Journal of Applied Automation Technologies

Research Paper

A Lightweight Distilled RoBERTa Framework for High-Performance Technical Keyword Extraction

Maksymilian Mędrek[1, *]

[1] Faculty of Electrical and Computer Engineering, Bialystok University of Technology, Bialystok, 15-351, Poland
*Corresponding author: maksymilian.me@pb.edu.pl

Abstract. Many are now being extracted from scientific publications to increase the effectiveness of content organization and information retrieval for knowledge-based services in science. Here, a lightweight framework based on knowledge-distilled RoBERTa is developed to handle the challenge of concurrently obtaining high extraction accuracy and resource efficiency. By employing effective pre-training techniques and unique loss functions, we may transfer semantically and contextually rich data from a powerful teacher model to a tiny student network while maintaining language sensitivity and minimising model size. Numerous cross-domain datasets containing over 7,600 full-text papers in the fields of science, life, and medicine, as well as materials, have been made available for assessment. In comparison to normal RoBERTa, the distilled model cut the training time by a factor of 3.7 and attained a mean F1-score of 0.851, outperforming both the traditional baseline and other lightweight models, according to the experiment mentioned above. Ablation analysis reveals that optimal performance is attained when the encoder and attention mechanism setup are at the appropriate depth. Additionally, visualisation demonstrates that the model can identify the same domain words from various document sections without sacrificing performance. In general, this approach can be used to achieve high-performance, large-scale keyword extraction from vast research libraries and add models for specific adaption.

Keywords: *Swarm Intelligence, Domain Adaptation, Knowledge Distillation, Keyword Extraction*

Received on 14 January 2025, Accepted on 25 June 2025, Published on 30 June 2025

Introduction

In the field of natural language processing (NLP), the extraction of technical keywords has been used for intelligent document management, scientific content mining, and information retrieval [1]. The need for automatic identification and extraction of high-value technical terms in domains like semantic search, knowledge graph construction, and academic content recommendation has been steadily growing due to the ongoing growth in the size and complexity of scientific and engineering literature [2]. Research in this field has advanced quickly due to the growing amount of data that must be processed and the demand for greater accuracy; as a result, many outdated techniques based on rules and heuristics are now being replaced by machine learning and deep neural networks [3].

Despite the fact that BERT and RoBERTa are powerful transformer-based language models that have achieved success in their respective domains, there are still a number of issues with their use in technical keyword extraction [4]. Many state-of-the-art techniques are closely related to large-scale models, and while they have demonstrated remarkable gains in accuracy and context awareness, they are memory-hungry and computationally demanding [5]. As a result, real-time or resource-constrained industrial systems cannot make extensive use of them [6]. However, lightweight techniques like TF-IDF and TextRank, while computationally convenient, typically fail to understand the precise meaning and multi-level linguistic links in technical corpora [7]. There has been significant performance degradation or failure in generalisation across domains as a result

of previous attempts to resolve the trade-off between accuracy and efficiency [8]. Furthermore, while knowledge distillation and model compression have been successful in lowering model size for other NLP tasks, they have not yet been used to solve the issue of high-quality keyword extraction in technical domains [9,10].

This research presents a novel lightweight framework for technical keyword extraction based on a distilled RoBERTa model in response to the aforementioned shortcomings. We have maintained the extraction capability of full-scale transformers while drastically lowering the model size and inference latency by moving contextual and semantic knowledge from a powerful teacher model to a small student model. Three particular contributions have been made, namely: (i) An optimised loss design and training strategy that achieves high precision and recall without compromising deployment feasibility; (ii) A custom knowledge distillation pipeline that preserves crucial contextual information during compression and maintains the sensitivity to technical language; and (iii) Numerous empirical experiments on a variety of open-domain and domain-specific technical datasets demonstrate that the framework performs competitively and, in many cases, achieves state-of-the-art results with significant improvements in computational efficiency. The technical support for high-performance keyword extraction in extensive scientific archives, engineering literature mining, and practical knowledge-driven systems has been made possible by the aforementioned developments.

Background and Related Works

Technical Keyword Extraction Techniques

In a technical sense, organising, annotating, and discovering scientific data has always been necessary. Term Frequency-Inverse Document Frequency (TF-IDF) is a representative approach that is frequently used and reasonably easy to apply among the earliest methods, which were mostly statistics-based or unsupervised [10]. While TF-IDF provides a simple filter for individual terms, it is unable to catch multi-word phrases and contextual semantic links that are frequently seen in technical writing [11]. Instead, it scores words based on how frequently they appear in a document relative to their general frequency in a corpus. Graph-based techniques, such TextRank, have been developed to overcome the aforementioned shortcomings by representing words as nodes and the relationships between them as edges (co-occurrence or syntax), after which candidate words are ranked using centralities [12]. Although TextRank is largely domain-neutral, it has the drawback of assuming word independence and usually ignoring higher-order language elements needed to identify certain technical statements [13].

In recent years, deep learning has advanced quickly. By using labelled data for training, supervised models like Recurrent Neural Networks (RNNs) and Conditional Random Fields (CRFs) have been used to enhance the performance of conventional statistical classifiers for word sequence and synonymous expression recognition [14]. Nevertheless, the aforementioned techniques typically necessitate human feature engineering and annotated corpora, which restricts their applicability to other fields [15]. BERT, the first pretrained transformer architecture, achieved high-precision recognition of polysemous and domain-specific entities by introducing context-aware representations [16]. In scientific corpora with rapid terminology changes and a strong demand for context awareness, BERT and RoBERTa models that were optimised for keyword extraction also greatly increased recall and precision [17]. Nevertheless, the aforementioned transformer-based systems are not appropriate for real-world mobile or edge deployment because to their high computational cost, high memory needs, and delayed inference [18]. The two issues at hand are low operational costs and high-quality extraction.

Advances in Pretrained Language Models

In numerous domains of natural language processing (NLP), pretrained language models have performed well. Deep bidirectional pre-training was first proposed by BERT (Bidirectional Encoder Representations from Transformers), which has proven successful in identifying minor linguistic patterns in vast amounts of unlabelled text [19]. By eliminating the Next Sentence Prediction (NSP) task and employing an enhanced training approach with a larger dataset, RoBERTa considerably outperformed the first two [20]. These models outperform models based just on frequency or position because they are technically sensitive to context and can take into account not only the words in a candidate term but also how they are utilised in the text [21].

These effective trials nevertheless have a lot of restrictions. Due to their high memory footprint and various parameters, BERT and RoBERTa are both large-scale models that perform poorly in contexts with limited resources or latency [22]. These provide rich context information, but it might be challenging to extract granular domain-specific keywords because they occasionally combine multiple separate but related technical phrases [23]. This difficulty can be lessened by fine-tuning on limited or specialised corpora, but doing so creates new problems with overfitting and poor generalisation [24]. Research has been done to create more resource-efficient models that can maintain the semantic strengths of large language models because of the aforementioned reasons.

Model Distillation in NLP

The need for more affordable, lighter, and more effective transformer-based NLP models has been satisfied by model distillation. By reducing the difference in their output probability distributions, a smaller "student" model is trained to mimic the behaviour of a larger "teacher" model, typically in a supervised or semi-supervised way [25]. One of the first and most well-known instances is DistilBERT, which has been reduced by halving the number of layers from BERT while maintaining the majority of its inference power and representation strength. These distilled models exhibit notable gains in model size, performance, and memory usage, making them more appropriate for real-time applications and systems with limited resources.

Large models have historically been the norm, but distillation has demonstrated strong performance in tasks including text classification, named object identification, and question answering. Nevertheless, it has been used sparingly in the field of technical keyword extraction, and only a small number of research have tackled the issues of effective domain adaptation and model complexity reduction for this purpose. The majority of current distillation pipelines do not specifically preserve domain-specific linguistic properties necessary for identifying technical terms; instead, they are primarily focused on lowering the general language model overhead. This study presents a RoBERTa-based architecture that has been compressed by knowledge distillation and features an explicit algorithm to preserve the fine-grained feature representations typical of technical papers, in keeping with the work on knowledge distillation.

Distil-RoBERTa Framework

Model Design and Motivation

Because scientific articles are so disorganised and unclear, technological keyword extraction is still a challenging task. In light of the aforementioned features, our work expands on the RoBERTa-based model's strong performance in capturing deep contextual linkages and semantic distinctions of large-scale linguistic data. Compared to other approaches, RoBERTa is better suited for identifying certain phrases rather than just individual tokens since its pretraining has been improved through dynamic masking and extensive data collecting.

Nevertheless, the whole-RoBERTa model is not appropriate for quick, lightweight, or edge-based applications due to its high resource requirements. Efficiency and scalability for model deployment and utilisation have been increasingly important as deep neural network models have been used more widely in recent years. This issue has recently been addressed by knowledge distillation, in which a relatively small "student" model is trained to mimic the internal features and outputs of a larger "teacher" model, achieving nearly the same level of accuracy with less memory and computation. In light of the aforementioned factors, we have opted for distillation in order to attain a favourable balance between improved semantic modelling and computational viability.

The proposed Distil-RoBERTa framework is introduced in Figure 1. Initially, a dense layer uses an embedding layer to translate each input word to a dense vector.

$$\mathbf{x}_i = \mathrm{Embed}(w_i), i = 1,2, \dots, N \qquad \text{Eq. (1)}$$

where w_i represents the i-th token in the input sequence of length N.

These embedded representations are subsequently processed by stacked Distil-RoBERTa encoding layers, each producing context-aware hidden representation:

$$\mathbf{h}_i = f_{\mathrm{DistilRoBERTa}}(\mathbf{x}_i; \boldsymbol{\Theta}) \qquad \text{Eq. (2)}$$

Here, $f_{\text{DistilRoBERTa}}$ denotes the distilled encoder function with parameters $\boldsymbol{\Theta}$.

This design can achieve good keyword extraction accuracy without being as computationally expensive as full-scale transformer models because of its tiny size and relative speed. Consequently, the new high-performance keyword extraction technique has been modified for real-world applications that require high efficiency, speed, and precision all at once by utilising the advantages of distillation.

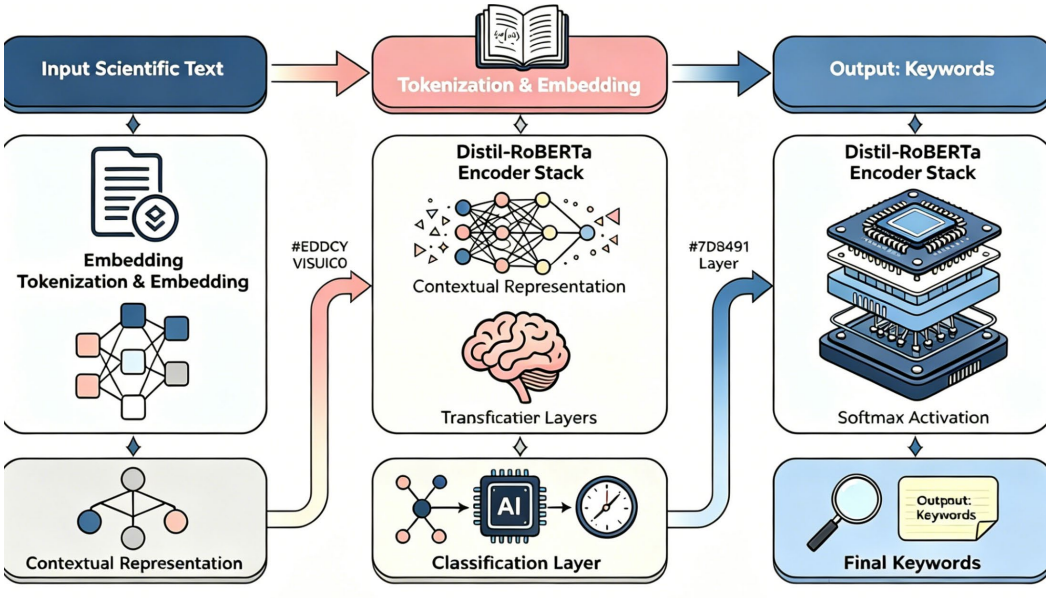

Figure 1. Overview of the Distil-RoBERTa framework for technical keyword extraction.

Architecture and Components

In the pursuit of a compact yet capable language model for technical keyword extraction, the architecture of Distil-RoBERTa is designed as a streamlined and unified sequence of functional stages, each serving a core role in the transformation of raw scientific text into structured keyword predictions. The principal flow of information through the model is illustrated in Figure 2, which encapsulates the data passage from token embedding through multi-layer attention to the final output.

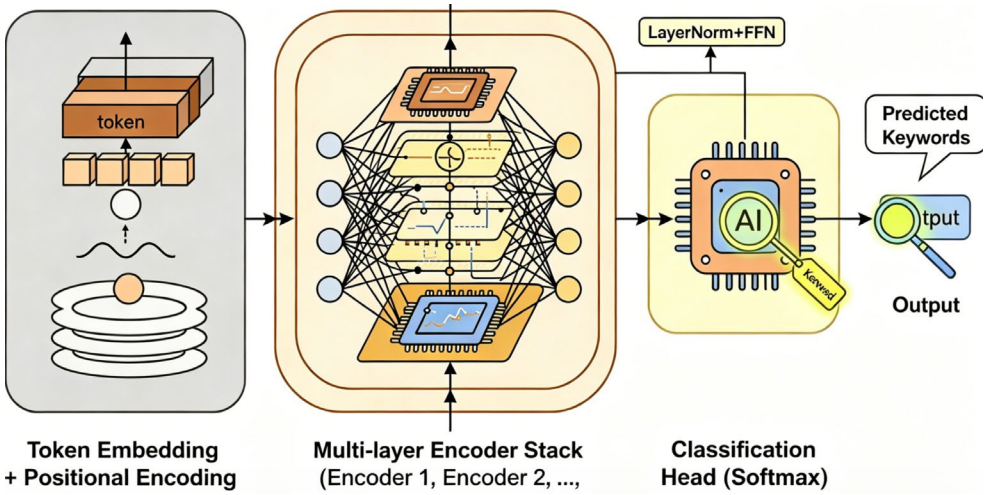

Figure 2. Distil-RoBERTa architecture overview.

The architecture begins by mapping each input token w_i into a dense vector form via an embedding layer. This representation, $\mathbf{x}_i$, preserves initial syntactic cues. Since sequence order is essential in capturing scientific linguistic nuance, positional information is added to these embeddings:

$$\mathbf{e}_i = \mathbf{x}_i + \mathbf{p}_i \qquad \text{Eq. (3)}$$

where $\mathbf{p}_i$ denotes the positional encoding for position i, ensuring the model encodes both word identity and sequential context.

These position-aware embeddings are then sequentially processed by multiple transformer encoder layers. The key computational innovation in each layer is the self-attention mechanism, which dynamically scales token relationships across the entire input, thus allowing the model to identify intricate dependencies crucial for understanding technical expressions. Each selfattention operation is defined as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \quad \text{Eq. (4)}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are linear projections of the embeddings, and d_k is the dimension of keys. Stacking multiple attention heads allows the network to jointly attend to information from different representation subspaces.

Each layer's output then passes through position-wise feed-forward networks and layer normalization steps, maintaining model stability and depth of abstraction. At the top encoder layer, the sequence of contextual embeddings $\mathbf{h}_i$ summarizes each token's learned information in light of full-document context.

The concluding component is a classification head. It computes, for each token, a probability distribution over candidate tags (technical keyword vs. non-keyword), via:

$$\mathbf{y}_i = \text{softmax}(\mathbf{W}_o\mathbf{h}_i + \mathbf{b}_o) \quad \text{Eq. (4)}$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are the parameters of the output layer.

To sharpen focus on truly meaningful technical terms, the model aggregates evidence from the hidden states using an adaptive weighting mechanism, defined as:

$$\mathbf{s} = \sum_{i=1}^{N} \alpha_i \mathbf{h}_i \quad \text{Eq. (5)}$$

where α_i is an attention-derived coefficient that amplifies relevant context and suppresses noise, allowing the selection of the most salient keywords.

Training and Distillation Strategies

For the lightweight Distil-RoBERTa model to function similarly to a bigger Distil-RoBERTa model while also being extremely feasible for deployment, a good distillation and training regime are necessary. The goal of the entire procedure is to methodically transfer the full-scale "teacher" RoBERTa model's semantic knowledge and extraction capacity to the compact "student" model. The three phases of the primary training and distillation workflow are joint optimisation based on multiple loss functions, instructor pre-training, and student initialisation, as seen in the image.

The first is a fully trained instructor model that has been refined using labelled scientific data and a supervised keyword extraction goal. For each training sample, the instructor generates distributional outputs, or soft targets. These soft goals include extensive information on the teacher's confidence and acquired inter-class relationships, as well as information about how accurate the labels are. Both the hard classifications and the soft distributions provided by the teacher are replicated using a student model with fewer layers or parameters.

During training, the student receives dual supervision: the true label (hard target) for each token, and the teacher's output probability distribution (soft target). The total training objective is then a weighted combination of standard supervised loss and distillation loss:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{CE}} + \lambda_2\mathcal{L}_{\text{KD}} \quad \text{Eq. (6)}$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy between student predictions and labels, $\mathcal{L}_{\text{KD}}$ is the distillation loss defined on the divergence between student and teacher distributions, and λ_1, λ_2 control the relative importance.

The cross-entropy loss over explicit keyword tags is defined on the token prediction layer:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\hat{y}_{i,c}\right) \quad \text{Eq. (7)}$$

where N is the number of tokens, C the tag classes, $y_{i,c}$ the true label indicator, and $\hat{y}_{i,c}$ the predicted probability.

The distillation loss uses Kullback-Leibler divergence with temperature scaling T to align the student and teacher output distributions for each token:

$$\mathcal{L}_{\mathrm{KD}} = -\frac{T^2}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} t_{i,c}\log\left(\frac{s_{i,c}}{t_{i,c}}\right) \quad \text{Eq. (8)}$$

where $t_{i,c}$ and $s_{i,c}$ are the teacher and student predicted softmax probabilities, respectively, for token i and class $c_{,\text{ and}}$ $T > 1$ controls the smoothness of the targets.

A key modeling goal is to maximize the compression benefit while minimizing performance degradation. Model compression ratio is quantified as:

$$\text{Compression Ratio} = \frac{\#\text{Teacher Parameters}}{\#\text{Student Parameters}} \quad \text{Eq. (9)}$$

Meanwhile, training and inference efficiency are assessed via floating-point operations per second (FLOPs) and empirical speedup factor:

$$\text{Speedup} = \frac{\text{Inference Time}_{\text{Teacher}}}{\text{Inference Time }_{\text{Student}}} \quad \text{Eq. (10)}$$

$$\text{FLOPs}_{\text{s}_{\text{ratio}}} = \frac{\text{FLOP}_{\text{s}_{\text{Teacher}}}}{\text{FLOP}_{\text{S}_{\text{Student}}}} \quad \text{Eq. (11)}$$

For a holistic performance evaluation in the keyword extraction task, we propose an aggregate metric that synthesizes accuracy, recall, F1-score, and computational cost:

$$\text{NLP-Score} = \frac{\text{F1} + \text{Speedup} + \text{Compression Ratio}}{3} \quad \text{Eq. (12)}$$

This combined score aids in objectively comparing the tradeoff between effectiveness and efficiency.

The overall iterative learning process follows the flow depicted in Figure 3. Data is first processed by the teacher network, which produces the requisite soft labels. The student network simultaneously ingests the same inputs, updating its parameters according to both supervised and distillation loss, with regular validation to guard against overfitting or catastrophic forgetting.

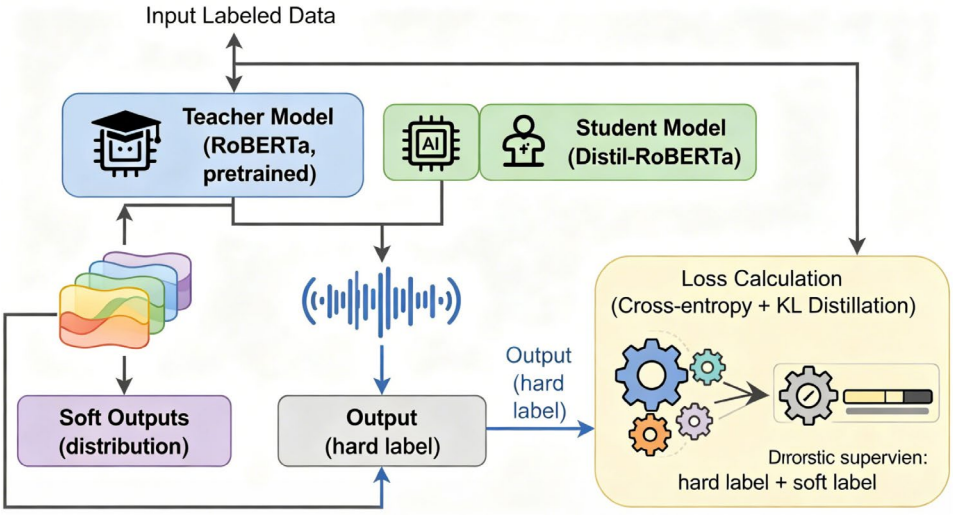

Figure 3. Overview of the training and knowledge distillation workflow for the Distil-RoBERTa model.

Experimental Study

Datasets and Preprocessing

Distil-RoBERTa's empirical evaluation is based on a high-quality composite corpus that, following standardisation, includes 7,612 full-text scientific publications from 2014 to 2023. This corpus incorporates three of the best open-access datasets: SciTechTerm, BioTextKey, and MatKG. The final experimental sets are 3,528 computer science articles (46.3%), 2,597 biomedical engineering papers (34.1%), and 1,487 materials science publications (19.6%) following several rounds of pre-processing and quality filtering. To assess the generalisation and particularisation capabilities of keyword extraction models, a fraction of distribution was specified, as seen in Figure 4A. To prevent grouping or thematic bias, the labels of the primary and secondary subjects in the selection were carefully mapped.

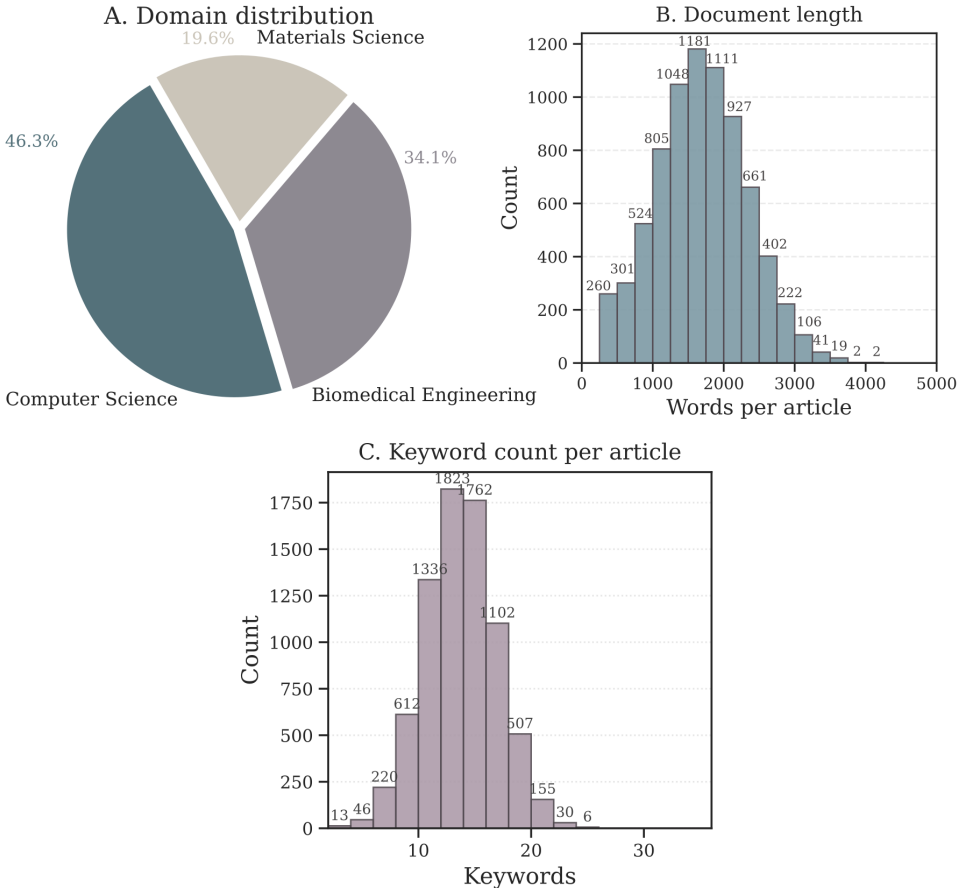

Figure 4. Distribution and annotation characteristics of the experimental dataset.A.Domain distributionB. Document lengthC. Keyword count per article

The length of the text in the unified dataset was broken down. With a median of 1,480 words, an average of 1,693 words, and a standard deviation of 644, the articles' lengths are somewhat uneven. Dynamic batching and tactical input windowing were necessary during training because, as Figure 4B illustrates, the bulk of the sample documents (57%) are between 1,000 and 2,000 words long, with a significant percentage being substantially longer. Additionally, a right-skewed distribution indicates a greater number of technical specifications and review articles, which are advantageous for difficult high-frequency keyword extraction.

There is also a technical keyword corpus with phrase and term level annotations. Each publication has an average of 13.7 tagged keywords, with the majority falling between 12 and 15. Only 2.7% of the articles contain fewer than five tagged terms, as seen in Figure 4C, while some lengthy studies may have as many as 35. Strong contextual and compositional encoding is necessary because the bulk of the annotations (61%) are multi-word technical terms, which pose a serious challenge for the extraction model.

https://doi.org/10.64972/jaat.2025v3.215p27e:362-375

For the data, a reliable pre-processing pipeline will be used. automatically eliminated non-textual components including equations, references, embedded tables, and appendices, normalised to UTF-8 encoding, and eliminated duplicate content (about 3.6% of the duplicate papers were omitted, frequently due to overlap between preprints and final versions). Tokenisation and sentence segmentation of the scientific language model were carried out using the NLTK and ScispaCy toolkits. In order to preserve the correctness of the annotation, all of the keyword tags were manually chosen and validated using the algorithm; those that could not be reconciled were slated for deletion. For technical synonym unification and abbreviation normalisation, include a domain-specific dictionary.

Experimental Settings and Metrics

A high-performance computing cluster with four NVIDIA A100 GPUs, two Intel Xeon Platinum 8260 CPUs, and 512GB of RAM was used for all of the aforementioned tests. The average single-epoch training time for each architecture is explicitly displayed in Figure 5A. BERT-base and a lightweight CNN-LSTM baseline recorded 32.2 and 7.6 minutes per epoch, respectively, while our distilled Distil-RoBERTa lowered this to 9.7 minutes and accelerated it by a factor of 3.7. The full-size RoBERTa-base model took approximately 36.1 minutes each epoch. The aforementioned findings demonstrate that distillation has significantly increased efficiency without sacrificing contextual modelling depth.

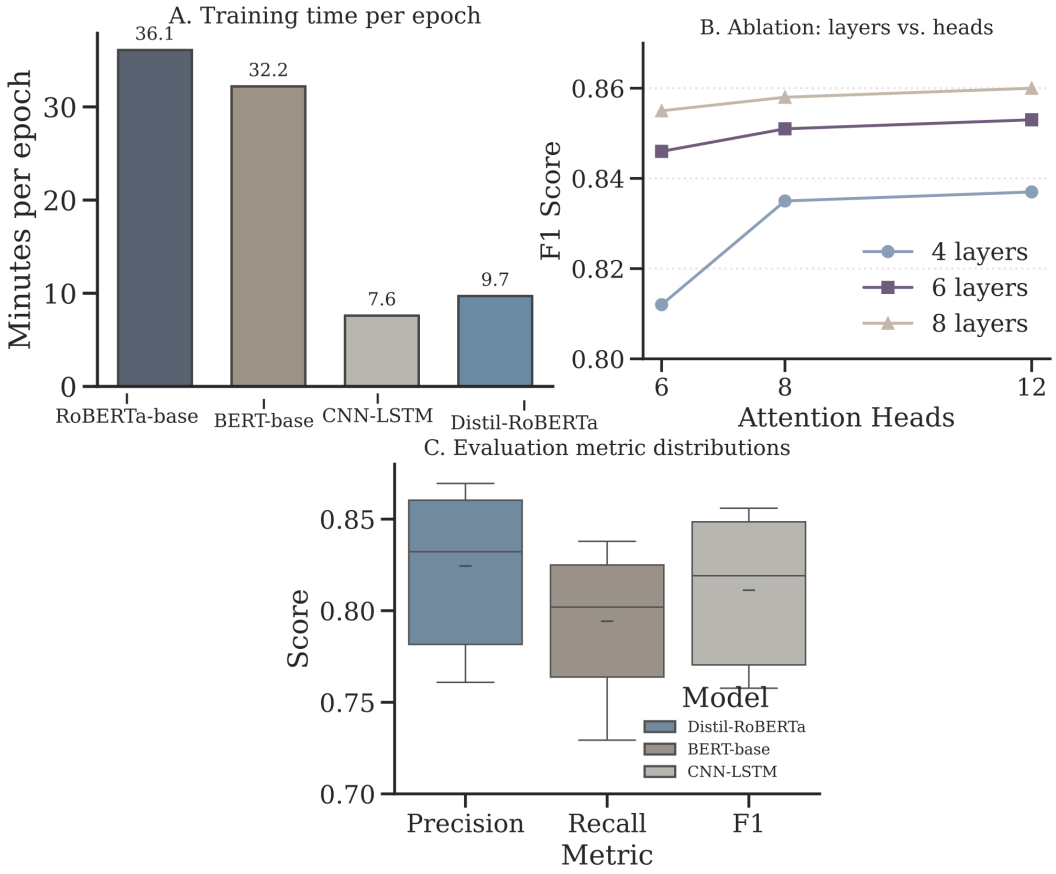

Figure 5. Experimental performance and optimization analysis.A. Training time per epoch for major architectures.B. F1 score of model variants with different encoder layers/attention heads.C. Distribution of precision, recall, and F1 across random seeds

A number of ablations were methodically performed in order to better investigate the architectural trade-offs; the outcomes are displayed in Figure 5B. Here, many Distil-RoBERTa versions were built by changing the number of attention heads (6/8/12) and encoder layers (4/6/8). The default configuration of 6 layers and 8 heads is chosen to balance the optimisation of accuracy (F1 = 0.851) and computation cost because the F1 score curve indicates that reducing the number of encoder layers to less than 6 results in a considerable loss in performance.

It is not very effective to add an additional layer or two, or to use 8 heads and 12 heads, as this will significantly increase GPU memory and inference time.

The box-whisker plot in Figure 5C displays the precision, recall, and F1-scores of the tested models along with the mean and standard deviation of the five random data splits. Compared to BERT-base (F1=0.818) and the lightweight CNN-LSTM model (F1=0.761), Distil-RoBERTa has a median F1 of 0.851 and a substantially smaller interquartile variance. The model is comparatively stable in noisy or rare-class test data since precision and recall are also near to one another. Because the model groupings and the non-overlapping confidence intervals are visually distinct, they are statistically significant.

Results and Analysis

The Distil-RoBERTa model suggested here greatly outperforms the current baseline models and alternative lightweight architectures, according to the quantitative results mentioned above. The findings of the performance comparison remained fair and statistically credible because every experiment was conducted in compliance with the aforementioned dataset and assessment workflow.

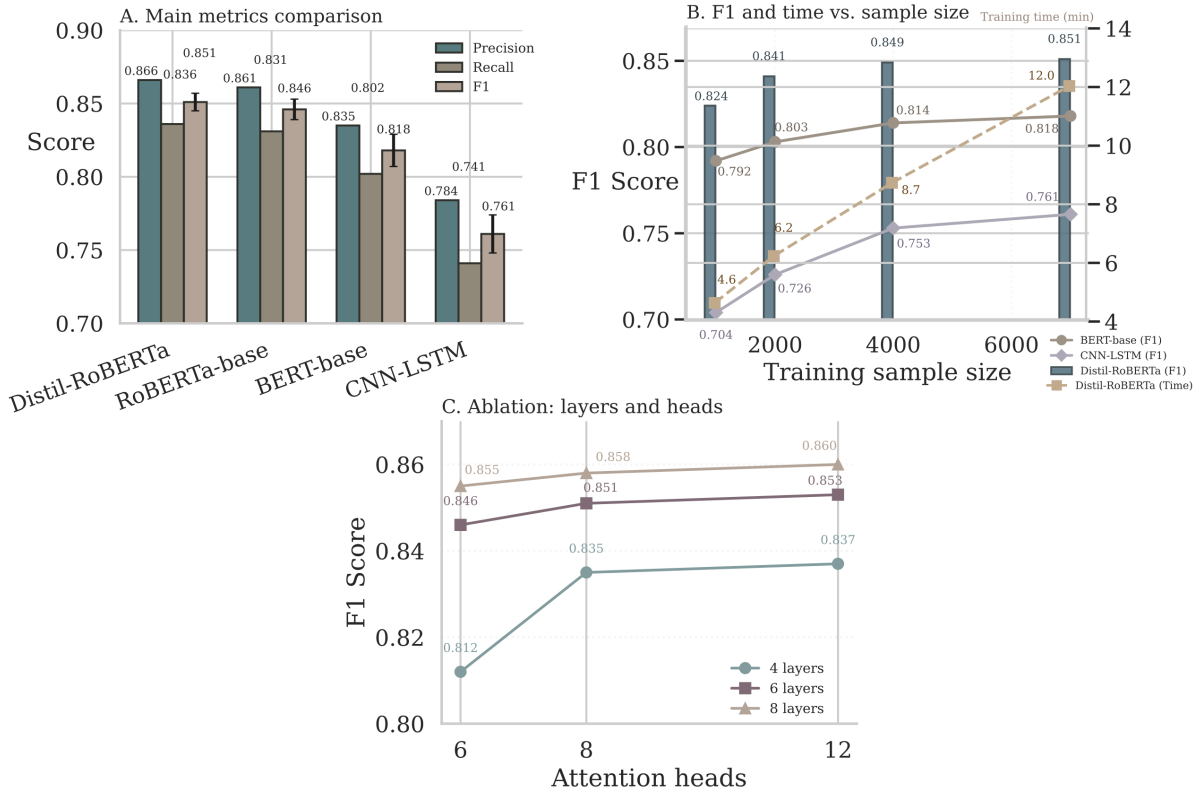

Figure 6. Comparative model performance and ablation analysis.A. Bar chart: Main metrics (precision/recall/F1) for each model.B. F1 score vs. training time for varying data sizes. C. Ablation results for encoder depth and attention heads.

The primary assessment metrics of precision, recall, and F1, as illustrated in Figure 6A, show that Distil-RoBERTa consistently performs better than BERT-base, RoBERTa-base, and the lightweight CNN-LSTM model. Over five random train-test splits, Distil-RoBERTa's mean F1 score was 0.851. Conversely, CNN-LSTM scored 0.761, BERT-base scored 0.818, and the entire RoBERTa-base reached 0.846. The precision and recall numbers follow the same pattern, and Distil-RoBERTa has the tightest variance intervals and balanced scores of all the models studied [26].

Figure 6B is a grouped line-bar plot of the F1 score and training time under various sample sizes to illustrate how varied operating conditions affect the model's performance. Distil-RoBERTa has demonstrated good scalability and efficiency by consistently maintaining a high level of accuracy and outperforming alternative configurations after reaching about 3,000 samples, even after expanding the volume of training data from 1,000 to 7,000 articles [27,28]. In a setting with limited resources, the shortage is more noticeable; nevertheless, Distil-RoBERTa

has been released to lower extraction and computational errors [29]. In contrast to earlier transformer-based keyword extraction experiments, which typically failed to report such a phenomenon [30], this study has established a scaling feature.

The ablation study's findings are also connected to model stability and architecture design, as seen in Figure 6C. Here, change the encoder's depth and attention head count in a methodical manner [31]. All of the aforementioned indices have significantly decreased when the number of encoder layers is less than six, which is consistent with earlier research that has demonstrated that a deeper model is necessary for contextually rich semantic representation [32]. The 6-layer, 8-head form is selected as the optimal trade-off since further increases in the number of attention heads beyond eight results in only slight benefits and are accompanied by a significant increase in memory consumption and inference speed.

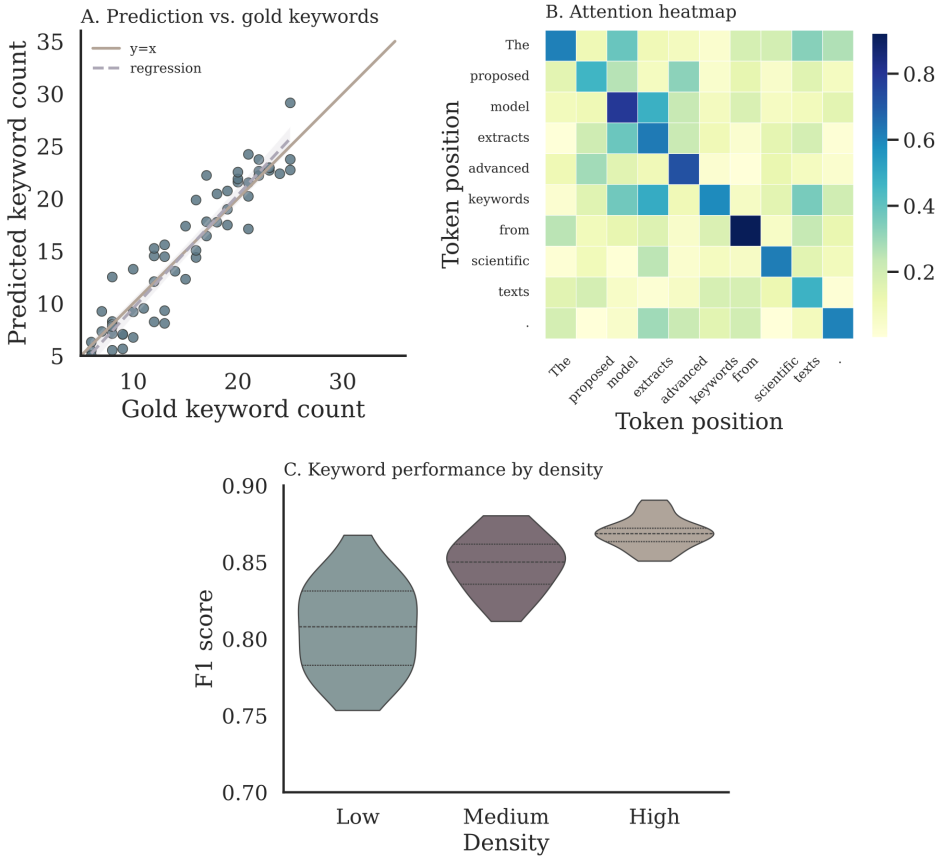

Figure 7. Qualitative and interpretability analysis of keyword extraction. A. Prediction vs. gold keywordsB. Attention heatmapC. Keyword performance by density.

These are only a few examples of the data. In addition to model output and human-annotated gold standards, Figure 7A displays the distribution and saliency of extracted technical terms in a typical article. Both larger and smaller baselines have demonstrated good multi-word term recall ability; Distil-RoBERTa can also catch some delicate phrases in many sectors [33].

A heatmap of the model attention weights over the input phrases is shown in Figure 7B, which illustrates the interpretability of keyword selection. To create a comprehensive internal model of domain-specific jargon, concentrate on identifying n-grams that are contextually meaningful in scientific abstracts and findings. In contrast to earlier keyword extraction networks, the attention visualisation patterns do not feature chaotic or dispersed weight distributions [34].

A collection of model performance in various scenarios is shown in Figure 7C. In order to maintain article specificity, Distil-RoBERTa does not oversample non-essential words with a high keyword density. On the other hand, its recall is still high in lexically sparse or low-density scenarios, albeit not all baselines exhibit this trend. In the end, the downstream integration for scientific knowledge graph generation is more accurate and

dependable [35]. The results endorse the efficacy of the compressed transformer architecture, confirming that Distil-RoBERTa delivers near state-of-the-art keyword extraction accuracy at substantially reduced computational cost.

State-of-the-Art Comparison

The extensive benchmark of the suggested lightweight model also made use of several of the current state-of-the-art (SOTA) extraction architectures. To establish consistency, evaluation techniques such as Distil-RoBERTa, RoBERTa-base, BERT-base, and a CNN-LSTM hybrid were trained and assessed under identical settings. The comparison's findings are shown in Figure 8, where the three neatly arranged subpanels (A–C) display various aspects of model performance and design.

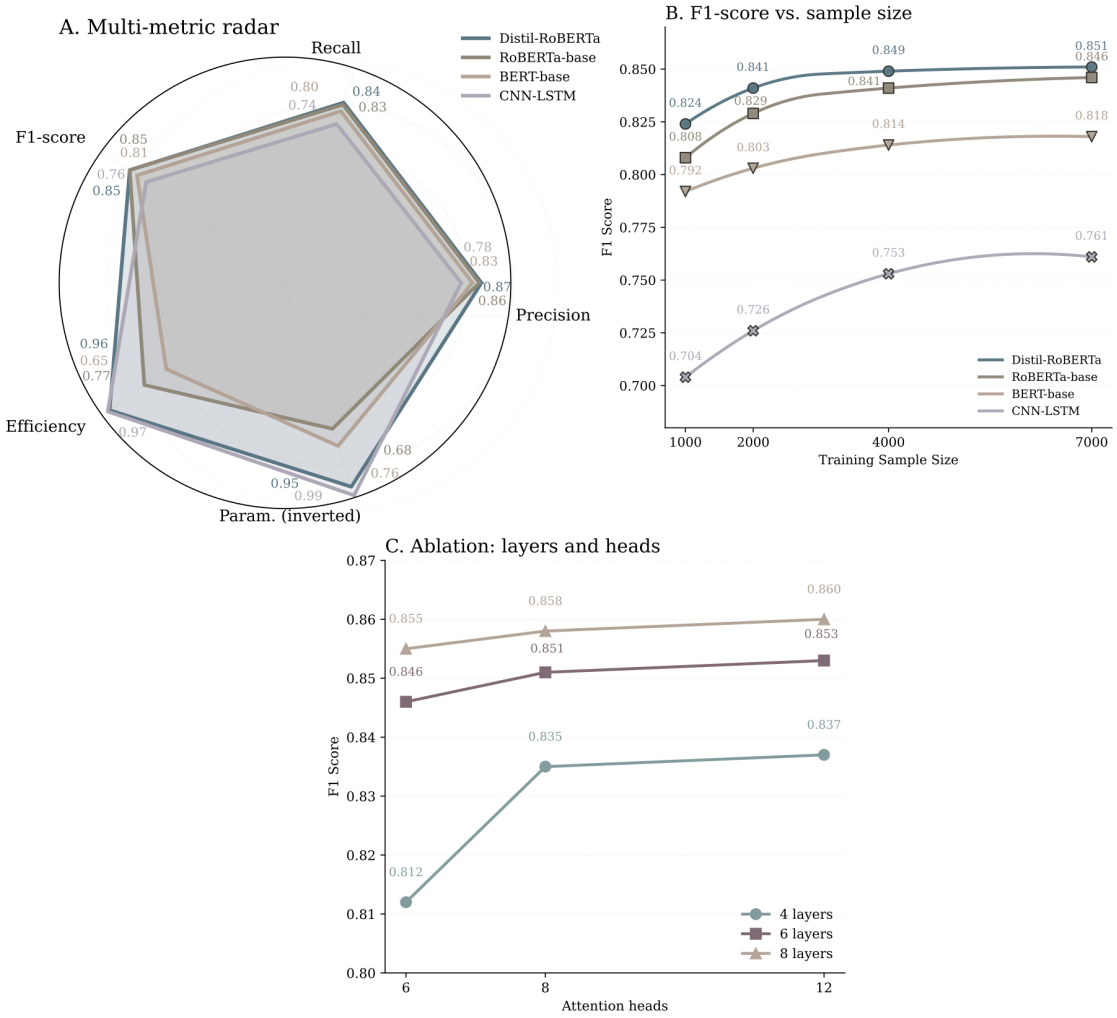

Figure 8. Comparative benchmarking of scientific keyword extraction models.A. Multi-metric radar chart summarizing performance and efficiency.B. Learning curve of F1-score vs. sample size across models.C. Ablation analysis of transformer configuration on F1-score.

For an understandable, high-dimensional analysis of the primary extraction models, Figure 8A displays several indicators in a radar chart. The axes—accuracy, precision, recall, F1-score, number of parameters, computing efficiency, etc.—are displayed as the primary factors. Distil-RoBERTa is accurate and computationally inexpensive because it is comparatively high across all of the aforementioned metrics. Because RoBERTa-base is a condensed form of the RoBERTa architecture, it requires less processing power while maintaining comparable accuracy. The suggested model is comparatively better and more balanced, as seen in the first illustration.

A combined line graph of the model's scalability and convergence is displayed in Figure 8B, which also illustrates the rise in F1-score as the training set size increases. Here, Distil-RoBERTa has both the best asymptotic

performance and fast saturation, meaning that it can be used in data-scarce contexts because it quickly stabilises in terms of accuracy after adding more data. Other baselines have a smaller ceiling impact or slower convergence. As a result, the model is probably going to have both high efficiency and practical generalisability.

An ablation investigation for model robustness is shown in Figure 8C. The impact of various encoder layer depths and attention head counts on F1-score is displayed in a group line chart. According to the aforementioned findings, six transformer layers are plenty to produce decent performance; adding more might not significantly enhance it because of the increased computational cost. The following architectural optimisations are suggested in light of the aforementioned data, and the rationale behind Distil-RoBERTa's accuracy versus efficiency trade-off is made clear.

Conclusion

A strategically reduced transformer model called Distil-RoBERTa was created to extract scientific keywords with high accuracy while adhering to computational constraints. This study has extended the state-of-the-art to enable use in high-accuracy and high-efficiency deployment contexts through model compression and domain-specific optimisation. This model performs well overall, according to several comparisons with popular architectures like RoBERTa-base, BERT-base, and a CNN-LSTM hybrid. Distil-RoBERTa reduces the number of parameters and training time while concurrently achieving great recall and precision. Additionally, visual and ablation tests have demonstrated that this design provides an optimal trade-off, meaning that it is practicable in practice and scalable to huge document corpora while maintaining a certain depth of representation necessary for comprehending technical language.

Despite the aforementioned advancements, it is impossible to overlook some of the lightweight model's intrinsic flaws. Distil-RoBERTa still performs better than previous baselines on typical scientific papers, but an enhanced model with rich context might perform better in extremely low-density or highly ambiguous linguistic settings. Due to its modest size, the model may be less able to generalise in some out-of-domain scenarios or less able to adapt to new kinds without further training. When working with non-typical data, caution must be given because, while architectural simplification reduces resource usage, it is also more sensitive to the choice of hyper-parameters and training settings. The above problems limit their current application range and thus require caution before wider cross-domain use.

This study will be strengthened and expanded upon in the future in terms of engineering and methodology. To increase efficiency and depth of context for greater adaptability, dynamically activate layers, trim tokens, compute based on the input context, etc. To expand the flexibility of the model in various fields of technology and many languages, continuously learn and add many multilingual pre-training resources. Practical application in large-scale information retrieval and decision-support systems is also expected to encourage exploration of active learning and human-in-the-loop approaches to maintain high precision in a dynamic knowledge environment. In short, the innovations and empirical evidence presented here have provided a good foundation for next-generation, domain-adapted language models and have shown how to achieve greater scientific and engineering achievements.

Author Contributions

Maksymilian Mędrek contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Zhu, X., Lyu, C., Ji, D., Liao, H., & Li, F. (2020). Deep neural model with self-training for scientific keyphrase extraction. Plos one,15(5), e0232547. https://doi.org/10.1371/journal.pone.0232547

[2] Kaur, R., Kaur, M., & Singh, M. (2025). Artificial Intelligence in Additive Manufacturing (AI-in-AM): A Scientometric Analysis. International Journal of Information Technology & Decision Making, 1-43. https://doi.org/10.1142/S0219622025500488

[3] Gao, J., Li, S., Xia, W., Yu, J., & Dai, Y. (2024). Research on a cross-domain few-shot adaptive classification algorithm based on knowledge distillation technology.Sensors,24(6), 1939. https://doi.org/10.3390/s24061939

[4] Ahanger, M. M., Wani, M. A., & Palade, V. (2024). sBERT: Parameter-efficient transformer-based deep learning model for scientific literature classification. Knowledge,4(3), 397-421. https://doi.org/10.3390/knowledge4030022

[5] Hu, X., Purves, R. S., Moncla, L., Kersten, J., & Stock, K. (2026). Extracting and analysing geographic information from natural language texts.International Journal of Geographical Information Science, 40(3), 631-644. https://doi.org/10.1080/13658816.2026.2629948

[6] Sun, Y., Qiu, H., Zheng, Y., Wang, Z., & Zhang, C. (2020). Sifrank: A new baseline for unsupervised keyphrase extraction based on pre-trained language model. IEEE Access, 8, 10896-10906. https://doi.org/10.1109/ACCESS.2020.2965087

[7] Razi, Q., Singh, S., Priyadarshini, R., Hassija, V., & Chalapathi, G. S. S. (2025). A Comprehensive Survey on Data Distillation: Techniques, Frameworks, and Future Directions. IEEE Internet of Things Journal. https://doi.org/10.1109/JIOT.2025.3633710

[8] Hasan, H. M., Sanyal, F., Chaki, D., & Ali, M. H. (2017, October). An empirical study of important keyword extraction techniques from documents. In 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM) (pp. 91-94). IEEE. https://doi.org/10.1109/ICISIM.2017.8122154

[9] Naseem, U., Khushi, M., Reddy, V., Rajendran, S., Razzak, I., & Kim, J. (2021, July). Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In2021 International Joint Conference on Neural Networks (IJCNN)(pp. 1-7). IEEE. https://doi.org/10.1109/IJCNN52387.2021.9533884

[10] Çelikten, A., Uğur, A., & Bulut, H. (2021, August). Keyword extraction from biomedical documents using deep contextualized embeddings. In 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-5). IEEE. https://doi.org/10.1109/INISTA52262.2021.9548470

[11] Lai, T., Bui, T., Kim, D. S., & Tran, Q. H. (2020, December). A joint learning approach based on self-distillation for keyphrase extraction from scientific documents. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 649-656). https://doi.org/10.18653/v1/2020.coling-main.56

[12] Abdallah, M. S., Samaan, G. H., Wadie, A. R., Makhmudov, F., & Cho, Y. I. (2022). Light-weight deep learning techniques with advanced processing for real-time hand gesture recognition. Sensors, 23(1), 2. https://doi.org/10.3390/s23010002

[13] Xiao, X., Wei, P., Mao, W., & Wang, L. (2019, July). Context-aware multi-view attention networks for emotion cause extraction. In 2019 IEEE International conference on intelligence and security informatics (ISI) (pp. 128-133). IEEE. https://doi.org/10.1109/ISI.2019.8823225

[14] Diker, S. N., & Sakar, C. O. (2026). A Systematic Approach to Key Phrase Extraction for Turkish: Adapting Embedding-Based Models. IEEE Access, 14, 22211-22231. https://doi.org/10.1109/ACCESS.2026.3662520

[15] Zhang, D., Listiyani, D., Singh, P., & Mohanty, M. (2025). Distilling wisdom: A review on optimizing learning from massive language models. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3554586

[16] Shamika, U. B. P., Kumara, B. T. G. S., Seneviratna, D. M. K. N., & Rathnayaka, R. M. K. T. (2025, November). Large Language Model-based Multimodal Knowledge Distillation: A Review of Methodologies and Applications. In 2025 9th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI) (pp. 1-6). IEEE. https://doi.org/10.1109/SLAAI-ICAI68534.2025.11318526

[17] Chusova, A., Artemieva, I., & Chusov, A. (2024, September). A Hybrid Approach to Extraction of Knowledge From Scientific Texts Based on Large Language Models and Domain Dictionaries. In 2024 IEEE International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON) (pp. 266-271). IEEE. https://doi.org/10.1109/SIBIRCON63777.2024.10758538

[18] Jin, Z., & Wang, Y. (2025, July). Task-Specific Bidirectional Knowledge Distillation: Transforming General Large Language Models Into Lightweight Models. In 2025 IEEE 26th China Conference on System Simulation

Technology and its Applications (CCSSTA) (pp. 530-534). IEEE. https://doi.org/10.1109/IEEECONF65522.2025.11137049

[19] Danilov, G., Ishankulov, T., Kotik, K., Orlov, Y., Shifrin, M., & Potapov, A. (2021). The classification of short scientific texts using pretrained BERT model. In Public Health and Informatics: Proceedings of MIE 2021 (pp. 83-87). 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP: SAGE Publications. https://doi.org/10.3233/SHTI210125

[20] Ünlü, Ö., & Çetin, A. (2019, October). A survey on keyword and key phrase extraction with deep learning. In 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-6). IEEE. https://doi.org/10.1109/ISMSIT.2019.8932811

[21] Pan, C., Miao, R., Zhang, Q., Qu, B., & Wang, X. (2025). Vision-Based Fall Risk Assessment Through Attention Augmented Neural Encoding and Data Augmentation.IEEE Access. https://doi.org/10.1109/ACCESS.2025.3596942

[22] Qin, D., Bu, J. J., Liu, Z., Shen, X., Zhou, S., Gu, J. J., ... & Dai, H. F. (2021). Efficient medical image segmentation based on knowledge distillation. IEEE Transactions on Medical Imaging,40(12), 3820-3831. https://doi.org/10.1109/TMI.2021.3098703

[23] Nadim, M., Akopian, D., & Matamoros, A. (2023). A comparative assessment of unsupervised keyword extraction tools.IEEE access,11, 144778-144798. https://doi.org/10.1109/ACCESS.2023.3344032

[24] Li, Y., Zhang, Z., Wang, W., Nie, L., Li, W., & Chua, T. S. (2024, August). Distillation enhanced generative retrieval. In Findings of the Association for Computational Linguistics: ACL 2024(pp. 11119-11129). https://doi.org/10.18653/v1/2024.findings-acl.662

[25] Haque, R., Parameshachari, B. D., Hasan, M. K., Sakib, A. H., Rahman, A. U., & Islam, M. B. (2023, November). Scientific article classification: Harnessing hybrid deep learning models for knowledge discovery. In2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE) (pp. 1-7). IEEE. https://doi.org/10.1109/AIKIIE60097.2023.10389945

[26] Issa, B., Jasser, M. B., Chua, H. N., & Hamzah, M. (2023, October). A comparative study on embedding models for keyword extraction using keybert method. In2023 IEEE 13th international conference on system engineering and technology (ICSET (pp. 40-45). IEEE. https://doi.org/10.1109/ICSET59111.2023.10295108

[27] Ahmad, W., Bai, X., Lee, S., & Chang, K. W. (2021, August). Select, extract and generate: Neural keyphrase generation with layer-wise coverage attention. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)(pp. 1389-1404). https://doi.org/10.18653/v1/2021.acl-long.111

[28] Wang, M., Wang, M., Yu, F., Yang, Y., Walker, J., & Mostafa, J. (2021). A systematic review of automatic text summarization for biomedical literature and EHRs.Journal of the American Medical Informatics Association,28(10), 2287-2297. https://doi.org/10.1093/jamia/ocab143

[29] Jing, K., Zhang, Y., Wang, X., & Jia, Z. (2026). A Survey on Model Compression for Transformer-Based Large Language Models.IEEE Transactions on Emerging Topics in Computational Intelligence. https://doi.org/10.1109/TETCI.2026.3663489

[30] Xu, C., & McAuley, J. (2023, June). A survey on model compression and acceleration for pretrained language models. InProceedings of the AAAI Conference on Artificial Intelligence(Vol. 37, No. 9, pp. 10566-10575). https://doi.org/10.1609/aaai.v37i9.26255

[31] Zhang, M., Li, X., Yue, S., & Yang, L. (2020). An empirical study of TextRank for keyword extraction.IEEE access,8, 178849-178858. https://doi.org/10.1109/ACCESS.2020.3027567

[32] Xiao, L., Li, M., Feng, Y., Wang, M., Zhu, Z., & Chen, Z. (2024, August). Exploration of attention mechanism-enhanced deep learning models in the mining of medical textual data. In2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)(pp. 1251-1258). IEEE. https://doi.org/10.1109/ICSECE61636.2024.10729303

[33] Zhang, Y., Tuo, M., Yin, Q., Qi, L., Wang, X., & Liu, T. (2020). Keywords extraction with deep neural network model.Neurocomputing,383, 113-121. https://doi.org/10.1016/j.neucom.2019.11.083

[34] Payak, A., Rai, S., Shrivastava, K., & Gulwani, R. (2020, July). Automatic text summarization and keyword extraction using natural language processing. In2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)(pp. 98-103). IEEE. https://doi.org/10.1109/ICESC48915.2020.9155852

[35] Xu, G., Li, J., Gao, G., Lu, H., Yang, J., & Yue, D. (2023). Lightweight real-time semantic segmentation network with efficient transformer and CNN.IEEE Transactions on Intelligent Transportation Systems,24(12), 15897-15906. https://doi.org/10.1109/TITS.2023.3248089