

Hybrid NLP And Deep Learning Techniques for Automated Knowledge Extraction from Scientific Literature

Dr. Deepak Kumar ^{*1}, Sumit Parmar ², Justin Mhel G. Vaquilar³, Shivam Upadhyay⁴, Dr. Rajinder Kumar⁵, Avijit Kumar Chaudhuri⁶

^{1*} Associate Professor, Department of Computer Science and Engineering, Surajmal University, Kichha Udham Singh Nagar, Uttarakhand, Orcid: 0000-0001-6946-2789, Email ID: deepakraj.377@gmail.com ss

² Assistant Professor, Department of IT, Specialization in Information Technology, Mumbai University, Mumbai, Email ID - sumit.parmar84@gmail.com

³ Doctor of Philosophy, Institute of Graduate and Professional Studies, Lyceum-Northwestern University, Dagupan, Pangasinan, Philippines, Orcid: 0009-0005-7177-4462, Email ID: vaquilar.justinmhel.g@gmail.com

⁴ Assistant Professor, Department of Electrical and Electronics Engineering, Mangalayatan University, Aligarh Email ID: shivam.upadhyay@mangalayatan.edu.in

⁵ Associate Professor, Department of Faculty of Computing, Specialization in Artificial Intelligence, IOT, Data Mining Guru Kashi University, Punjab, Talwandi Sabo, Bathinda, 151302 Punjab, Orcid: <https://orcid.org/0009-0001-4129-0388>, Email ID: drrajinder1983@gmail.com

⁶ Professor & HOD, Department of Computer Science & Engineering, Brainware University, Barasat, Kolkata – 125, West Bengal, Orcid: 0000-0002-5310-3180, Email ID: c.avijit@gmail.com

HOW TO CITE:

Deepak Kumar , Sumit Parmar ,
Justin Mhel G. Vaquilar, Shivam
Upadhyay, Rajinder Kumar, Avijit
Kumar Chaudhuri (2026). Hybrid
NLP And Deep Learning
Techniques for Automated
Knowledge Extraction from
Scientific Literature.
International Journal of Special
Education, 41(13s), 1546-1557

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT:

The rapid expansion of scientific literature has created a critical need for intelligent systems capable of extracting structured knowledge from unstructured textual data. This study proposes a hybrid framework integrating Natural Language Processing (NLP) and deep learning techniques for automated knowledge extraction from scientific literature. The proposed approach combines contextual semantic representations with engineered linguistic and structural features using a Self-Attention-based Hybrid Feature Fusion Network. The framework processes scientific abstracts at the sentence level, enabling classification into rhetorical sections and facilitating structured knowledge reconstruction. By leveraging both deep contextual learning and statistical feature analysis, the model captures semantic dependencies and document structure effectively. Experimental evaluation demonstrates that the proposed hybrid model outperforms traditional baseline methods in terms of accuracy and F1-score, highlighting its robustness and generalization capability. Furthermore, the system enables the extraction of meaningful information segments and key domain-specific terms, supporting applications such as information retrieval, document summarization, and knowledge discovery. The results validate the effectiveness of integrating NLP and deep learning within a unified framework for scalable scientific text analysis. Overall, this work contributes to the advancement of automated knowledge extraction systems for large-scale scientific literature.

Keywords: Hybrid NLP, Deep Learning, Knowledge Extraction, Scientific Literature, Data Analytics

1. Introduction

The proliferation of scientific literature in a wide range of disciplines has posed serious problems in the effective extraction, organization and exploitation of knowledge contained within huge unstructured bodies of text. Automated systems are becoming more popular among researchers and practitioners of scientific documents as they seek to handle large amounts of data because manual analysis is time-consuming and highly inconsistent. In this regard, the use of sophisticated data analytics and artificial intelligence has proved a strong tool in converting unstructured scientific text into formatted and signifying knowledge. Recent works have stressed the need to combine Natural Language Processing (NLP) and deep learning procedures to solve the problems of scientific text mining and knowledge extraction that are rather complicated [1]. NLP is an important aspect that allows machines to comprehend and interpret human language, especially in a limited field of science like literary text. Such methods as named entity recognition, relation extraction and text classification were popular to extract the information of interest in research papers. Indicatively, [2] established the usefulness of NLP methods in mining metadata in environmental scientific journals with references to the possibilities of a computerized system in organizing domain-related information. On the same note, [3] investigated the application of deep learning-based NLP models to extract full-text data elements and provided evidence that hybrid solutions are much more effective in facilitating systematic literature review processes. The ability of NLP systems has been advanced by deep learning models that make it possible to perceive and represent complex textual data in a context. Recurrent neural networks, graph neural networks and transformer-based models are advanced architectures that have been used successfully in tasks of extracting scientific knowledge. The [4] proposed a hybrid system of ontology models, graph neural networks, with large language models to enhance the accuracy in the extraction of structured knowledge in scientific publications. Similarly, [5] applied the

large language models to extract information through medical case reports and showed better results at extracting domain-specific information through complex textual tasks. Besides semantic interpretation, structural features of scientific texts are also very important in successful knowledge extraction. Scientific literature generally has a very clear format, with the sections being background, methodology and results. It can greatly enhance the extraction of these structural patterns. The paper by [6] initiated a layout sensitive method of gathering structured information in scientific PDFs that emphasizes the need to use document structure and visual layout in NLP models. These practices emphasize the necessity of hybrid systems with both textual and structural characteristics to ensure higher performance. Moreover, the previous studies have identified the relevance of combining various methods of analyses within document analytics and text mining. A method of artificial intelligence and machine learning applied to intellectual property analytics was discussed by Aristodemou and [7], who showed the ability of more sophisticated analytics to bring valuable insights on large document repositories. On the same note, [8] surveyed methods of deep learning-based sentiment analysis, showing that using the combination of multiple representations of features improves the performance of models in document-level analysis. All these studies imply that the combination of various methodologies may give results that are more robust and accurate. Although these achievements are significant, the current approaches are usually constrained to manage heterogeneous scientific texts and contextual dependencies, as well as combine various types of features to an adequate degree. Numerous methods are based only on deep learning models or standard NLP methods, and they cannot take full advantage of the complementary benefits of both. Also, issues like data imbalance, writing styles and domain peculiarity still serve as perplexing factors on model performance and generalization.

In a bid to address these deficiencies, the present paper proposes a state-of-the-art data analytics

that will integrate hybrid NLP, as well as deep learning, in automated knowledge mining of scientific publications. The provided framework is a mix of contextual embeddings, the linguistic features, which are designed and self-attention mechanisms, which improve the classification performance and extraction performance. Both semantic and structural data will be employed in the model and will make the systems of knowledge extraction more accurate and powering. The proposed methodology may be considered one of the additions to the constantly growing scope of smart document processing and a sufficient way of exploring large-scale scientific literature.

Objectives of the Study

1. To develop a hybrid NLP and deep learning-based framework for automated knowledge extraction from scientific literature, integrating semantic, statistical, and structural features.
2. To enhance classification accuracy of scientific text sections by utilizing advanced data analytics techniques, including self-attention mechanisms and feature fusion strategies.
3. To generate structured and meaningful knowledge representations from unstructured scientific documents for applications such as information retrieval, summarization, and decision support systems.

2. Literature Review

Ranjeeth et al. (2020) examined predictive models within the learning analytics field, where the authors have examined different machine learning methods of deriving meaningful information out of educational data. They argued that predictive modelling is essential in identifying the trends and enhancing the decision-making process, and that data-driven methods are critical to large-scale text and data analysis[9]. Wątróbski (2020) concentrated on text-based knowledge-based ontology learning. The paper discussed the strategies of converting unstructured text to structured knowledge representations, which is needed in automated knowledge extraction systems. The work offers a solid base of semantic processing and text-based

data generation for structured data[10]. Vlah et al. (2022) researched empirical engineering design by means of a scient metric analysis method. Their study focused on the effectiveness of analyzing large amounts of engineering data and literature to glean trends, patterns and insights. The research shows the importance of data analytics in deriving knowledge in scientific papers[11].

Dwivedi et al. (2023) explored the history of artificial intelligence studies, including research trends, topics and future research. The authors stressed the growing importance of AI in data analytics, decision support systems and knowledge discovery, which is in line with the growing necessity of intelligent systems in scientific literature reading[12]. Collins et al. (2021) researched the use of artificial intelligence in information systems by conducting a systematic literature review. Their contribution discussed the improvement of information retrieval, processing and management using AI techniques and showed the possibilities of AI-based systems in dealing with large amounts of textual data[13]. Munappy et al. (2022) explored the problem of data management in the creation of deep learning models of production quality. The authors have put forward the problems associated with data preprocessing, scalability, and quality, and have offered solutions to enhance performance of the model, which is of paramount importance in constructing reliable NLP and deep learning systems[14]. Balla et al. (2022) researched the use of deep learning algorithms to detect intrusion. Their study revived the potential of deep learning models to detect complex patterns in large datasets and provided an example of the power of deep learning methods in various directions[15]. Digamberrao and Prasad (2018) emphasized on machine learning-based author identification of textual information with decision trees that are rule-based. Their contribution indicated the efficiency of mixed strategies when it comes to text classification tasks, and thus it can be free to use mixed methodologies in NLP-based systems[16]. Liu et al. (2017) researched the use of machine learning in the discovery and design

of materials. The study indicated the possibility of knowledge extraction and innovation in scientific areas and the use of machine learning in scientific data analysis acceleration[17]. Deng et al. (2020) investigated how we can combine machine learning with human knowledge to enhance model interpretability and performance. The paper has emphasized the need to integrate computational intelligence with domain knowledge which is applicable in the creation of powerful systems of knowledge extraction[18]. Abdullah et al. (2023) carried out a systemic review of information retrieval methods of text information. Their article was about the latest techniques, implementations, trends, and issues of NLP-based extraction systems, focusing on the increased relevance of hybrid strategies in text analytics[19]. Asaad et al. (2022) researched hybrid big data analytics through the mergence of structured and unstructured data to predictive intelligence. Their study showed that integrating various data sources and methods of analysis results in better performance of knowledge extractors and decision system[20].

3. Methodology

3.1 Proposed Framework Overview

An elegant data analytics model that combines Natural Language Processing (NLP) and deep learning is suggested to extract knowledge automatically out of scientific literature in this research. The structure is an abstract engine that works with unstructured scientific abstracts and converts them into structured information that is readable by a machine. The pipeline of the whole process comprises of the dataset processing, sentence level reconstruction, hybrid feature engineering, deep learning classification and extracting structured knowledge. The originality of the suggested solution is the integration of semantic representations of context and statistical linguistic representations on the basis of a hybrid deep network architecture augmented with self-attention modules.

3.2 Dataset Preparation and Preprocessing

The experimental assessment is made of a complete scientific literature dataset of 15,000 abstracts, which are further discarded into

182,583 labelled sentences. Every sentence falls in one of the five rhetorical sections: BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS[21]. First, the data is subjected to various preprocessing procedures such as noise and special characters removal, lowercasing text normalization and tokenization. Missing values in one or more sections are dealt with by omitting unfinished entries or re-creating existing sections. Sentence-level reconstruction is also done so that the contextual flow of each abstract can be preserved, and therefore, sequential dependencies are observed during the model training. This data is subsequently split into training, validation and test data in order to have quality performance analysis.

3.3 Hybrid Feature Engineering

A hybrid feature engineering method is used to allow capturing of both semantic and structural features of scientific text effectively. Semantic features are obtained based on text sequences (which are tokenized) and transformed into numerical representations with embedding methods and padded to match homogenous input dimensions. With them, a set of handcrafted statistical and linguistics features is obtained, such as the number of tokens, the average token length, the ratio of digits to punctuation, the ratio of uppercase characters to total characters, and the availability of domain-specific markers, such as percentages and p-values. In addition, positional features expressing the relative position of a sentence in an abstract are added. This combination helps the model to exploit the deep contextual knowledge as well as explicit structure hints, which are present in scientific writing.

3.4 Hybrid Deep Learning Model Architecture

The suggested model, which is called Self-Attention-based Hybrid Feature Fusion Network (SA-HFFN), is aimed at learning simultaneously with textual sequences and computed features. The architecture is built upon an embedding layer which converts input tokens to dense vectors. These embeddings are then fed to a Bidirectional

Long Short-Term Memory (Bi-LSTM) network which is used to learn bidirectional contextual dependencies. To further boost significant information, a mechanism of self attention is used, which enables the model to be more significant on the relevant words and phrases in a sentence. A feature fusion layer is then the combination of the output of the sequence model with the engineered statistical features. The fused representation is then propagated by fully connected dense layers where the dropout regularization is used to avoid overfitting. Lastly, the softmax output layer is employed in order to execute multiplex classification of the five rhetorical categories.

3.5 Model Training and Optimization

The categorical cross-entropy loss function is used to train the model, and the Adam optimizer is used to optimize the model. The batch size is 64, and a small number of training epochs is employed to make sure convergence is efficient. Because of the skewed distribution of the sentence classes, weighting of classes is added in the training process to enhance learning of underrepresented classes. Such regularization techniques as dropout and early stopping on the loss of the validation are applied to prevent overfitting and ensure generalization. The precision, accuracy, recall, and F1-score are the standard measures that are applied in the analysis of the model performance.

3.6 Baseline Model for Comparative Analysis

To demonstrate the performance of the proposed hybrid structure, a baseline model is used using the conventional machine learning techniques. Specifically, the TF-IDF vectorization is employed in order to convert a textual data into a numerical one, and then the latter is classified by using logistic regression. This metric is employed to show the improvements of the application of deep learning and hybrid feature engineering in the suggested method.

3.7 Knowledge Extraction and Output Generation

Once the classification has been performed, the relied labels are rebuilt to reconstruct the

structured representations of scientific abstracts. Sentences are also classified with reference to their predicted categories, and this is what constitutes sections that contain predicted background, objective, methods, results and conclusions. This enables automatic classification of scientific content into meaningful portions. In addition to that, the frequency-based analysis of the classified parts is also applied to obtain the keywords that are relevant to the domain. We can have applications of such process in the context of summarization, indexing and knowledge mining which is the practical usefulness of the proposed structure.

3.8 Summary of Methodological Contributions

The provided methodology proposes a novel hybrid approach that is the fusion of deep learning-aware contextual modelling with manually structured linguistic features in studying scientific texts. The inclusion of the self-attention mechanisms boosts the attention of the model to significant information to a greater extent. The framework provides an entire pipeline of classification and knowledge extraction, and hence may be regarded as useful in converting unstructured scientific literature into structured and analyzable data and, therefore, may be applied to the more sophisticated applications of data analytics.

4. Results

4.1 Dataset Reconstruction and Section Distribution

The suggested framework was tested by using PubMed 20k RCT abstract data, consisting of 15,000 structured scientific abstracts. Following sentence-level reconstruction, the corpus had 182,583 labelled sentences, each of which was a part of one of five rhetorical sections: BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS. Figure 1 displays the distribution of the classes. The dataset is also evidently unbalanced with the largest parts of corpus represented by RESULTS (60,956 sentences) and METHODS (59,376 sentences), with OBJECTIVE (13,824 sentences) and BACKGROUND (21,591 sentences) being

significantly smaller classes. With 26,836 sentences it is placed in the middle between CONCLUSIONS and other articles. Such imbalance shows that class weighting is justified in model training as there are areas of rhetoric that are much more common in scientific

literature than others. Figure 1, thus, validates the expectation that the data is realistic in terms of abstract scientific structure, but it is also characterized by a problematic multi-class classification problem because of dissimilar frequencies of labels.

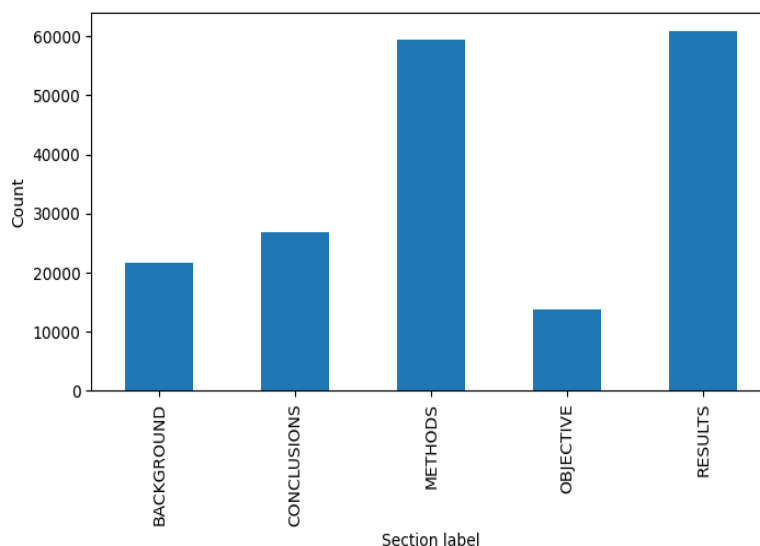


Fig 1. Class Distribution of Labelled Sentence Categories in the Scientific Abstract

4.2 Training Behaviour of the Proposed Hybrid Model

Self-Attention based Hybrid Feature Fusion Network (SA-HFFN) was trained on sentence sequences with the help of engineered positional and statistical features. The model has about 3.93

million trainable parameters, which means it has a moderately deep architecture that can be trained to learn semantic as well as structural patterns in scientific text. Figure 2 displays the curves of training and validation accuracy, whereas Figure 3 displays the curves of loss.

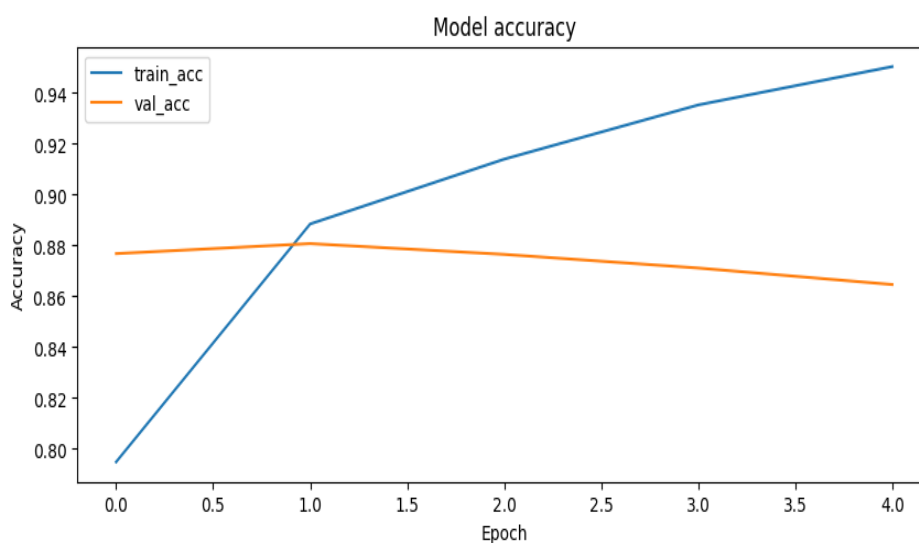


Fig 2. Accuracy Convergence Curve of the Hybrid NLP-Deep Learning Model During Training

Based on Figure 2, it is possible to note that the model converged very quickly in the early epochs. Accuracy during training also went up by 79.49% in Epoch 1 to 88.85% in Epoch 2 and kept on increasing until it was measured to be over 95% in Epoch 5. In the first epoch, validation reached

87.68%, and in Epoch 2 it had a highest value of 88.08, then it began to decrease slightly in later epochs. The observed pattern indicates that the model acquired discriminatory sentence representations very fast and had high validation performance at the early stages of training.

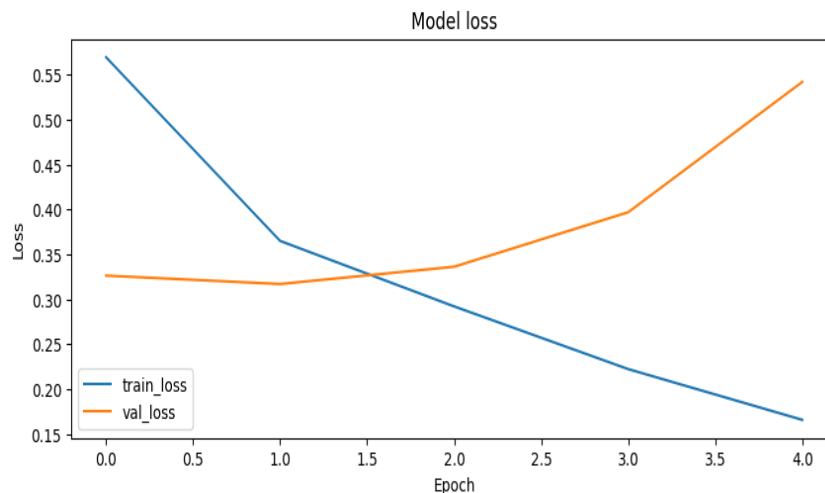


Fig 3. Loss Convergence and Overfitting Behaviour of the Hybrid NLP–Deep Learning Model During Training

As Figure 3 below shows, the training loss steadily decreased with each epoch, the validation loss reached its lowest point around the second epoch, and then it began to grow. Such a differentiation between training and validation loss is evidence of overfitting on the early convergence, and this is why early stopping is useful in the context of choosing the most efficient model checkpoint. On the whole, Figures 2 and 3 prove that the proposed hybrid architecture can efficiently learn on the basis of the dataset, whereas regularization and the choice of checkpoints is also significant.

4.3 Quantitative Classification Performance

The chosen model that was finally selected was tested on the held-out test set of 27,477 sentences. The hybrid model proposed attained:

- Test Accuracy: 0.8771
- Macro F1-score: 0.8277
- Weighted F1-score: 0.8779

These findings point to a good general classification performance, particularly given the skew in the number of rhetorical categories.

The class-wise performance is summarized below:

Table 1. Precision, Recall, and F1-Score Analysis Across Rhetorical Sections in Scientific Literature

Section	Precision	Recall	F1-score	Support
BACKGROUND	0.78	0.68	0.73	3329
CONCLUSIONS	0.89	0.94	0.92	4046
METHODS	0.92	0.93	0.93	8985
OBJECTIVE	0.59	0.73	0.65	2046
RESULTS	0.94	0.90	0.92	9071

Out of all categories in Table 1, METHODS and RESULTS made the best performance, both having an F1-score of higher than 0.92. This implies that these parts have rather stable lexical and structural arrangements, which are easier to distinguish by the model. There was also good performance in CONCLUSIONS with 0.92 F1-score. Conversely, OBJECTIVE produced the lowest F1-score (0.65), which was primarily because of its lower precision (0.59) though its recall was also relatively high (0.73). It means that the model may frequently identify statements of the type of an objective, yet also use them interchangeably with semantically similar syntax

in adjoining rare. BACKGROUND was also less recollected (0.68) than associated with a comparative level, whereas introductory context sentences are more heterogeneous and thus more difficult to classify.

4.4 Confusion Matrix Analysis

Figure 4 presents the confusion matrix that gives a more detailed picture of the class-level prediction behaviour. The predominant cells are the diagonal cells, which affirms that most of the sentences were categorized appropriately in all the five rhetorical groupings

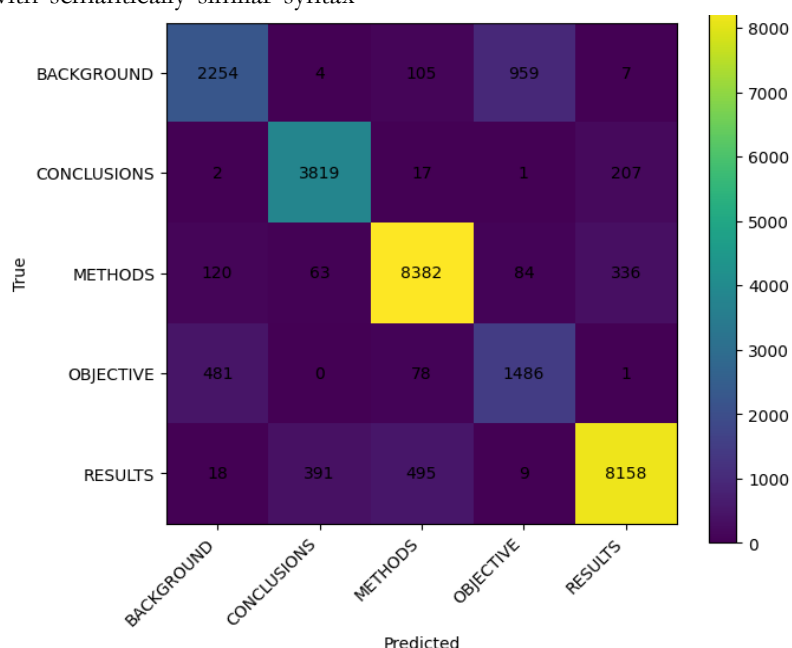


Fig 4. Confusion Matrix Representing Class-wise Prediction Performance Across Rhetorical Sections

The lowest error predictions are observed in the METHODS, RESULTS, and CONCLUSIONS, as shown in Figure 4, where the diagonal intensities are the highest. The precision and recall of these categories are high, hence this observation. But when one looks at the non-diagonal entries, one can see that the most common confused classes are OBJECTIVE and BACKGROUND. Specifically, some of the objective statements seem to overlap with background or methodological statements, which is quite natural since these sections tend to use much of the same scientific terminology: aim,

evaluate, study, and assess. Therefore, Figure 4 validates that the proposed model is robust on the whole, but also it correlates the rhetorically subtle classes that are difficult to deal with.

4.5 Comparative Evaluation with Baseline Model

In order to confirm the usefulness of the given hybrid framework, the performance of the hybrid framework was performed in comparison with a traditional baseline model that relies on TF-IDF features and logistic regression. The results of comparison are as follows:

Table 2. Comparative Performance Analysis of Baseline and Proposed Hybrid Models

Model	Accuracy	Macro F1
Baseline Model	0.8076	0.7465
Proposed Hybrid Model	0.8771	0.8277

The proposed model improved the baseline by:

- 6.95 percentage points in accuracy
- 8.12 percentage points in Macro F1-score

This advantage shows that there is a need to integrate deep contextual sequence modelling in conjunction with engineered linguistic and positional features. Although the baseline uses mostly shallow lexical statistics, the suggested framework is able to capture deeper semantic links and rhetorical structures, which are vital in automatic knowledge extraction on scientific literature that is automatic.

4.6 Structured Knowledge Extraction Output

In addition to the classification of sentences, the suggested framework also provided structured knowledge representations of scientific abstracts. Outputs of the notebook indicate that the system is able to recreate the predicted sections of the system, including predbackground, predobjective, predmethods, predresults and pred-conclusions of each paper. That proves that the model does not rely only on categorical prediction but can also enable practical pipelines of knowledge extraction used to arrange the scientific content in a machine-readable manner. Moreover, relevant content-bearing terms were identified with the help of the keyword extraction of the outputs of the generated section. In demonstration, some of the keywords based on the objectives were weeks, systemic and sustained, whereas those based on the method were pain, scores, outcome and included. These results suggest that the domain-relevant information is contained in the extracted section-wise text and can be used in downstream applications, including summarization, indexing, evidence retrieval, and literature mining.

5. Discussion

The experimental findings prove that the suggested hybrid framework is rather effective in dealing with the dilemmas linked to the automated extraction of knowledge in scientific literature. The model can classify the test with a test accuracy of 87.71% and a macro F1-score of 82.77, which is much better than the baseline approach obtained by using deep learning and engineered linguistic features. The results emphasize the significance of semantic learning of context and analysis of structural features, which is consistent with the increased focus on data-centric and hybrid learning in the field of artificial intelligence [22]. Among the important findings of the findings is the fact that the model performs better when it comes to distinguishing structurally coherent sections, i.e. METHODS and RESULTS, with F1-scores exceeding 0.92. This is explained by the comparatively uniform linguistic patterns used in these sections in scientific literature. Particularly, document-based deep learning studies have demonstrated similar results with structured textual pieces being easier to model owing to their predictable format and use of terminology [23]. Conversely, the reduced scores in the OBJECTIVE and BACKGROUND sections indicate the fact that these categories are subject to variations and semantic overlaps. This issue is similar to the previous research in document analysis and sentiment classification, where ambiguous or context-sensitive texts reduce the accuracy of the classification[8]. It is also observed that the suggested model converges very fast and this is another indication of the applicability of hybrid feature fusion strategy as indicated in the training curves. The positional indicators and the statistical properties that are custom made contribute to the learning process by providing clear hints that deep learning models do not

necessarily choose. This helps argue the point that standalone end-to-end models can be augmented with additional engineered solutions, particularly with domain-specific use cases such as scientific text analysis [24]. In addition, self-attention mechanisms can be used to concentrate on the key keywords and phrases, which enhances the performance and interpretability of the model. This is in line with the recent developments in interpretable deep learning models, where attention-based models are becoming more popular as a way of improving accuracy and explainability[25]. The other significant issue in this study is the fact that structured knowledge (in unstructured abstracts) has been successfully reconstructed. The practical usefulness of the model even in areas not related to classification is illustrated by the fact that it is able to divide the text into meaningful parts and extract keywords that are related to the domain. This is especially useful with huge amounts of scientific knowledge in management, such as an automated knowledge management system needs to be created to store and retrieve information efficiently. Intellectual property analytics and research data mining Simultaneously with machine learning in natural language processing, similar applications have been investigated to derive information out of large document repositories[7]. The effectiveness of the proposed approach is also confirmed by the comparative analysis with the base model. It is confirmed that hybrid models offer a large benefit rate when compared to traditional machine learning methods, as the difference between the accuracy and the macro F1-score has improved by almost 7 and more than 8 percent, respectively. The results are in line with the recent reports that highlight the combination of NLP and deep learning to improve its functionality in complicated document analysis problems [26]. The findings also confirm the idea that analyses based on several analytical dimensions- including semantic, statistical and positional features- result in more solid and generalizing models. Although such encouraging outcomes can be observed, some limitations should be mentioned. The dataset used in the present research is also limited to one specific field, which is biomedical

literature, and this may not be extendable to other areas of science. Also, the overfitting behaviour observed in subsequent training epochs suggests that other optimization strategies might be viewed to improve model stability, including data augmentation techniques and regularization techniques. These problems are generic flaws with the application of artificial intelligence in other sectors, where scalability and adaptation are still the areas of critical concern [27]. Overall, the proposed hybrid model demonstrates great promise in the creation of automated knowledge extraction in the scientific literature. The model, which intertwines the NLP and the deep learning method into one data analytics model, not only enhances the performance of the classification but also provides a representation of useful knowledge. Aware document processing is an emerging research tradition, and these results can be used to point out the importance of hybrid treatments to real-world data challenges.

6. Conclusion

This study presented a hybrid NLP and deep learning-based framework for automated knowledge extraction from scientific literature. The proposed approach effectively integrates contextual embeddings, engineered linguistic features, and self-attention mechanisms to capture both semantic and structural characteristics of scientific text. The model demonstrated strong performance in classifying sentence-level information and reconstructing structured knowledge from unstructured abstracts. Compared to traditional baseline methods, the hybrid framework achieved significant improvements in accuracy and F1-score, confirming the effectiveness of combining multiple analytical techniques. Beyond classification, the system enables meaningful knowledge extraction by organizing scientific content into structured sections and identifying key domain-specific information. This highlights its potential for practical applications such as information retrieval, summarization, and large-scale scientific data analysis. The results indicate that hybrid approaches provide a more comprehensive understanding of scientific documents compared to standalone NLP or deep

learning methods. However, the study is limited by domain-specific data and minor overfitting during training. Future work may focus on extending the framework to diverse scientific domains, incorporating transformer-based architectures, and enhancing knowledge

representation through entity and relation extraction. Overall, the proposed framework contributes to the development of intelligent and scalable knowledge extraction systems for scientific literature.

References

- Z. Hong, L. Ward, K. Chard, B. Blaiszik, and I. Foster, "Challenges and advances in information extraction from scientific literature: a review," *JOM*, vol. 73, no. 11, pp. 3383–3400, 2021.
- G. Blanchy, L. Albrecht, J. Koestel, and S. Garré, "Potential of natural language processing for metadata extraction from environmental scientific publications," *Soil*, vol. 9, no. 1, pp. 155–168, 2023.
- J. Du, D. Wang, B. Lin, L. He, L. C. Huang, J. Wang, et al., "Use of deep learning-based NLP models for full-text data elements extraction for systematic literature review tasks," *Scientific Reports*, vol. 15, no. 1, p. 19379, 2025.
- T. V. Ivanisenko, P. S. Demenkov, and V. A. Ivanisenko, "An accurate and efficient approach to knowledge extraction from scientific publications using structured ontology models, graph neural networks, and large language models," *International Journal of Molecular Sciences*, vol. 25, no. 21, p. 11811, 2024.
- V. Sciannameo, D. J. Pagliari, S. Urru, P. Grimaldi, H. Ocagli, S. Ahsani-Nasab, et al., "Information extraction from medical case reports using OpenAI InstructGPT," *Computer Methods and Programs in Biomedicine*, vol. 255, p. 108326, 2024.
- Z. Shen, K. Lo, L. L. Wang, B. Kuehl, D. S. Weld, and D. Downey, "VILA: Improving structured content extraction from scientific PDFs using visual layout groups," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 376–392, 2022.
- L. Aristodemou and F. Tietze, "The state-of-the-art on Intellectual Property Analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data," *World Patent Information*, vol. 55, pp. 37–51, 2018.
- F. Alshuwaier, A. Areshey, and J. Poon, "Applications and enhancement of document-based sentiment analysis in deep learning methods: Systematic literature review," *Intelligent Systems with Applications*, vol. 15, p. 200090, 2022.
- S. Ranjeeth, T. P. Latchoumi, and P. V. Paul, "A survey on predictive models of learning analytics," *Procedia Computer Science*, vol. 167, pp. 37–46, 2020.
- J. Wątróbski, "Ontology learning methods from text-an extensive knowledge-based approach," *Procedia Computer Science*, vol. 176, pp. 3356–3368, 2020.
- D. Vlah, A. Kastrin, J. Povh, and N. Vukašinić, "Data-driven engineering design: A systematic review using scientometric approach," *Advanced Engineering Informatics*, vol. 54, p. 101774, 2022.
- Y. K. Dwivedi, A. Sharma, N. P. Rana, M. Giannakis, P. Goel, and V. Dutot, "Evolution of artificial intelligence research in Technological Forecasting and Social Change: Research topics, trends, and future directions," *Technological Forecasting and Social Change*, vol. 192, p. 122579, 2023.
- C. Collins, D. Dennehy, K. Conboy, and P. Mikalef, "Artificial intelligence in information systems research: A systematic literature review and research agenda," *International Journal of Information Management*, vol. 60, p. 102383, 2021.
- A. R. Munappy, J. Bosch, H. H. Olsson, A. Arpteg, and B. Brinne, "Data management for production quality deep learning models: Challenges and solutions," *Journal of Systems and Software*, vol. 191, p. 111359, 2022.
- A. Balla, M. H. Habaebi, M. R. Islam, and S. Mubarak, "Applications of deep learning algorithms for Supervisory Control and Data Acquisition intrusion detection system," *Cleaner Engineering and Technology*, vol. 9, p. 100532, 2022.
- K. S. Digamberrao and R. S. Prasad, "Author identification using sequential minimal optimization with rule-based decision tree on Indian literature in Marathi," *Procedia Computer Science*, vol. 132, pp. 1086–1101, 2018.

-
- Y. Liu, T. Zhao, W. Ju, and S. Shi, "Materials discovery and design using machine learning," *Journal of Materiomics*, vol. 3, no. 3, pp. 159–177, 2017.
- C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu, "Integrating machine learning with human knowledge," *iScience*, vol. 23, no. 11, 2020.
- M. H. A. Abdullah, N. Aziz, S. J. Abdulkadir, H. S. A. Alhussian, and N. Talpur, "Systematic literature review of information extraction from textual data: Recent methods, applications, trends, and challenges," *IEEE Access*, vol. 11, pp. 10535–10562, 2023.
- R. Asaad, R. I. Ali, and S. Almufti, "Hybrid Big Data Analytics: Integrating Structured and Unstructured Data for Predictive Intelligence," *Qubahan Techno Journal*, vol. 1, no. 2, 2022.
- M. Jansen, "PubMed 200k RCT," *Kaggle Dataset*. [Online]. Available: <https://www.kaggle.com/datasets/matthewjansen/pubmed-200k-rtc>.
- P. Singh, "Systematic review of data-centric approaches in artificial intelligence and machine learning," *Data Science and Management*, vol. 6, no. 3, pp. 144–157, 2023.
- P. Atandoh, F. Zhang, D. Adu-Gyamfi, P. H. Atandoh, and R. E. Nuhoho, "Integrated deep learning paradigm for document-based sentiment analysis," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 7, p. 101578, 2023.
- K. Lazović, F. Madeira, E. Zdravevski, L. A. Silva, P. J. Coelho, and I. M. Pires, "An analytical review of optimization techniques in information retrieval for enhanced decision support," *Decision Analytics Journal*, p. 100657, 2025.
- A. Asemi, M. Houshang, and N. Houshang, "An analytical deep learning framework for interpretable author similarity and research collaboration," *Decision Analytics Journal*, p. 100645, 2025.
- K. A. Shastry and A. Shastry, "An integrated deep learning and natural language processing approach for continuous remote monitoring in digital health," *Decision Analytics Journal*, vol. 8, p. 100301, 2023.
- X. Chen, H. Xie, D. Zou, and G. J. Hwang, "Application and theory gaps during the rise of artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 1, p. 100002, 2020.