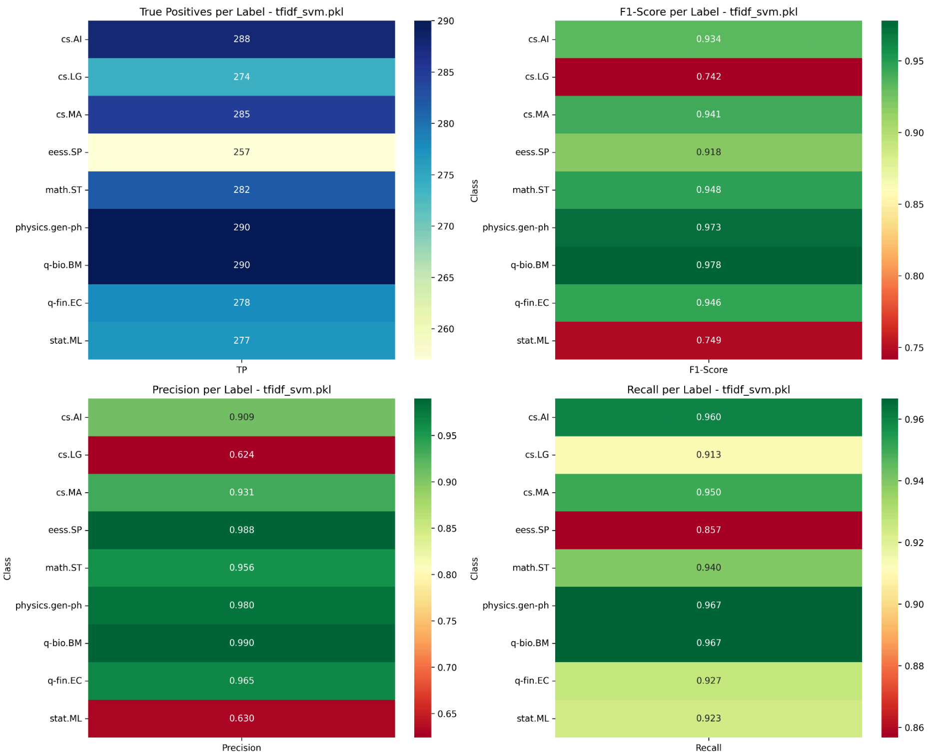
- **Model Comparison :**



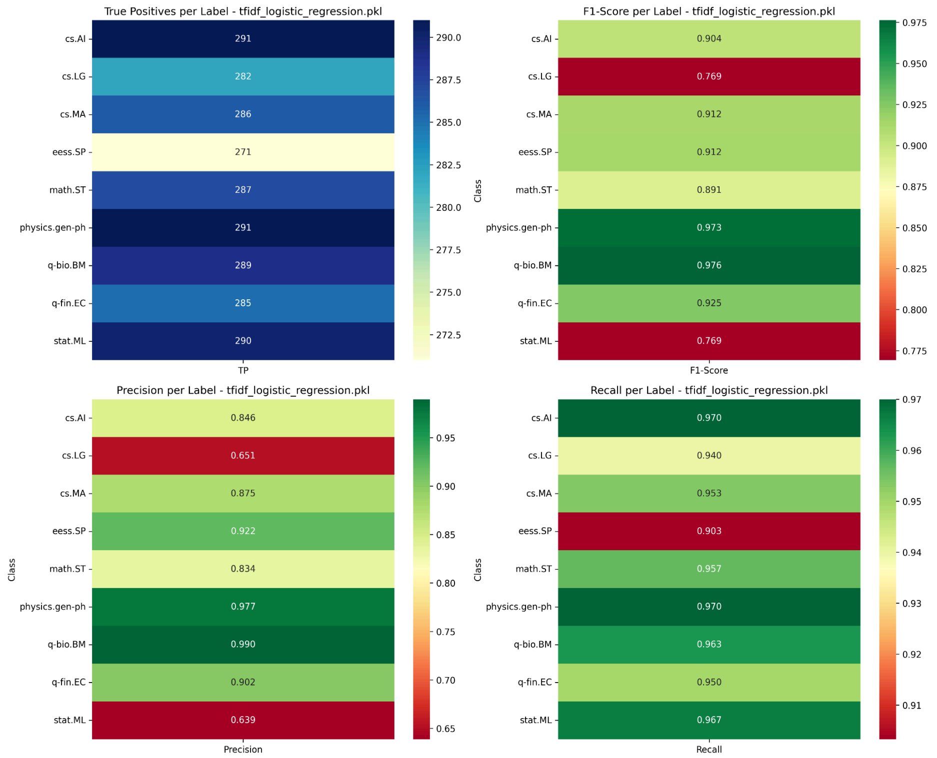
- **SBERT MLP Heatmaps :**



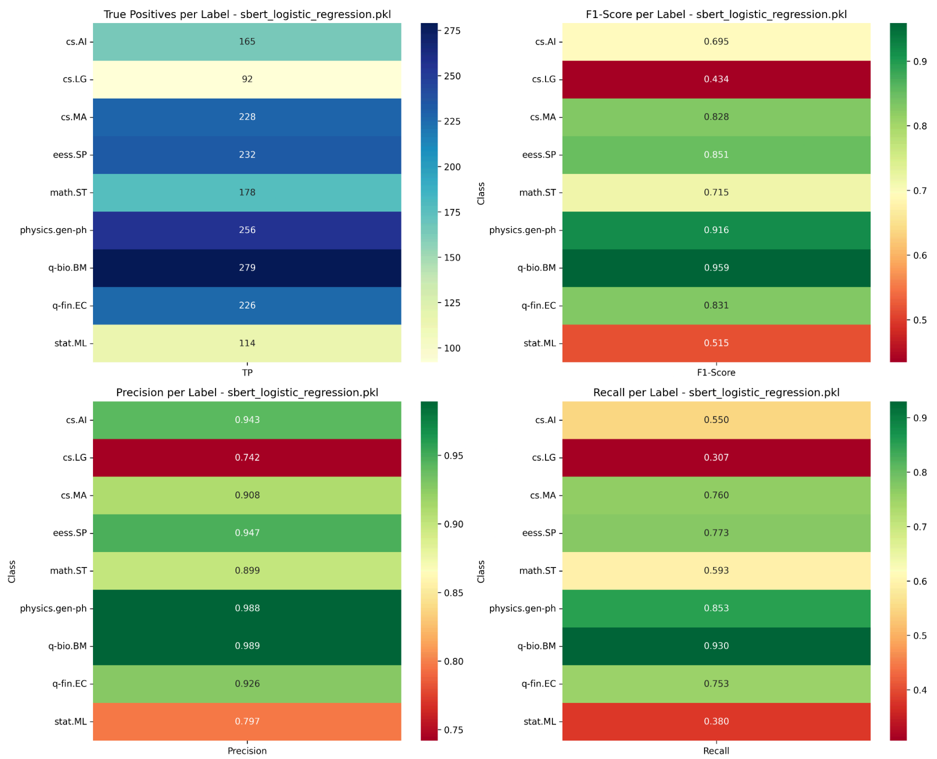
- **TF-IDF SVM Heatmaps :**



- ## TF-IDF Logistic Regression Heatmaps :



- ## SBERT Logistic Regression Heatmaps :



Explainability & Interpretability

TF-IDF Models :

- **SHAP explanations** are word-based and human-interpretable.
- **Example**:
    - Positive features:
        cosmology: 0.12
        scalar: 0.09
    - Negative features:
        quantum:-0.07
        neural:-0.06

SBERT Models :

- **Production Solution:** LIME is used to provide word-level explanations for SBERT models, ensuring interpretability for all model choices.
- **Example:** For a given abstract, LIME highlights the most influential words for the predicted category.

Analysis

- **SBERT models** provide superior performance, especially for nuanced, semantic distinctions.
- **TF-IDF models** are faster and more interpretable, making them suitable for scenarios where the transparency is critical.
- **Explainability:** The system ensures that all model choices provide human-interpretable explanations, critical for client-facing and production deployments

---

Conclusion

This system delivers robust, accurate, and interpretable scientific paper classification. It is ready for production deployment, with strong performance, comprehensive evaluation, and human-friendly explanations for all model choices

---

References

- arXiv API
- Sentence Transformers (SBERT)
- scikit-learn
- LIME, SHAP
- spaCy, NLTK