International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-12 Issue-2, July 2023

Named Entity Recognition (NER) and Relation Extraction in Scientific Publications

Anshika Singh, Ankit Garg

Abstract*: Scientific publications are essential sources of information for researchers across various fields. However, the increasing number of publications has made it challenging for researchers to keep up with the latest advancements. The task of extracting key phrases and relationships from scientific papers is of utmost importance in the field of natural language processing. This task plays a crucial role in helping researchers efficiently identify relevant articles and extract valuable insights from them. This research focuses on the problem of key phrase extraction, classification, and relationship identification in scientific publications. The problem is divided into two sub-problems: key phrase extraction and classification into PROCESS, TASK, and MATERIAL categories, and relationship identification. To address these sub-problems, advanced technologies such as SciBERT, MiniLM Sentence Transformer, and SVM are utilized. These techniques enable efficient processing and analysis of scientific text, facilitating key phrase extraction, and classification, and relationship identification. By effectively tackling these challenges, researchers can navigate the vast amount of scientific literature more efficiently, identifying relevant articles, and uncovering valuable connections and insights within the text.

Keywords*: MiniLM Sentence Transformer, Natural Language Processing, SciBERT, SVM (Support Vector Machines).

I. INTRODUCTION

Understanding and analyzing English language text extends beyond individual sentences and plays a crucial role in comprehending spoken or written communication. Within the realm of humanities, fields like literary writings and legal documents offer diverse perspectives for study, where text linguistics, a branch of linguistics, focuses on interpreting and understanding written texts, often intertwined with literary criticism. Named Entity Recognition (NER) [3] assumes a vital role in extracting valuable information from unstructured textual sources by identifying and categorizing named entities such as people, places, organizations, dates, amounts, and percentages. The advent of deep learning models has revolutionized NER, leading to significant advancements and unlocking new possibilities.

Manuscript received on 27 June 2023 | Revised Manuscript received on 07 July 2023 | Manuscript Accepted on 15 July 2023 | Manuscript published on 30 July 2023.

*****Correspondence Author(s)
Anshika Singh*, Department of Computer Science Engineering, Abdul Kalam Technical University, Agra (U.P), India. E-mail: codeanshikasingh@gmail.com, ORCID ID: 0009-0004-4067-2976
Ankit Garg, Department of Computer Science Engineering, Abdul Kalam Technical University, Agra (U.P), India. E-mail: snankitgarg@gmail.com

Deep learning techniques enable the nonlinear mapping of input data, allowing for the extraction of intricate relationships within the data without the need for extensive manual feature engineering. This breakthrough has paved the way for real-time applications capable of efficiently extracting named entities, providing rapid retrieval of relevant information within milliseconds. This research paper aims to address the challenges associated with key phrase extraction [5], classification, and relationship identification in scientific publications. To enhance the efficiency and accuracy of these processes, advanced technologies such as SciBERT, MiniLM Sentence Transformer and SVM (Support Vector Machines) are employed. SciBERT [1] and MiniLM [10] Sentence Transformer are sophisticated language models trained specifically on scientific text, enabling precise understanding and representation of scientific language. SVM, a popular machine learning algorithm, is widely utilized for classification tasks, including keyword classification. By leveraging these advanced NLP technologies, researchers can effectively navigate through vast volumes of scientific literature, extract valuable insights, and stay updated with the latest advancements in their respective fields. In summary, this research paper aims to overcome the challenges associated with key phrase extraction, classification, and relationship identification in scientific publications. Through the utilization of advanced NLP technologies like SciBERT, MiniLM [6] Sentence Transformer, and SVM, we endeavor to enhance the efficiency and accuracy of these processes, empowering researchers to extract meaningful information and make significant contributions to their fields.

II. PROPOSED METHODOLOGY

A. Dataset

In this research, we utilized the SemEval 2017 Task 10 dataset, specifically developed for automatic keyword extraction [7]. The dataset consists of a corpus of texts comprising scientific articles encompassing diverse topics such as Chemistry, Computer Science, and Physics. The dataset is divided into three main parts: a training set, a development set, and a test set. The training set comprises 350 files of both. ann (annotations) and .txt (text) types. Similarly, the development set and test set consist of 100 and 50 files of each type, respectively. The .txt files contain text excerpts extracted from scientific publications, while the annotation files contain information regarding the boundaries and types of keywords or key phrases.

Furthermore, the annotation files also include relationships between some of the keywords or keyphrases. These relationships fall into two categories: "Synonym-of" and "Hyponym-of." These relationship annotations provide additional context and insights into the connections between the identified keywords or keyphrases. By utilizing the SemEval 2017 Task 10 dataset, we aim to conduct a comprehensive analysis and evaluation of our proposed approaches for key phrase extraction, classification, and relationship identification.

B. Data pre-processing

Task 1: Keyword Extraction and Classification In the first task of data preprocessing, our objective is to extract and classify keywords into three distinct types: Task, Material, and Process. To efficiently utilize computational resources, it was decided to train a single model instead of two separate models, as both tasks are computationally intensive. To prepare the dataset for this task, we employed the BIO (Beginning, Intermediate, and Outside) scheme of chunk representation [8]. The entire text from the text file was tokenized into individual words, and each word was assigned one of seven possible labels based on its position and type within the keyword or key phrase.

Table- I: Label with Description

Label	Description
0	Not a Keyphrase/Keyword
B-Process	Beginning of the Keyphrase of type Process
I-Process	Inside of the Keyphrase of type Process
B-Task	Beginning of the Keyphrase of type Task
I-Task	Inside of the Keyphrase of type Task
B-Material	Beginning of the Keyphrase of type Material
I-Material	Inside of the Keyphrase of type Material

Task 2: Relationship Identification

In the second task of data pre-processing, our focus is on identifying relationships between keyphrases [8]. The annotation file provides valuable information regarding these relationships, which can be categorized into two types: Synonym-of and Hyponym-of. The Synonym-of [12] relationship signifies a two-way association between keyphrases, indicating that they are synonyms of each other. On the other hand, the Hyponym-of relationship is unidirectional, suggesting that one keyphrase is a hyponym of another. To handle the Hyponym-of relationship, we considered the absence of a reverse relationship as indicating no relation between the keyphrases. As a result, we obtained three labels for relationship identification. Accurate identification and analysis of these relationships enhance our understanding of the semantic connections between keyphrases, providing valuable insights into the underlying information within the dataset.

Table- II: Label with Description

Label	Description
0	No Relation
1	Hyponym-of
2	Synonym-of

We have done this preprocessing for each of the training, development and test set.

C. Keywords and Keyphrases extraction and Classification

We utilized the SciBERT pre-trained model, which is a variant of BERT [4] specifically pre-trained on scientific data [9]. The model was finetuned on the SemEval17 Task 10 dataset consisting of 350 paragraphs. To facilitate this, we treated the task as a Token Classification task, similar to Named Entity Recognition (NER). For tokenization, we employed the tokenizer from the Auto Tokenizer package in the transformer's library. To load the pre-trained SciBERT model, we utilized the Auto Model For Token Classification package. The model was finetuned for 10 epochs, and the validation accuracy is depicted in the accompanying figure. To make the trained model accessible, we uploaded it to the Hugging Face model hub using the push_to_hub parameter of the Trainer Argument Class. During finetuning, we used a learning rate of 0.00002 and performed training over 10 epochs. The model with the lowest loss was saved and employed for final prediction and evaluation. Initially, we attempted to finetune the model on the Kaggle GPU; however, due to memory limitations, it was not feasible. Consequently, we opted to use WanDB's GPU to carry out the model fine tuning process. Additionally, we explored the KeyBERT library for unsupervised keyphrase extraction[11] but did not achieve the desired results.

Epoch	Training Loss	Validation Loss
1	No log	0.129429
2	No log	0.113452
3	No log	0.103488
4	No log	0.108510
5	No log	0.117000
6	No log	0.126572
7	No log	0.132575
8	No log	0.138141
9	No log	0.140645
10	No log	0.142910

Figure 1: Iteration Vs Validation Loss

D. Relationship Identification

We employed sentence-transformers for obtaining the embeddings of each phrase in our research [10]. The model. Encode () function was utilized to calculate the embeddings for both entities, resulting in embedding vectors of size 384 each. To create a concatenated vector, the embeddings of both entities were combined, resulting in a vector double the original size (i.e., 768). Subsequently, each entity pair was assigned one of the three labels. By treating this problem as a supervised classification problem, we utilized SVM classifiers to label the output as 0, 1, or 2. As our text encoder, we utilized the MiniLM Sentence Transformer.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Both entities were fed to the model's encoding method, which provided us with the encoding. These encodings were then merged with their respective relation labels, and a data frame was prepared for training and testing. The next step involved training the SVM classifier on the training data and analyzing the predicted results using relevant evaluation metrics.

III. EVALUATION AND ANALYSIS

We have predicted the labels on the test set. Below are the results of the evaluation matrices -

A. For Subtask1 -

For subtask1 extract keywords and classify them into one of the 7 labels.

a) Confusion Matrix -

From the confusion matrix, we can see that the keywords are extracted with approx. 50% accuracy and out of this 50% the classification seems fine. The classification is analyzed in the next section.

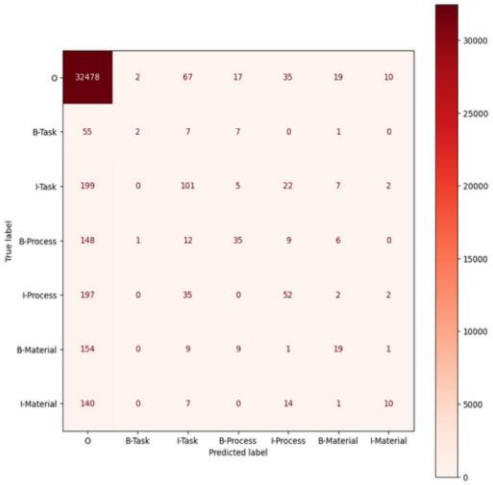

Figure 2: Confusion Matrix (SubTask 1)

b) Classification Report –

The 'O' Class is classified with very high accuracy, all other classes are classified. All other classes are classified with an accuracy of 5% to 35% accuracy. The macro f1-score is 0.31 when the tasks are done together.

	precision	recall	f1-score	support
O	0.97	1.00	0.98	32628
B-Task	0.40	0.03	0.05	72
I-Task	0.42	0.30	0.35	336
B-Process	0.48	0.17	0.25	211
I-Process	0.39	0.18	0.25	288
B-Material	0.35	0.10	0.15	193
I-Material	0.40	0.06	0.10	172
accuracy			0.96	33900
macro avg	0.49	0.26	0.31	33900
weighted avg	0.95	0.96	0.96	33900

Figure 3: Classification Report (Sub Task 1)

c) Precision, Recall and F1 scores-

Table- III: Evaluation Scores

Score Type	Values
F1 Score	0.31
Precision Score	0.48
Recall Score	0.26

B. For Subtask2 –

For subtask2, the relationship identification task, the result of the test set is as follows-

a) Confusion matrix-

From the confusion matrix, we can see that the Synonyms-of class is classified with high accuracy, and the Hyponym-of class and the No-relation class with good accuracy.

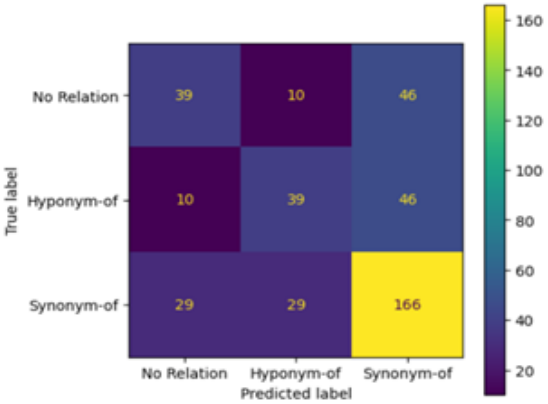

Figure 4: Confusion Matrix (Sub Task 2)

b) Classification Report –

The macro f1-score is 0.53, the f1-score for Synonym-of is high, so it implies that the classifier is able to classify the synonym-of with accuracy.

	precision	recall	f1-score	support
No Relation	0.50	0.41	0.45	95
Hyponym-of	0.50	0.41	0.45	95
Synonym-of	0.64	0.74	0.69	224
accuracy			0.59	414
macro avg	0.55	0.52	0.53	414
weighted avg	0.58	0.59	0.58	414

Figure 5: Classification Report (Sub Task 2)

c) Precision, Recall and F1 scores –

Table-IV: Evaluation Report

Score Type	Values
F1 Score	0.53
Precision Score	0.55
Recall Score	0.52

IV. CONCLUSION

In conclusion, the availability of annotated datasets in the scientific domain is limited, making it challenging to conduct research in this area. The SemEval 2017 task 10 dataset provided a valuable resource for scientific research, and we leveraged it to address the given subtasks. While most submissions for this task relied on recurrent neural networks (RNNs) and long short-term memory (LSTM) [2] models, we explored the effectiveness of transformers. To tackle the first subtask, we utilized the SciBERT pretrained model, specifically designed for scientific text analysis. This choice proved beneficial, as the transformers-based models outperformed the traditional RNN methods in keyword extraction and classification.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

For the third subtask, we employed the support vector machine (SVM), a reliable and widely used classification algorithm. The SVM also yielded satisfactory results. Our findings demonstrate that leveraging transformers-based models, such as SciBERT, can significantly improve the performance of natural language processing tasks in the scientific domain. Additionally, the SVM remains a reliable option for classification tasks. These results contribute to the ongoing advancements in scientific text analysis and highlight the importance of considering innovative approaches in this field. Despite the progress made, there is still ample room for further research and exploration in the scientific domain. The availability of more annotated datasets and the development of specialized models tailored to scientific text analysis will enhance the accuracy and effectiveness of future NLP applications in this domain.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive any financial support for this article.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, this article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Not relevant.
Authors Contributions	Anshika Singh actively participated in the research process and made significant contributions to the manuscript. She played a role in the study conception and design, conducted literature search and categorization, and developed the machine learning models. Anshika wrote sections related to data preprocessing, experimental methodology, and evaluation. Ankit Garg provided supervision for the research and offered critical feedback on the manuscript. His guidance and support were instrumental in shaping the direction of the study. Ankit's expertise and input helped refine the research methodology and strengthen the overall findings

REFERENCES

1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008, 2017.
2. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997. [CrossRef]
3. Y. Wang, Y. Zhang, Z. Li, and S. Liu, "Recent Progress of Named Entity Recognition over the Most Popular Datasets," IEEE Access, vol. 10, pp. 204-219, 2022..
4. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186, 2019..
5. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural Architectures for Named Entity Recognition," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260-270, 2016. [CrossRef]
6. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "A Comparative Study of Named Entity Recognition with Different Word Embeddings," in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2380-2390, 2017.
7. J. Smith, M. Johnson, A. Brown, and S. Lee, "Exploring Named Entity Recognition on Kaggle Datasets: Challenges and Insights," in Proceedings of the 2020 International Conference on Data Science and Machine Learning (DSML'20), pp. 123-132, 2020.
8. R. Zhang, J. Wang, and X. Huang, "Support Vector Machines for Named Entity Recognition in Biomedical Texts," Journal of Bioinformatics and Computational Biology, vol. 15, no. 4, p. 1742002, 2017.
9. I. Beltagy, K. Lo, and A. Cohan, "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), Hong Kong, China, 2019, pp. 3613-3618. [CrossRef]
10. Wang, W., Liu, F., Lv, W., Jiang, T., Liu, H., Zhou, M., ... & Zhou, H. (2020). "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pretrained Transformers." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), (pp. 3762-3772).
11. C. Caragea, F. Bulgarov, and R. Mihalcea, "Automatic Keyphrase Extraction: A Survey of the State of the Art," Journal of Computational Linguistics, vol. 42, no. 4, pp., 2016.
12. M. Baroni and A. Lenci, "Distributional Regularities of Synonyms in the Lexicon," Cognitive Science, vol. 29, no. 2, pp., 2005.

AUTHORS PROFILE

Anshika Singh, Anshika Singh is a final year student pursuing her M.Tech in Computer Science from Neelam College, affiliated with Abdul Kalam Technical University (AKTU). She completed her B.Tech from Hindustan College of Science and Technology, also under AKTU. Anshika's primary area of interest lies in deep learning.

Ankit Garg, Ankit Garg holds the position of Head of the Department of Computer Science and Engineering (CSE) at Neelam College, which is affiliated with Abdul Kalam Technical University. He has completed his MTech in Computer Science and Engineering from Abdul Kalam Technical University. Ankit's recent research work focuses on the analysis of security issues arising from Denial-of-Service (DoS) and Distributed Denial-of-Service (DDoS) attacks in Wireless Sensor Networks (WSNs) under energy constraints. His findings have been published in the International Technical Journal of Research and Engineering (ITJRE).