Scientific Article Section Classification (SASC) Dataset

Nicolau Duran-Silva[1,2], **Julian Moreno-Schneider**[3], **César Parra-Rojas**[1], **Georg Rehm**[3]
[1]SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain
[2]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain
[3] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany
{nicolau.duransilva, cesar.parra}@sirisacademic.com, {julian.moreno_schneider, georg.rehm}@dfki.de

Abstract

We introduce a novel, publicly available dataset of scientific publications specifically designed to focus on the structural and semantic analysis of their full texts. This collection comprises 4,896 scholarly articles processed using GROBID and self-defined parsers for its segmentation and section parsing. To ensure broad utility and diversity, the dataset includes (≈1,000) papers from 4 specialized research areas: Energy, Cancer, Neuroscience, and Transport, supplemented by an additional ≈1,000 papers randomly selected from general scientific domains. This dataset is annotated using a newly-defined hierarchical taxonomy comprising 2 levels: the first level contains 9 semantic classes (coarse-grained), while the second level contains 47 semantic classes (fine-grained). All source documents were ethically and legally sourced via OpenAIRE, and the corpus is restricted exclusively to content available under open licenses. License verification was performed through cross-referencing publisher metadata, landing pages, and the Unpaywall database. This curated dataset provides a robust and domain-diverse resource, ideal for developing and evaluating NLP models that require training on hierarchical structure of scientific literature.

Keywords: Text Classification, Scientific NLP, Dataset Generation

1. Introduction

The endless growth of scientific literature has created an interesting challenge for researchers trying to efficiently navigate, comprehend, and synthesize knowledge. Every year many new academic papers are published, making the task of document understanding and information extraction from scientific texts a critical area of research in Natural Language Processing (NLP).

The identification and classification of its structural components (e.g., title, sections, tables, figures or references) can be considered as a fundamental step in understanding a scientific document. This structural and semantic organization in sections (Introduction, Method, Results, etc.), often referred to as logical structure, is useful for various downstream tasks, such as (1) section specific information retrieval: allowing researchers to search for all "Method" sections related to a specific technique, instead of searching the entire document text; (2) Fine-grained natural language processing: enabling systems to apply concrete models or techniques in specific areas of the content; or (3) Scientific knowledge graph construction: providing structured information that connects research contributions to their context.

While significant progress has been made in document layout analysis and text classification, the problem of semantically classifying sections in scientific article (Scientific Article Section Classification – SASC) remains challenging. Current benchmark datasets often fall short in providing the fine-grained, hierarchical structural annotations required to train robust models. Specifically, existing resources tend to focus on broad document classification (e.g., by subject or venue), physical layout analysis (e.g., identifying bounding boxes for text blocks, figures, and tables) and shallow semantic annotation that does not fully capture the nested, multi-level nature of a paper's structure, for example 2. Methods, 2.1. Dataset, 2.1.1. Preprocessing.

To address this critical gap, we introduce the Scientific Article Section Classification (SASC) Dataset, a collection of scientific papers [1] parsed and segmented by section, primarily designed for research in the development and evaluation of NLP models. This dataset contains ≈4,000 full-text papers from various scientific domains, namely Neuroscience, Cancer, Transport, and Energy, along with an additional ≈1,000 random papers from general scientific domains obtained from a stratified dataset by scientific discipline (Pàmies et al., 2023). This novel resource offers a meticulously (semi-automatically) annotated collection of scientific articles, with each section explicitly marked by its: Section Title and Text Content, Hierarchical Level and Explicit Section Number (if available).

The main contributions of this work are:

- A collection of scientific documents, semi-automatically annotated for its structural and logical (semantical) information, which will be made publicly available.

[1]Datasets available at https://github.com/sirisacademic/sasc

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7415–7422
11–16 May 2026.

- An NLP pipeline that processes scientific articles in PDF and automatically annotates the structural and logical information.
- A manual correction process that generates a curated, reviewed and completed annotated dataset for further NLP technologies.

The primary objective of this work is to introduce the dataset, describe its construction methodology, and provide a detailed analysis of its structural properties. While the dataset is designed to support machine learning experiments for section classification and document understanding, comprehensive benchmarking experiments are beyond the scope of the present paper and will be presented in future work focusing specifically on model evaluation.

2. Related Work

The analysis of the structure of scientific articles has been studied from different points of view, but we will focus only on two of them: physical layout analysis (identifying elements by their position and visual features) and logical structure analysis (identifying elements by their semantic role).

Several large-scale datasets have been created to drive research in Document Layout Analysis (DLA) and Document Structure Analysis (DSA). PubLayNet (Zhong et al., 2019) is one of the largest datasets for document image analysis, classifying elements into five coarse categories: *text*, *title*, *list*, *figure*, and *table*. DocBank (Li et al., 2020) is another benchmark, containing 500K document pages from arXiv articles. DocBank expands the set of semantic units to 12 categories, including *Abstract*, *Reference*, and most notably, *Section*. More recent efforts, such as DocLayNet (Pfitzmann et al., 2022) and its successor GraphDoc (Chen et al., 2025), incorporate scientific articles alongside patents, financial reports, and legal texts. GraphDoc, in particular, aims to model hierarchical structure by introducing relation annotations like "Parent Relation" between a section header and its subsection headers. While this relational approach is closer to our goal, its annotations are primarily concerned with the geometric and logical *relationships* between bounding boxes on a page, rather than providing a structured, text-based dataset of section content and its explicit semantic properties.

The automatic understanding of scientific documents has long relied on classifying textual units (sentences, paragraphs, or entire sections) into categories that denote their rhetorical function or structural role. The standard model for the logical structure of scientific papers is the IMRaD format (Introduction, Methods, Results, and Discussion) (Sollaci and Pereira, 2004). This de-facto taxonomy has been the basis for many text classification tasks focused on identifying the semantic role of paragraphs or sentences within a paper (e.g., identifying purpose, method, or result sentences). Other datasets target section classification from the explicit structural organization of the full scientific paper. Datasets operating at this level aim to assign a semantic role (e.g., method, discussion) to a title or a block of text, which is crucial for tools that extract structured information (Ronzano and Saggion, 2015).

Some works focus on classifying individual sentences within abstracts or full papers according to their rhetorical status (e.g., goal, method, result). Some efforts utilized rule-based systems (Liakata et al., 2010) and simpler statistical methods (Kim et al., 2010). More recent advances leverage deep learning models, often employing sequential classification architectures to capture inter-sentence dependencies (PubMed 200k RCT (Dernoncourt and Lee, 2017), (Jin and Szolovits, 2018), (Hu et al., 2023), CoreSC/ART (Liakata et al., 2010) and Emerald 110k (Stead et al., 2019)).

Several of these datasets are domain-specific while others allow cross-domain transfer. Domain-specific datasets derived from fields like Biomedical Sciences (BioMed), Computer Science (CS), and Natural Language Processing (NLP), often extracted from collections like the ACL Anthology Reference Corpus (Bird et al., 2008) and the ACL Anthology Network (Radev et al., 2013), are used to check performance within a homogeneous field. Cross-domain transfer datasets, such as CoreSC/ART (Liakata et al., 2010) or Emerald 110k (Stead et al., 2019), are commonly employed to test the transferability of structural understanding across different scientific disciplines (Brack et al., 2022).

In addition to considering the objective of each dataset or approach, it is also important to consider the size of the datasets, because it dictates the viability of training large-scale deep learning models. There are several large datasets. The PubMed 200k RCT (Dernoncourt and Lee, 2017), containing a large, unstructured subset collection of the arXiv corpus (Cohan et al., 2018), is one of the most expansive resources used to pre-train discourse-aware models (Cohan et al., 2018). Academic Section Classification Dataset (Hoepner, 2025) is also a large dataset, containing hundreds of thousands of section labels. Others are smaller but more focused datasets, such as BioRC800 (Lan et al., 2024), which facilitates multi-label sequential sentence classification, and the aforementioned SPACE-IDEAS (Garcia-Silva et al., 2024).

By integrating explicit numerical and hierarchical features alongside the title and text content, SASC provides the necessary components to develop more accurate and context-aware models for the

logical analysis of scientific documents.

Existing datasets present several limitations for section-level structural analysis. Many resources focus on document layout detection rather than the semantic role of sections, while others operate at the abstract level instead of full-text structure. Furthermore, most datasets do not explicitly model hierarchical section relationships or provide fine-grained taxonomies of section types. These limitations motivated the creation of the SASC dataset, which combines explicit hierarchical information with both coarse- and fine-grained semantic section labels across diverse scientific domains.

3. Dataset Construction and Annotation

To address the need for a dataset focused on the explicit hierarchical and semantic structure of scientific article sections, we constructed the Scientific Article Section Classification (SASC) Dataset. Figure 1 illustrates an example of a publication's structure with annotated section type and hierarchical level. This section details the data source, the extraction pipeline, and the subsequent process used to generate the final structured dataset.

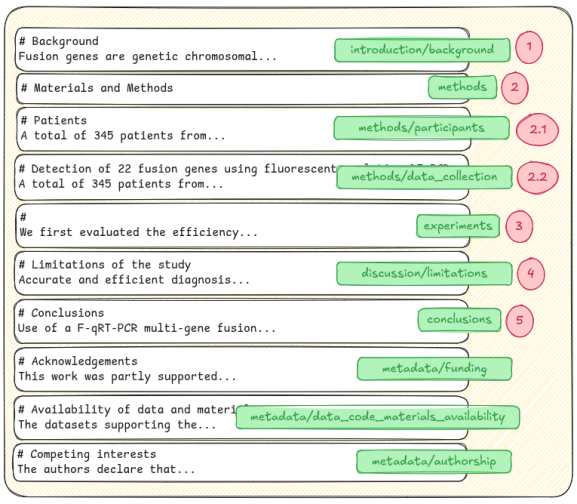

Figure 1: Example of a publication structure annotated with section label and hierarchical level.

3.1. Data Source and Initial Extraction

The documents for the SASC dataset were sourced through OpenAIRE (Manghi et al., 2019), a publicly available collection of scholarly articles. The documents were selected randomly to ensure diverse content.

The first criterion for document selection is the domain of the scientific article. In our work we are focused on 4 specific scientific domains (Energy, Cancer, Transport and Neuroscience); apart from that, we also wanted to get an extra split focused on general domain (or domain-independent) papers in order to cover different disciplines. Therefore we end up with 5 splits of documents.

3.2. Document Pre-Processing

We have used GROBID (GROBID) to extract the information from the PDF documents. License validation was conducted to ensure that all documents could be legally used for text-mining and dataset redistribution. Specifically, only papers under permissive licenses (e.g., CC-BY or Public Domain) were included[2].

We utilized a custom parser to process the structured XML (TEI) files. Regarding the front matter, the parser extracts `Title` and `Abstract`. For every `Section` detected in the document's main body (excluding front matter and references) the parser was designed to extract three primary components:

- `Section Title`: The verbatim text of the section header (e.g., *"Experimental Setup"*).
- `Section Number`: The explicit numerical code (e.g., *"3.2"*) if available.
- `Section Content`: The full text body of the section, starting immediately after the header and ending just before the next header or the end of the document.

In addition, we combine sections from `body` and `back` to create a reliable and generalizable representation of document structure across heterogeneous sources, which are not identified properly due to layout inconsistencies. By standardizing the extraction of section-level text and metadata, we enable downstream models to generalize across different scientific domains and publishing formats.

The annotation protocol started by first examining the `Section Title` (e.g., *"Data Preprocessing"*). For sections with canonical titles, the label was assigned directly. For ambiguous or non-standard titles (e.g., *"Observations"* or *"The Proposed System"*), the Section Content was reviewed to infer its function within the taxonomy from Section 3.3. For instance, a section titled *"Results and Analysis"* was labeled as `Experiment`.

Among all extracted `Section Names`, 31.78% matched to predefined first-level taxonomy classes (e.g., *Introduction*, *Methods*, *Results*) and some defined variants, enabling initial normalization without model inference or human review.

[2]License information was verified through cross-referencing publisher landing pages, OpenAIRE metadata, and the Unpaywall database (https://unpaywall.org/).

The last step in the processing pipeline is the normalization of the `Section Numbers`, using the format $N_1.N_2.\ \cdots\ .N_L$, i.e., a set of (Arabic) numbers concatenated by dots, where every dot represents a new sub-level. However, the 64,7% of publications of the dataset do not contain numbered sections. Therefore they must be identified manually from sections names, position in the paper and content.

3.3. Defining the 2-Level Section Taxonomy

A critical step was defining a manageable and semantically meaningful set of class labels, moving beyond the simple `Section` label found in datasets like DocBank. We adopted an expanded version of the "IMRaD" model to reflect modern scientific publishing practices, resulting in a taxonomy of 9 level-1 (coarse-grained) classes and 47 level-2 (fine-grained) classes. This coarse-grained classes are described in Table 1 (the fine-grained classes are only listed). The taxonomy was derived through an iterative analysis of the most frequent section titles and subsection patterns observed across the collected corpus. Several rounds of exploratory inspection and normalization were conducted to consolidate semantically equivalent section names into a consistent taxonomy. Section titles that did not fit the initial taxonomy were analyzed and grouped into additional categories. The resulting schema reflects common structural patterns found in scientific publications across multiple domains. For instance, at level 1, we have added `Metadata` and `Other` sections.

3.4. Annotation Process

Annotation consists of a semi-supervised procedure combining pseudo-labelling with a language model and human refinement. The language model is used only to provide initial pseudo-labels, while final labels are validated and corrected by human annotators. In the first stage, for each article, a tailored prompt containing `section titles`, `numbers` (when available), and truncated `content`, and the overall structural order of the sections, was provided to a large language model (GPT-4o-mini) following detailed instructions with our proposed 2-level taxonomy. The model predicts normalized section and subsection names, and section numbers. In the second stage, two annotators with experience in scientific writing are employed to assign the semantic class label and hierarchy to each extracted section, introducing new categories when necessary. During annotation, the section `title`, `number` (when provided), and `content` were considered in the context of the overall paper structure. Annotators also have access to the full-text PDF to verify layout and contextual elements when required.

We developed a set of detailed annotation guidelines describing the taxonomy definitions, decision rules for ambiguous section titles, and examples of typical section mappings.

3.4.1. Hierarchical Level Derivation

The Hierarchical Level (L) was automatically derived from the extracted Section Number. For a number $N_1.N_2.\ \cdots\ .N_L$, the level L is the count of numerical components. For example, Section 3.1 has $L = 2$, and Section 4 has $L = 1$.

3.4.2. Inter-Annotator Agreement

To evaluate the reliability of the annotation protocol and taxonomy proposed, we conducted an inter-annotator agreement (IAA) study on a pilot subset of the dataset, measured using Cohen's κ coefficient. Two annotators independently labeled sections from scratch from a sample of 14 scientific articles selected from different domains included in the corpus. The sample contained 213 sections.

The results show a high level of agreement for the coarse-grained taxonomy, with $\kappa = 0.8776$ and a raw agreement of 90.6%, which corresponds to *almost perfect agreement* according to the commonly used interpretation scale. For the fine-grained taxonomy ($L = 2$), the agreement is lower but still substantial, with $\kappa = 0.5329$ and a raw agreement of 59.6%. This reduction is expected due to the larger number of fine-grained categories and the presence of semantically similar labels.

Additionally, we evaluate the agreement between the human annotators and the model used for pseudo-label generation. The average human–LLM agreement reached $\kappa = 0.8122$ for the coarse-grained taxonomy and $\kappa = 0.4202$ for the fine-grained taxonomy, indicating that the automatically generated labels provide a useful starting point for the human annotation process.

3.4.3. Evaluation of Hierarchical Numbering

Since 64.7% of the articles in the dataset do not contain explicit section numbering in the layout, we evaluated the reliability of manually reconstructed hierarchical numbers. We conducted an experiment on a subset of 14 articles (218 sections) where the original section numbers were available. Annotators assigned hierarchical numbers without inspecting the ground-truth numbering and relying only on section titles and document structure.

The predicted numbers were compared against the original ones using several metrics. Annotators achieved 86.7% strict accuracy in average. More importantly, structural consistency remained high,

ID	L=1 Class Label	L=2 Class Label	Description
1	Introduction	background, problem statement, objectives, contributions, outline	Context, problem statement, research gap, and paper goals. Corresponds to I in IMRaD.
2	Related Work	literature review, theoretical framework, gaps identified	Review of prior literature and state-of-the-art methods.
3	Method	study design, participants, data collection, procedures, statistical analysis, ethical considerations, materials and instruments, data preprocessing, evaluation, case study	Description of the system, materials, models, or procedures used. Corresponds to M in IMRaD.
4	Experiment	descriptive statistics, main findings, secondary findings, tables figures, statistical tests, evaluation	Details on experimental setup, data, metrics, and results. Closely tied to R in IMRaD.
5	Discussion	interpretation, comparison literature, implications, limitations, strengths, future work	Interpretation, implications, limitations, and future work. Corresponds to D in IMRaD.
6	Conclusion	summary, key findings, contributions, final remarks, future directions recommendations	Final summary of findings and concluding remarks.
7	Other	-	Unclassified sections.
8	Appendix/Misc.	supplementary data, additional analyses, technical details, code and materials, ethics and consent, compliance statement	Supplementary or procedural material directly related to the study: extended analyses, code, data, technical documentation, and ethical or compliance statements to the research process.
9	Metadata	acknowledgments, authorship, funding, license, data code materials availability, publisher note	Sections providing administrative or provenance information about the publication itself: author roles, funding sources, legal licenses, and data-sharing declarations. Not part of the scientific argument but essential for transparency and reproducibility.

Table 1: 2-level taxonomy defined for the SASC dataset. Class labels from levels 1 and 2 are included.

with depth accuracy of 94.2%, and parent accuracy of 90.9%, respectively. These results indicate that the hierarchical structure can be reliably inferred even when exact numbering differs.

3.5. Final Dataset Structure

The final SASC dataset consists of 117,876 unique sections extracted from 4,896 full-text articles from 5 different domains (Energy, Cancer, Transport, Neuroscience and Random). Each instance in the dataset is structured as an object, with the sections identified as `Section Name`, `Section Number` and `Content`, and labels are provided as `Class Label` and `Hierarchy Level`.

4. Dataset Analysis and Statistics

This section provides a detailed description (statistical overview) of the Scientific Article Section Classification (SASC) Dataset. It focuses on its scale, the distribution of the defined semantic classes, and the relationship between content length and hierarchical level.

4.1. Overall Dataset Dimensions

The SASC dataset comprises a total of 117,876 unique sections extracted from 4,896 scientific articles. The data split adheres to a strict article-level separation, ensuring no single paper contributes sections to more than one set.

Split	# Arts	%	# Secs	%
train	3,918	80.02%	86,919	80.08%
val	489	9.99%	10,409	9.59%
test	489	9.99%	11,208	10.33%
Total	4,896	100%	108,536	100%

Table 2: Distribution of sections and articles across the SASC dataset splits.

4.2. Dataset Languages

Language identification was performed on all 117,876 extracted sections to characterize the language distribution of the SASC corpus. The vast majority of sections were written in English (94.31%), while 5.69% were in other languages. The most frequent non-English languages were Portuguese (1.8%), Spanish (1.5%), Turkish (0.6%), Italian (0.5%), Russian (0.4%), and French (0.2%), with smaller proportions in more than twenty additional languages. The corpus is predominantly English but includes a diverse set of multilingual scientific texts.

4.3. Semantic Class Distribution

Table 3 presents the counts and relative frequencies of the distribution of sections across the 8 semantic classes of level 1 in our taxonomy.

Looking at the table, it is surprising that the appearance of several classes is bigger than the number of papers in the dataset. It makes no sense that one paper has several sections from the same

ID	Semantic Class	Count	Freq (%)
1	Introduction	8,469	7.8
2	Related Work	2,004	1.9
3	Method	30,410	28.0
4	Experiment	20,611	19.0
5	Discussion	7,978	7.3
6	Conclusion	3,586	3.3
7	Other	10,410	9.6
8	Appendix	4,918	4.5
9	Metadata	20,150	18.6
	Total	108,536	100.0

Table 3: Distribution of sections across the 8 semantic classes of level $L = 1$.

class, but we are only looking at level 1 sections, and we are covering all subsections inside the parent class for this table. Therefore, it is possible that a semantic class has much bigger appearance numbers than the number of papers.

The observed distribution shows a notable skew, with `Method`, `Experiment` and `Metadata` being the most prevalent classes. On the contrary, `Related Work` represents the minority class, having a too low number. That is happening because many papers use non-normalized names for this section, e.g., "Background" or "Previous Work". `Introduction` seems to match almost perfectly the number of papers in the dataset, similarly to `Conclusion` and `Discussion`. That is congruent with the fact that normally those sections do not contain subsections (only several in the latter case).

4.4. Analysis of Hierarchical Structure

A unique feature of the SASC dataset is the explicit inclusion of the Hierarchical Level (L). This analysis describes the distribution of sections based on these levels. The distribution of sections per level (L) is shown in Table 4.

Hierarchical Level (L)	Count	Freq (%)
$L = 1$ (Primary)	34,847	41.5
$L = 2$ (Subsection)	45,680	54.4
$L = 3$ (Sub-subsection)	3,288	3.9
$L \geq 4$ (Deeply Nested)	123	0.1
Total	83,938	100.0

Table 4: Distribution of sections by their hierarchical depth (L).

As expected, the majority of sections reside at the first (41.5%) and second levels (54.4%).

In summary, the SASC dataset provides two unique, non-trivial features for section classification: Semantic Class (Y), a label representing the section's semantic label; and Hierarchical Level (L), explicit numerical metadata defining structural depth. This multi-faceted structure makes SASC a robust benchmark for developing multi-modal models that combine text features with explicit structural metadata.

4.5. Discussion on the Utility of the SASC Dataset

The creation of the Scientific Article Section Classification (SASC) Dataset constitutes a significant step toward developing structurally aware models for scientific document understanding. Our work addresses the critical limitations of existing datasets, which fail to capture hierarchical roles of indexed sections.

Furthermore, due to the growth of enriched textual files to represent scholarly articles, such as Markdown, this plain text format can be useful and represent a wide range of input textual formats.

5. Conclusions

We have successfully constructed the Scientific Article Section Classification (SASC) Dataset, a new benchmark explicitly designed to facilitate the training of models that understand the hierarchical structure of scientific documents. By providing explicit labels for the semantic class and the numerical hierarchical level, SASC bridges the gap between document layout analysis and rhetorical structure analysis. We anticipate that models trained on this multi-faceted data will significantly advance the state of the art in automatic scientific document comprehension, ultimately enabling more accurate information retrieval and knowledge graph construction from scholarly texts. The dataset is publicly released to encourage further research and innovation in this crucial domain.

Future work will explore benchmarking experiments using both traditional classifiers and recent transformer-based models to evaluate the impact of hierarchical structural information on section classification performance.

6. Limitations

While the SASC dataset offers a novel approach to explicit structure modeling, its limitations guide future research:

- Manual Annotation Process: the creation of SASC dataset relies on manual human annotation for semantic labels and hierarchical numbers. Although this process has already started, it is not finished yet. Therefore, finishing this manual correction of the automatically generated semantic labels is the first, most important priority of the future work. If everything goes as planned, the complete manually

corrected dataset should be available for the camera ready version of the paper. Besides, a full updated analysis of both semantic classes levels would be included.

- Scope of taxonomy and hierarchical (fine-grained) labeling: The current taxonomy includes 9 coarse-grained and 47 fine-grained labels designed to capture common structural patterns in scientific articles. While the taxonomy already covers the majority of observed section types, future work may explore minor refinements as additional scientific domains are incorporated. However, the current version provides a stable and sufficiently expressive structure for training and evaluating section classification models.

7. Acknowledgements

This work was supported by the Industrial Doctorates Plan of the Departament de Recerca i Universitats de la Generalitat de Catalunya (grant reference 2022/DI /00017). This work was co-funded by the EU HORIZON SciLake (Grant Agreement 101058573). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders.

8. Ethical considerations

Our dataset is composed of scientific publications, in which we have already tested the licensing and the availability to be used for our purposes. They have open licenses and are freely available for usage. Therefore, we found no ethical considerations applicable to this work.

9. Bibliographical References

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Arthur Brack, Anett Hoppe, Pascal Buschermöhle, and Ralph Ewerth. 2022. Cross-domain multi-task learning for sequential sentence classification in research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, New York, NY, USA. Association for Computing Machinery.

Yufan Chen, Ruiping Liu, Junwei Zheng, Di Wen, Kunyu Peng, Jiaming Zhang, and Rainer Stiefelhagen. 2025. Graph-based document structure analysis. In *The Thirteenth International Conference on Learning Representations*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Franck Dernoncourt and Ji Young Lee. 2017. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Andres Garcia-Silva, Cristian Berrio, and Jose Manuel Gomez-Perez. 2024. SPACE-IDEAS: A dataset for salient information detection in space innovation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15087–15092, Torino, Italia. ELRA and ICCL.

Niklas Hoepner. 2025. Dataset for classification of sections of academic papers. Hugging Face. URL: https://huggingface.co/datasets/nhop/academic-section-classification. Accessed: 2025-09-01. CC BY 4.0.

Yan Hu, Yong Chen, and Hua Xu. 2023. Towards more generalizable and accurate sentence classification in medical abstracts with less data. *Journal of Healthcare Informatics Research*, 7(4):542–556.

Di Jin and Peter Szolovits. 2018. Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3100–3109, Brussels, Belgium. Association for Computational Linguistics.

Su Nam Kim, David Martinez, and Lawrence Cavedon. 2010. Automatic classification of sentences

for evidence based medicine. In *Proceedings of the ACM Fourth International Workshop on Data and Text Mining in Biomedical Informatics*, DTMBIO '10, page 13–22, New York, NY, USA. Association for Computing Machinery.

Mengfei Lan, Lecheng Zheng, Shufan Ming, and Halil Kilicoglu. 2024. Multi-label sequential sentence classification via large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16086–16104, Miami, Florida, USA. Association for Computational Linguistics.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Paolo Manghi, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Jochen Schirrwagen, Pedro Principe, Michele Artini, Amelie Becker, Michele De Bonis, et al. 2019. The openaire research graph data model. *Zenodo*.

Marc Pàmies, Joan Llop, Francesco Multari, Nicolau Duran-Silva, César Parra-Rojas, Aitor Gonzalez-Agirre, Francesco Alessandro Massucci, and Marta Villegas. 2023. A weakly supervised textual entailment approach to zero-shot text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 286–296.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3743–3751, New York, NY, USA. Association for Computing Machinery.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Lang Resources & Evaluation*, 47(4):919–944.

Francesco Ronzano and Horacio Saggion. 2015. Dr. inventor framework: Extracting structured information from scientific publications. In *Discovery Science*, pages 209–220, Cham. Springer International Publishing.

Luciana Sollaci and Maurício Pereira. 2004. The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. *Journal of the Medical Library Association : JMLA*, 92:364–7.

Connor Stead, Stephen Smith, Peter Busch, and Savanid Vatanasakdakul. 2019. Emerald 110k: A multidisciplinary dataset for abstract sentence classification. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 120–125, Sydney, Australia. Australasian Language Technology Association.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. Publaynet: Largest dataset ever for document layout analysis. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022.

10. Language Resource References

GROBID. 2008–2026. Grobid. https://github.com/kermitt2/grobid.