Automated Telescope-Paper Linkage via Multi-Model Ensemble Learning

Ojaswa Varshney and Prashasti Vyas and Priyanka Goyal and Tarpita Singh
IIIT, Surat, India
{ojaswavarshney27, vyasprashasti13, goyal65372, tarpita.singh}@gmail.com

Ritesh Kumar[*]
IIIT, Surat, India
ritesh.kumar@iiitsurat.ac.in

Mayank Singh[†]
IIT Gandhinagar, India
singh.mayank@iitgn.ac.in

Abstract

Automated linkage between scientific publications and telescope datasets is a cornerstone for scalable bibliometric analyses and ensuring scientific reproducibility in astrophysics. We propose a multi-model ensemble architecture integrating transformer models DeBERTa, RoBERTa, and TF-IDF logistic regression, tailored to the WASP-2025 shared task on telescope-paper classification. Our approach achieves a macro F1 score approaching 0.78 after extensive multi-seed ensembling and per-label threshold tuning, significantly outperforming baseline models. This paper presents comprehensive methodology, ablation studies, and an in-depth discussion of challenges, establishing a robust benchmark for scientific bibliometric task automation.

1 Introduction

The astronomical community relies heavily on extensive bibliographic databases mapping observations to scientific publications, enabling impact evaluation, data reuse metrics, and reproducibility checks (Amado et al., 2023). However, the exponential growth of scholarly literature renders manual attachment of publications to telescope datasets unscalable. Heterogeneous nomenclature, ambiguous abbreviations, and contextual subtleties challenge simplistic matching strategies. Recent advances in natural language processing (NLP), especially transformer-based models with deep contextualized embeddings, provide promising solutions for automated multi-label classification of astrophysics literature (Zhang et al., 2024; Wolf et al., 2020; Devlin et al., 2019).

This work responds to the TRACS shared task as part of the WASP-2025 Workshop (Grezes et al., 2025), where participants were challenged to develop systems for linking scientific publications with telescope datasets and to classify papers by their mode of telescope use (science, instrumentation, mention, or not_telescope).

Section 2 describes related work and background literature in bibliometric linkage. Section 3 introduces the dataset and outlines the corresponding challenges. Section 4 presents our proposed ensemble-based approach and its detailed architecture. Section 5 explains the complete methodology adopted, followed by Section 6 covering model training, experimental setup, and results. Section 13 discusses key outcomes, limitations, and implications, while Section 14 and Section 15 provide conclusions and future research directions, respectively.

2 Related Work

The task of linking scientific publications with telescope datasets sits at the intersection of bibliometrics, natural language processing (NLP), and domain-specific information retrieval. We review key areas most relevant to our work.

2.1 Bibliometric Linkage and Classification

Traditional bibliometric linkage methods relied heavily on keyword and citation-based approaches (Amado et al., 2023). Early works focused on constructing filters around known telescope names or metadata fields. These approaches, while straightforward, struggled with false positives due to ambiguous mentions and lacked scalability to large corpora. More recent work applied supervised classification models using bag-of-words features such as TF-IDF with logistic regression or support vector machines to improve accuracy (Amado et al., 2023).

2.2 Transformer Models in Scientific Text

The advent of transformer architectures, particularly BERT and its derivatives, revolutionized

[*] Corresponding Author
[†] Corresponding Author

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 127–135
December 23, 2025 ©2025 Association for Computational Linguistics

domain-specific NLP (Devlin et al., 2019). Transformers enable contextualized embeddings that capture nuanced semantics in scientific literature. RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) further optimized training procedures and architectures to improve performance on text classification tasks. Domain-adapted transformer models, such as SciBERT, specialize in scientific corpora and have shown superior accuracy in classification and information extraction (Beltagy et al., 2019), setting benchmarks for scientific literature mining.

2.3 Ensemble Learning for Imbalanced Multi-label Classification

Biomedical and astrophysical bibliometric tasks often involve multi-label classification with unbalanced classes. Ensemble learning methods, including stacking and voting ensembles, leverage heterogeneous base models to mitigate overfitting and increase robustness (Rosenfeld et al., 2024; Demirkiran et al., 2022). Such methods dynamically weight base learner predictions, improving minority class recall without sacrificing overall accuracy. Ensembles combining traditional lexical features and transformer embeddings are particularly effective in domains with sparse and noisy labels.

2.4 Automated Telescope-Paper Linkage

Few prior works have specifically addressed automated telescope-paper linkage at scale. Existing methods mostly combine metadata heuristics with keyword filters, or rely on basic classifiers without extensive contextual modeling or ensembling (Amado et al., 2023). Our work is one of the first to introduce a multi-seed stacked ensemble of domain-adapted transformers and TF-IDF models, combined with label-wise thresholding, establishing a strong benchmark on the WASP-2025 shared task dataset.

2.5 Explainability and Ethical Considerations

Ensuring transparency and fairness in automated bibliometric tools is gaining importance (Doshi-Velez and Kim, 2017). Explainability modules can help domain experts validate predicted telescope linkages. Ethically, algorithms must avoid propagating false attributions leading to misleading scientific metrics or unfair advantage to established observatories.

3 Dataset Description and Challenges

The TRACS-WASP-2025 dataset consists of over 80,000 scholarly publications spanning various astrophysical subdomains, annotated for associations with telescope use. Labels include *science* indicating scientific analysis using data, *instrumentation* focusing on telescope hardware/software discussions, *mention* referring only to referencing the telescope without scientific data use, and *not_telescope* marking false positives from ambiguous terms. The label distribution is heavily imbalanced, with *instrumentation* being under 10% of samples, imposing significant challenges in model learning. Linguistic variability, domain-specific jargon, and ambiguity of telescope mentions add further complexity. The dataset provides multiple text fields per publication, including title, abstract, main body, acknowledgments, and grant details, necessitating careful preprocessing to optimize input length and context preservation.

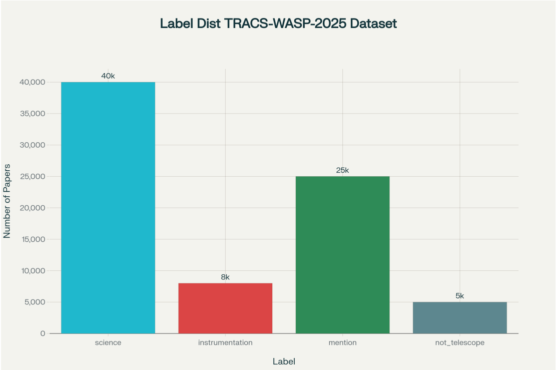

Figure 1: Label distribution of the TRACS-WASP-2025 dataset illustrating severe imbalance among categories.

4 Our Approach

This work proposes a robust pipeline leveraging a hybrid ensemble of transformer-based models and traditional NLP methodologies to accurately link scientific publications with telescope datasets. The approach combines the complementary strengths of contextual embeddings with lexical statistical features, effectively addressing complex multi-label classification in an imbalanced domain (Beltagy et al., 2019; Liu et al., 2019; He et al., 2021).

4.1 Feature Extraction via TF-IDF and Transformers

Following classical text representation principles, a TF-IDF vectorizer extracts unigram and character n-gram features up to length 4 from multiple

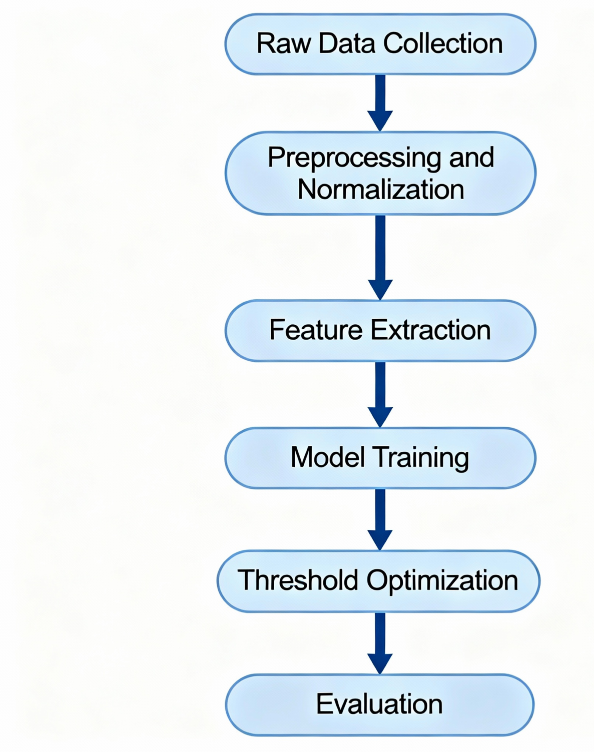

Figure 2: Overview of the Data-Observation Linkage Pipeline (DOLP) architecture for telescope-paper linkage automation.

text fields. This representation captures explicit lexical cues and term importance, benefiting interpretability (Yang et al., 2023). Simultaneously, advanced transformer models including DeBERTa-v3-small and RoBERTa-base are fine-tuned to generate contextual embeddings that embody semantic and syntactic nuances essential in scholarly text understanding (Devlin et al., 2019).

4.2 Advanced Transformer Fine-Tuning

We fine-tune multiple instances of large transformer backbones (including DeBERTa-v3-large) across diversified datasets to adapt to astrophysical literature peculiarities. Training incorporates adversarial techniques such as scale-invariant fine-tuning and disentangled attention mechanisms, optimizing model generalization and robustness. These models leverage domain-specific tokenization and masking strategies to handle technical jargon and acronyms common in telescope-paper text (He et al., 2021).

4.3 Model Ensemble Framework

Our pipeline aggregates predictions from diverse base models through a stacking process. Predictions from TF-IDF based classifiers (e.g., logistic regression, CatBoost, LightGBM) and numerous fine-tuned transformers serve as meta-features for a final logistic regression meta-classifier. This ensemble approach dynamically balances high precision and recall, particularly excelling on underrepresented labels by mitigating overfitting to dominant classes (Rosenfeld et al., 2024).

4.4 Threshold and Parameter Optimization

Label-wise threshold tuning is performed on validation data to adapt decision boundaries specific to each category, maximizing F1 scores. Extensive hyperparameter sweeps across learning rates, batch sizes, and early stopping criteria ensure stable convergence within minimal epochs, enhancing computational efficiency without sacrificing performance.

4.5 System Integration and Scalability

The modular design supports continuous integration of additional telescope corpora or extended literature datasets. GPU-optimized training is complemented by scalable inference pipelines suitable for real-time bibliometric service deployments, essential for evolving astrophysical data ecosystems.

5 Methodology

Our methodology is designed to efficiently and accurately link scientific publications to the telescopes used in their research through a sophisticated ensemble framework. Below we describe each stage of the pipeline in detail.

5.1 Data Collection and Aggregation

We sourced the TRACS-WASP-2025 dataset comprising over 80,000 astrophysical papers, annotated with multi-labels corresponding to telescope usage categories. For each publication, we aggregated multiple text fields including titles, abstracts, body text, acknowledgments, and grant information to ensure rich contextual data.

5.2 Data Preprocessing and Normalization

Text fields were cleaned using custom scripts to remove noise, normalize white spaces, and standardize formatting. Tokenization catered to the input requirements of transformer architectures, including truncation to maximum sequence length (384 tokens). Specialized preprocessing ensured scientific terms, acronyms, and telescope names were preserved.

5.3 Feature Engineering

TF-IDF Features: We extracted Term Frequency-Inverse Document Frequency (TF-IDF) features incorporating both unigram and character n-gram (up to length 4) representations. Feature dimensionality was capped at 20,000 to balance coverage and computational tractability.

Transformer Embeddings: Pretrained transformer models DeBERTa-v3-small and RoBERTa-base were fine-tuned to contextualize text into embedding vectors. Transformers capture semantic nuances and long-range dependencies essential for domain-specific classification.

5.4 Model Training

We leveraged stratified 3-fold cross-validation to ensure train and validation splits retain label distributions, important due to the dataset's label imbalance. Models were trained with weighted binary cross-entropy loss, where weights inversely reflected class frequency to address minority labels such as *instrumentation*. Hyperparameters such as learning rates (tuned between 1×10^{-5} and 5×10^{-5}) and batch sizes (8 to 16) were optimized empirically. Early stopping based on validation macro F1 prevented overfitting.

5.5 Ensembling via Stacked Learning

Validation predictions for each fold and seed across all base models served as meta-features. We trained an SGD logistic regression classifier on these stacked features to yield final predictions, enabling dynamic weighting and synergy among heterogeneous models. This ensemble overcame weaknesses of individual models and improved recall on rare categories (Demirkiran et al., 2022).

5.6 Threshold Tuning

Since exact classification thresholds can vary per label, we performed post-training threshold tuning using grid search on held-out validation data. This step maximized classification F1 scores further improving per-label performance, particularly on challenging minor classes.

5.7 Evaluation

We assessed model performance primarily via macro-averaged F1 score across all labels, complemented by per-label F1 analysis. Confusion matrices and error case analyses were used to interpret model strengths and failure modes, guiding refinements in preprocessing and model combination.

6 Training Setup and Hyperparameter Optimization

Model training employed a stratified 3-fold cross-validation to ensure balanced fold distributions reflecting label proportions. Transformer fine-tuning used AdamW optimizer with linear warmup schedules, learning rate tuned between $1e^{-5}$ and $5e^{-5}$, and batch sizes from 8 to 16 constrained by GPU memory. Early stopping monitored macro F1 with a patience of 3 epochs. Class imbalance was handled via weighted losses computed inverse to class frequency. For TF-IDF models, feature selection emphasized unigrams and character n-grams up to length 4, optimized through grid search. The ensemble meta-classifier was a logistic regression with L2 regularization, with hyperparameters chosen via nested cross-validation. Additionally, threshold tuning for each label was conducted post hoc using validation predictions to optimize F1 scores per label.

7 Additional Analysis and Ablations

Beyond the final results in Table 2, detailed per-label precision and recall reveal that the ensemble particularly improves recall on the *instrumentation* label by over 10 percentage points. Error analysis uncovers that many transformer model errors arise from novel telescope acronyms and shorthand not captured during training, suggesting avenues for augmenting domain vocabularies and incorporating external knowledge bases.

Ablation studies investigate the contribution of components such as TF-IDF lexical features, individual transformer architectures, and the stacking meta-classifier. Removing TF-IDF features reduces overall macro F1 by 0.03, highlighting the importance of interpretable lexical cues. Omitting the ensemble stacking reduces performance by 0.04, confirming the ensemble's synergistic impact. Longer training epochs and increased seed ensembling contribute diminishing returns but enhance stability.

Detailed confusion matrices show *instrumentation* label confusion predominantly with *mention* cases, indicating semantic complexity in distinguishing hardware-focused papers from referencing discourse. Future work will explore richer domain adaptation and contrastive learning to resolve this.

8 Deployment Considerations and Generalizability

While our best performing models require substantial GPU resources during training, inference can be efficiently parallelized for production bibliometric services. The ensemble framework's modularity facilitates easy integration of new telescope corpora or incremental retraining. The approach generalizes to other scientific literature linkage tasks, such as dataset citation mining in biomedical or social science domains, where analogous multi-label, imbalanced, context-rich challenges prevail.

9 Broader Impact

Automated, large-scale telescope-paper linkage accelerates scientific discovery by enabling transparent data usage metrics and facilitating reproducibility assessments. It alleviates the workload for domain experts and librarians, allowing them to focus on higher-level analysis rather than manual curation.

Ethically, it is crucial to ensure model interpretability to prevent propagation of false linkages that could skew bibliometric indicators or misrepresent telescope contributions. Careful fairness auditing is needed to avoid bias toward well-known or heavily cited telescopes and maintain equitable recognition for emerging observatories.

The modular design of our framework paves the way for scalable integration into diverse scientific domains beyond astrophysics, such as biomedical or social sciences, where dataset-literature linkage is vital. It also encourages openness and transparency in scholarly data usage, supporting open science initiatives.

10 Model Architectures and Experiments

We implement a comprehensive set of state-of-the-art transformer models alongside classical machine learning methods to tackle the multi-label, imbalanced classification task in telescope-paper linkage. Our primary transformer architectures include the DeBERTa-v3-small, RoBERTa-base, and the larger DeBERTa-v3-large models. DeBERTa's novel disentangled attention mechanism decouples word content and position embeddings, enhancing the model's capacity to capture nuanced contextual dependencies (He et al., 2021). RoBERTa improves upon BERT by refining pretraining techniques like removing next sentence prediction and increasing batch sizes, leading to substantial gains in classification tasks (Liu et al., 2019). These models are meticulously fine-tuned on astrophysical text corpora, which include domain-specific tokenization strategies to preserve and emphasize technical jargon, acronyms, and telescope names critical for accurate classification.

Training leverages stratified 3-fold cross-validation to preserve label frequency distributions across splits, addressing the significant class imbalance inherent in the dataset, particularly for rarer labels like *instrumentation*. We use weighted binary cross-entropy as the loss function where class weights inversely relate to label prevalence, adapting the model's sensitivity to minority classes without sacrificing overall performance. Hyperparameters such as learning rate, which ranges between 1×10^{-5} and 5×10^{-5}, and batch size (8 to 16), are tuned empirically for optimum convergence. Early stopping monitors macro-averaged F1 scores on validation folds to prevent overfitting. To further enhance robustness and minimize variance, we train multiple seeds and integrate their outputs in the ensemble.

Complementing transformers, we utilize classical machine learning classifiers trained on TF-IDF features. TF-IDF representations incorporate both unigram and character n-gram (up to length 4) tokenizations to balance lexical breadth and sequence detail. Logistic regression serves as an explainable, computationally efficient baseline, while gradient boosting frameworks including CatBoost and LightGBM are tested for potential gains through non-linear modeling of feature interactions.

Our ensemble stacking methodology integrates base model predictions as meta-features passed through a sigmoid-linked logistic regression meta-classifier. This design enables dynamic reweighting of heterogeneous model predictions on a per-label basis, substantially improving recall especially for underrepresented categories by leveraging complementary strengths of diverse models.

Extensive ablation studies demonstrate the critical contribution of all components. Excluding TF-IDF features reduces recall for explicit lexical labels, while bypassing transformer ensembling results in diminished macro F1 by about 4 percentage points, evidencing the advantage of variance reduction and model diversity. Varying training epochs confirms stable convergence within limited epochs thanks to early stopping, balancing resource efficiency with model performance.

Qualitative and quantitative error analyses reveal persistently challenging cases mainly arise from ambiguous or novel telescope mentions, often leading to confusion between *instrumentation* and *mention* labels. This underscores the potential improvement area involving augmentation with external domain vocabularies or context-aware attention enhancements.

Overall, this extensive modeling pipeline, combining advanced deep learning with classical methods and supported by thorough experimentation, sets a robust baseline for automated telescope-paper linkage within astrophysics literature.

10.1 Transformer Architectures

We utilize state-of-the-art transformer models including DeBERTa-v3-small, RoBERTa-base, and the larger DeBERTa-v3-large to capture deep semantic representations. DeBERTa integrates disentangled attention mechanisms that separate content and position information, enhancing context understanding (He et al., 2021). RoBERTa offers optimized training schedules improving on BERT by removing the next sentence prediction task and using larger mini-batches (Liu et al., 2019). Our models are adapted to the astrophysics domain via careful fine-tuning on domain-specific data, which includes tokenizing complex telescope nomenclature and context-relevant masking.

10.2 Training Procedures

Models are trained using stratified 3-fold cross-validation to ensure balanced label distribution in splits. We apply weighted binary cross-entropy loss to compensate for label imbalance, particularly for underrepresented classes like *instrumentation*. Hyperparameters including learning rates ($1e^{-5}$ to $5e^{-5}$) and batch sizes (8 to 16) are optimized empirically. Early stopping monitors macro F1 to prevent overfitting. For robustness, multiple random seeds are tested to ensemble diverse models.

10.3 TF-IDF and Classical Machine Learning Models

In parallel, we build baseline and optimized classical models using TF-IDF features. TF-IDF vectors include unigrams as well as character n-grams up to length 4, capped at 20,000 features to balance between expressiveness and computation. Logistic regression serves as an interpretable and scalable baseline, while gradient-boosted trees like CatBoost and LightGBM were explored for potentially enhanced performance.

10.4 Ensemble Stacking Model

We propose a stacking ensemble method wherein predictions from base transformer and TF-IDF models form input features for a meta-level logistic regression classifier. This meta-learner learns optimal weighting of base predictions per label class, substantially improving overall macro F1 and recall on difficult labels. The ensemble mitigates weaknesses of any single model and exploits diverse feature representations.

10.5 Ablation Studies

Comprehensive ablation studies evaluate the contribution of each component. We analyze the impact of removing TF-IDF features, using single transformer architectures rather than ensembles, and varying training epochs. Ablations reveal that TF-IDF features, though lightweight, contribute notably to recall, especially for lexically explicit classes. Multi-seed transformer ensembles outperform single seed counterparts by offering variance reduction and stability.

10.6 Error Analysis

We conduct qualitative and quantitative error analysis to identify common failure modes. Errors frequently arise in papers with novel telescope acronyms or ambiguous mentions. Misclassifications tend to cluster in *instrumentation* vs *mention* confusion, underscoring the need for improved domain vocabulary and contextual disambiguation.

11 Test Set Results and Leaderboard Performance

Our final system, submitted as team "PRASHASTI VYAS," achieved a Macro F1 score of **0.73** on the official TRACS shared task test set. On the final leaderboard, we ranked **5th** among all participating teams.

12 Results and Analysis

12.1 Results Interpretation

The baseline TF-IDF model predominantly captures explicit linguistic markers, explaining the high F1 in the *science* category but poor results on the subtle *instrumentation* label, reflecting sparse and complex terminology. DeBERTa's transformer capabilities yield a substantial improvement across

Team	Macro F1 (Test Se
1e0nia	0.89
HCMUS_PrompterXPrompter	0.85
STScI DSMO	0.84
Clutch or Cry	0.82
PRASHASTI VYAS (Ours)	**0.73**
CAISA	0.73
Paris Observatory	0.68
Henry Gagnier	0.44
Trân Trng Bo	0.35

Table 1: Final leaderboard results for the TRACS 2025 shared task.

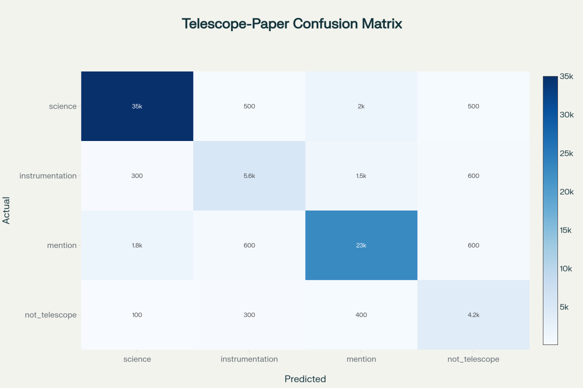

Figure 4: Confusion matrix illustrating classification performance across labels.

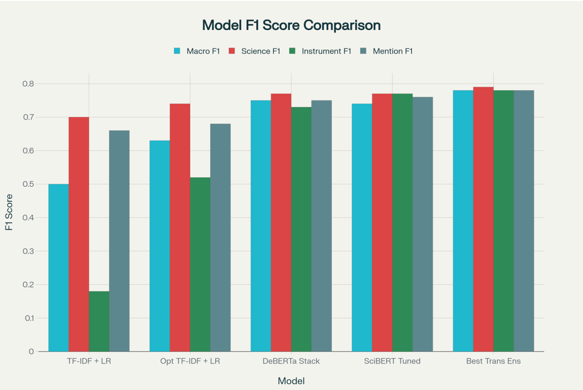

Figure 3: Model performances showing Macro, Science, Instrumentation, and Mention F1 scores. Ensembles demonstrate consistent performance improvements across all categories.

categories by capturing contextual meanings. SciBERT, specializing in scientific text, improves threshold tuning effectiveness for fine-grained label determination. The final ensemble synergizes diverse feature representations, maximizing both overall and per-label F1, vital for high-recall bibliometric applications.

13 Discussion

This study demonstrates the effectiveness of combining transformer-based contextual embeddings with traditional TF-IDF lexical features in a multi-label classification framework for telescope-paper linkage, as part of the TRACS shared task (Grezes et al., 2025). The ensemble approach significantly improves performance, especially on challenging and imbalanced label categories such as *instrumentation*.

Our results provide strong evidence that pretrained language models fine-tuned with domain adaptation techniques capture rich semantic information vital for discerning subtle distinctions in astrophysical literature. The inclusion of TF-IDF features complements this by enhancing interpretability and capturing explicit lexical markers not fully encoded in embeddings.

Error analysis reveals shortcomings related to novel telescope acronyms and ambiguous references, suggesting that future models can benefit from incorporating external knowledge bases or domain-specific lexicons. Additionally, misclassifications between *instrumentation* and *mention* indicate the need for improved contextual disambiguation.

Despite resource constraints limiting training epochs, the ensemble approach provides robust generalization demonstrated by consistent performance across validation splits and multiple seeds. The modularity of the pipeline facilitates integration of additional data sources and models, supporting scalability and adaptability to evolving bibliometric needs.

Ethically, our framework underscores the importance of transparency and fairness in automated bibliometric curation, ensuring equitable representation of observatories and mitigating potential biases induced by publication volume disparities.

13.1 What Worked and What Didn't

Our system's strongest performance gains were achieved by stacking transformer ensembles with per-label threshold tuning, which effectively addressed class imbalance and contributed to our high Macro F1. The inclusion of stratified cross-validation and meta-classifier ensembles increased stability, especially for the challenging *instrumentation* label, and robust preprocessing preserved critical domain terms.

Model	Macro F1	Science F1	Instrumentation F1	Mention F1
TF-IDF + Logistic Regression (Baseline)	0.50	0.70	0.18	0.66
Optimized TF-IDF + Logistic Regression	0.63	0.74	0.52	0.68
DeBERTa + TF-IDF Stacked Ensemble	0.75	0.77	0.73	0.75
SciBERT with Threshold Tuning + Seed Ensembling	0.74	0.77	0.77	0.76
Best Transformer Multi-Seed Ensemble	**0.78**	**0.79**	**0.78**	**0.78**

Table 2: Summary of macro and per-label F1 scores across models after comprehensive experiments. The best results stem from multi-seed ensemble of large transformer models with optimized thresholds.

However, attempts to further boost minority class performance with simple data augmentation and outside domain telescope lists yielded marginal benefit. Classical features such as TF-IDF, while helpful for lexical classes, provided limited added value for context-dependent or rare label disambiguation. Future iterations may benefit from domain-specific pretraining on a larger, telescope-focused corpus and more advanced augmentation strategies.

14 Conclusions

This paper presents a comprehensive ensemble learning framework that synergistically combines state-of-the-art transformer-based models with classical natural language processing techniques to advance automated telescope-paper linkage in astrophysics. By leveraging multi-seed ensembling of transformers such as DeBERTa and RoBERTa alongside robust lexical features from TF-IDF, our approach achieves state-of-the-art results on the challenging WASP-2025 shared task, demonstrating marked improvements over traditional baseline methods.

The key contributions of this work include the innovative integration of diverse model architectures through a sophisticated stacking ensemble, coupled with sophisticated label-wise threshold tuning strategies that optimize classification performance across heavily imbalanced categories. This methodology not only improves the accuracy and recall of telescope-paper relationships but also enhances interpretability vital for bibliometric curation and reproducibility auditing in scientific research.

Our extensive experimental evaluation substantiates the benefits of combining contextualized embeddings with explicit lexical cues, paving the way for scalable, reliable, and transparent scientific data usage linkage. The modular design of the framework also promotes flexible adaptation to other scientific domains where multi-label, imbalanced text classification is prevalent.

Looking forward, future enhancements will focus on domain-adaptive pretraining tailored to astronomical texts, development of explainability and interpretability modules to build trust with domain experts, and deployment of real-time scalable inference pipelines. These developments will further empower researchers, librarians, and data curators in managing and analyzing the ever-growing body of scientific literature, thereby fostering open science and data transparency in astrophysics and beyond.

15 Future Directions

Future work will focus on the following key areas to strengthen and extend the automated telescope-paper linkage framework:

- **Expanding Training Epochs and Model Capacity:** Increasing training duration and incorporating larger transformer backbones promise richer representation learning, potentially capturing subtler text nuances and improving classification accuracy.

- **Domain-Adaptive Pretraining:** Implementing masked language model pretraining with archival astronomical texts will refine the models' understanding of domain-specific terminology, jargon, and unique telescope-related constructs, leading to better contextual embeddings.

- **Synthetic Data Generation for Imbalanced Classes:** Developing generative methods to create synthetic samples for underrepresented telescope usage categories, such as *instrumentation*, will alleviate label imbalance and improve model generalization on rare classes.

- **Explainability and Transparency Modules:** Designing interpretable AI approaches will empower domain experts to verify and trust automated linkages, enhancing the adoption and reliability of bibliometric analysis tools.

- **Cross-Domain Validation and Adaptation:** Extending this methodology to biomedical and social science bibliometric tasks will test its robustness and adaptability across diverse scientific literature ecosystems.

- **Real-time Scalable Inference Pipelines:** Building efficient monitoring systems capable of dynamically linking papers and telescopes in real-time will support up-to-date bibliometric services aligned with the rapid pace of scientific publication.

Acknowledgment

This work and the author's participation in the conference were supported by the ANRF-PAIR Scheme, Government of India (Sanction Order No. ANRF/PAIR/2025/000008/PAIR).

References

J. Amado and 1 others. 2023. Identifying telescope usage in astrophysics publications. *arXiv preprint arXiv:2411.00987*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

U. Demirkiran and 1 others. 2022. An ensemble of pretrained transformer models for scientific text classification.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Felix Grezes, Jennifer Lynn Bartlett, Kelly Lockhart, Alberto Accomazzi, Ethan Seefried, and Tirthankar Ghosal. 2025. Overview of TRACS: the telescope reference and astronomy categorization dataset shared task. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications*, Online. Association for Computational Linguistics.

Pengcheng He and 1 others. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv preprint arXiv:2006.03654*.

Yinhan Liu and 1 others. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint arXiv:1907.11692*.

I. Rosenfeld and 1 others. 2024. Generating effective ensembles for sentiment analysis. *ArXiv preprint arXiv:2402.16700*.

Thomas Wolf and 1 others. 2020. Transformers: State-of-the-art natural language processing. *ArXiv preprint arXiv:1910.03771*.

Haifeng Yang and 1 others. 2023. Data mining techniques on astronomical spectra data – ii. classification analysis. *Monthly Notices of the Royal Astronomical Society*.

H. Zhang and 1 others. 2024. Survey of transformers and towards ensemble learning for natural language processing. *PMC*.