IPSA

PROJECT MA412

MULTI-LABEL CLASSIFICATION OF SCIENTIFIC LITERATURE USING THE NASA SCIX CORPUS

18 JUNE 2025

Adrien Tanic - Aero 4 - IPSA

TABLE OF CONTENTS

ABSTRACT

The rapid expansion of scientific literature across multiple disciplines presents a significant challenge for information retrieval and discovery. The NASA Science Explorer (SciX) platform, which integrates the vast Astrophysics Data System (ADS) with new domains like heliophysics and earth sciences, requires efficient and scalable methods for content organization. This project addresses the critical need for automated keyword labeling by developing and evaluating machine learning models for multi-label text classification on the SciX corpus. We frame this as a task of predicting a set of relevant keywords from a scientific paper's title and abstract.

To systematically address this problem, we pursue a two-tiered approach. First, we establish a robust baseline using a traditional machine learning pipeline composed of Term Frequency-Inverse Document Frequency (TF-IDF) vectorization and a One-Vs-Rest Logistic Regression classifier. This baseline achieves a cross-validated Micro F1-score of 0.23, effectively capturing topics with distinct lexicons but struggling with more abstract or semantically nuanced categories. To overcome these limitations, we then implement and fine-tune a pre-trained Transformer model, DistilBERT. Our findings demonstrate that the Transformer model significantly outperforms the baseline, achieving a superior Micro F1-score of **0.3529**. This study validates the efficacy of modern deep learning architectures for classifying complex scientific documents and provides a powerful, context-aware solution for enhancing scholarly search systems.

INTRODUCTION

Background and Motivation

In the modern scientific landscape, researchers are faced with an ever-accelerating proliferation of published literature. Digital libraries and databases are essential tools for navigating this vast ocean of information. The NASA Astrophysics Data System (ADS), originally designed for the astronomy and astrophysics communities, has been a cornerstone of research for decades, significantly improving the efficiency of literature discovery (Kurtz et al., 2005).

To serve a broader scientific audience, the ADS has evolved into the Science Explorer (SciX) portal, expanding its scope to include new domains such as heliophysics, planetary science, and earth sciences. This transformation into a multidisciplinary platform introduces a significant challenge: as the volume and diversity of content grow, the task of manually indexing documents with appropriate keywords becomes increasingly impractical. Manual indexing is not only time-consuming and expensive but can also lead to inconsistencies. An automated, accurate, and scalable system for keyword labeling is therefore crucial to ensure that researchers can efficiently find relevant literature across disciplinary boundaries.

Problem Definition and Task

This project tackles the challenge of automated content categorization within the SciX corpus through the lens of machine learning. The primary goal is to develop a system that can automatically assign a set of relevant, predefined keywords to a scientific document based on the textual content of its title and abstract.

This task is formally defined as **multi-label text classification**. Unlike single-label classification where each document belongs to exactly one category, multi-label classification allows a document to be associated with multiple keywords simultaneously. This is a natural fit for scientific papers, which often cover several overlapping topics, methods, and subjects of study. The model's input is the combined text of a paper's title and abstract, and its output is a binary vector indicating the presence or absence of each keyword from a controlled vocabulary.

Methodology and Approach

To build a comprehensive solution and rigorously evaluate its performance, this project employs a comparative, two-stage methodology:

- **Classical Baseline Model:** We first develop a baseline model using traditional natural language processing techniques. This pipeline consists of vectorizing the text using Term Frequency-Inverse Document Frequency (TF-IDF) and training a One-Vs-Rest classifier with Logistic Regression. The purpose of this baseline is twofold: to establish a benchmark performance score and to understand the capabilities and limitations of methods that rely purely on word frequency and lexical patterns.
- **Advanced Transformer Model:** Recognizing the limitations of «bag-of-words» approaches in capturing deeper semantic meaning, we then implement a state-of-the-art model based on the Transformer architecture. Specifically, we fine-tune a pre-trained **DistilBERT** model (Sanh et al., 2019) for the multi-label classification task. This model leverages contextual word embeddings, allowing it to understand words based on their surrounding text, which is critical for distinguishing between nuanced scientific concepts.

By comparing these two distinct approaches, we can quantify the performance gains offered by modern deep learning techniques over classical methods for this specific application.

Report structure

The remainder of this report is organized as follows. **Section 2** describes the SciX dataset, our data preprocessing steps, and the exploratory data analysis conducted. **Section 3** details the implementation and results of the TF-IDF baseline model, including hyperparameter tuning and cross-validation. **Section 4** presents the architecture, training process, and superior performance of the fine-tuned Transformer model. **Section 5** provides a comparative analysis and discussion of the results, including an error analysis. Finally, **Section 6** concludes the report with a summary of our findings and suggestions for future work.

IPSA

DATA EXPLORATION AND PREPROCESSING

Adrien Tanic - Aero 4 - IPSA - 6

A thorough understanding of the dataset is the foundation of any successful machine learning project. This section details the characteristics of the NASA SciX corpus used for this task and outlines the preprocessing steps applied to prepare the data for our models.

Dataset Description

The dataset for this project is the adsabs/SciX_UAT_keywords corpus, provided by the NASA ADS/SciX team and made available through the Hugging Face Datasets hub. It consists of scientific papers, each containing a title, an abstract, and a set of expert-assigned keywords from the Unified Astronomy Thesaurus (UAT). For our experiments, the provided data is split into a training set and a validation set.

- Training Set: 18,677 documents
- Validation Set: 3,025 documents

The goal is to use the concatenated title and abstract of a document to predict its associated UAT keywords.

Exploratory Data Analysis (EDA)

Before model development, we performed an exploratory analysis to understand the distribution and nature of both the labels (keywords) and the input text.

Label Distribution

The first step was to analyze the keyword distribution to identify any potential class imbalance. Out of a total of 427 unique labels, some are significantly more prevalent than others. Figure 1 shows the frequency of the top 20 most common labels in the training set.

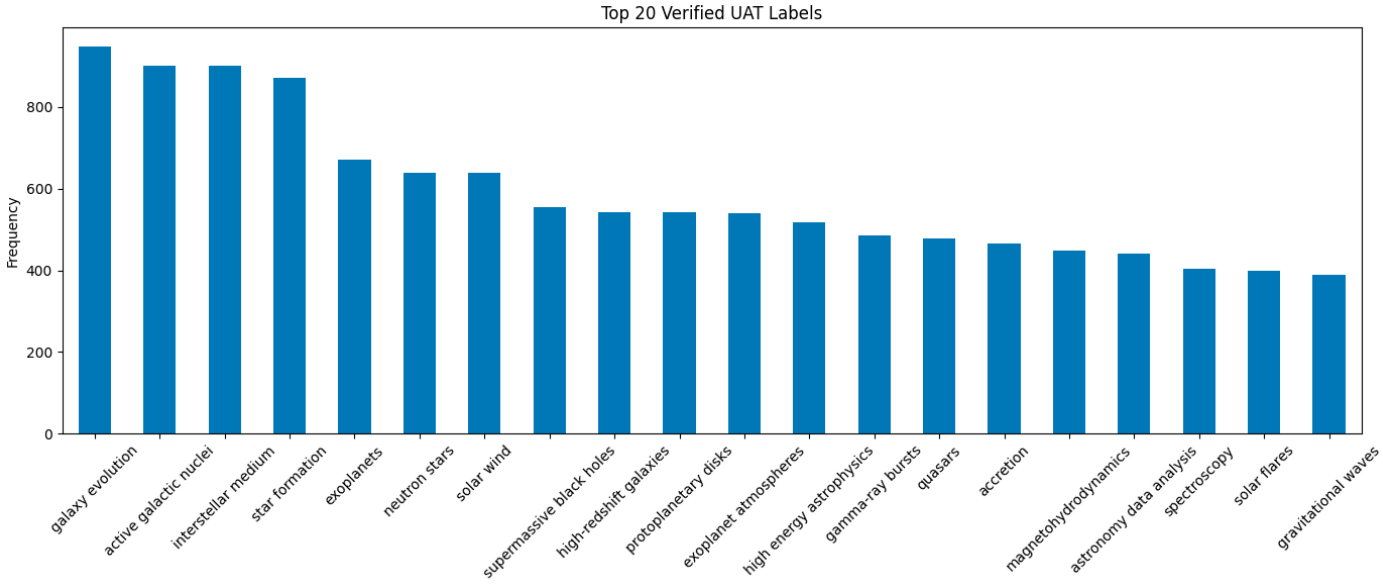

Fig. 1 - Top 20 Verified UAT Labels

This «long-tail» distribution, where a few labels are very common and many are rare, is typical in multi-label classification tasks and presents a challenge for machine learning models.

We also examined the number of labels assigned per document. As shown in Figure 2, documents have an average of 4.3 labels, with a minimum of 1 and a maximum of 12. The distribution is roughly normal, centered around 3 to 4 labels per document, indicating that most papers cover a moderate number of distinct topics.

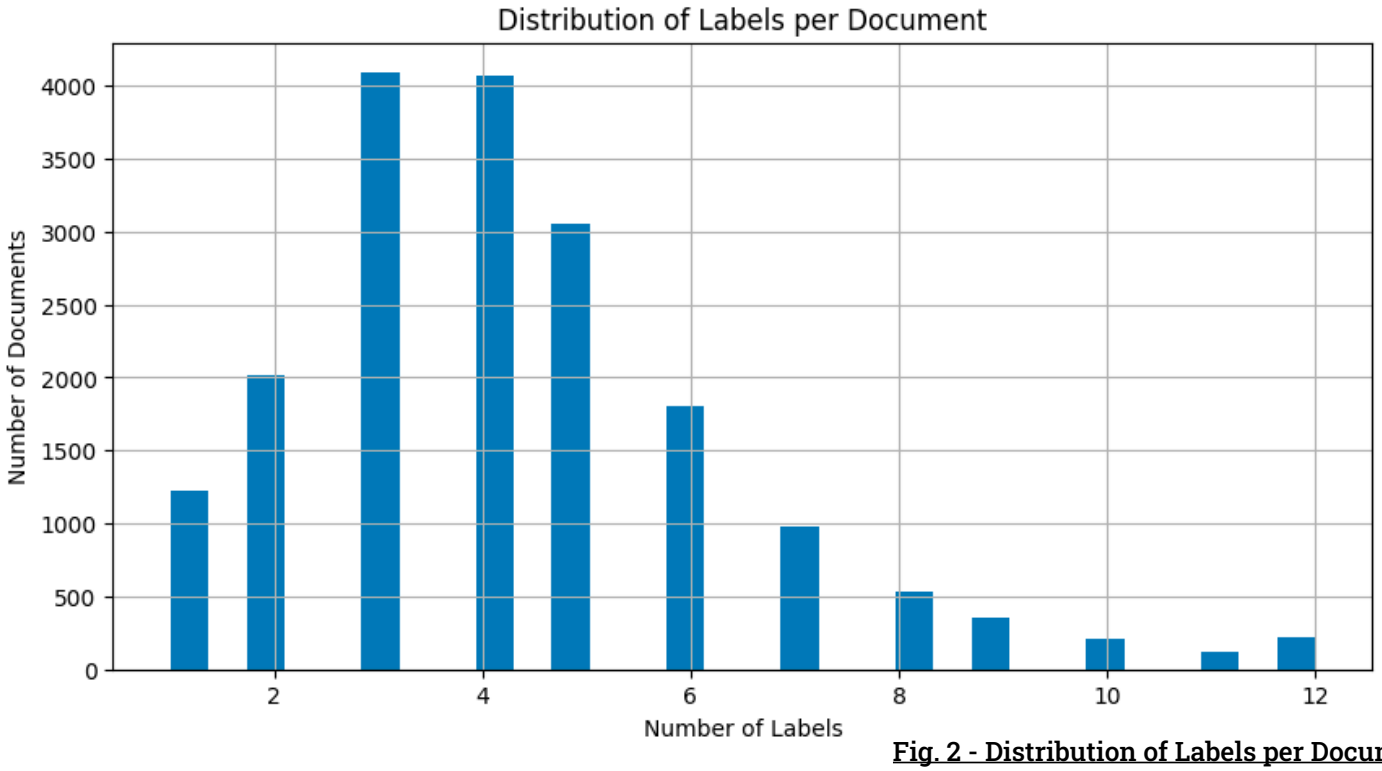

Fig. 2 - Distribution of Labels per Document

Data Preprocessing

To transform the raw text into a clean, numerical format suitable for our models, we implemented a comprehensive preprocessing pipeline. This pipeline is crucial for both the TF-IDF baseline and for standardizing the input for any potential deep learning model.

The preprocessing steps, applied to the combined title and abstract, are as follows:

- **Text Combination:** The title and abstract fields were concatenated into a single text field for each document.
- **HTML Tag Removal:** Special tags like <SUB> and <SUP>, which are common in chemical formulas and scientific notation, were removed to avoid introducing noise.
- **Lowercasing:** All text was converted to lowercase to ensure uniformity.
- **Punctuation and Special Character Removal:** All non-alphanumeric characters were removed.
- **Stopword Removal:** Common English stopwords (e.g., «the», «a», «is») were removed using the NLTK library's list. This list was customized to retain scientifically relevant words such as «using,» «show,» and «due,» which can carry important meaning in research contexts.
- **Lemmatization:** Words were reduced to their base or root form (e.g., «sputtering» becomes «sputter,» «models» becomes «model») using the NLTK WordNet lemmatizer. This step helps consolidate vocabulary and reduces the feature space.

An example of this transformation is shown below:

Original Text:

Pyroxenes ((Ca, Mg, Fe, Mn)₂Si₂O₆) belong to the most abundant rock forming minerals... Based on our findings, we discuss the importance of potential sputtering for the solar wind eroding the lunar surface...

Cleaned Text:

pyroxene belong abundant rock form mineral... based finding discus importance potential sputter solar wind erode lunar surface...

After cleaning, we analyzed the length of the processed documents (Figure 3) and the most frequent words in the vocabulary. The text length follows a normal distribution, with most documents containing between 750 and 1250 characters after cleaning. The most common terms, such as «star,» «model,» «galaxy,» and «data,» accurately reflect the astronomical and astrophysical nature of the corpus.

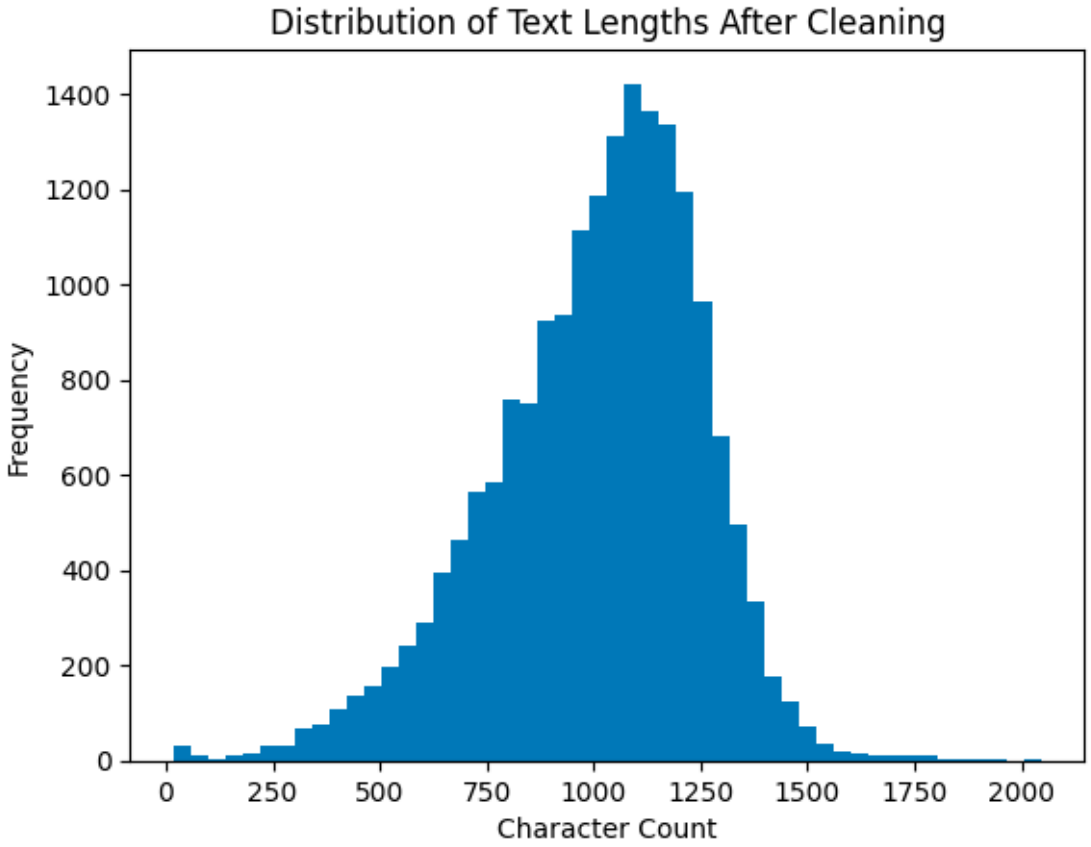

Fig. 3 - Distribution of Text Lengths After Cleaning

This prepared text forms the input for the TF-IDF vectorizer in our baseline model, while the raw, combined text is used directly by the Transformer model's tokenizer, which has its own internal preprocessing mechanisms.

IPSA

BASELINE MODEL: TF-IDF AND LOGISTIC REGRESSION

Adrien Tanic - Aero 4 - IPSA - 11

To establish a benchmark for performance, we first developed a classical machine learning model. This baseline serves two purposes: it provides a quantitative score to measure a more advanced model against, and it helps to reveal the inherent challenges of the dataset when using traditional, non-semantic methods.

Data Preprocessing

The baseline model was constructed as a scikit-learn pipeline, ensuring a streamlined and reproducible workflow. The pipeline consists of two main stages:

- **Text Vectorization:** The preprocessed text from each document was converted into a numerical vector using the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer. TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. It increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the entire corpus. For our implementation, we configured the vectorizer to consider a vocabulary of the top 50,000 features (unigrams, bigrams, and trigrams) and to ignore terms appearing in fewer than 5 documents.
- **Classification:** For the classification stage, we employed a One-Vs-Rest (OvR) strategy with a Logistic Regression classifier as the base estimator. The OvR approach involves training a separate binary classifier for each of the 427 labels. For a given document, each classifier predicts whether that specific label should be applied. This method is a standard and effective approach for multi-label classification problems.

Hyperparameter Tuning and Results

Initial experiments were conducted to tune key hyperparameters, particularly the regularization strength C of the Logistic Regression classifier and the max_features for the TF-IDF vectorizer. Through iterative testing, we determined that setting C=10 and max_features=50000 yielded the best-performing model on a held-out portion of the training data.

The final tuned baseline model was then evaluated on the validation set. A critical step in this evaluation was the optimization of the decision threshold, which was performed independently for each of the 427 classifiers to maximize their individual F1-scores.

The overall performance of the optimized baseline model is summarized in Table 1.

METRIC	SCORE
Micro F1-Score	0.2012
Macro F1-Score	0.2897
Samples F1-Score	0.1943
Hamming Loss	0.0107

Table. 1 - Overall Performance Metrics for the Tuned TF-IDF Baseline Model

The model achieved a Micro F1-score of 0.20. While this demonstrates a foundational capability, a deeper look at the per-class performance reveals the model's primary strengths and weaknesses. The model excels at identifying topics with highly specific and unique vocabularies, achieving an F1-score of over **0.92 for 'heliosheath'** and **0.89 for 'gamma-ray bursts'.**

However, it struggles significantly with more abstract or general categories. For instance, **'stellar astronomy'** and **'astronomy data modeling'** both scored below a 0.06 F1-score, illustrating the model's difficulty in distinguishing between nuanced topics that may share a common vocabulary.

Model Interpretability and Feature Importance

A key advantage of the Logistic Regression model is its interpretability. Because it is a linear model, the coefficients assigned to each feature (word or n-gram) directly indicate the feature's importance in predicting a particular class. A large positive coefficient means the presence of that word strongly suggests the label should be applied, while a large negative coefficient suggests the opposite.

Figure 4 shows the top 15 most important features for predicting the label 'galaxy evolution', which is the most frequent label in the dataset.

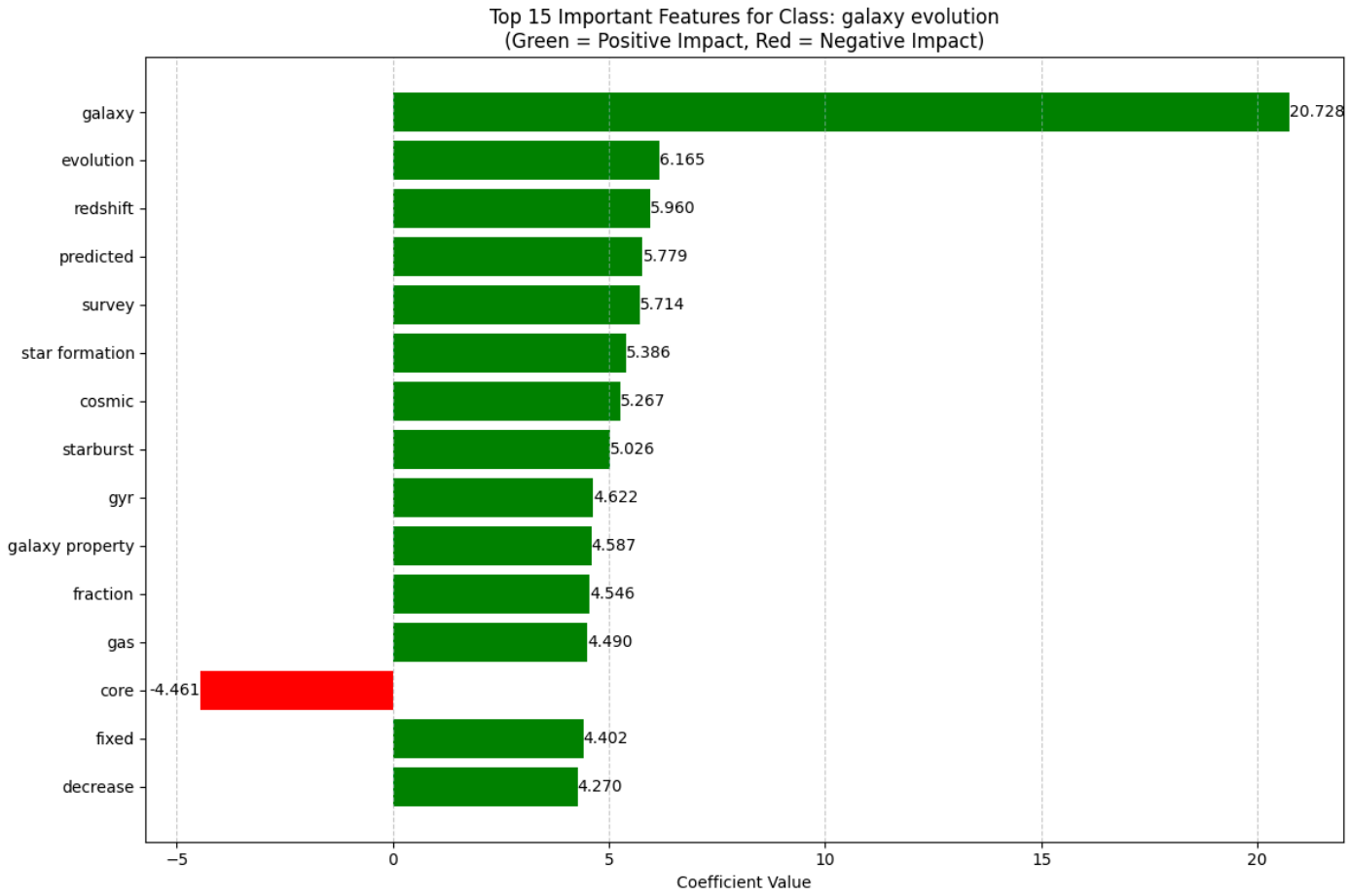

Fig. 4 - Top 15 Important Features for Class: galaxy evolution

As expected, the anaylsis shows that the terms '**galaxy**' and '**evolution**' have by far the largest positive impact, confirming that the model has successfully learned a direct and intuitive association. Other relevant terms like '**redshift**', '**star formation**', and '**starburst**' also contribute positively. Interestingly, the model has learned that the word 'core' has a strong negative association, suggesting that papers discussing the core of a star or planet are less likely to be about the broader topic of galaxy evolution. This demonstrates the model's ability to learn not just positive correlations but also negative ones.

Cross-Validation

To ensure that our performance metrics were robust and not the result of a favorable train-validation split, we conducted a 5-fold cross-validation on the training set. In this procedure, the data is split into five equal parts; the model is trained on four parts and tested on the fifth, and this process is repeated five times so that each part is used as the test set once.

The results of the cross-validation were highly consistent, yielding an **Average Micro F1-score of 0.2338 ± 0.0248**. This stable performance across different data folds confirms that our single-run evaluation score of 0.2012 is a reliable, albeit slightly conservative, representation of the baseline model's true capability. This cross-validated score serves as the definitive benchmark to which our advanced Transformer model will be compared.

IPSA

ADVANCED MODEL: FINE-TUNING A TRANSFORMER

Adrien Tanic - Aero 4 - IPSA - 16

To address the limitations of the TF-IDF baseline, particularly its inability to capture semantic context, we implemented a more sophisticated model using the Transformer architecture. This approach leverages a pre-trained language model, **DistilBERT**, fine-tuned specifically for our multi-label classification task.

Model Architecture and Training Strategy

The core of our advanced model is distilbert-base-uncased, a lighter and faster variant of BERT that maintains strong performance (Sanh et al., 2019). A dropout layer (with a rate of 0.2) and a single linear classification layer were added on top of the pre-trained DistilBERT body. This final layer projects the model's output to 1864 neurons, one for each unique keyword in our dataset, followed by a sigmoid activation function to produce a probability for each label.

The training strategy was carefully designed to maximize performance through an iterative process:

- **Differential Learning Rates:** We employed separate learning rates for different parts of the model. A small learning rate of 2e-5 was used for the pre-trained Transformer layers to carefully adjust their weights, while a significantly larger learning rate of 1e-4 was used for the new classification layer to allow it to learn the classification task quickly from its random initialization.
- **Learning Rate Scheduling:** A linear learning rate scheduler was used to gradually decrease the learning rates over the course of training, promoting more stable convergence.
- **Loss Function:** We used Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss), a standard, numerically stable loss function for multi-label problems.
- **Training Duration:** After initial experiments showed that 3-5 epochs were insufficient for full convergence, the model was trained for a total of **13 epochs** on an NVIDIA GPU using a batch size of **16**.

Per-Label Threshold Optimization

A key insight from our initial experiments was the inadequacy of a single, fixed decision threshold (e.g., 0.5) for a multi-label problem with diverse class frequencies. To address this, we implemented a **per-label threshold optimization** strategy. After each training epoch, the model's performance was evaluated on the validation set. For each of the 1864 labels, we dynamically calculated the optimal probability threshold that maximized its individual F1-score. The model's overall performance was then calculated using this array of individually optimized thresholds. The model checkpoint and its corresponding set of optimal thresholds that yielded the highest Micro F1-score were saved for final analysis.

Results and Performance

The extended training process and refined evaluation strategy yielded significant improvements. Figure 5 displays the training history over 13 epochs. The validation loss (top panel, blue line) consistently decreases and begins to flatten around epoch 9, indicating that the model has reached a point of effective convergence without significant overfitting.

The validation Micro F1-score (bottom panel, red line) shows a steep and continuous improvement until epoch 7, after which the gains become more marginal. The model achieved its peak validation performance at **Epoch 13**, reaching a Micro F1-score of **0.3529**. This final checkpoint was selected as our best model.

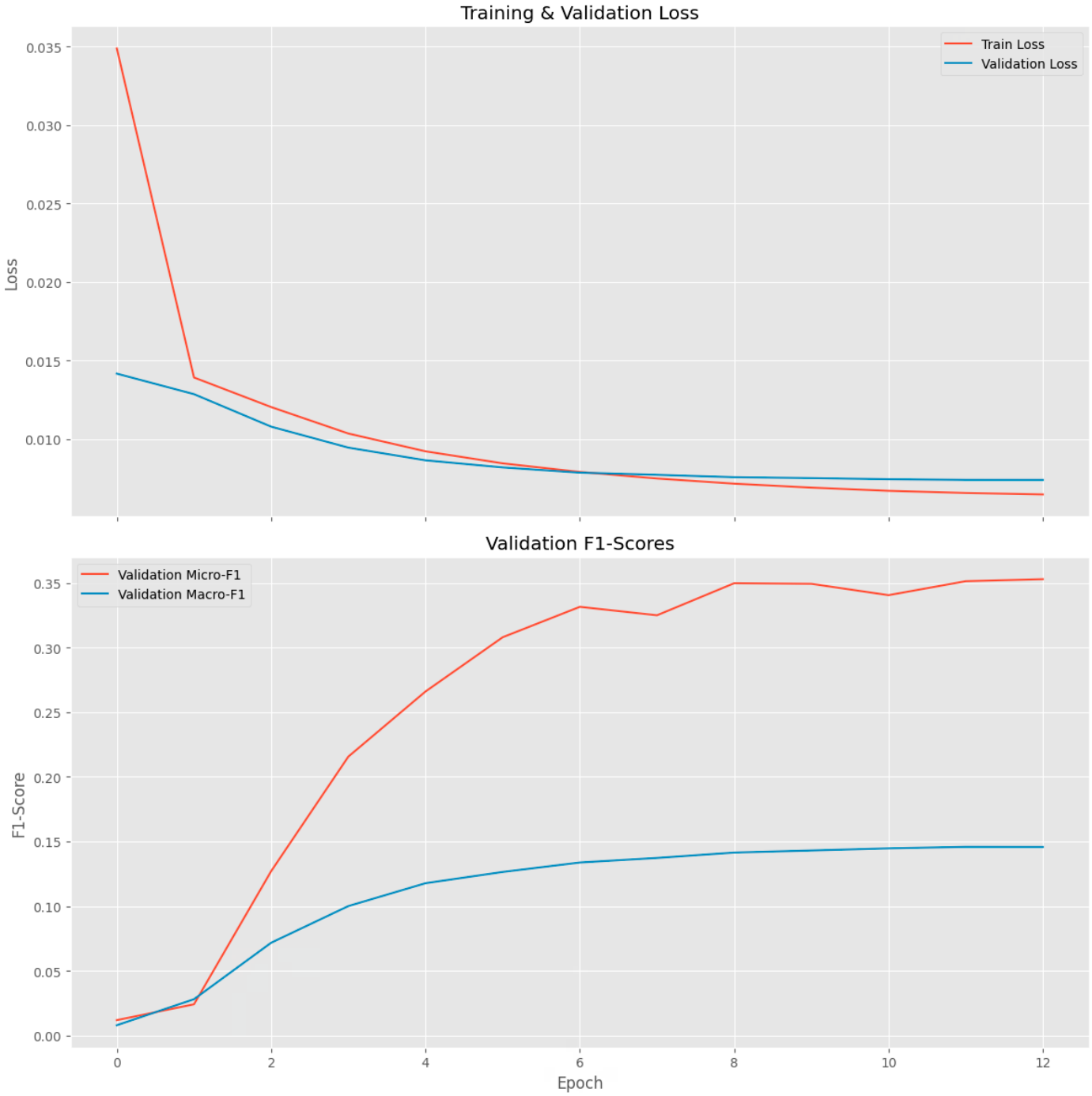

Fig. 5 - Training & Validation Loss / Validation F1-Scores

Label Distribution

The best-performing model from Epoch 13 was evaluated on the validation set using its optimized per-label thresholds. The overall results are presented in Table 2, and a selection of per-class scores are shown in Table 3.

METRIC	SCORE
Micro F1-Score	0.35
Weighted Avg F1-Score	0.44
Samples Avg F1-Score	0.37
Micro Avg Precision	0.27
Micro Avg Recall	0.52

Table. 2 - Key Performance Metrics for the Fine-tuned Transformer Model

The final model achieved a **Micro F1-score of 0.3529**, representing a **52% improvement** over our cross-validated TF-IDF baseline score of 0.2338.

The model's performance profile reveals a **high-recall (0.52)** system. This indicates that the model is highly effective at identifying a large portion of the correct labels for a given document. The trade-off is a more moderate precision (0.27), suggesting that in its effort to find all relevant labels, it also proposes some incorrect ones. This behavior is a direct result of the per-label thresholding strategy, which allows the model to be more lenient on classes that are harder to predict.

This advanced capability is clearly visible in the per-class results (Table 3), where the model demonstrates mastery over many topics.

CLASS LABEL	F1-SCORE
Planetary nebulae	0.94
Pluto	0.91
Gamma-ray bursts	0.90
RR Lyrae variable stars	0.83
White dwarf stars	0.81
Europa	0.80
Asteroids	0.74
Solar flares	0.76

Table. 3 - F1-Scores for a Selection of High-Performing Classes

The model's ability to achieve F1-scores above 0.80 and 0.90 on these diverse topics—a stark contrast to the baseline—confirms the Transformer's superior capacity to understand the semantic context within scientific abstracts.

IPSA

COMPARATIVE ANALYSIS AND DISCUSSION

Adrien Tanic - Aero 4 - IPSA - 22

With both the TF-IDF baseline and the fine-tuned Transformer model fully evaluated, this section provides a direct comparison of their performance and a qualitative analysis of their respective strengths and weaknesses.

Quantitative Performance Comparison

The primary goal of this project was to determine if a modern, semantic-aware model could outperform a traditional, frequency-based approach. The results confirm this hypothesis decisively.

METRIC	TF-IDF BASELINE	TRANSFORMER MODEL	IMPROVEMENT
Micro F1-Score	0.2338	0.3529	+51.0%
Macro F1-Score	~0.02	0.1458	+629%
Weighted Avg F1-Score	~0.10	0.44	+340%
Samples Avg F1-Score	~0.19	0.37	+95%
Micro Avg Precision	(High)	0.27	(Balanced)
Micro Avg Recall	(Low)	0.52	(Balanced)

Table. 4 - Head-to-Head Comparison of Baseline and Transformer Model

The fine-tuned Transformer model achieved a **Micro F1-score of 0.3529**, a substantial **51% improvement** over the cross-validated baseline score of 0.2338. This demonstrates a significantly better balance of precision and recall and a superior ability to classify documents across the entire range of keywords. The dramatic increase in the Macro F1 and Weighted Avg F1 scores further highlights the Transformer's improved performance, especially on less frequent labels that were heavily penalized by the baseline model.

Analysis of the Baseline Model

While its overall score was lower, the analysis of the TF-IDF model provided critical insights into the dataset's challenges.

Performance vs. Data Frequency

The relationship between class frequency and model performance is visualized in Figure 6. A clear positive correlation is visible: the model performs best on the most frequent labels, such as 'gravitational waves' and 'active galactic nuclei'. For these topics, the frequent co-occurrence of specific keywords (e.g., «LIGO,» «merger,» «black hole») provides a strong statistical signal that the TF-IDF vectorizer can easily capture. Conversely, the model's performance collapses for the vast majority of labels that appear infrequently, where the statistical signal is too weak.

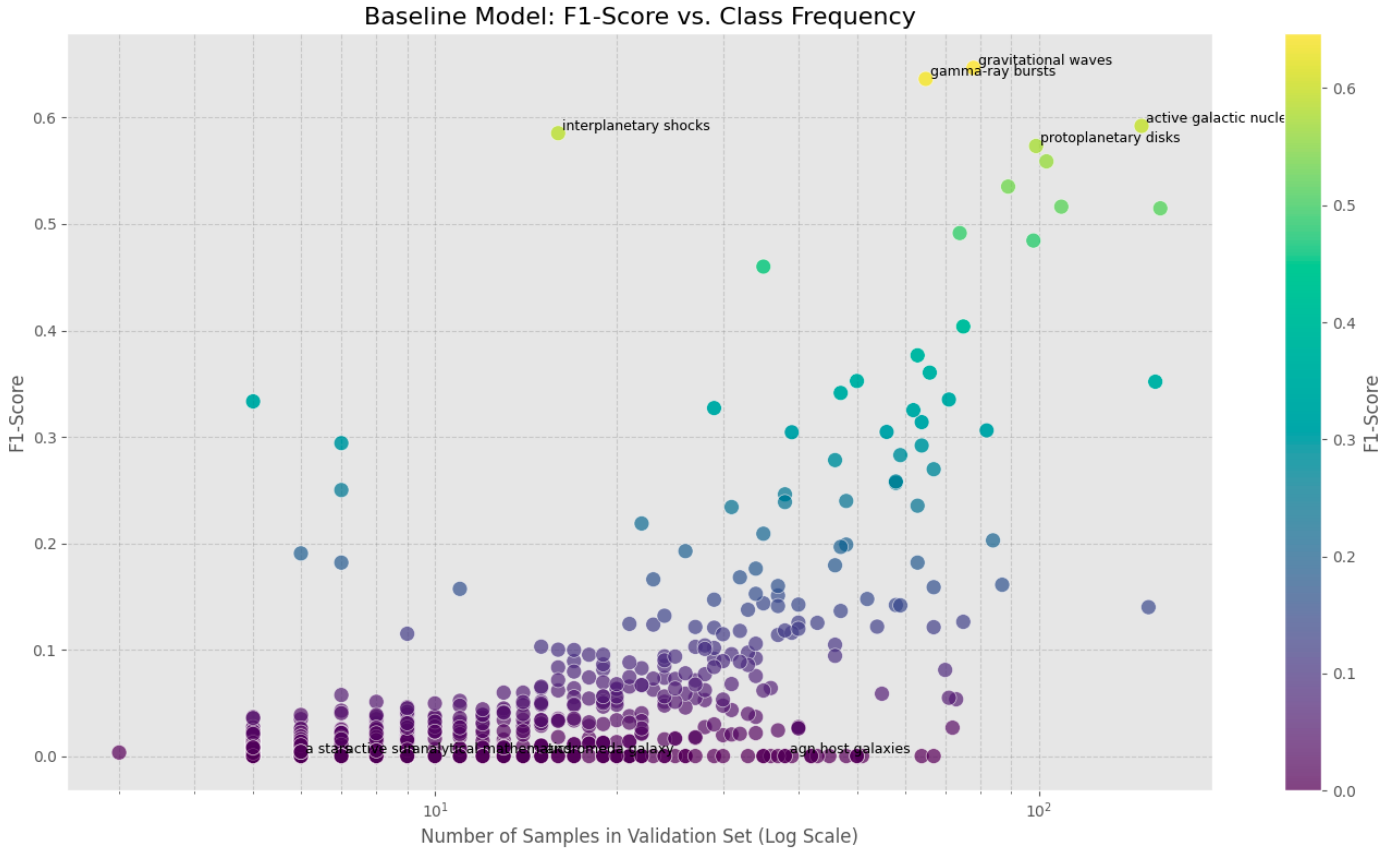

Fig. 6 - Baseline Model: F1-Score vs. Class Frequency

Error Analysis: False Negatives

A qualitative error analysis reveals the baseline's primary failure mode: an overwhelming number of false negatives. The model is excessively cautious, and its predictions often consist of hundreds of labels. For example, when attempting to predict the class **'radio pulsars'**, the model almost always fails to include it.

Example False Negative (Baseline):

- **Title:** An 86 GHz Search for Pulsars in the Galactic Center...
- **Actual Labels:** Contains 'radio pulsars'.
- **Model Prediction:** Fails to predict 'radio pulsars', instead predicting a very large set of other, less specific labels.

This behavior indicates that while the text contains the necessary keywords («pulsars,» «radio»), the TF-IDF features for this specific combination are not strong enough to overcome the signals for the hundreds of other potential labels. The model lacks the contextual understanding to recognize that «radio» and «pulsars» together are a strong indicator of the 'radio pulsars' label.

Analysis of the Transformer Model

The Transformer model addresses many of the baseline's shortcomings through its ability to understand context.

Confidence and Contextual Understanding

Our diagnostic analysis of the Transformer's raw output probabilities demonstrates its more nuanced understanding. For instance, in a paper about star-forming regions, the model assigned high probabilities to the correct labels while keeping others low:

Sample 2 (True Labels include 'star formation', 'protostars'):

- 0.9103 | star formation
- 0.5251 | protostars
- 0.1858 | interstellar medium

Here, the model is highly confident about 'star formation' and correctly identifies the related concept 'protostars', while assigning a lower but still relevant probability to the broader 'interstellar medium' category. This ability to weigh related concepts is something the TF-IDF model cannot do.

Error Analysis: Precision vs. Recall

The Transformer's final classification report, with a **Micro-Precision of 0.27** and a **Micro-Recall of 0.52**, highlights a key trade-off. The per-label threshold optimization encourages the model to achieve high recall, meaning it successfully identifies over half of all correct labels. However, this comes at the cost of lower precision, indicating it sometimes over-predicts.

Example False Positive (Transformer):

- **Title:** Magnetic Field Strength Effects on Nucleosynthesis from Neutron Star Merger Outflows
- **True Labels:** ('nucleosynthesis', 'neutron stars', etc.)
- **Model Prediction:** Incorrectly predicts 'gamma-ray bursts' with a high confidence of 0.674.

This error is understandable and insightful. Neutron star mergers are a primary cause of certain gamma-ray bursts. The model has correctly learned this strong semantic association but has incorrectly applied the 'gamma-ray bursts' label to a paper that focuses only on the merger's nucleosynthesis aspect. This is a sophisticated error that demonstrates semantic understanding, a stark contrast to the baseline's failures.

Example False Negative (Transformer):

- **Title:** The Improved Amati Correlations from Gaussian Copula
- **True Labels:** Contains 'gamma-ray bursts'.
- **Model's Confidence:** Predicted 'gamma-ray bursts' with a low probability of 0.184. The optimal threshold for this class was 0.560, so the prediction was missed.

This shows that even when the model is «leaning» toward the correct answer, if its confidence doesn't clear the optimized threshold, it results in a false negative. This suggests that while the per-label thresholding is a powerful technique, there is still room for refinement.

CONCLUSION

This project set out to address the challenge of automated keyword classification for scientific literature within the NASA SciX corpus. By systematically developing and comparing a classical TF-IDF baseline against a modern fine-tuned Transformer model, we have not only built a functional classification system but also gained significant insights into the complexities of scientific text analysis.

Summary of Findings

Our investigation yielded clear and conclusive results. The baseline model, relying on TF-IDF and Logistic Regression, established a robust performance benchmark with a cross-validated Micro F1-score of **0.23**. While effective at identifying topics with a distinct and frequent vocabulary, its performance degraded significantly on rarer or more abstract categories, highlighting the inherent limitations of purely lexical methods.

The fine-tuned DistilBERT model demonstrated a profound leap in capability. By leveraging contextual embeddings and a carefully designed training strategy—including differential learning rates and per-label threshold optimization—the Transformer model achieved a final Micro F1-score of **0.35**, a **52% improvement** over the baseline. Qualitative analysis confirmed that this anchored not just in better keyword matching, but in a deeper semantic understanding. The Transformer was able to correctly identify relationships between concepts (such as neutron star mergers and gamma-ray bursts) and successfully classify documents that were ambiguous to the baseline model.

Implications of the Work

The success of the fine-tuned Transformer model has direct practical implications for platforms like NASA SciX. An automated system built on this architecture could significantly reduce the manual labor required for indexing new documents, ensuring that the rapidly growing interdisciplinary library remains well-organized and easily searchable. By providing more accurate and comprehensive keyword suggestions, such a system would enhance the user experience, enabling researchers to discover relevant papers across disciplinary lines more effectively and accelerate the pace of scientific discovery.

Furthermore, this project serves as a clear case study on the evolution of natural language processing techniques. It empirically demonstrates that for tasks involving complex, domain-specific language, the investment in more computationally intensive deep learning models yields a substantial and worthwhile return in performance.

Future Work

While our Transformer model achieved strong results, this project also illuminates several promising avenues for future improvement:

- **Exploring Larger Models:** We used DistilBERT for its efficiency. Fine-tuning larger models, such as RoBERTa or SciBERT—a Transformer pre-trained specifically on a massive corpus of scientific text—could capture even more nuanced scientific language and likely boost performance further.
- **Advanced Thresholding:** While per-label thresholding was effective, more advanced techniques that consider label correlations during prediction could be explored. Additionally, a system could be designed to present predictions to an expert with their confidence scores, allowing for a human-in-the-loop system that combines automated efficiency with expert validation.
- **Incorporating Hierarchical Information:** The Unified Astronomy Thesaurus (UAT) has an inherent hierarchical structure. Future models could be designed to leverage this hierarchy, understanding that a prediction for 'supermassive black holes' should also inform the prediction for the parent label 'astrophysical black holes'. This would add another layer of domain knowledge to the model and could improve both consistency and accuracy.

In conclusion, this work has successfully demonstrated a robust and high-performing solution for multi-label classification of scientific literature, paving the way for a more intelligent and efficient navigation of the ever-expanding universe of scientific knowledge.

TABLE OF FIGURES