Artificial Intuition: Efficient Classification of Scientific Abstracts

Harsh Sakhrani, Naseela Pervez, Anirudh Ravi Kumar, Fred Morstatter
Information Sciences Institute, Viterbi School of Engineering, University of Southern California
Alexandra Graddy Reed
Sol Price School of Public Policy, University of Southern California
Andrea Belz
Information Sciences Institute, Viterbi School of Engineering, University of Southern California

Abstract

It is desirable to coarsely classify short scientific texts, such as grant or publication abstracts, for strategic insight or research portfolio management. These texts efficiently transmit dense information to experts possessing a rich body of knowledge to aid interpretation. Yet this task is remarkably difficult to automate because of brevity and the absence of context. To address this gap, we have developed a novel approach to generate and appropriately assign coarse domain-specific labels. We show that a Large Language Model (LLM) can provide metadata essential to the task, in a process akin to the augmentation of supplemental knowledge representing human intuition, and propose a workflow. As a pilot study, we use a corpus of award abstracts from the National Aeronautics and Space Administration (NASA). We develop new assessment tools in concert with established performance metrics.

1 Introduction

Analyzing technical documents is a crucial strategic task, enabling the management of research portfolios, tracking investment trends, and exploring scientific advancement. On a more tactical level, it can aid the preliminary screening of scientific abstracts in systematic reviews (Buchlak et al., 2020; Rios and Kavuluru, 2015; Ambalavanan and Devarakonda, 2020).

Several approaches are possible. First, authors can label their own work, but this presents several challenges: (1) authors that self-label their own texts may make idiosyncratic decisions, (2) authors in close disciplines may use different terms for related concepts, such as "robotics" and "autonomy", and (3) multidisciplinary projects may require novel or multiple labels.

A second method is to impose an external ontology. However, these schemes often have both fine- and coarse-granularities (e.g., "networks" versus "ad-hoc networks"). Another concern is that the scheme simply lacks appropriate labels, especially for emerging fields.

Automated processes do exist. Those with a large number of parameters are now customizable at lower computational cost (Hu et al., 2021; Ben Zaken et al., 2022). Although dedicated pretrained models can yield robust results, they incur significant expenses in manual annotation due to reliance on supervised learning (Beltagy et al., 2019; Chang et al., 2008; Cohan et al., 2020).

In summary, we face two distinct needs in the analysis of scientific documents: (1) a unified, coarse-grained, non-overlapping taxonomy, tailored to uniquely classify a set of documents; and (2) an unsupervised methodology that circumvents the reliance on manual annotation while effectively managing the peculiarities of scientific text. These challenges are exacerbated for abstracts.

In manual labeling, an expert's rapid progress often hinges on integrating prior knowledge, crucial for effective comprehension (Reid Smith and Hammond, 2021). In so doing, the expert rapidly identifies the phrases conveying the most information and uses those for classification. Importantly, this process is not a simple frequency or statistical analysis; indeed, the most important phrase may appear only once. Moreover, multigrams carrying high semantic value may not appear systematically in the same place in a sentence or paragraph.

Here we describe "artificial intuition," a method mimicking the expert's process to execute two objectives: generating an optimal label space and producing accurate predictions within this new space. We integrate tools into a novel workflow to identify important terms, augment them with relevant background information, then aggregate these enhanced documents into clusters for classification purposes.

As a pilot case to evaluate our methodology, we analyze award abstracts of federally funded projects from the National Aeronautics and Space Administration (NASA) Small Business Innova-

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 191–201

tion Research (SBIR) Program. We obtain domain knowledge by extracting and ranking the abstract's keywords / keyphrases (which we will collectively refer to as "keywords"). We generate metadata for these keywords in a zero-shot setting and derive embeddings for the keyword-metadata concatenations using a pre-trained Sentence Transformer.

For label space generation, we implement a clustering process that represents the task of organizing awards into funding themes. This method not only clarifies the thematic organization of the documents but also reveals the hierarchical relationships between different topics. We introduce a novel evaluation scheme to assess whether the label set comprehensively spans the document space and can serve as a set of basis vectors.

To predict labels, we reinterpret the multilabel-classification problem as a semantic matching challenge wherein the document space is characterized by the keyword-metadata concatenation and the label space is described by the element closest to the centroid for each cluster. This retrieval-based perspective allows for flexibility in adapting to new label spaces without the need for retraining.

This framework accommodates various levels of parsimony, which we explore extensively in our experiments. Finally, using our test sample, we demonstrate the efficacy of our prediction methodology and quantify the performance.

2 Related Work

Various methodologies have been proposed for text classification. Bayesian approaches (Tang et al., 2016) classify the text by extracting features. One method is to first select document features with discriminative power, then compute the semantic similarity between features and documents (Zong et al., 2015), but this becomes more difficult as the number of features grows. Support Vector Machines (SVMs) can be used for document classification (Cai and Hofmann, 2004). However, these approaches are constrained by the requirement for manual feature engineering, limiting their ability to capture the complexity of natural language.

New deep learning techniques have advanced scientific document classification. Neural network-based architectures (Lee and Dernoncourt, 2016), particularly Convolutional Neural Networks (CNNs) (Sun et al., 2019) and Recurrent Neural Networks (RNNs) (Xun et al., 2019; Liu et al., 2016), outperform some traditional machine learning methods. These models automatically learn feature representations from data, capturing both the semantic and syntactic nuances of text.

These methods presume that documents are related to only one label. Newer approaches (e.g., (Liu et al., 2017; Song et al., 2022; Xiao et al., 2019; Blanco et al., 2019; Chang et al., 2020)) classify documents with multiple labels, and one alternative attempts to map 10,000 fine-grained labels for scientific documents (Zhang et al., 2022a) although most methods consider 10-50 coarse labels. These models are incompletely validated because many real-world datasets will have limited or poorly labeled data.

Weakly supervised learning and zero-shot learning (ZSL) models do not use annotated data. Some pre-trained language models demonstrate impressive performance in zero-shot document classification (Devlin et al., 2019; Beltagy et al., 2019; Liu et al., 2019) and can be used to assign multiple labels to a given document (Yin et al., 2019). On the other hand, hierarchical multi-class methods can use just class names - without training examples - as supervision (Shen et al., 2021; Zhang et al., 2022b). Large language models trained on scientific data, such as Galactica (Taylor et al., 2022) and SciNCL (Ostendorff et al., 2022), can be used to assign labels to a scientific document. Many approaches use metadata, such as generic descriptions, as supervision for further classification (Zhang et al., 2023). However, these methods are still potentially subject to noise. Here we describe a method to identify keywords and derive context-specific metadata to improve classification accuracy, particularly for short abstracts.

3 Approach

3.1 Problem Formulation

The scientific literature tagging task can be conceptualized as a multi-label classification (each paper can be relevant to more than one label) problem, where all candidate tags (e.g., "Aerodynamics," "Superconductance/Magnetics") constitute a label space Y of arbitrary size. We seek to:

- Construct a new label space Y comprising coarse-grained labels and aggregating correlated labels (e.g., merging "Optics" and "Photonics" into "Optical technologies").
- Develop an unsupervised multi-label classifier that can effectively map an abstract to the new

label space Y.

A simplistic approach would utilize a pre-trained language model to encode each document and label, generate their embeddings, and then conduct a nearest neighbor search in the embedding space. However, this method encounters two primary challenges: (1) the existing language models are largely trained on general English text that does not discern technical terms, and (2) analogous labels (e.g., "networking" and "ad-hoc networks") confound the results. One might augment the label embedding process with generic metadata, such as a brief description from Wikipedia or using solutions like Positive Instance Feature Aggregation (PIFA) (Yu et al., 2022).

Instead, we seek to generate a context-specific glossary. This has the added advantage that the labels can be fine-tuned, converting a multi-label problem into a simpler system. For instance, a thermal protection system (TPS) consists of materials suited to handle extremely high temperatures. In a conventional classification scheme, this might require two labels, such as "materials" and "temperature." In contrast, we create a system such that "thermal protection system" is itself sufficient to serve as the only label. This is possible only with a label space customized to the knowledge domain.

3.2 Implementation Components

- Yet Another Keyword Extractor (YAKE) (Campos et al., 2020) is a lightweight, unsupervised keyword extraction algorithm that uses statistical properties and contextual information.

- Mistral 7B is a Large Language Model (LLM) with strong performance of Llama-2 13B on key benchmarks (Jiang et al., 2023).

- Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) iteratively selects candidate items that simultaneously maximize their relevance to the query and their novelty compared to previously selected items.

- Sentence Transformer (S-Transformer)[1] (Reimers and Gurevych, 2019) constructs dense vector representations of sentences to enable efficient comparison of text semantics.

[1]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

3.3 Document Corpus

The NASA SBIR program publishes abstracts of funded projects. We used 1,230 abstracts from 2010 to 2015 extracted online from the publicly available archive[2]. The average abstract length is about 450 words. All abstracts were pre-processed by removing stop words, which were variations and combinations of: "NASA", "space", "mission(s)", "research", "SBIR", "spacecraft", "future", and "science". These words and multigrams comprising these words appeared in a large number of the abstracts, and therefore they provided little information to assist in classification. We randomly drew 100 abstracts (roughly 10%) for manual classification, described below.

3.4 Label Space Generation

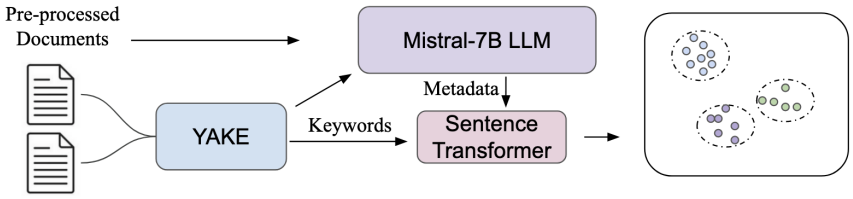
We generate the label space as illustrated in Figure 1. Initially, pre-processed abstracts are submitted to YAKE to extract keywords. One hyperparameter of our workflow is the number of keywords $\hat{c}$. We estimated that $\hat{c}$ should be approximately 5 as it represents 1-2% of the abstract length. We confirmed that the F1 results, described in more detail below, showed a general lack of sensitivity to this parameter (Figure 2), and therefore we set $\hat{c} = 5$ in our main analyses.

We sought to supplement these keywords with contextual definitions to form metadata. We used Mistral-7B Instruct v0.2 with hyperparameters set at default values and submitted the following prompt:

> *Given the scientific abstract and the keywords that have been extracted for the document, provide a concise meta data/prior information for every keyword in context of the document. Incorporate any extra knowledge that can help classify the document to relevant topics.*

This combined data-keyword concatenation is processed using the S-Transformer model to produce embeddings. A critical aspect of this process is that the metadata generated for each keyword is tailored specifically to the context of the related document, ensuring that the embeddings are context-specific rather than generic. We use k-means clustering (Habibi and Popescu-Belis, 2015) to partition these embeddings into clusters, represented by the keyword closest to their centroids and

[2]sbir.gov



Cluster	Label
1	Optics
2	Flight System
3	Technology
$\hat{k}$	Remote Sensing

Figure 1: Label Space Generation flowchart. The clusters are named with the keyword closest to the cluster centroid.

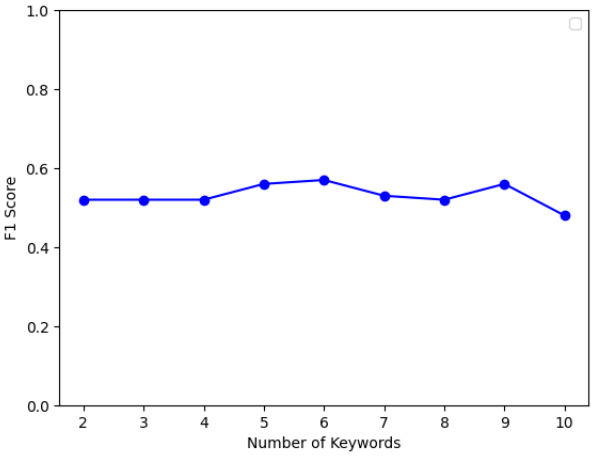


Figure 2: Variation of F1 score with the number of keywords at the threshold of top 1%.

effectively summarizing each cluster's thematic focus. This method approximates the scheme by which such abstracts would be sorted in a funding portfolio.

Unlike $\hat{c}$, the number of clusters $\hat{k}$ requires closer examination. We seek a parsimonious model that minimizes the number of labels per document. In practice, we seek to organize approximately 1,000 documents into approximately 10-20 classes. In addition to making this a tractable problem, it adequately represents the portfolio management process.

3.5 Annotation Task Design

We conducted a manual annotation task to label the test set of the NASA SBIR abstracts. We presented the annotator, a NASA expert, with a scientific abstract and the generated label set. The annotator was instructed to assign a label to the scientific abstract only if one of the presented labels was appropriate, and to leave it unlabeled otherwise. The same documents were labeled for each configuration for consistency.

4 Results

4.1 Label Space Orthogonality: Redundancy

Our first task is estimate the degree of overlap within the label space. To do so, we define the redundancy, $\mathcal{R}$, as a measure of the orthogonality between labels. This figure of merit (FOM) is intrinsic to the label space and assessed independently of individual document projections.

The labels are transformed into normalized embeddings using the S-Transformer model, resulting in a label matrix $\mathcal{L}$ of dimensions $\hat{k} \times v$ (in our case, $v = 768$). Each element $\mathcal{L}_{ij}$ represents the j-th dimension of the i-th label embedding.

To measure the orthogonality, we calculate the cosine similarity between each pair of distinct label embeddings. If the labels are orthogonal and distinct, the cosine similarity should approach 0; on the other hand, two labels capturing closely related ideas will give a cosine similarity that approaches 1. Formally, for normalized label vectors $\mathcal{T}_i$ and $\mathcal{T}_j$ in $\mathcal{L}$, we define redundancy $\mathcal{R}$ as the maximum cosine similarity among all pairs:

$$\mathcal{R} = \max_{i \neq j}(\text{cosine similarity}(\mathcal{T}_i, \mathcal{T}_j)) \quad (1)$$

where

$$\text{cosine similarity}(\mathcal{T}_i, \mathcal{T}_j) = \frac{\mathcal{T}_i \cdot \mathcal{T}_j}{\|\mathcal{T}_i\|\|\mathcal{T}_j\|}$$

A value of $\mathcal{R}$ close to 0 is desirable because orthogonal label embeddings suggest that each label contributes unique information without redundancy. Conversely, a value of $\mathcal{R}$ approaching 1 shows that at least one pair of labels shares a high degree of overlap. Overlap implies that multiple labels may be describing similar features within the documents, thus complicating the interpretability and utility of the label space. Our goal is to represent each key concept with a unique label.

To understand the redundancy in our basis vector set, we executed the label space generation process

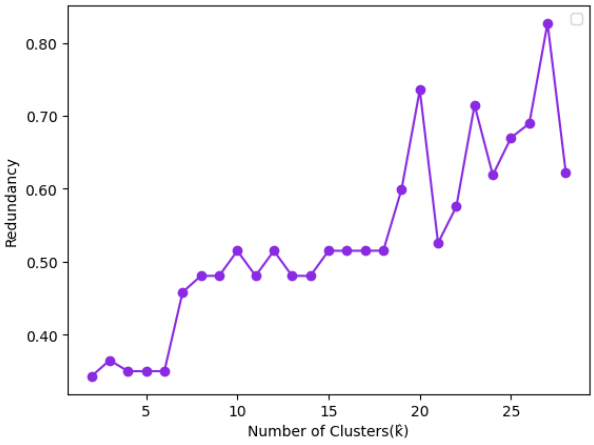


Figure 3: Variation of redundancy $\mathcal{R}$ with the number of clusters $\hat{k}$.

but systematically varied $\hat{k}$. We then evaluated $\mathcal{R}$ for each label space. $\mathcal{R}$ increased with $\hat{k}$ (Figure 3), as expected. Notably, we identified three general regimes: At low $\hat{k}$, $\mathcal{R}$ was fairly flat and low. The labels do not overlap. At approximately $\hat{k} = 8$, the redundancy increased to a new plateau. At much higher values of $\hat{k}$ (18 and higher), this FOM entered a regime in which the value dramatically oscillated.

We therefore conclude that at very low cluster numbers ($\hat{k} < 6$), the severely reduced $\mathcal{R}$ indicates that the labels are probably insufficient to describe the document set. At higher values of $\hat{k} > 18$, the risk of overlapping labels increases substantially, but the likelihood that each document is at least minimally described also increases.

4.2 Spanning the Document Space: Coverage

We defined the redundancy $\mathcal{R}$ to characterize the orthogonality of our proposed label space basis vectors. Next, we study how comprehensively these labels describe the documents, essentially determining if our labels can span the document space.

We architected a second workflow (Figure 4). Again we begin with YAKE usage for a single document. We submit these keywords to Mistral-7B for document-specific contextual definitions as supplementary metadata. Both the document itself and the keyword-metadata concatenations are subsequently processed through the S-Transformer model to generate their individual embeddings, refined using MMR. This forms a new keyword embedding matrix $\mathcal{C}$ of dimensions $v \times \hat{c}$, where v (768 in our case) represents the embedding dimension, and the extracted keywords are still parameterized by $\hat{c}$.

Likewise, we still have the label embedding $\mathcal{L}$ of dimensions $\hat{k} \times v$. As our goal is to understand the overlap between the labels and the corpus embeddings, we define a new matrix, termed "coverage", $\mathcal{W}$ with elements w_{ij}:

$$w_{ij} = \sum_{v} L_{iv} C_{vj}, \tag{2}$$

The coverage matrix $\mathcal{W}$ has the resulting dimension $\hat{k}$ x $\hat{c}$, where $\hat{k}$ represents the number of labels and $\hat{c}$ represents the number of keywords. In other words, $\mathcal{W}$ is the projection of the keywords onto the label space. (Strictly speaking, the S-Transformer embeddings of length v can be understood as creating a coordinate system to facilitate projections.) Each w_{ij} element ranges from -1 to 1.

A high value of any element w_{ij} indicates that a label and keyword are highly aligned. Therefore, finding the maximum value that appears in this matrix $\mathcal{W}$ will signify how well the label space describes the keywords of an individual document in the best case. Consequently, we define the coverage $\mathcal{S}$ for a given document d (where d is a member of the document corpus $\mathcal{D}$):

$$\mathcal{S}^d = \max(w_{ij}^d) \tag{3}$$

The coverage for the corpus D is simply the average of the individual documents' coverage:

$$\mathcal{S}^D = \frac{\sum_D S^d}{D} \tag{4}$$

This proposed figure of metric, coverage, provides critical validation that the new label space is pertinent to the knowledge domain encompassed by the documents. One would expect for coverage to be small if the label space is not large enough - namely, for small values of $\hat{k}$. An intermediate regime would appear if each new label adds significant new information. Eventually, a final regime would be reached wherein the new information provided by an additional label is marginal as the segmentation becomes finer, such as comparing 'Chemical Propulsion Technologies' and 'Electronic Propulsion Technologies'. In other words, a general analytical form for coverage should start near 0, then experience rapid growth until the space is largely covered and it tapers off. The corpus coverage is bounded by 1 because the individual documents' coverage is given by a cosine similarity of two normalized vectors, thus limited to 1.

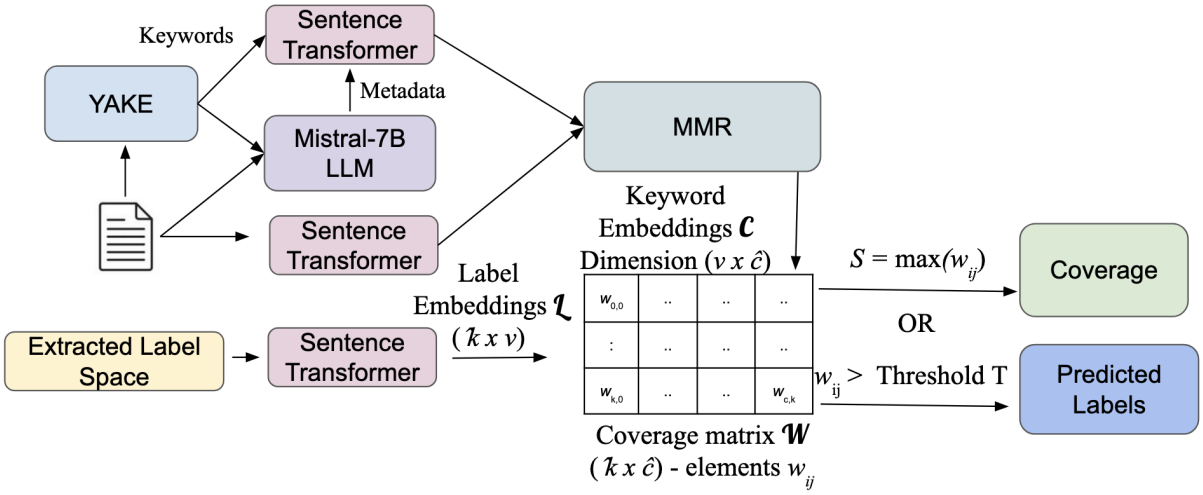


Figure 4: Analysis workflow and use of the coverage matrix $\mathcal{W}$. In one application (final step in green), the element with the maximum value is used to generate the Coverage. The second usage (blue final step) is to extract those values exceeding a specific threshold T for the label prediction task.

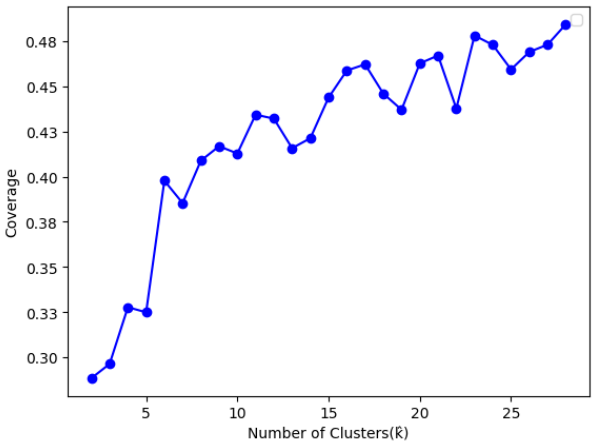


Figure 5: Variation of coverage S with $\hat{k}$.

Table 1: Labels at $\hat{k} = 15$

Advanced Optical Systems
Advanced Photovoltaic Systems
Aeroservoelastic Analysis and Aircraft Systems Analysis
Aeroservoelastic Analysis Tools
Electric Propulsion Systems
Electrolyzers
High Energy Density Electronics
LIDAR Remote Sensing
Multifunctional Composite Materials
Optical Communications Technology
Radiation-Hardened Electronics
Robotic Science Missions
Technologies Fault Management
Thermal Protection Systems
Unmanned Aircraft Systems

We tested this concept by varying the number of clusters $\hat{k}$ from 2 to 28 and evaluating coverage $\mathcal{S}$ for each newly developed label space. As $\hat{k}$ increased, the labels did indeed relate well to the documents, as represented by keywords (Figure 5). In addition, the variation revealed a generally asymptotic form, as expected.

4.3 Label Assignment: Precision and Recall

We seek to create a label space with high coverage, indicating relevance; and low redundancy or overlap. However, these measures act in opposition as higher coverage naturally can lead to greater redundancy. That is, these two measures form a trade space in which we strive to optimize $\hat{k}$.

We revisited the workflow generating the coverage matrix (Figure 4) and developed a prediction pipeline, mirroring the initial process through the creation of the coverage matrix $\mathcal{W}$.

In the coverage study, we took the maximum w_{ij} value to characterize the space. Here, we seek to find *all* relevant values of w_{ij}. To operationalize this, we analyze the distribution of all w_{ij} values and establish a threshold T, which defines the minimum percentile to be used as a filter for the w_{ij} values, effectively distinguishing between significant and negligible overlaps. For instance, setting $T = 1\%$ means we retain only the top 1% of the w_{ij} values, which is more restrictive than setting $T = 10\%$. In practical terms, for a system of 5 keywords and 15 labels, a 1% threshold would retain just one label (top 1% of 5x15 = 75 matrix elements results in one). On the other hand, a 10% threshold retains seven elements that could be distributed in

Table 2: Labels at $k = 25$

Advanced Aeroservoelastic Analysis and Rotorcraft Aeromechanics
Advanced Composite and Ceramic Matrix Materials
Advanced ESR Technologies for Space Exploration
Advanced Energy Storage and Power Systems
Advanced Fluid and Thermal Management Technologies
Advanced Laser and Optical Communication Technologies
Advanced Manufacturing Technologies for Aerospace
Advanced Microwave and Remote Sensing Technologies
Advanced Optical Systems for Scientific Missions
Advanced Structural Sensors and NDE Technologies
Advanced Thermal Protection Systems
Airborne Measurement and Sensing Systems
Automation and Control in Robotic Science Missions
Fault Management Technologies
High-End Computing and Data Handling
Highly Capable Propulsion Systems
Innovative Aerospace Structural Design
Innovative Fiber-Optic and Navigational Technologies
International Space Station
LIDAR Remote Sensing Technologies
Mars Sample Return Missions
Radiation-Hardened Electronics and Sensors
Regenerative Life Support Systems
Solar Power Technologies for Advanced Energy Solutions
Unmanned Aircraft Systems Operations

various ways. For instance, all five keywords might describe label 1, with two of those keywords linked to label 2; or only one keyword could be associated with each of seven labels. As the threshold T gets larger, the variability in possible outcomes increases.

For each document, we select labels associated with the values of w_{ij} that exceed the threshold T. However, to accurately evaluate the classification, a set of 'true' labels is required. While NASA maintains its own taxonomy of approximately 200 labels that could theoretically serve this purpose, the inconsistency in this taxonomy year-to-year and the excessive number of labels compared to our needs complicate its use. Instead, as noted in Section 3.5, we manually aligned the abstracts with our new labels.

Using the three regimes of Figure 3 as a guide, we considered three values for $\hat{k}$ - 4, 15, and 25 - and estimated the usual classification measures of precision, recall, and F1. Moreover, we varied the threshold T, hypothesizing that at low restrictive values of T, these measures should improve as only the most significant overlaps in the coverage matrix would be retained.

At $\hat{k} = 4$, the labels were: Propulsion Technologies, Remote Sensing Technologies, Thermal Protection Systems, and Unmanned Aircraft Systems. However, the manual classification task failed because the labels simply did not describe the abstracts.

At $\hat{k} = 15$, the labels consisted of words generally associated with space technologies (Table 1). Similarly, the $\hat{k} = 25$ generated labels related to space (Table 2); however, in this case the word "Advanced" preceded nearly half the technical topics, suggesting that the semantic content of that word decreases in this context. (Notably, the word "advanced" has been linked to other technical contexts where its semantic content is diluted (Belz et al., 2023)).

To evaluate our method's quality, we set aside $\hat{k} = 4$ and considered differences between $\hat{k} = 15$ and $\hat{k} = 25$. We focused on the F1 score and found that $\hat{k} = 15$ consistently yielded higher scores than the overdetermined space represented by $\hat{k} = 25$ (Figure 6). As a result, we concluded that $\hat{k} = 15$ represented a better set of labels to describe this space.

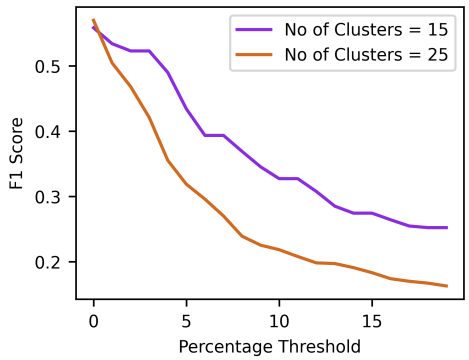


Figure 6: Variation of F1 scores for assigned labels with weights w exceeding the percentile threshold T, as defined in the text.

Our final task was to demonstrate the advantage of augmenting the abstract with the metadata extracted from the additional analysis of the keywords. Using the $\hat{k} = 15$ label space described above, we evaluated the performance of our model with and without the metadata generated by the LLM. We found that the LLM consistently improved the F1 score (Figure 7) for all tested values of the threshold T. This was due to improvement primarily in the precision (Table 3).

5 Discussion and Future Research

Scientific communication is designed to efficiently carry rich information between experts. The abstract of a grant or publication is perhaps the most striking example, wherein sophisticated concepts

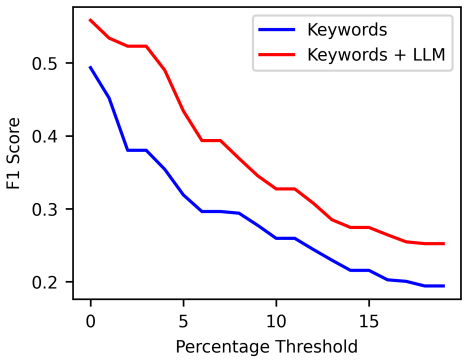


Figure 7: Variation of F1 scores for assigned labels with weights w exceeding the percentile threshold T, as defined in the text, for $\hat{k} = 15$.

Table 3: Precision, recall and F1 scores for $k = 15$ at varying thresholds (T).

T (%)	Precision	Recall	F1
1	0.56	0.56	0.56
5	0.35	0.79	0.49
10	0.22	0.81	0.35
15	0.16	0.86	0.27
20	0.15	0.91	0.25

are conveyed in a relatively dense, short vehicle. Years of study generate a large body of knowledge to guide the expert in a classification process. Indeed, this additional material and the associated judgments underpin the rapid decision-making characteristic of human intuition.

We have sought to replicate that process in an automated methodology. Our unsupervised approach is robust and flexible, enabling its use in various domains. Its independence from specific label sets underscores its adaptability and broad applicability. Our contributions range from applied text processing tasks to economics and public policy, with several interesting directions ahead.

First, we have tested this approach on a relatively narrow set of abstracts by selecting a NASA corpus of documents as the first test case. This exercise should be conducted on benchmark datasets such as Maple (Metadata-Aware Paper colLEction) [3]. This would demonstrate the generalizability of our approach.

Second, a different validation would be to compare these results with those of longer documents. For instance, one could analyze both publication abstracts and the full text. It is not clear if the publications would contain more noise; or perhaps the complete text would carry the metadata such that the LLM task would be less necessary.

In addition, here the manual classification exercise assigned only one label to each abstract as a rigorous test. We have not explored the opportunity to generate multiple labels for a single abstract. Indeed, the $\hat{k} = 25$ data set points to this, as some of the labels (such as "Advanced Thermal Protection Systems") addressed the technology itself, while others described the intended application (e.g., "Mars Sample Return missions"). In the future, we can develop a new weighting scheme addressing this complex classification.

Finally, our method opens lines of inquiry in business or public policy, as we could use this labeling method to generate metadata for the abstracts themselves. In this fashion, the labels could form a variable to be used in further assessment, such as patterns in funding, research direction, patents, or other corpora where scientific documents are condensed in short summaries. Using this method with public company reports could create entirely new industry categories, updating existing schemes (Shweta and Belz, 2021; Hoberg and Phillips, 2010, 2016). Moreover, these data could be combined with other tags, such as principal investigator, institution, or other bibliometric characteristics to create a complex profile. Such a data set could be used to track a number of interesting trends.

6 Conclusion

For labeling short scientific documents, such as abstracts, pre-existing domain-specific taxonomies are ambiguous. Defining a label space spanning the set of documents is an important task that humans execute easily. In this paper, we demonstrate that the text of the documents is insufficient to either define the label space or predict the labels. We present evidence that an LLM can provide critical metadata to address this gap, forming the basis for artificial intuition. Additionally, we propose both an architecture to address this and two novel measures to evaluate the constructed label spaces. Testing our model with a corpus of NASA award abstracts, we demonstrate a workflow that integrates the LLM's supplemental data successfully.

[3] https://github.com/yuzhimanhua/MAPLE/tree/master/

References

Ashwin Karthik Ambalavanan and Murthy V. Devarakonda. 2020. Using the contextual language model bert for multi-criteria classification of scientific articles. *Journal of Biomedical Informatics*, 112:103578.

Soumya Banerjee, Debarshi Kumar Sanyal, Samiran Chattopadhyay, Plaban Kumar Bhowmick, and Partha Pratim Das. 2020. Segmenting scientific abstracts into discourse categories: a deep learning-based approach for sparse labeled data. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 429–432.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Andrea Belz, Alexandra Graddy-Reed, FNU Shweta, Aleksandar Giga, and Shivesh Meenakshi Murali. 2023. Deterministic bibliometric disambiguation challenges in company names. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*. IEEE.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Alberto Blanco, Arantza Casillas, Alicia Pérez, and Arantza Diaz de Ilarraza. 2019. Multi-label clinical document classification: Impact of label-density. *Expert Systems with Applications*, 138:112835.

Buchlak, Quinlan, and Leveque JC. 2020. Machine learning applications to clinical decision support in neurosurgery. *an artificial intelligence augmented systematic review.*

Lijuan Cai and Thomas Hofmann. 2004. Hierarchical document categorization with support vector machines. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, page 78–87, New York, NY, USA. Association for Computing Machinery.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pages 830–835.

Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

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

Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. 2018. Science of science. *Science*, 359(6379):eaao0185.

Maryam Habibi and Andrei Popescu-Belis. 2015. Keyword extraction and clustering for document recommendation in conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):746–759.

Gerard Hoberg and Gordon Phillips. 2010. Product market synergies and competition in mergers and acquisitions: A text-based analysis. *Review of Financial Studies*, 23(10):3773–3811.

Gerard Hoberg and Gordon Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5):1423–1465.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank adaptation of large language models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 515–520, San Diego, California. Association for Computational Linguistics.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 115–124, New York, NY, USA. Association for Computing Machinery.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings.

Tanya Serry Reid Smith, Pamela Snow and Lorraine Hammond. 2021. The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, 42(3):214–240.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Anthony Rios and Ramakanth Kavuluru. 2015. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '15, page 258–267, New York, NY, USA. Association for Computing Machinery.

Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical multi-label text classification using only class names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4239–4249, Online. Association for Computational Linguistics.

FNU Shweta and Andrea Belz. 2021. Computational linguistic analysis of submitted sec information (classi). In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE.

Dezhao Song, Andrew Vold, Kanika Madan, and Frank Schilder. 2022. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems*, 106:101718.

Xingping Sun, Yibing Li, Hongwei Kang, and Yong Shen. 2019. Automatic document classification using convolutional neural network. *Journal of Physics: Conference Series*, 1176(3):032029.

B. Tang, H. He, P. M. Baggenstoss, and S. Kay. 2016. A bayesian classification approach using class-specific features for text categorization. *IEEE Transactions on Knowledge & Data Engineering*, 28(06):1602–1606.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 466–475, Hong Kong, China. Association for Computational Linguistics.

Ran Xu, Yue Yu, Joyce Ho, and Carl Yang. 2023. Weakly-supervised scientific document classification via retrieval-augmented multi-stage training. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Guangxu Xun, Kishlay Jha, Ye Yuan, Yaqing Wang, and Aidong Zhang. 2019. MeSHProbeNet: a self-attentive probe net for MeSH indexing. *Bioinformatics*, 35(19):3794–3802.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *CoRR*, abs/1909.00161.

Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S Dhillon. 2022. Pecos: Prediction for enormous and correlated output spaces. *Journal of Machine Learning Research*, 23(98):1–32.

Yu Zhang, Bowen Jin, Xiusi Chen, Yanzhen Shen, Yunyi Zhang, Yu Meng, and Jiawei Han. 2023. Weakly supervised multi-label classification of full-text scientific papers.

Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. 2022a. Metadata-induced contrastive learning for zero-shot multi-label text classification. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3162–3173, New York, NY, USA. Association for Computing Machinery.

Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. 2022b. Metadata-induced contrastive learning for zero-shot multi-label text classification.

Wei Zong, Feng Wu, Lap-Keung Chu, and Domenic Sculli. 2015. A discriminative and semantic feature selection method for text categorization. *International Journal of Production Economics*, 165:215–222.