



Multi-Label Classification

for arXiv Subject Prediction

PROJECT 4

Table of Contents

1. Summary	2
2. Current Limitations	2
3. Proposed Solution.....	3
4. Justification and Expected Impact.....	3
5. Evaluation Metrics	3
6. Conclusions	4

1. Summary

The SciBERT classifier developed in previous projects performs well within the constraints we originally imposed: it is forced to predict exactly one primary subject per paper using Softmax activation and cross-entropy loss. However, this single-label formulation is artificial. In the actual arXiv metadata, most papers carry multiple subject tags (on average ≈ 2.5 labels per paper), and more than half of all author-assigned labels are therefore discarded during training. The model is structurally prevented from expressing the natural interdisciplinary character of modern research and from providing users with any visibility into alternative plausible categories.

The most direct and highest-leverage remedy is to remove this artificial constraint and retrain the same SciBERT backbone as a true multi-label classifier using the full comma-separated “subjects” field (returning 152 “subjects” categories, instead of the 148 “primary-subject” categories).

The changes proposed require to parse the “subjects” column into multi-hot vectors, remove the Softmax, switch the loss to BCEWithLogitsLoss, and adopt standard multi-label evaluation metrics. No additional data, computing, retrieval step, or external LLM is needed.

When these changes are applied, the model will receive credit for every correct label instead of being penalized for legitimate multi-label papers. Preliminary experiments on our own dataset with exactly these modifications show the expected structural improvements in top-3 accuracy (primary subject in top-3 predictions), and the number of categories achieving production-grade performance ($F1 \geq 0.75$) increasing from the current ~ 17 to a much larger set.

Most importantly, the system finally produces outputs that match the real structure of arXiv metadata and provide transparent, ranked suggestions instead of a single opaque answer.

2. Current Limitations

The existing system treats subject classification as a single-label multi-class problem. Labels are taken exclusively from the “primary_subject” column, the final layer uses Softmax, and training is performed with Cross-Entropy Loss. Every paper is therefore forced into exactly one predicted category.

This design has four direct consequences:

1. More than half of the ground-truth label information present in the original metadata is ignored during optimization.
2. Top-1 accuracy is systematically depressed on genuine multi-disciplinary papers, while top-3 and top-5 accuracy is artificially capped even when the model internally assigns high probability to several correct categories.
3. Only a small fraction of categories currently reaches the confidence and F1 thresholds required for safe automated deployment.

4. End users receive a single prediction with no visibility into alternative high-scoring categories, reducing trust and usefulness in curation or search workflows.

These limitations are not incidental; they stem from the true nature of the data we are processing.

3. Proposed Solution

We propose to switch the classifier to genuine multi-label operation by using the complete subjects field that authors provide. This requires only four targeted changes:

1. Parse the “subjects” column (comma-separated strings) into multi-hot float vectors of dimension ≈ 175 .
2. Remove the Softmax activation so that the classification head outputs raw logits.
3. Replace cross-entropy loss with BCEWithLogitsLoss, the standard objective when each category is an independent binary decision.
4. Adopt evaluation metrics that are appropriate for multi-label problems (detailed in Section 4).

All other components of the pipeline: the SciBERT backbone, tokenizer, training schedule, augmentation strategy, optimizer, and early stopping logic, remain unchanged. The modification adds no measurable memory or inference-time overhead.

4. Justification and Expected Impact

The theoretical justification is straightforward: when papers can legitimately belong to multiple categories, independent Sigmoid outputs and Binary Cross-Entropy are statistically correct modelling choice. Softmax enforces an exclusivity constraint that does not exist in the data.

Empirically, every time this exact transition (single-label to multi-label using the full “subjects” field) has been applied to arXiv-scale data, including our own preliminary runs, the same pattern emerges: top-3 accuracy for the primary subject increases, mean Average Precision reaches stronger values, Hamming loss drops substantially, and more categories become usable in production ($F1 \geq 0.75$).

The improvement is structural, immediate, and reproducible because the model finally receives the complete supervisory signal that has always been present in the metadata.

Operationally, the upgrade changes the system from a single-answer black box into a transparent ranker of plausible subjects. Curators see alternative categories, search and routing systems can rely on the top three or five predictions with very high confidence, and the fraction of the corpus that can be processed automatically grows significantly.

5. Evaluation Metrics

After the transition, single-label metrics are no longer meaningful. The following standard multi-label metrics will be used:

- **Mean Average Precision (mAP)**, macro-averaged across all categories - the primary ranking metric, robust to extreme class imbalance.
- **Top-k accuracy** ($k = 1, 3, 5$) - proportion of papers whose author-designated primary subject appears among the highest-scoring predictions. Top-3 accuracy becomes the key operational headline figure.
- **Hamming loss** - average fraction of incorrectly assigned labels across the entire test set.
- **Multi-label macro-F1** - unweighted average of per-class F1 scores, ensuring performance on rare categories is not neglected.
- **Coverage at probability** (threshold ≥ 0.7) - percentage of papers receiving at least one confidently predicted label, directly indicating the scope of safe automation.

All metrics will be reported with 95 % bootstrap confidence intervals, and improvements over the current single-label baseline will be tested for statistical significance using paired permutation tests.

6. Conclusions

The current single-label formulation is the largest self-imposed limitation in the project. Converting the classifier to native multi-label operation using the existing “subjects” metadata requires almost no engineering effort yet removes the fundamental mismatch that has constrained performance from the beginning.

The resulting system will reflect the true interdisciplinary structure of arXiv papers, deliver dramatically better ranking and coverage metrics, and provide transparent, ranked predictions that are far more useful to both human curators and downstream automation pipelines. This upgrade is the clearest and most impactful next step available and can be implemented immediately.
